Channel Simulation: Theory and Applications to Lossy Compression and Differential Privacy

Cheuk Ting Li
Department of Information Engineering,
The Chinese University of Hong Kong,
Hong Kong, China
ctli@ie.cuhk.edu.hk

Abstract

One-shot channel simulation (or channel synthesis) has seen increasing applications in lossy compression, differential privacy and machine learning. In this setting, an encoder observes a source X, and transmits a description to a decoder, so as to allow it to produce an output Y with a desired conditional distribution $P_{Y|X}$. In other words, the encoder and the decoder are simulating the noisy channel $P_{Y|X}$ using noiseless communication. This can also be seen as a lossy compression scheme with a stronger guarantee on the joint distribution of X and Y. This monograph aims at giving an overview of the theory and applications of the channel simulation problem. We will present a unifying review of various one-shot and asymptotic channel simulation techniques that have been proposed in different areas, namely dithered quantization, rejection sampling, minimal random coding, likelihood encoder, soft covering, Poisson functional representation, and dyadic decomposition.

The final publication is available from now publishers via http://dx.doi.org/10.1561/0100000141

Cite this monograph as: Cheuk Ting Li (2024), "Channel Simulation: Theory and Applications to Lossy Compression and Differential Privacy", Foundations and Trends® in Communications and Information Theory: Vol. 21: No. 6, pp 847-1106. http://dx.doi.org/10.1561/0100000141

Contents

Preface

In this monograph, we give an overview of the theoretical results on channel simulation and related settings, as well as their applications in lossy compression, differential privacy and machine learning. We collect various channel simulation schemes appearing in different fields of research. Many of them are not referred to as "channel simulation" in their respective fields. Nevertheless, they fit within the same setting of simulating a noisy channel through communications, and can therefore be analyzed and compared under a unified framework. Our goal is to gather these channel simulation techniques, and present them as a common toolbox that different lines of research can utilize.

Although this monograph is intended to be accessible to researchers outside of information theory, familiarity with basic notions such as entropy, mutual information and channel coding is necessary. Readers may consult textbooks such as Cover and Thomas, 2006, Chapters 1-10; MacKay, 2003, Chapters 1-11; Yeung, 2008, Chapters 1-11; or Csiszár and Körner, 2011, Chapters 1-7.

In Section 1, we will give an intuitive description of the channel simulation setting, and present several motivations for this setting. Readers using this monograph as a reference book may jump directly to the overview of various channel simulation schemes in Section 2.2, and the comparison in Table 2.1.

Notations

Throughout this monograph, probability spaces are assumed to be Polish equipped with the Borel sigma algebra. Entropy and mutual information are in bits. The notations used are listed as follows.

Symbol	Meaning
X, Y	Random variables (RVs)
x, y	Non-random variables
$\mathbf{X}, \mathbf{Y}$	Random vectors (sometimes non-random matrices)
$\mathbf{x}, \mathbf{y}$	Non-random vectors
$\mathcal{X}$	Set in which RV X takes values
$\mathbb{N}^+$	Positive integers $\{1, 2, \ldots\}$
$\mathbb{N}_0$	Nonnegative integers $\{0, 1, \ldots\}$
$[n]$	$\{1, \ldots, n\}$
$[a..b]$	$\{a, a+1, \ldots, b\}$
$\{0,1\}^*$	Set of bit sequences of any length $\bigcup_{\ell=0}^{\infty}\{0,1\}^\ell$
$a \| b$	Concatenation of $a, b \in \{0,1\}^*$
X^n	Sequence $(X_1, \ldots, X_n)$

Symbol	Meaning
$X \sim P$	X is an RV with distribution P
$Y\|X \sim Q$	The conditional distribution of Y given X is $Q(y\|x)$
$X \perp\!\!\!\perp Y$	RV X is independent of RV Y
$X \leftrightarrow Y \leftrightarrow Z$, or $X \perp\!\!\!\perp Z\|Y$	RVs X, Y, Z forms a Markov chain, or X, Z are conditionally independent given Y
P_X	Probability distribution of X
$P_{Y\|X}$	Conditional distribution of Y given X
$P_X P_Y$	Product of P_X and P_Y, i.e., distribution of (X, Y) if $X \sim P_X$, $Y \sim P_Y$, $X \perp\!\!\!\perp Y$
$P_X P_{Y\|X}$	Semidirect product of P_X and $P_{Y\|X}$, i.e., distribution of (X, Y) if $X \sim P_X$, $Y\|X \sim P_{Y\|X}$
P_X^n	n-fold product distribution, i.e., distribution of $X^n = (X_1, \ldots, X_n)$ if $X_i \sim P_X$ i.i.d.
$P_{Y\|X}^n$	Conditional distribution $P_{Y^n\|X^n}$, Y^n is the output when X^n is passed through memoryless channel $P_{Y\|X}$ (for discrete RVs, $P_{Y\|X}^n(y^n\|x^n) = \prod_{i=1}^n P_{Y\|X}(y_i\|x_i)$)
$\mathrm{Unif}(S)$	Uniform distribution over the set S
$\mathrm{Unif}(a, b)$	Uniform distribution over the interval $[a, b]$
$\mathrm{Bern}(a)$	Bernoulli distribution $P(0) = 1 - a$, $P(1) = a$
$\mathrm{Geom}(a)$	Geometric distribution $P(x) = (1-a)^{x-1}a$, $x = 1, 2, \ldots$
$\mathrm{Zipf}(s)$	Zipf distribution $P(x) \propto x^{-s}$, $x = 1, 2, \ldots$ (Appendix A)

Symbol	Meaning
$\mathbf{1}\{E\}$	Indicator of event E ($\mathbf{1}\{E\} \in \{0,1\}$ is 1 iff E occurs)
$\mathbf{1}_S(x)$	Indicator of event $x \in S$
$\mu \ll \nu$	Measure μ is absolutely continuous with respect to ν, i.e., $\nu(S) = 0 \Rightarrow \mu(S) = 0$ for measurable set $S \subseteq \mathcal{X}$
$\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$	Radon-Nikodym derivative between measures $\mu \ll \nu$
$H(X)$	Entropy (in bits) $\mathbb{E}[-\log_2 P_X(X)]$
$H(X\|Y)$	Conditional entropy $\mathbb{E}[-\log_2 P_{X\|Y}(X\|Y)]$
$I(X;Y)$	Mutual information $\mathbb{E}\big[\log_2 \frac{\mathrm{d}P_{X,Y}}{\mathrm{d}P_X P_Y}(X,Y)\big]$
$h(X)$	Differential entropy $\mathbb{E}[-\log_2 f_X(X)]$ (f_X is pdf of X)
$D_{\mathrm{KL}}(P\|Q)$	Kullback-Leibler divergence $\mathbb{E}_{X\sim P}[\log_2 \frac{\mathrm{d}P}{\mathrm{d}Q}(X)]$ (needs $P \ll Q$)
$D_\infty(P\|Q)$	Max divergence $\operatorname{ess\,sup}_{X\sim Q} \log_2 \frac{\mathrm{d}P}{\mathrm{d}Q}(X)$ $= \inf\left\{t : \mathbb{P}_{X\sim Q}\big(\log_2 \frac{\mathrm{d}P}{\mathrm{d}Q}(X) > t\big) = 0\right\}$ (needs $P \ll Q$)
$\delta_{\mathrm{TV}}(P,Q)$	Total variation distance $\sup_E \lvert P(E) - Q(E)\rvert$ (Section 5.3)
$\gamma\mathcal{A}$	$\{\gamma\mathbf{y} : \mathbf{y} \in \mathcal{A}\}$ (for $\mathcal{A} \subseteq \mathbb{R}^n$, $\gamma \in \mathbb{R}$)
$-\mathcal{A}$	$\{-\mathbf{y} : \mathbf{y} \in \mathcal{A}\}$ (for $\mathcal{A} \subseteq \mathbb{R}^n$)
$\mathcal{A} + \mathbf{x}$	$\{\mathbf{y} + \mathbf{x} : \mathbf{y} \in \mathcal{A}\}$ (for $\mathcal{A} \subseteq \mathbb{R}^n$, $\mathbf{x} \in \mathbb{R}^n$)
$\mathcal{A} + \mathcal{B}$	$\{\mathbf{x} + \mathbf{y} : \mathbf{x} \in \mathcal{A}, \mathbf{y} \in \mathcal{B}\}$ (for $\mathcal{A}, \mathcal{B} \subseteq \mathbb{R}^n$)
$\mathrm{Vol}(\mathcal{A})$	Lebesgue measure of measurable set $\mathcal{A} \subseteq \mathbb{R}^n$

Notation for channel simulation results (Section 2.1)

E.g., G/1/E/VL/KAS/UCR

D	Discrete channels (X, Y finite discrete)
C	Continuous output channels ($P_{Y\|X}(\cdot\|x)$ continuous)
1DAC	1D additive continuous noise channels ($Y = X + Z \in \mathbb{R}$)
nDAC	n-D additive continuous noise channels ($\mathbf{Y} = \mathbf{X} + \mathbf{Z} \in \mathbb{R}^n$)
G	General channels (over Polish space)
1	One-shot (simulating channel $P_{Y\|X}$)
F	Finite-blocklength (memoryless channel $P_{Y^n\|X^n}$)
∞	Asymptotic ($n \to \infty$)
A	Approximate ($P_{\tilde{Y}\|X}$ is approximately $P_{Y\|X}$)
E	Exact ($Y\|X \sim P_{Y\|X}$ exactly)
FL	Fixed-length description ($M \in \{0,1\}^\ell$)
VL	Variable-length description ($M \in \{0,1\}^*$)
KS	Known source distribution (P_X known)
AS	Arbitrary source
KAS	Known or arbitrary source
NCR	No common randomness W
LCR	Limited common randomness W
UCR	Unlimited common randomness W

1 Introduction and Motivations

1.1 Overview

A channel is a conditional distribution $P_{Y|X}$, or in a more operational sense, a mechanism which takes an input symbol X and produces an output symbol Y which depends on X in a random manner. Common examples include the binary symmetric channel (where $X, Y \in \{0, 1\}$, $Y = X$ with probability $1 - p$, or $Y \neq X$ with probability p), the binary erasure channel (where $X \in \{0, 1\}$, $Y \in \{0, 1, e\}$, $Y = X$ with probability $1 - p$, or $Y = e$ with probability p), and the additive Gaussian noise channel (where $X \in \mathbb{R}$, $Y = X + Z$ where $Z \sim N(0, 1)$ is an independent Gaussian noise).

In noisy channel coding (Shannon, 1948), we are given a channel $P_{Y|X}$, and we have to transmit a message M almost noiselessly (i.e., with error probability close to 0) through the channel. In other words, we are trying to simulate a noiseless bit channel using a noisy channel. A channel code converts a noisy channel into a more readily usable almost-noiseless bit channel where reliable communication can be conducted.

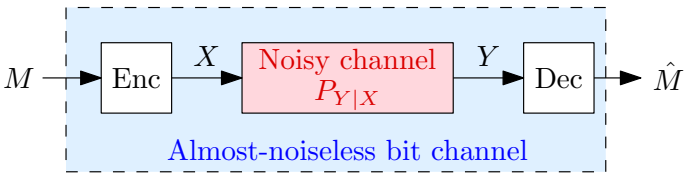
Channel simulation (Bennett *et al.*, 2002; Winter, 2002; Cuff, 2013; Bennett *et al.*, 2014) works the other way around. We are given a noiseless bit channel, and we want to simulate a noisy channel $P_{Y|X}$. More precisely, the encoder observes the input X, and sends a description M noiselessly to the decoder. The decoder will then produce the output Y. The goal is to have Y following a prescribed conditional distribution $P_{Y|X}$ given X, using a description M as short as possible. The encoder and the decoder are also sometimes allowed to access a common randomness source W. If we look at this setting from the outside, and put the encoder, the decoder and the description inside a box with input X and output Y, then this box would behave as if it is a noisy channel $P_{Y|X}$. Refer to Figure 1.1 for an illustration.

In channel coding, the capacity of a channel

$$C := \max_{P_X} I(X; Y)$$

is the number of noiseless bits one can transmit per use of the channel, in the asymptotic setting where the number of uses of the memoryless channel tends to infinity, as shown by Shannon's channel coding theorem (Shannon, 1948). If one must design a code with an empirical distribution of the input sequence close to a given distribution P_X, then the number of bits per channel use is $I(X; Y)$. A central result in channel simulation, called the *reverse Shannon theorem* (Bennett *et al.*, 2002; Bennett *et al.*, 2014), shows that one can perform conversion in the other way around at the same exchange rate. Simulating the channel $P_{Y|X}$ takes C noiseless bits per channel simulated, in the sense that if we want to simulate n copies of the memoryless channel $P_{Y|X}$, we require $\approx nC$ noiseless bits asymptotically as $n \to \infty$ (assuming unlimited common randomness; see Theorems 32 and

Channel coding:



Channel simulation:

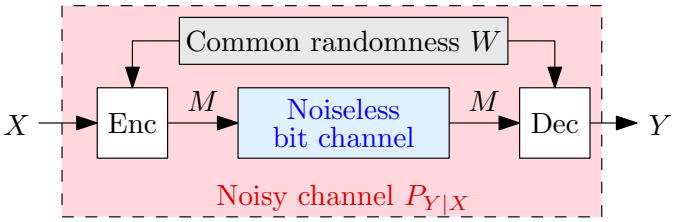


Figure 1.1: Channel coding (top) and channel simulation (bottom).

38). If the distribution of X is known, then it takes $I(X;Y)$ bits per channel simulated. This shows that one can convert a noisy channel to a noiseless bit channel and vice versa without any loss of asymptotic efficiency.

The asymptotic memoryless nature of the result imposes a limitation on its applications. What if we want to simulate channels with memory? What if the input to the channel is an image instead of a sequence? What if we want to simulate a finite number of copies of the channel? What if we want to simulate just *one* copy of the channel (which can take a finite sequence, an image, or *anything* as input)? Interestingly, the channel simulation result holds approximately even when only one copy of the channel is simulated. More precisely, one copy of the channel $P_{Y|X}$ can be simulated exactly, using

$$C + \log_2(C+2) + 3 \,\text{bits} \tag{1.1}$$

on average with a prefix-free description, in the presence of unlimited common ranomness. If the distribution of X is known, the channel can be simulated using

$$I(X;Y) + \log_2(I(X;Y)+2) + 3 \,\text{bits} \tag{1.2}$$

on average. This result (in a weaker form) was first studied by Harsha *et al.* (2010), and later improved by Braverman and Garg (2014), and Li and El Gamal (2018b) (which gave the form stated in (1.1) with a slightly larger constant).[1] See Theorem 4. This general "one-shot" result only has a logarithmic gap compared to the more restrictive asymptotic

[1] Li and El Gamal (2018b) gave the bound $C + \log_2(C+1) + 5$. The bound $C + \log_2(C+2) + 3$ was given in (Li, 2024) using a slightly improved analysis.

result. In this regard, channel simulation is similar to lossless source coding, where Huffman code (Huffman, 1952) can compress one source symbol X to at most $H(X)+1$ bits on average, close to the optimal asymptotic rate $H(X)$. This is in stark contrast to channel coding, where the number of bits that can be conveyed reliably by one channel use is often far less than C, and can even be 0. While asymptotically one can convert a noisy channel to a noiseless bit channel and vice versa equally efficiently, in the nonasymptotic setting, it is much easier to convert a noiseless bit channel to a noisy channel (channel simulation) than the other direction (channel coding).

But why do we want to simulate a noisy channel? Isn't a noiseless bit channel more useful than a noisy channel? To answer these questions, one should not compare channel simulation to channel coding, but to lossy source coding, which is another "dual" of channel coding. The lossy source coding (or lossy compression) setting looks almost identical to channel simulation. The only difference is that the goal is not to have Y following $P_{Y|X}$, but to have a small distortion $d(X,Y)$ between the input and the output, i.e., the output Y should be "close" to the input X under some distortion measure. If we can guarantee a precise joint distribution of (X,Y), then we can characterize the expected distortion $\mathbb{E}[d(X,Y)]$. Therefore, a channel simulation scheme can serve as a lossy source coding scheme, with a stronger guarantee on the conditional distribution of Y given X. Moreover, this stronger guarantee does not necessarily impose any penalty on the asymptotic rate.[2] The reason for simulating a noisy channel is the same as the reason for lossy compression—because the noise allows us to have a smaller compression size.

But why do we want this stronger guarantee? Why do we care about the distribution of Y? Isn't it enough that Y is close to X? To answer these questions, we list several reasons for wanting a precise distribution of Y given X:

- *Because noisy channels provide better worst-case performance.* Consider a quantization $Y=\lfloor X+1/2\rfloor$ of the signal $X\in\mathbb{R}$ to the nearest integer. As a lossy compression of X, the quantization Y has zero error when X is an integer, but not as good when X is a half-integer like 1.5, which gives an error $|Y-X|=1/2$. One may argue that if the distribution of X is "sufficiently spread out", then the expected error should be around $1/4$, though there is often no guarantee that X is indeed spread out in the way we want. Compare this to the output of the additive noise channel $Y'=X+Z$ where $Z\sim\mathrm{Unif}(-1/2,1/2)$. The expected error $\mathbb{E}[|Y'-X|]$ is *always* $1/4$, regardless of the distribution of X. The channel $P_{Y'|X}$ has a better worst-case performance than

[2] Since the optimal rate of asymptotic lossy source coding under a distortion constraint is characterized by the rate-distortion function $R(D)=\min_{P_{Y|X}:\mathbb{E}[d(X,Y)]\le D} I(X;Y)$, which is the same description rate needed to simulate (with unlimited common randomness) the channel $P_{Y|X}$ that attains the minimum in $R(D)$, the stronger guarantee on the conditional distribution of Y given X does not come with any penalty on the rate (Winter, 2002).

the deterministic mapping $X \mapsto Y$. This is investigated in the work on *universal quantization* by Ziv (1985) and Zamir and Feder (1992), showing that simulating an additive noise channel gives a universally good lossy compression scheme. See Sections 1.3, 1.5 and 3.6.

Randomization improving the worst-case performance is a common occurence in different fields of mathematics. In game theory, the optimal minimax strategy is often a mixed strategy, where a player chooses the action at random. Channel simulation can be applied to allow players to generate their actions from the correct distribution using a small amount of communication. See Section 1.10. In interactive protocols for function computation (Yao, 1979), we often require the protocol to produce the correct value of the function with high probability, for every choice of inputs. Randomized protocols can give a better guarantee on the worst-case error probability, and channel simulation can be applied to compress the messages in such protocols. See Section 1.9.

- *Because noisy channels are nicer objects.* Recall the "quantization versus channel simulation" example. The mapping $X \mapsto Y = \lfloor X + 1/2 \rfloor$ behaves poorly—it is flat almost everywhere, and jumps discontinuously. In comparison, in the additive noise channel $Y' = X + Z$ (where $Z \sim \mathrm{Unif}(-1/2, 1/2)$), Y' changes continuously with X. Such niceness makes channel simulation appealing to compression tasks in machine learning (Havasi *et al.*, 2019; Choi *et al.*, 2019; Choi *et al.*, 2020; Agustsson and Theis, 2020). For example, in neural compression where we train a neural network for a compression task, channel simulation can be more suitable than quantization (Choi *et al.*, 2019; Agustsson and Theis, 2020; Yang *et al.*, 2023). The quantization Y varies with X in a discontinuous and nondifferentiable manner, and almost always has zero gradient $\mathrm{d}Y/\mathrm{d}X$; whereas the channel simulation output Y' varies with X in a differentiable manner, allowing gradient-based optimization algorithms to be applied via the reparameterization trick (Kingma and Welling, 2013). Another popular channel to simulate is the additive white Gaussian noise channel (Havasi *et al.*, 2019; Flamich *et al.*, 2020), with desirable theoretical properties. See Section 1.4.

- *Because noisy channels help preserving privacy.* Suppose a user wants to convey the data X to the server, but does not want the server to know X exactly. One common method is the *additive noise mechanism* (Dwork *et al.*, 2006), where the user instead transmits $X + Z$ to the server, where Z is a noise (e.g. Gaussian or Laplace) which masks the least significant digits of X, while still allowing the server to know a rough estimate of X. Channel simulation can be applied to convey $X + Z$ with a small amount of communications (Feldman and Talwar, 2021; Amiri *et al.*, 2021; Triastcyn

et al., 2021; Lang *et al.*, 2023; Shahmiri *et al.*, 2024; Hasırcıoğlu and Gündüz, 2024; Hegazy *et al.*, 2024).[3] See Section 1.6.

- *Because random output looks/sounds better.* In image compression, dithering is a technique where random noises are added to the pixels in order to improve perceptual quality by reducing undesirable effects such as color banding (Roberts, 1962). Dithering is also used in audio compression to eliminate undesired patterns in the quantization error (Jayant and Rabiner, 1972). Another example is neural image compression where, in case the rate is too low to accurately convey the source image, we might prefer a random realistic-looking image similar to the source image as the output (Tschannen *et al.*, 2018; Blau and Michaeli, 2018; Blau and Michaeli, 2019). Another related technique is perceptual noise substitution for audio coding, where the encoder does not encode the noise-like components of the audio, and the decoder fills in the missing components by generating noises that mimick those components (Sayood, 2018). The desire of having a nice-looking/sounding output distribution is the basis of the works on distribution preserving compression (Li *et al.*, 2010; Li *et al.*, 2011; Tschannen *et al.*, 2018) and the rate-distortion-perception tradeoff (Blau and Michaeli, 2018; Blau and Michaeli, 2019). See Section 1.7.

- *Because we indeed want to simulate a random phenomenon.* One example is quantum measurement, which is inherently random. Studying the amount of classical communication needed to simulate the outcomes of some quantum measurements can allow us to understand the "conversion rates" between quantum and classical resources. See Section 1.8.

We remark that this monograph is focused on the simulation of classical channels. For the simulation of quantum channels, readers are referred to (Barnum *et al.*, 2001; Bennett *et al.*, 2014; Bennett *et al.*, 2002; Pirandola *et al.*, 2018). We also remark that this monograph is not about software or hardware simulation of physical communication channels (e.g., (Mezzavilla *et al.*, 2015; Sun *et al.*, 2017)), though a related setting will be briefly discussed in Section 9.3.

[3] Technically, channel simulation allows the server to know $X + Z$, but does not guarantee that the server knows $X + Z$ *only*. Additional efforts may be needed to ensure that a channel simulation scheme is differentially private. See Section 1.6.

1.2 History

The formal study of channel simulation was started by Bennett *et al.* (2002). Earlier related works include the work on distributed source simulation by Wyner (1975a) (to be discussed in Section 9.2.2), the work on simulating a noisy channel at a single terminal (unlike Figure 1.1 which is a distributed setting) with the minimum amount of randomness by Steinberg and Verdú (1994) (to be discussed in Section 9.3.2), the work on quantum coding by Barnum *et al.* (2001), and the work on the simulation of quantum entanglement by classical communication by Steiner (2000). It is also known by different names, such as distributed channel synthesis (Cuff, 2013), compression of sources of probability distributions (Winter, 2002), communication of probability distributions (Barnum *et al.*, 2001; Kramer and Savari, 2007), communication complexity of correlation (Harsha *et al.*, 2010), generic transformation (in the context of local privacy protocols) (Bassily and Smith, 2015; Bun *et al.*, 2019),[4] relative entropy coding (Flamich *et al.*, 2020; Flamich *et al.*, 2022; Flamich *et al.*, 2024),[5] and reverse channel coding (Theis and Yosri, 2022; Flamich *et al.*, 2022). In this monograph, we adopt the name "channel simulation" from (Bennett *et al.*, 2002; Cuff, 2008; Bennett *et al.*, 2014).

Although channel simulation is often referred by different names in different lines of works, they can all fit within the setting in Figure 1.1 where samples following a desired conditional distribution are communicated. The purpose of this monograph is to place these different techniques arising in different lines of research under a unifying theoretical framework. In the remainder of this section, we will briefly review the many places where the channel simulation setting has appeared.

1.3 Dithered Quantization

Lossy compression generally refers to a method which compresses the input X into a discrete description, from which we can recover the output Y that is close to X. For the sake of concreteness, before we proceed to the abstract setting in Section 1.5, we first discuss a simple form of lossy compression—quantization (Gray and Neuhoff, 1998), where the input signal $X \in \mathbb{R}$ is mapped to the quantized signal

[4] Generic transformation (Bassily and Smith, 2015; Bun *et al.*, 2019) refers to methods that transform a general privacy protocol (which is essentially a noisy channel) to a protocol with a smaller amount of communication. In addition to the communication constraint in channel simulation, generic transformation is also required to be differentially private.

[5] Relative entropy coding specifically refers to channel simulation schemes with a communication cost close to the relative entropy (Kullback-Leibler divergence) between the target distribution and the reference distribution.

$$Y = Q(X) := \Delta \left\lfloor \frac{X}{\Delta} + \frac{1}{2} \right\rfloor,$$

where $\Delta > 0$ is the quantization step. The quantizer maps a continuous input X to a discrete output $Q(X) \in \{\ldots, -2\Delta, -\Delta, 0, \Delta, 2\Delta, \ldots\}$. The problem of such a quantizer is that it may introduce a systematic bias to the signal. For example, if $\Delta = 1$, and the input signals are usually concentrated around $X \approx 0.7$, then the output will usually be 1, with a bias upward.

To eliminate this bias, a dither signal is sometimes added to the input signal before quantization (Gray and Stockham, 1993). Consider a dither signal W uniformly distributed over $[-1/2, 1/2]$. The quantized signal is taken to be

$$Y = Q(X + W\Delta).$$

This method is called *nonsubtractive dithering* (Gray and Stockham, 1993; Wannamaker *et al.*, 2000) to distinguish it from another form of dithering discussed later. One can show that $\mathbb{E}[Y \mid X] = X$, and the quantization error $Y - X$ is uncorrelated with X.

Nevertheless, the distribution of the quantization error $Y - X$ still depends on X. The best case is when X is a multiple of Δ, where there is no error. The worst case is $X = (k + 1/2)\Delta$ for some $k \in \mathbb{Z}$, where the error is uniform over $\{-\Delta/2, \Delta/2\}$. For the sake of simplicity of analysis and the worst-case performance, it is desirable to have a quantization error that is independent of X. This can be achieved by a technique called *subtractive dithering* (Roberts, 1962; Schuchman, 1964; Ziv, 1985), where we subtract the dither signal from the output of the quantizer, i.e., the final reconstruction is

$$Y = Q(X + W\Delta) - W\Delta.$$

Note that Y is the closest point to X among the reconstruction levels in $\{\ldots, (-2 - W)\Delta, (-1 - W)\Delta, -W\Delta, (1 - W)\Delta, \ldots\}$. It can be checked that the quantization error $Y - X$ is uniformly distributed over $[-\Delta/2, \Delta/2]$, independent of the input signal X. In other words, the channel $X \to Y$ is an additive white noise channel with a uniformly distributed noise. Refer to Figure 1.2 for an illustration. This can be regarded as the earliest form of channel simulation. For generalizations to additive noise channels with nonuniform noises, refer to Section 3.6.

Compared to quantization without dithering, or non-subtractive dithering, now we have a tractable model of the channel $X \to Y$ that does not depend on how X aligns with the quantization levels. This allows subtractively dithered quantization to perform uniformly well regardless of the input X. In (Ziv, 1985; Zamir and Feder, 1992), it was shown that subtractively dithered quantization has a *universal quantization* property, that is, it is an almost optimal lossy compression scheme for every input distribution. This shows that channel simulation is not only "lossy compression with a stronger distributional guarantee

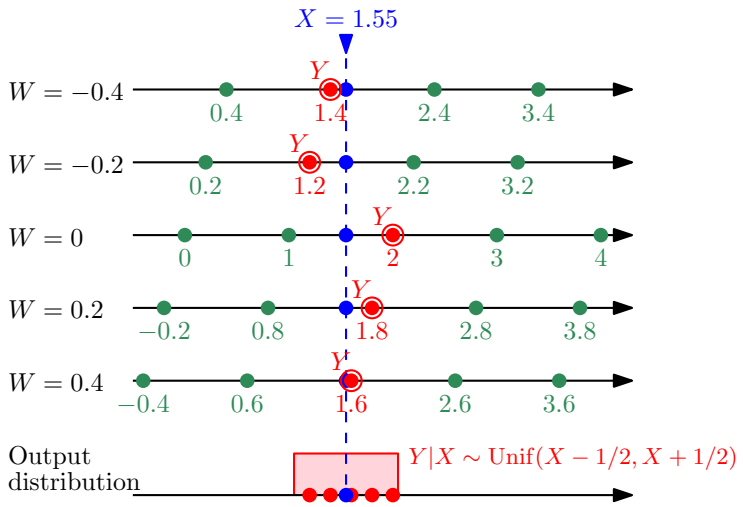


Figure 1.2: Subtractive dithering with $\Delta = 1$, $X = 1.55$. We first generate the dither signal $W \sim \mathrm{Unif}(-1/2, 1/2)$, and then find the reconstruction level among $\{\ldots, -2 - W, -1 - W, -W, 1 - W, \ldots\}$ that is closest to X, and output it as Y. The figure shows the values of Y under different values of W. The conditional distribution of the output given the input is $Y|X \sim \mathrm{Unif}(X - 1/2, X + 1/2)$.

on the output", but can also be desirable as a lossy compression scheme even if we ignore the stronger distributional guarantee, due to its better distortion guarantee for the worst-case input distribution. Refer to Section 1.5 for further discussion.

1.4 Machine Learning

Nonlinear transform coding. In neural image compression via nonlinear transform coding (Ballé *et al.*, 2020), the image is mapped to a latent representation $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ via a neural network (the *analysis transform*), which can then be used to reconstruct the image via another neural network (the *synthesis transform*). Since real numbers cannot be encoded into finitely many bits, a quantization must be performed on $\mathbf{X}$ (e.g., 16 or 32-bit floating point numbers, or 8-bit integers) before it can be stored. A simple strategy is to quantize each entry of $\mathbf{X}$ using the quantization function $Q(x) := \Delta\lfloor x/\Delta + 1/2 \rfloor$ with step size $\Delta > 0$. Compressing $\mathbf{X}$ into $\hat{\mathbf{X}} := (Q(X_i))_{i=1,\ldots,d}$ introduces a distortion to the latent representation, so we should take such distortion into account when we train the networks. Nevertheless, $Q(x)$ has zero derivative almost everywhere, making gradient-based optimization methods impossible. It has been proposed by Ballé *et al.* (2017) that we should use a proxy loss function during training, which treats the quantization output $Q(X_i)$ as if it is the output $Y_i = X_i + Z_i$ of an additive noise channel with uniform noise $Z_i \sim \mathrm{Unif}(-\Delta/2, \Delta/2)$, making the dependency between the input and the output differen-

tiable. When computing the gradient using a batch of datapoints, we generate and fix the noises $\mathbf{Z} = (Z_1, \ldots, Z_d)$, turning the quantization layer into a layer $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$ that merely adds a bias $\mathbf{Z}$, where the gradient can be readily propagated. This is an example of the reparameterization trick (Kingma and Welling, 2013), a technique for gradient computation on a layer given by a randomized function.

However, if we use additive noise channels during training, but use quantization for the actual compression, this again introduces a discrepancy between training and the actual performance (Yang *et al.*, 2020). Minimizing the training loss under the additive noise model may not result in a satisfactory compression under quantization. For example, when X_i lies around a quantization boundary, the loss will be underestimated.

It turns out that we can completely eliminate the discrepancy, by performing channel simulation for the additive noise channel $Y_i = X_i + Z_i$. One method is to apply universal quantization (Ziv, 1985; Zamir and Feder, 1992) described in the previous section to simulate this additive uniform noise channel, which is the basis of the neural compression algorithm by Choi *et al.* (2019) and Agustsson and Theis (2020). Other techniques for channel simulation (also known as *relative entropy coding* and *reverse channel coding* in the machine learning literature), with additive Gaussian noise channels being popular channels to be simulated, have been proposed for neural compression (Havasi *et al.*, 2019; Flamich *et al.*, 2020; Flamich *et al.*, 2022; Flamich *et al.*, 2024). Refer to Figure 1.3 for an illustration.

Implicit neural representation. Another approach to image compression via neural network is *implicit neural representation* (Stanley, 2007; Sitzmann *et al.*, 2020; Tancik *et al.*, 2020; Dupont *et al.*, 2021), which treats an image $\mathbf{D}$ as a function mapping the coordinates (x, y) to the RGB values of the pixel at (x, y), and trains a neural network to fit this function. The weight vector $\mathbf{w}$ of the network is then quantized to give the encoding of the image. This method is flexible and can be applied to other kinds of data, such as audio and video (Dupont *et al.*, 2022; Guo *et al.*, 2023; He *et al.*, 2024a), signed distance functions of 3D shapes (Park *et al.*, 2019), and radiance field representations of scenes (Mildenhall *et al.*, 2021). In order to optimize the encoding function, the idea by Guo *et al.* (2023) and He *et al.* (2024a) is to train a distribution $P_{\mathbf{W}|\mathbf{D}=\mathbf{d}}$ of weight vectors (the *model posterior distribution*, taken to be a Gaussian distribution) using the image $\mathbf{D} = \mathbf{d}$, instead of a single weight vector. Relative entropy coding (Havasi *et al.*, 2019; Maddison *et al.*, 2014; Flamich *et al.*, 2022), a channel simulation technique, is then applied to simulate the channel $P_{\mathbf{W}|\mathbf{D}}$, so the encoder observing $\mathbf{d}$ can produce a description M so as to allow the decoder observing M to recover a weight vector $\mathbf{W}$ following the distribution $P_{\mathbf{W}|\mathbf{D}=\mathbf{d}}$, and then use a neural network with weights $\mathbf{W}$ to recover the image. To apply relative entropy coding, it is also necessary to estimate the model prior distribution $P_{\mathbf{W}}$ using the training data $\mathbf{d}_1, \ldots, \mathbf{d}_m$.

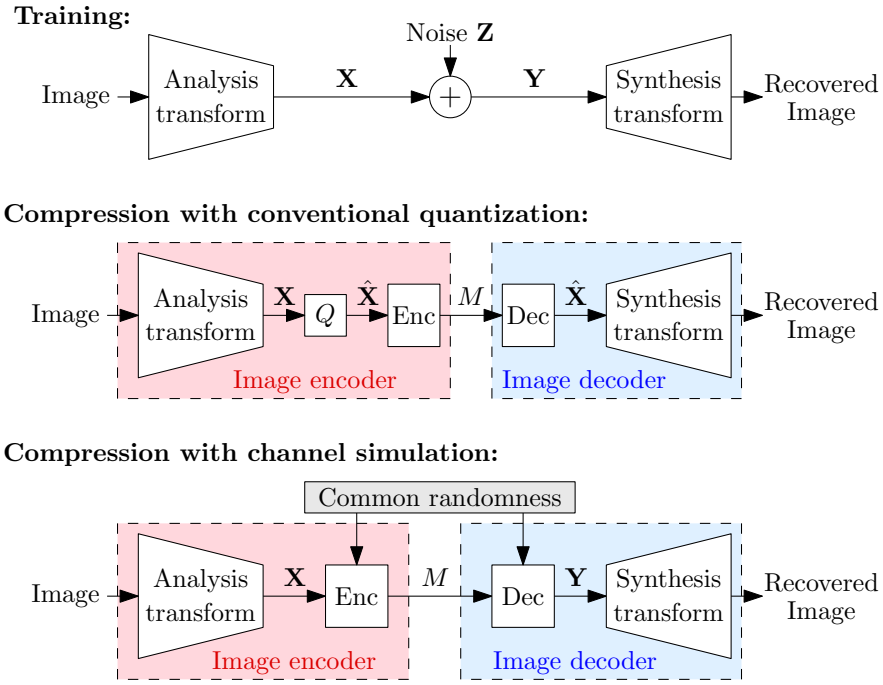


Figure 1.3: Top: Training of the analysis transform and the synthesis transform neural networks, where the effect of compression is mimicked by adding a noise $\mathbf{Z}$ to the latent representation $\mathbf{X}$ to form the noisy representation $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$ (Ballé *et al.*, 2017). Middle: The deployed image encoder and decoder using conventional quantization, where $\mathbf{X}$ is quantized into $\hat{\mathbf{X}}$ and encoded into a sequence of bits M (Ballé *et al.*, 2017). Bottom: The deployed image encoder and decoder using channel simulation (Agustsson and Theis, 2020), where the decoder can recover $\mathbf{Y}$ distributed exactly as if it is corrupted by the same noise $\mathbf{Z}$ as in the training.

Neural estimator of the rate-distortion function. In the work by Lei *et al.* (2022), channel simulation is employed in a neural compressor capable of compressing data to a size close to the rate-distortion function (Section 1.5). It utilizes a property of sampling-based channel simulation schemes (such as minimal random coding (Havasi *et al.*, 2019), Poisson functional representation (Li and El Gamal, 2018b; Li and Anantharam, 2021) and ordered random coding (Theis and Yosri, 2022); see Section 3.5) that these schemes can operate when given a sequence of candidate outputs $Y_1, Y_2, \ldots$ and the conditional distribution $P_{Y|X}$ of the channel to be simulated, both of which can be provided by the neural network proposed by Lei *et al.* (2022).

Interested readers are also referred to (Choi *et al.*, 2020) for the use of universal quantization in the compression of the weights of a neural network, (Theis *et al.*, 2022) for the application of channel simulation ideas to diffusion generative models, (Isik *et al.*, 2024) for the use of channel simulation in the compression of model updates in federated learning, and (Yang *et al.*, 2023) for a comprehensive overview of channel simulation and other techniques in neural compression.

1.5 Lossy Compression

Consider the lossy source coding setting, where the encoder compresses the source $X \in \mathcal{X}$ into a description $M = f(X)$ (which may be fixed-length or variable-length), and the decoder uses M to recover the reconstruction $Y = g(M) \in \mathcal{Y}$ with distortion $d(X, Y)$, where $d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a distortion function. The precise formulation depends on whether the source distribution is known:

- **Known source distribution.** If we know $X \sim P_X$, there are several distortion constraints we can impose, e.g., *small excess distortion probability*

$$\mathbb{P}(d(X, Y) > D) \leq \epsilon, \tag{1.3}$$

  or the *expected distortion constraint*

$$\mathbb{E}[d(X, Y)] \leq D. \tag{1.4}$$

- **Arbitrary source.** If the distribution of X is unknown,[6] we usually impose the *worst-case distortion constraint*

$$\sup_{x \in \mathcal{X}} d(x, g(f(x))) \leq D. \tag{1.5}$$

[6] Here the source is completely arbitrary. Another line of work is *universal compression* (Ziv and Lempel, 1977; Wyner and Ziv, 1994), where the precise distribution of the source sequence $X_1, \ldots, X_n$ is unknown, but some assumptions on the distribution are imposed (e.g. X_i are i.i.d., or stationary ergodic), though we do not consider this line of work here.

For the case where the source distribution is known, it was traditionally studied in the asymptotic setting, where the source is $X_1, \dots, X_n \overset{iid}{\sim} P_X$, the reconstruction is $Y_1, \dots, Y_n \in \mathcal{Y}$, the distortion is $d(X^n, Y^n) = n^{-1} \sum_{i=1}^n d(X_i, Y_i)$, and we take $n \to \infty$. In this setting, both choices of the distortion constraint (1.3) and (1.4) have the same optimal description rate given by the *rate-distortion function* (Berger, 2003)[7]

$$R(D) := \inf_{P_{Y|X}: \mathbb{E}[d(X,Y)] \le D} I(X;Y).$$

The worst-case distortion constraint is the stronger condition, which guarantees $d(X, Y) \le D$ regardless of the source distribution, making it suitable for situations where we do not know the source distribution, or when the source is chosen by an adversary. The optimal description length can be given in terms of the *epsilon-entropy* $\log_2 \min\{k : \exists y_1, \dots, y_k \in \mathcal{Y} : \forall x \in \mathcal{X} : \exists i : d(x, y_i) \le D\}$ (Kolmogorov, 1956). In the asymptotic setting where $X_1, \dots, X_n \in \mathcal{X}$ is arbitrary, the optimal description rate is given by $\max_{P_X} R(D)$, where the maximization is over source distributions P_X over the support $\mathcal{X}$ (Berger, 1971; Csiszár and Körner, 2011). Intuitively, if we do not know the distribution of X, then we have to cater for the worst distribution P_X that maximizes $R(D)$.

It has been noted by Winter (2002) that channel simulation schemes can be regarded as lossy source coding schemes. Assume that the common randomness $W \sim P_W$ is available to the encoder and the decoder. The encoding function is $M = f(W, X)$, and the decoding function is $Y = g(W, M)$. By simulating the channel $P_{Y|X}$, the expected distortion $\mathbb{E}[d(X, Y)]$ can be controlled. A channel simulation scheme with common randomness can be converted to a lossy source coding scheme without common randomness by fixing a value of W that gives the smallest expected distortion. It has been shown in (Li and El Gamal, 2018b) as a corollary of (1.1) (slightly improved in (Li, 2024)) that when the source distribution P_X is known, there exists a one-shot lossy source coding scheme (without common randomness, but possibly with local randomness) with prefix-free description $M \in \mathcal{C}$ (where $\mathcal{C} \subseteq \{0,1\}^*$ is a prefix-free codebook), achieving an expected distortion $\mathbb{E}[d(X, Y)] \le D$ and expected length[8]

$$\mathbb{E}[|M|] \le R(D) + \log_2(R(D) + 2) + 4.01 \text{ bits}.$$

[7]For vanishing excess distortion probability ($\epsilon \to 0$ as $n \to \infty$) and the expected distortion constraint, we can use fixed-length codes for M to attain the rate-distortion function (Berger, 2003). We can also have the almost-sure distortion constraint where $\epsilon = 0$ (i.e., *d-semifaithful code* (Ornstein and Shields, 1990)), where prefix-free codes for M are required to attain the rate-distortion function (Zhang *et al.*, 1997). If we require $\epsilon = 0$ and a fixed-length code, we will generally require a larger rate given by $\max_{\tilde{P}_X} R(D)$, where the maximization is over distributions $\tilde{P}_X$ with a support contained in the support of P_X (this is the same rate as the case of worst-case distortion constraint) (Berger, 1971; Csiszár and Körner, 2011). We require some regularity conditions for d, which are omitted here.

[8]To convert a channel simulation scheme to a lossy source coding scheme without common randomness, we have to choose w with small $\mathbb{E}[d(X, Y)|W = w]$ and $\mathbb{E}[|M| \,|\, W = w]$. Carathéodory's theorem is invoked in (Li and El Gamal, 2018b) to show that we can average over two values w_1, w_2 to keep both the distortion and the expected length small, imposing a 1 bit penalty on the expected length.

This can be regarded as a lossy analogue of Huffman coding (Huffman, 1952), in the sense that they both show that the expected length of one-shot variable-length source coding is within a small gap from the rate of asymptotic source coding.

Furthermore, if the common randomness W is allowed, the guarantee provided by channel simulation that Y follows the conditional distribution $P_{Y|X}$ can be valuable. By designing the channel $P_{Y|X}$ carefully, we can have the following *worst-case expected distortion constraint*

$$\sup_{x\in\mathcal{X}} \mathbb{E}[d(X,Y) \,|\, X = x] \le D. \tag{1.6}$$

This constraint provides a similar guarantee as the worst-case distortion constraint for lossy source coding with arbitrary source, in the sense that the expected distortion is small regardless of the source distribution P_X.[9] Nevertheless, channel simulation under the worst-case expected distortion constraint often requires a shorter description compared to lossy source coding with arbitrary source.

One example is the aforementioned universal quantization setting (Ziv, 1985; Zamir and Feder, 1992) with $X \in [0,t]$ for a moderately large t, where a quantization with step size Δ can be used to simulate the additive noise channel $Y = X + Z$, $Z \sim \mathrm{Unif}(-\Delta/2, \Delta/2)$, guaranteeing an average distortion $\Delta/4$ under the absolute error distortion $d(x,y) = |x-y|$, or an average distortion $\Delta^2/12$ under the squared error distortion $d(x,y) = (x-y)^2$, regardless of the source distribution. This is the same performance as lossy source coding with known source distribution $X \sim \mathrm{Unif}(0,t)$ via a fixed quantizer with step size Δ. In comparison, for lossy source coding with arbitrary source, we require a quantization step $\Delta/2$ to achieve the same worst-case distortion $\Delta/4$ under the absolute error distortion (requiring approximately twice as many quantization levels), or a quantization step $\Delta/\sqrt{3}$ to achieve the same worst-case distortion $\Delta^2/12$ under the squared error distortion (approximately $\sqrt{3}$ times as many quantization levels). In sum, lossy source coding with known source distribution gives a short description, whereas lossy source coding with arbitrary source gives a longer description but has a stronger guarantee. Channel simulation achieves the best of both worlds in this example since it has a description about as short as lossy source coding with known source distribution, and has a worst-case guarantee that holds for every source distribution.

Intuitively, the randomization in channel simulation allows us to "average over" different values of x in order to produce a compression uniformly good for all values of x, similar to how a mixed strategy provides a better worst-case expected payoff in game theory.

[9] We remark that (1.6) is weaker than (1.5) since (1.5) requires that the distortion is always small, whereas (1.6) only requires that the expectation of the distortion is small. If we cannot allow the distortion to be large due to some hard constraints of the system, then (1.5) is more suitable. However, if we are imposing a worst-case distortion constraint merely because the source distribution P_X is unknown, then (1.6) would also be reasonable.

1.6 Privacy

Local differential privacy. Suppose a user holds the data X, and wants to reveal some, but not too much information about X to an untrusted server. For example, if X represents the location of the user, he/she may want to convey an approximation location to the server in order to obtain some restaurant recommendations in the district, but not the precise address due to privacy concern (Andrés *et al.*, 2013). The user would apply a randomized algorithm, called a *privacy mechanism* (Dwork *et al.*, 2006; Dwork and Roth, 2014), on X to produce a noisy version Y, and send Y to the server instead. For instance, if $X \in \mathbb{R}$, one may apply an *additive noise mechanism* (Dwork and Roth, 2014) by adding a noise Z to produce $Y = X + Z$. Two popular examples are the Gaussian mechanism (where Z is a Gaussian random variable) and the Laplace mechanism (where Z is a Laplace random variable) (Dwork and Roth, 2014). Ignoring the computational aspect, a privacy mechanism can be regarded as a noisy channel $P_{Y|X}$, and additive noise mechanisms correspond to additive noise channels. A popular mathematical criterion for privacy is *local differential privacy* (Evfimievski *et al.*, 2003; Kasiviswanathan *et al.*, 2011).

Definition 1 ((ϵ, δ)-local differential privacy (Kasiviswanathan *et al.*, 2011))**.** We say that the conditional distribution $P_{Y|X}$ is (ϵ, δ)-*locally differentially private*, $\epsilon, \delta \geq 0$, if for every pair $x_1, x_2 \in \mathcal{X}$, and measurable set $\mathcal{S} \subseteq \mathcal{Y}$, we have

$$\mathbb{P}(Y \in \mathcal{S} \,|\, X = x_1) \leq e^{\epsilon} \cdot \mathbb{P}(Y \in \mathcal{S} \,|\, X = x_2) + \delta, \tag{1.7}$$

where Y follows $P_{Y|X}$ given X.

Intuitively, for two different values of the data x_1, x_2, the output Y given $X = x_1$ should have a similar distribution as the output Y given $X = x_2$, so it is hard for a party observing only Y to tell x_1 and x_2 apart. The parameter ϵ is called the *privacy budget*, whereas the parameter δ is called the *privacy leakage*. If $\delta = 0$, we omit δ and simply call (1.7) ϵ-*local differential privacy* (also referred to as *pure local differential privacy*, in contrast to the case $\delta > 0$ called *approximate local differential privacy*).

But how should the user send Y to the server, preferably using as few bits as possible? If Y is a real number, it cannot be compressed into finitely many bits. In this situation, a quantization is often performed on Y before transmission, which retains the privacy, but introduces an additional distortion, and may destroy the desirable statistical properties of Y.

Observant readers would notice that this is the problem channel simulation aims to solve. Channel simulation allows the user to send a description M of finitely many bits, for the server (which might share a common randomness W with the user) to be able to generate Y following the conditional distribution $P_{Y|X}$. For example, the *generic* 1-*bit*

protocol is a popular method for approximately simulating a privacy-preserving channel $P_{Y|X}$ using only *one* bit of communication (Bassily and Smith, 2015). Intuitively, a good privacy-preserving channel should be highly noisy, and a highly noisy channel with a small capacity costs a small number of bits to simulate. The goal of privacy (to ensure Y reveals little information about X) aligns nicely to the goal of channel simulation (to compress the description needed to simulate $P_{Y|X}$ as much as possible).

There is one caveat—channel simulation preserves the conditional distribution $P_{Y|X}$, but it might not retain the privacy property of $P_{Y|X}$. The server not only knows Y, but also the description M and the common randomness W which may reveal more information about X (Shah *et al.*, 2022; Shahmiri *et al.*, 2024). For example, one simple channel simulation scheme is to have the user send $M = X$ and have the server generate $Y|X \sim P_{Y|X}$ itself, revealing all information about X, and failing completely at privacy.

To solve this issue, we may design the channel simulation scheme with privacy taken into account, making sure that the conditional distribution $P_{W,M|X}$ from the data X to the decoder's observation (W, M) is differentially private as well, i.e., $P_{W,M|X}$ satisfies Definition 1. Refer to Figure 1.4. Examples include minimal random coding (Havasi *et al.*, 2019; Shah *et al.*, 2022) (Section 3.4),[10] generic 1-bit protocol (Bassily and Smith, 2015) (Section 3.2.3), local pseudo-randomizer (Feldman and Talwar, 2021) (Section 3.5.3), GenProt (Bun *et al.*, 2019) (Section 3.5.3), dyadic quantized Laplace mechanism (Shahmiri *et al.*, 2024) (Section 3.6.4), and Poisson private representation (Liu *et al.*, 2024) (Section 3.3.5).[11] These schemes (except dyadic quantized Laplace mechanism) are applicable to general privacy mechanisms $P_{Y|X}$ with a small privacy budget. Differentially-private channel simulation schemes that can be applied to general $P_{Y|X}$ are also called *generic transformation* in the differential privacy literature (Bassily and Smith, 2015; Bun *et al.*, 2019).

Distributed mean estimation and federated learning. We then discuss how channel simulation can be applied to distributed mean estimation (Duchi *et al.*, 2013). There are n users with data $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^d$, respectively, and would like to communicate to the server so as to allow it to compute an estimate $\hat{\boldsymbol{\mu}}$ of the mean $n^{-1} \sum_i \mathbf{X}_i$. There are two ways to impose privacy constraints—local differential privacy where the users do not want the server to know the $\mathbf{X}_i$'s precisely, and *central differential privacy* where the goal is only to disallow a party who observes $\hat{\boldsymbol{\mu}}$ to learn individual $\mathbf{X}_i$'s precisely (Dwork and Roth, 2014). A prominent application is *federated learning* (McMahan *et al.*, 2017; Kairouz *et al.*, 2021; Abadi *et al.*, 2016; McMahan *et al.*, 2018), where the users would like to allow the server to train a machine learning model using the users' data, but due to privacy concerns,

[10] Interestingly, minimal random coding is a channel simulation scheme with privacy guarantee, without the need of modifications.

[11] Poisson private representation is the only known exact channel simulation scheme for general privacy mechanisms with differential privacy guarantees.

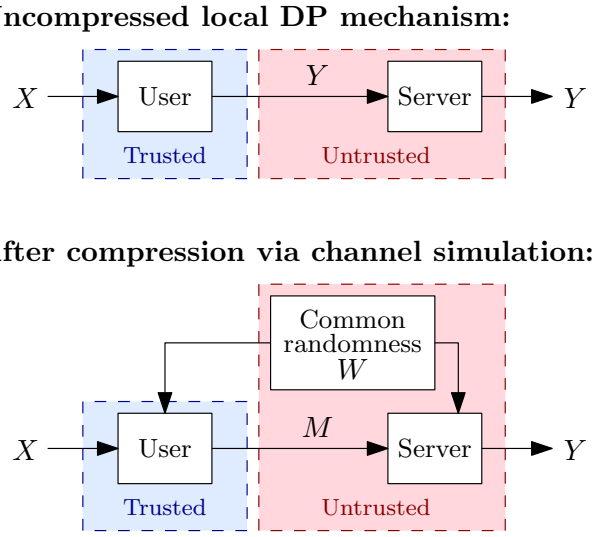


Figure 1.4: Top: An uncompressed local differential privacy mechanism, where the user applies the privacy mechanism $P_{Y|X}$ on the data X to produce Y and send it to the server. Privacy depends on the conditional distribution $P_{Y|X}$ from the data X to the observation Y of the untrusted server. Bottom: The mechanism after being simulated via channel simulation. Privacy depends on the conditional distribution $P_{W,M|X}$ (not $P_{Y|X}$) from the data X to the observation (W, M) of the untrusted server.

they do not want the server to learn too much about their data, and/or they do not want the resultant model to reveal the data of individual users. Channel simulation has been applied to distributed mean estimation and federated learning in order to compress the communication between the users and the server, while guaranteeing local and/or central differential privacy (Shah *et al.*, 2022; Lang *et al.*, 2023; Hasırcıoğlu and Gündüz, 2024; Hegazy *et al.*, 2024; Yan *et al.*, 2023; Liu *et al.*, 2024).

1.7 Distribution Preserving Quantization and Rate-Distortion-Perception Tradeoff

Lossy compression concerns the compression of a source X, such that the reconstruction $\hat{X}$ is close to X in the sense that the distortion $d(X, \hat{X})$ is small, where d is the distortion function. However, it has been noted by Li *et al.* (2011) and Blau and Michaeli (2018) that a small distortion does not equate a better perceptual quality of the reconstruction. In addition to having a small distortion, we should also ensure that the reconstruction $\hat{X}$ looks *natural*. In the works on distribution preserving quantization (Li *et al.*, 2010; Li *et al.*, 2011) and distribution preserving lossy compression (Tschannen *et al.*, 2018), it was argued that a natural-looking $\hat{X}$ should have the same distribution as the source X, so that it is

impossible for an observer to tell the output $\hat{X}$ apart from natural samples of X.

To capture the notion of perceptual quality, it was proposed by Blau and Michaeli (2018) and Blau and Michaeli (2019) that we should control the divergence $\delta(P_X, P_{\hat{X}})$ between the distribution of X and the distribution of $\hat{X}$, where δ is a divergence between two distributions (e.g., KL divergence, total variation distance or Wasserstein distance). A small $\delta(P_X, P_{\hat{X}})$ means it is hard to tell the output $\hat{X}$ apart from natural samples of X. If we require $\delta(P_X, P_{\hat{X}}) = 0$, it reduces to the aforementioned distribution preserving quantization setting.

Although conventional lossy compression schemes usually give a deterministic mapping from X to $\hat{X}$, we need to utilize randomness to guarantee perceptual quality. Consider an example where X is a random image belonging to one of two classes: cat images and dog images. A deterministic lossy compression scheme that compresses the image to one bit may map every cat image to the "average" of the cat images, and map every dog image to the "average" of the dog images. This gives a large divergence $\delta(P_X, P_{\hat{X}})$, and it would be easy to tell the outputs of the scheme apart from natural images by looking at their diversity (the outputs of the scheme will keep repeating the same two images). Furthermore, the "average" of the cat images might not even look like a natural cat image. On the other hand, if the encoder compresses an image to its class, and the decoder generates a random cat image upon receiving the class "cat", or generates a random dog image upon receiving the class "dog", then the output will have perfect perceptual quality. Therefore, the mapping from X to $\hat{X}$ needed for a good perceptual quality is generally not a deterministic mapping, but a noisy channel. The compression scheme to create the best noisy channel can be regarded as a channel simulation scheme.

To study the tradeoff between compression rate, distortion and perceptual quality, the rate-distortion function can be generalized to the following rate-distortion-perception function (Blau and Michaeli, 2019; Matsumoto, 2018; Matsumoto, 2019):

$$R(D, P) := \min_{P_{\hat{X}|X}: \mathbb{E}[d(X,\hat{X})] \le D,\, \delta(P_X, P_{\hat{X}}) \le P} I(X; \hat{X}).$$

One particularly interesting property is that if $X, \hat{X} \in \mathbb{R}^n$ and we are using the squared error distortion $d(x, \hat{x}) := \|x - \hat{x}\|^2$, then $R(D, 0) \le R(D/2, \infty)$, i.e., we can achieve perfect perceptual quality if we are willing to have a mean squared error twice as large (Blau and Michaeli, 2019).

Operationally, a one-shot lossy compression scheme with unlimited common randomness shared between the encoder and the decoder (Theis and Wagner, 2021) consists of a common randomness source $W \sim P_W$, a (possibly stochastic) encoder $P_{M|W,X}$ that emits the description M (in a discrete set, e.g., $M \in \mathbb{N}$) given W and the source symbol X, and a (possibly stochastic) decoder $P_{\hat{X}|W,M}$ that emits the reconstruction $\hat{X}$ given W, M. The goal is to minimize the conditional entropy $H(M|W)$ (which is approximately how many bits are

needed to compress M conditional on W), under the constraint that $\mathbb{E}[d(X, \hat{X})] \leq D$ and $\delta(P_X, P_{\hat{X}}) \leq P$. Invoking the channel simulation result (1.2) in (Li and El Gamal, 2018b), it has been shown in (Theis and Wagner, 2021) that there exists a scheme with

$$H(M|W) \leq R(D, P) + \log_2 \left(R(D, P) + 1\right) + 4,$$

and hence giving an operational meaning to the rate-distortion-perception function.

Interested readers are referred to (Zhang *et al.*, 2021; Chen *et al.*, 2022; Wagner, 2022) for further discussions on the rate-distortion-perception tradeoff, and (Saldi *et al.*, 2013; Saldi *et al.*, 2014; Saldi *et al.*, 2015) for generalizations of distribution preserving quantization, where the marginal distribution of the output is constrained to be another given distribution.

Another related line of research is *semantic communications* (Weaver, 1953; Bao *et al.*, 2011; Gündüz *et al.*, 2022; Shao *et al.*, 2022; Erdemir *et al.*, 2023), which concerns the transmission of the semantic information (loosely speaking, the meaning) of the source. Channel simulation has been applied to semantic communications in (Gündüz *et al.*, 2022; Pase *et al.*, 2023).

1.8 Simulating Quantum Measurements via Classical Communication

Quantum mechanics presents a different view to particles and information compared to classical physics. But *how different* are they? Are they different in the same sense of how an object's length in meters is different from its length in feet (technically different numbers, but effortlessly convertible)? Or are they really fundamentally different, with conversion between them being generally impossible? To investigate the difference between classical and quantum physics, we can study how classical resources can be converted to quantum resources and vice versa.

A *quantum bit* (*qubit*) is a unit of quantum information, referring to a system with two orthonormal basis states $|0\rangle$ and $|1\rangle$. Unlike a classical bit which can be either 0 or 1, a (pure state) qubit is in the form $\alpha|0\rangle + \beta|1\rangle$, where $\alpha, \beta \in \mathbb{C}$ with $|\alpha|^2 + |\beta|^2 = 1$.[12] This extra generality suggests that a qubit may be "worth more than" a classical bit. How many classical bits is a qubit worth? Quantum teleportation (Bennett *et al.*, 1993) shows that Alice can transfer a qubit of information to Bob by transmitting two classical bits (assuming Alice and Bob have prior shared entanglement). On the other hand, superdense coding (Bennett and Wiesner, 1992) demonstrates that Alice can transmit two classical bits

[12] Generally, a pure state of n entangled qubits is in the form $\sum_{i \in \{0,1\}^n} \alpha_i |i\rangle$ where $\alpha_i \in \mathbb{C}$ with $\sum_{i \in \{0,1\}^n} |\alpha_i| = 1$.

to Bob by sending one qubit (assuming Alice and Bob have prior shared entanglement). This seems to suggest a "1 qubit = 2 classical bits" conversion rate.[13]

Nevertheless, one cannot simply treat one qubit as two classical bits (or else quantum information theory would become so much easier). Quantum teleportation and superdense coding require prior shared entanglement in the form of the *Bell state* (Bell, 1964): Alice and Bob each holds one qubit, where the two qubits are entangled with a state $|\Phi^+\rangle = (1/\sqrt{2})(|00\rangle + |11\rangle)$. Shared entanglement is very different from shared classical information. Bell's theorem (Bell, 1964) shows that no amount of classical shared randomness (i.e., sharing a classical random variable W among Alice and Bob; also called *local hidden variables*) can simulate the Bell state. No conversion in the form "1 pair of entangled qubits $= n$ shared classical random bits" is possible. This is a fundamental result in quantum mechanics, which opens up the avenue of experimentally verifying, through conducting measurements on Bell states, that we are indeed living in a quantum reality.

We briefly describe Bell's experiment. Assume Alice and Bob have a pair of entangled qubits with the Bell state. Alice and Bob will measure their qubits with respect to the operators $R(x)$ and $R(y)$ respectively ($x, y \in [0, 2\pi)$), where $R(x)$ denotes the operator for the von Neumann measurement

$$R(x) := \begin{pmatrix} \cos x & \sin x \\ \sin x & -\cos x \end{pmatrix}.$$

Let $A, B \in \{-1, 1\}$ be the outcomes of Alice's and Bob's measurements respectively. The joint probability matrix of A, B is given by (Brassard *et al.*, 1999)

$$\begin{pmatrix} \frac{1}{2}\cos^2(\frac{x-y}{2}) & \frac{1}{2}\sin^2(\frac{x-y}{2}) \\ \frac{1}{2}\sin^2(\frac{x-y}{2}) & \frac{1}{2}\cos^2(\frac{x-y}{2}) \end{pmatrix}. \tag{1.8}$$

Bell's theorem (Bell, 1964) states that there does not exist a classical random variable W and functions $A = f(W, x)$ and $B = g(W, y)$ such that the joint distribution of A, B is (1.8) for every x, y.

While Bell's theorem states that it is impossible to simulate the Bell state via classical shared randomness, it does not forbid the simulation of the Bell state via classical *communication*. If we also allow Alice to send a classical message M to Bob that may depend on her measurement x, i.e., $(A, M) = f(W, x)$ and $B = g(W, M, y)$, then it is possible to ensure that the joint distribution of A, B is (1.8) for every x, y (Maudlin, 1992; Brassard *et al.*, 1999).

Characterizing how much communication and common randomness is needed to simulate the Bell state can shed light on the gap between quantum and classical information. To this end, various protocols have been proposed. If we allow Alice and Bob to share unlimited

[13] Without prior shared entanglement, a qubit can only convey one bit of classical information (Holevo, 1973; Frenkel and Weiner, 2015).

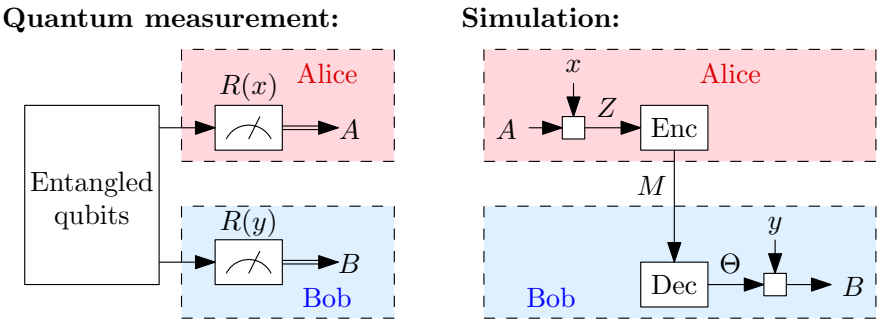


Figure 1.5: Left: Measurement of entangled qubits. Right: Simulating the measurement via channel simulation applied on the channel $f_{\Theta|Z}$ (1.9).

classical common randomness / local hidden variables W before the protocol commences, and Alice can send a fixed number of bits to Bob, then a protocol using 4 bits was given by Brassard *et al.* (1999),[14] and a protocol using 3 bits was given by Csirik (2002). If we allow unlimited common randomness, and Alice can send a variable number of bits to Bob, then a protocol using an expected 1.485 bits was given by Steiner (2000), and a protocol using an expected 1.19 bits was given by Cerf *et al.* (2000).[15] For the situation without common randomness W between Alice and Bob, Massar *et al.* (2001) gave a protocol with expected ≤ 11 bits of interactive communications;[16] and Li and El Gamal (2018a) gave a protocol with expected ≤ 9 bits of one-way communication.

The protocols in (Steiner, 2000; Li and El Gamal, 2018a) are based on the following observation by Feldmann (1995). Assume there is a noisy channel $P_{\Theta|Z}$ ($\Theta, Z \in [0, 2\pi)$) from Alice to Bob with a conditional density function

$$f_{\Theta|Z}(\theta|z) := \frac{1}{2} \max\left\{\cos(z-\theta),\, 0\right\}. \tag{1.9}$$

Alice generates $A \sim \mathrm{Unif}(\{\pm 1\})$ and sends Z through the channel, where $Z = x$ if $A = 1$, or $Z = x + \pi \bmod 2\pi$ if $A = -1$. Bob receives Θ and outputs $B = \mathrm{sgn}(\cos(y - \Theta))$. It can be checked that (A, B) follows the correct joint distribution in (1.8). Therefore, any channel simulation scheme for the channel $P_{\Theta|Z}$ can give a protocol for simulating the Bell state. For example, Steiner (2000) applied the rejection sampling scheme for channel simulation (see Section 3.2), whereas Li and El Gamal (2018a) applied the dyadic decomposition scheme (see Section 4.2). Refer to Figure 1.5 for an illustration.

It is also of interest to generalize this setting to the simulation of a system of n Bell states. In this case, it is insufficient to simply run the protocol n times, and an exponential

[14] Brassard *et al.* (1999) also gave a protocol using 8 bits for general von Neumann measurements, not only those parametrized by $x \in [0, 2\pi)$.

[15] The bound 1.19 by Cerf *et al.* (2000) applies also to general von Neumann measurements.

[16] Massar *et al.* (2001) also gave a bound ≤ 20 for general positive-operator-valued measurements.

amount of communication is needed (Brassard *et al.*, 1999; Massar *et al.*, 2001).

Another setting is to simulate a measurement on a quantum system held by Alice, where the measurement result is given to Bob. We would like to simulate this measurement by allowing Alice and Bob to share common randomness, and allowing Alice to perform classical communication to Bob. It has been shown by Berta *et al.* (2014) that the asymptotic amount of communication needed can be given in terms of the quantum mutual information.

1.9 Communication Complexity and Message Compression

Consider a two-party interactive protocol for computing a function (Yao, 1979). Alice holds a random variable X, and Bob holds another independent random variable Y. They are allowed k rounds of interactive communications, i.e., Alice sends message M_1 (as a function of X) to Bob, and then Bob sends message M_2 (as a function of Y, M_1) to Alice, and then Alice sends message M_3 (as a function of X, M_1, M_2) to Bob, and so on up to M_k. Their goal is to allow Alice and Bob to compute the function $f(X, Y)$ at the end of the protocol with an error probability at most ϵ, using the smallest amount of interactive communications $M^k = (M_1, \ldots, M_k)$. The smallest amount of communications is called the (deterministic ϵ-error) *communication complexity* of f, which we denote as $C_\epsilon(f)$.[17] Channel simulation (usually referred to as *message compression* in the communication complexity literature) can be applied to compress the interactive communications to an amount approximately given by the mutual information, stripping away unnecessary parts of the message (Jain *et al.*, 2003; Harsha *et al.*, 2010; Barak *et al.*, 2010). We briefly describe the idea below.

A central problem in communication complexity is the *direct sum problem* (Chakrabarti *et al.*, 2001), which asks whether computing n copies of f requires $\Theta(n)$ times the amount of communication needed for computing one copy of f. Suppose Alice holds an i.i.d. sequence $X^n = (X_1, \ldots, X_n)$, and Bob holds an independent i.i.d. sequence $Y^n = (Y_1, \ldots, Y_n)$. The direct sum problem asks whether the communication complexity for computing $f(X_1, Y_1), \ldots, f(X_n, Y_n)$ with an error probability at most ϵ, denoted as $C_\epsilon(f^n)$, has to be $\Theta(n)$ times the communication complexity of computing one copy $f(X, Y)$, i.e., whether $C_\epsilon(f^n) = \Theta(nC_\epsilon(f))$.

To prove a direct sum result, one would start with the optimal protocol for computing n copies of f with an amount of communications $C_\epsilon(f^n)$, and construct a new protocol that computes only one copy $f(X, Y)$ when Alice and Bob hold X and Y respectively (Chakrabarti *et al.*, 2001; Jain *et al.*, 2003; Harsha *et al.*, 2010). If we can construct such a

[17] We take $C_\epsilon(f)$ to be worst-case (over values of X and Y) total length of $M_1, \ldots, M_k$, which are variable-length codewords, under the constraint that Alice and Bob can output $f(X, Y)$ with an error probability at most ϵ (averaged over the distribution of X, Y). Here $C_\epsilon(f)$ is only meant to represent the approximate amount of communication, and hence we will omit the precise definition.

new protocol that only uses an amount of communications $n^{-1}C_\epsilon(f^n)$, then we have shown that simulating one copy of f cannot be less efficient than simulating n copies. One way to construct this new protocol is to have Alice take $X_i = X$ (where i is a fixed index) and generate the other entries $(X_j)_{j\neq i}$ i.i.d. at random, Bob take $Y_i = Y$ and generate $(Y_j)_{j\neq i}$ i.i.d. at random, apply the optimal n-fold protocol to compute $f(X_1, Y_1), \dots, f(X_n, Y_n)$, and use the i-th entry $f(X_i, Y_i) = f(X, Y)$ only. However, this will only construct a 1-fold protocol with an amount of communication the same as the n-fold protocol, and show the obvious fact that $C_\epsilon(f) \leq C_\epsilon(f^n)$. In order to reduce the amount of communication, note that a lot of the communications $M_1, M_2, \dots$ are for the calculation of irrelevant entries $(f(X_j, Y_j))_{j\neq i}$, and we should strip away the irrelevant parts of $M_1, \dots, M_k$ and retain only the parts relevant to (X_i, Y_i). Using channel simulation (1.2) on the channel $X_i \to M_1$, Alice can compress the message M_1 into a new message V_1 of expected length $\approx I(X_i; M_1)$, so that Bob can recover M_1 using V_1. Technically, Bob does not recover M_1, but another random variable $\tilde{M}_1$ that has the same conditional distribution given X_i as M_1 (i.e., $P_{\tilde{M}_1|X_i} = P_{M_1|X_i}$), which is all that matters since we are not trying to preserve the information in M_1 about $(X_j)_{j\neq i}$ anyway. Next, using channel simulation (1.2) on the channel $Y_i \to M_2$ conditional on M_1 (i.e., applying the channel simulation result on the conditional distribution $P_{Y_i,M_2|M_1=m_1}$ if $M_1 = m_1$), Bob can compress M_2 into a new message V_2 of expected length $\approx I(Y_i; M_2|M_1)$. Refer to Figure 1.6 for an illustration.

Repeating this argument for each M_a, the expected total length would be

$$\begin{aligned}
&\approx I(X_i; M_1) + I(Y_i; M_2|M_1) + I(X_i; M_3|M^2) + \cdots \\
&\leq \sum_{a=1}^{k} I(X_i, Y_i; M_a|M^{a-1}) \\
&= I(X_i, Y_i; M^k).
\end{aligned}$$

Since

$$\sum_{i=1}^{n} I(X_i, Y_i; M^k) \leq I(X^n, Y^n; M^k) \leq H(M^k) \leq C_\epsilon(f^n),$$

there must exist an i such that the expected total length of the protocol is $\lesssim n^{-1}C_\epsilon(f^n)$. Hence, we have obtained a 1-fold protocol with expected total length approximately upper-bounded by n^{-1} times the total length of the n-fold protocol. This is the main idea of the direct sum result by Harsha *et al.* (2010) (also see (Chakrabarti *et al.*, 2001; Jain *et al.*, 2003)).[18] Interested readers are referred to (Barak *et al.*, 2010; Braverman and Garg, 2014; Brody *et al.*, 2016; Rao and Yehudayoff, 2020) for more discussions on message compression in communication complexity settings.

[18] Note that $C_\epsilon(f)$ is the worst-case total length instead of the expected total length, and hence Markov's inequality is needed in (Harsha *et al.*, 2010) to bound the worst-case total length.

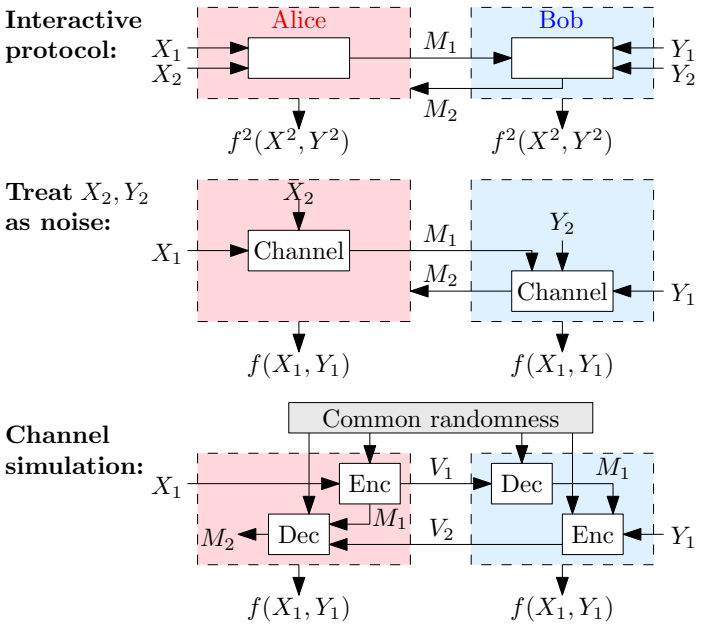


Figure 1.6: Top: The interactive protocol for computing two copies of f on X_1, X_2 and Y_1, Y_2 with $k = 2$ rounds. Middle: If we are only interested in computing $f(X_1, Y_1)$, Alice can treat X_2 as a noise to the channel $P_{M_1|X_1}$, and Bob can treat Y_2 as a noise to the channel $P_{M_2|Y_1,M_1}$ (where M_1 is a state known to all parties). Bottom: Apply channel simulation on these channels to compress the messages to V_1, V_2.

1.10 Coordination Problems

Consider the task assignment problem in (Cuff *et al.*, 2010) described as follows. There are k tasks labeled $1, \dots, k$. In each time slot $i = 1, \dots, n$, one uniformly randomly chosen task $X_i \sim \text{Unif}(\{1, \dots, k\})$ is assigned to Alice. After observing $X_1, \dots, X_n$, Alice sends a message $M \in \{1, \dots, \lfloor 2^{nR} \rfloor\}$ at a rate R bits per time slot to Bob, allowing Bob to decide on a sequence of tasks $Y_1, \dots, Y_n$, such that Alice and Bob will not work on the same task at the same time, i.e., $X_i \neq Y_i$ for all i. One strategy is to take Y_i to be the task after X_i in the cyclic order (i.e., $Y_i = (X_i \bmod k) + 1$), which requires a communication rate $R > \log_2 k$ for large n. Another strategy is to take $Y_i = 1$ if $X_i \neq 1$, or $Y_i = 2$ if $X_i = 1$, which requires $R > \frac{1}{k} \log_2 k + \frac{k-1}{k} \log_2 \frac{k}{k-1}$.

The optimal strategy (Cuff *et al.*, 2010) is not to choose Y_i as a function of X_i, but to make $X_i \to Y_i$ a noisy channel. Applying channel simulation on the channel $P_{Y|X}(y|x) = 0$ if $y = x$, or $P_{Y|X}(y|x) = 1/(k-1)$ if $y \neq x$, we have a scheme that requires a rate $R > I(X;Y) = \log_2 \frac{k}{k-1}$. This scheme also has the benefit that (X_i, Y_i) is uniformly distributed over $\{(x, y) \in \{1, \dots, k\}^2 : x \neq y\}$, i.i.d. over $i = 1, \dots, n$. As a result, the empirical joint distribution of $(X_i, Y_i)_{i=1,\dots,n}$ tends to $\text{Unif}(\{(x, y) \in \{1, \dots, k\}^2 : x \neq y\})$ as $n \to \infty$, ensuring that the tasks are evenly assigned.

This setting is an example of the general study on *coordination capacity* (Cover and Permuter, 2007; Cuff *et al.*, 2010), which concerns the amount of communication needed for the nodes in a network to simulate the desired joint distribution. There are two kinds of distribution requirements. First, *strong coordination* requires the joint distribution of the outputs of the nodes to approach the prescribed distribution (e.g., $(X_i, Y_i) \sim \text{Unif}(\cdots)$ i.i.d. in the task assignment problem). Second, *empirical coordination* only requires the empirical joint distribution of the output sequences to approach the prescribed distribution (e.g., the empirical joint distribution of $(X_i, Y_i)_{i=1,\dots,n}$ tends to $\text{Unif}(\cdots)$ in the task assignment problem).

Coordination settings can also be found in cooperative game theory, where a team of players can cooperate through rate-limited communications in order to decide on their actions, with a goal of maximizing their expected payoff against their opponent's action (Cuff, 2008). In game theory, the optimal minimax strategy is often a mixed strategy where players select their actions at random. Hence, the players should not waste the communication bandwidth to divulge their actions noiselessly, but should instead convey the minimal amount of information necessary to establish the desired joint distribution given by the mixed strategy. Empirical coordination is insufficient against an opponent who knows the strategy,[19] and strong coordination is necessary to ensure the worst-case payoff. Refer

[19] For example, the optimal mixed strategy for rock-paper-scissors (numbered $1, 2, 3$) is $\text{Unif}(\{1, 2, 3\})$. The strategy $X_1 = 1$, $X_2 = 2$, $X_3 = 3$ satisfies the empirical distribution constraint, but will lose every match against an opponent who knows the strategy. We need $X_1, X_2, \dots$ to be truly random.

to (Gossner *et al.*, 2006; Anantharam and Borkar, 2007; Cuff, 2013; Satpathy and Cuff, 2014; Le Treust and Tomala, 2018; Li and El Gamal, 2018a) for studies on coordination strategies for cooperative games.

This monograph focuses almost exclusively on strong coordination, which implies empirical coordination. For channel simulation under the empirical coordination constraint, readers are referred to the work on the joint empirical distribution between the source and reconstruction sequences in lossy source coding in (Kramer and Savari, 2007; Weissman and Ordentlich, 2005).

1.11 Other Applications

For applications to statistical learning, techniques in channel simulation (more specifically, the rejection sampling scheme (Harsha *et al.*, 2010; Braverman and Garg, 2014)) have been employed to study learners that use a limited amount of information from the sample (Bassily *et al.*, 2018). Channel simulation is applied to a minimax learning setting with limited communication in (Li *et al.*, 2018). The application of channel simulation via rejection sampling to smoothed online learning was investigated by Block and Polyanskiy (2023). Channel simulation was used by Sefidgaran *et al.* (2024) to reduce the communication cost of statistical learning.

Channel simulation is related to *coupling from the past* (Propp and Wilson, 1996; Propp and Wilson, 1998), a method for the exact sampling from the stationary distribution of a Markov chain. The *layered multishift coupler* (Wilson, 2000), one of the earliest channel simulation schemes for additive noise channels, was initially conceived as a technique for coupling from the past. Refer to (Hegazy and Li, 2022) for discussions on the connections between coupling from the past and channel simulation.

1.12 Preliminaries—Prefix-free Codes

In this section, we review some basic concepts about fixed-length and variable-length codes, which are needed for the encoding of the description in channel simulation. For source coding or channel simulation settings where the encoder can send a description M to the decoder, one may impose a strict limit on the length of the description, resulting in a *fixed-length* description. If we require that the description M must fit within ℓ bits, then we would restrict $M \in \{0,1\}^\ell$ or $M \in \{1, \dots, 2^\ell\}$. For one-shot settings, we often allow a slightly finer granularlity and impose that the cardinality of the description does not exceed a fixed limit N, i.e., $M \in \{1, \dots, \mathsf{N}\}$, where N does not need to be a power of 2.

Alternatively, one may allow the description to be *variable-length*. Here we restrict attention to *prefix-free codes* (Cover and Thomas, 2006). Write $\{0,1\}^* = \bigcup_{\ell=0}^{\infty} \{0,1\}^\ell$ for the

set of bit sequences of any length. A codebook $\mathcal{C} \subseteq \{0,1\}^*$ satisfies the *prefix-free condition* if $a, b \in \mathcal{C}$ and $a \neq b$ implies a is not a prefix of b. We require M to be an element in a fixed prefix-free codebook $\mathcal{C}$. Prefix-free codes are also called *instantaneous codes* (Cover and Thomas, 2006) since the decoder can read the bits of M one by one, and stop reading when the sequence of bits read so far is a codeword in $\mathcal{C}$, without the need of looking into the future bits since there cannot be another longer codeword. This property makes prefix-free codes useful for the design of decoding algorithms and communication protocols. Popular prefix-free codes include the Huffman code (Huffman, 1952), the Shannon code (Shannon, 1948), and Elias' codes for integers (Elias, 1975).

The performance metric of a fixed-length code is the fixed length ℓ or the cardinality N, whereas the performance metric of a variable-length code is the expected description length $\mathbb{E}[|M|]$. The use of the expected length is justified by the law of large numbers. If we repeat n times a scheme with expected length r, the total length will be close to nr with high probability, and we can truncate or pad it to a $(n + o(n))r$-bit fixed-length code with low error probability.

The same also holds when we are concatenating different variable-length schemes. The prefix-free condition ensures that we can concatenate two (possibly different) prefix-free codebooks $\mathcal{C}_1, \mathcal{C}_2$, and the resultant codebook $\mathcal{C} = \{a\|b : a \in \mathcal{C}_1, b \in \mathcal{C}_2\}$ will still be prefix-free.[20] If we concatenate n prefix-free codewords of expected lengths $r_1, \dots, r_n$, then the total length will likely be close to $\sum_{i=1}^{n} r_i$. Therefore, there is no need for every component in a file format, or every packet in a communication protocol, to have a fixed limit on the length. As long as each component has a small expected length, the total memory usage or traffic over the network will average out due to the law of large numbers.

Examples of prefix-free codes include the Shannon code (Shannon, 1948) and the Huffman code (Huffman, 1952), both of which encode a random variable X into a prefix-free codeword $f(X) \in \mathcal{C}$ with an expected length $\mathbb{E}[|f(X)|]$ that satisfies

$$H(X) \leq \mathbb{E}[|f(X)|] \leq H(X) + 1 \text{ bit}.$$

This provides a convenient way to bound the expected length, showing that it is often good enough to bound the entropy $H(M)$ instead of $\mathbb{E}[|M|]$ since it is off by at most 1 bit.

The *Shannon code* (Shannon, 1948), also known as the *Shannon-Fano code* due to a related construction by Fano (1949), encodes x into a codeword of length $|f(x)| = \lceil -\log_2 P_X(x) \rceil$, and generally does not attain the smallest possible $\mathbb{E}[|f(X)|]$. Its advantage is that the expected length can be bounded even when the distribution of the actual input deviates from the distribution P_X that we design the code for. When $X \sim Q_X$ follows another distribution, we have

$$\mathbb{E}_{X \sim Q_X}[|f(X)|] \leq H(Q_X, P_X) + 1 \text{ bit},$$

[20] The same may not hold for other kinds of variable-length codes, such as uniquely decodable codes in general (Cover and Thomas, 2006).

where $H(Q_X, P_X) := -\sum_{x \in \mathcal{X}} Q_X(x) \log_2 P_X(x)$ is the cross entropy.

The Huffman code (Huffman, 1952) is proved to achieve the minimum $\mathbb{E}[|f(X)|]$, making it the optimal choice when the distribution of X is known. Nevertheless, it does not generally satisfy $|f(x)| \leq -\log_2 P_X(x) + 1$,[21] and hence does not guarantee that $\mathbb{E}_{X \sim Q_X}[|f(X)|] \leq H(Q_X, P_X) + 1$. Therefore, Huffman code is less suitable for theoretical guarantees when the precise actual distribution is unknown.

Prefix-free codebooks are not only useful for encoding descriptions sent by an encoder, but also for generating random variates, for example, via the *discrete distribution-generating (DDG) tree* (Knuth and Yao, 1976). This will be elaborated in Section 9.1.1.

[21] An example (taken from (Cover and Thomas, 2006) with a slight variation) is $P_X(1) = 0.35$, $P_X(2) = 0.33$, $P_X(3) = 0.31$, $P_X(4) = 0.01$. The Huffman code gives $|f(1)| = 1$, $|f(2)| = 2$, $|f(3)| = |f(4)| = 3$. We have $|f(3)| > -\log_2 P_X(3) + 1$.

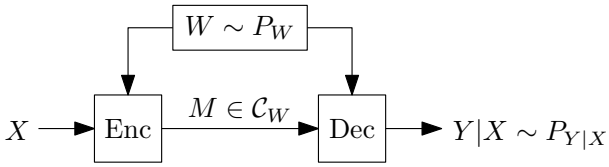


Figure 2.1: One-shot channel simulation with common randomness.

2 The Channel Simulation Setting

2.1 Definition and Notations

We start with the one-shot channel simulation setting with unlimited common randomness depicted in Figure 2.1. We first give an informal description, and then state a more precise definition later in Definition 2. Suppose we want to simulate the channel $P_{Y|X}$ with input symbol $X \in \mathcal{X}$. There is a common random source $W \sim P_W$ available to the encoder and the decoder. The encoder observes W and the input $X \sim P_X$, and sends a description M to the decoder. The decoder uses W and M to produce Y. The goal is to have a Y that follows the prescribed distribution $P_{Y|X}$ given X, while minimizing the expected amount of communication.

There are two options for the assumptions. If we assume a *known source distribution* (KS), we assume $X \sim P_X$ for a known P_X, and we are allowed to design the scheme according to P_X. The scheme should have a short expected description length averaged over $X \sim P_X$. If we assume *arbitrary source* (AS), then no P_X is given, and the scheme needs to have a short expected description length for every $X \in \mathcal{X}$. We now define the setting more precisely.

Definition 2 (One-shot variable-length exact channel simulation with unlimited common randomness)**.** Consider a general (discrete/continuous) channel $P_{Y|X}$ and a general source distribution P_X if the source distribution is known. A one-shot variable-length channel simulation scheme with unlimited common randomness is characterized by a tuple $(P_W, (\mathcal{C}_w)_{w\in\mathcal{W}}, P_{M|W,X}, P_{Y|W,M})$ described below:

- **Common randomness.** There is a common random source $W \in \mathcal{W}$, $W \sim P_W$ available to the encoder and the decoder, where we can choose an arbitrary distribution P_W as a part of the coding scheme.

- **Codebook.** Let $(\mathcal{C}_w)_{w\in\mathcal{W}}$ be a family of prefix-free codebook which we can design as a part of the coding scheme, i.e., each $\mathcal{C}_w \subseteq \{0,1\}^*$ is a prefix-free codebook (see Section 1.12).

- **Encoder.** The encoder observes W and a source symbol $X \in \mathcal{X}$ ($X \sim P_X$ for known source distribution), and sends a description $M \in \mathcal{C}_W$ produced by passing W, X through a conditional distribution $P_{M|W,X}$ from $\mathcal{W} \times \mathcal{X}$ to $\{0,1\}^*$ (called the *encoding Markov kernel*). We require that $M \in \mathcal{C}_W$ with probability 1. The encoding Markov kernel $P_{M|W,X}$ represents a stochastic (randomized) encoding function. In case if the encoding is deterministic, we can have $M = f(W, X)$ where $f : \mathcal{W} \times \mathcal{X} \to \{0,1\}^*$ is the encoding function.

- **Decoder.** The decoder then outputs $Y \in \mathcal{Y}$ produced by passing W, M through a conditional distribution $P_{Y|W,M}$ from $\mathcal{W} \times \{0,1\}^*$ to $\mathcal{Y}$ (called the *decoding Markov kernel*). The decoding Markov kernel $P_{Y|W,M}$ represents a stochastic decoding function. In case if the decoding is deterministic, we can have $Y = g(W, M)$ where $g : \mathcal{W} \times \{0,1\}^* \to \mathcal{Y}$ is the decoding function.

- **Requirement.** We require $Y|X \sim P_{Y|X}$ exactly.[1]

- **Performance metric.**

  - For known source distribution, we are interested in the smallest expected length $\mathbb{E}[|M|]$ of the prefix-free description M. Let

    $$L^* := \inf \mathbb{E}\big[|M|\big]$$

    be the infimum of the expected lengths among all schemes satisfying the requirement $Y|X \sim P_{Y|X}$ ($L^* = \infty$ if there is no such scheme).

  - For arbitrary source, we are interested in the worst-case expected length $\sup_{x \in \mathcal{X}} \mathbb{E}[|M| \mid X = x]$. Now, the quantity of interest L^* would be the infimum of the set of achievable worst-case expected lengths among all schemes satisfying the requirement, i.e.,

    $$L^* = \inf \sup_{x \in \mathcal{X}} \mathbb{E}\big[|M| \,\big|\, X = x\big]. \tag{2.1}$$

    be the infimum of the expected lengths among all schemes satisfying the requirement $Y|\{X = x\} \sim P_{Y|X}(\cdot|x)$ for all $x \in \mathcal{X}$ ($L^* = \infty$ if there is no such scheme).

[1] This can be relaxed in approximate settings, where Y only follows $P_{Y|X}$ approximately.

The requirement that M is a prefix-free codeword in $\mathcal{C}_W$ ensures that the encoder can send M through a noiseless bit channel, and the decoder would know the position where M ends without reading any bits further than M, and hence the bit channel can be reused for other purposes. This is because both the encoder and the decoder know W, they know which codebook $\mathcal{C}_W$ to use, and hence they are able to synchronize and know where M ends. Nevertheless, most schemes (e.g., the construction used in Theorem 4 later) require only a single codebook $\mathcal{C}$ that does not depend on W. We allow the flexibility of choosing different codebooks for different values of W in the problem setting only for the sake of full generality. The synchronization of the description M and the common randomness W will be discussed in detail in Section 2.3.

We describe two alternative ways to describe the channel simulation setting.

Ensemble of lossy compression schemes. Another way of understanding Definition 2 is to think of a channel simulation scheme as an ensemble of lossy compression schemes indexed by $w \in \mathcal{W}$ (Li *et al.*, 2011). For each w, we have an encoding function $x \mapsto f(w, x) \in \mathcal{C}_w$ (assuming a deterministic encoding function), and a decoding function $m \mapsto g(w, m)$ (assuming a deterministic decoding function). Each pair of encoding and decoding function can have their own prefix-free codebook $\mathcal{C}_w$. The requirement is that the conditional distribution of the output Y given the input X, when averaged over the ensemble $W \sim P_W$, is the prescribed channel $P_{Y|X}$.

Remote generation. Yet another way to interpret Definition 2 is to consider the following remote generation/sampling problem. The encoder and the decoder share a common random source $W \sim P_W$. The encoder observes a distribution P in a class of distributions $\mathcal{P}$, and sends a description $M \in \mathcal{C}_W$ to the decoder (who knows the class $\mathcal{P}$ but not the precise distribution P prior to receiving M), in order to allow the decoder to produce a sample $Y \sim P$. Unlike the usual sampling (or random number generation) setting where a single party knows the distribution and produces a sample from that distribution, here the sampling process is distributed among two parties—the encoder who knows the distribution, and the decoder who must produce a sample. The channel simulation setting in Definition 2 can be regarded as a remote generation setting with $\mathcal{P} = \{P_{Y|X}(\cdot|x)\}_{x\in\mathcal{X}}$, and $P = P_{Y|X}(\cdot|X)$, so having $Y \sim P_{Y|X}(\cdot|X)$ would mean that the channel $P_{Y|X}$ is simulated successfully.

The setting in Definition 2 is very general, and many other settings in this monograph are special cases of this setting where additional constraints are imposed. We now describe several categories of channel simulation settings and constraints (with their abbreviations in parentheses).

Discrete (D), Continuous output (C) or General (G) channels. The default assumption is that $P_{Y|X}$ is a general channel defined in an arbitrary Polish space. Some constructions in this monograph only applies to discrete channels $P_{Y|X}$, where X and Y lie in finite discrete spaces. Some other constructions only applies to channels with continuous output, i.e., $P_{Y|X}(\cdot|x)$ is a continuous distribution for every x. There are also constructions that work specifically for additive continuous noise channels, i.e., $Y = X + Z$ where Z is a continuous noise. The 1D case ($X, Y, Z \in \mathbb{R}$) is denoted as "1DAC", whereas the n-D case ($\mathbf{X}, \mathbf{Y}, \mathbf{Z} \in \mathbb{R}^n$) is denoted as "$n$DAC".

One-shot (1), Finite-blocklength (F) or Asymptotic (∞). In Definition 2, the encoder only observes one symbol X, and the decoder only outputs one symbol Y. Such a setting is referred to as *one-shot*. However, one should not treat X as just one bit or one small input symbol to a channel. Instead, X, Y can be *anything*, from numbers to sequences to images to any random object one can define. This is why one-shot is regarded as the most general setting. Nevertheless, there are situations where one may want to restrict attention to the case where $X = (X_1, \ldots, X_n)$ is an i.i.d. sequence, $Y = (Y_1, \ldots, Y_n)$ is also a sequence, and the channel to be simulated $P(y_1, \ldots, y_n|x_1, \ldots, x_n) = \prod_{i=1}^n P(y_i|x_i)$ is memoryless. This is referred to as *finite-blocklength* when n is finite, and *asymptotic* when $n \to \infty$. The benefit of the asymptotic setting is that one can often apply the law of large numbers to obtain simpler results (e.g., the reverse Shannon theorem (Bennett *et al.*, 2002) stated in terms of the channel capacity). The asymptotic setting will be discussed in Sections 5, 6 and 7.

Approximate (A) or Exact (E). In Definition 2, we require Y to follow the conditional distribution $P_{Y|X}$ exactly. While we focus on the exact setting in this section, it is also possible to relax this condition, and only require Y to approximately follow $P_{Y|X}$ under a certain metric. One-shot approximate settings (e.g. likelihood encoders (Cuff, 2013; Watanabe *et al.*, 2015; Song *et al.*, 2016) and minimal random coding (Havasi *et al.*, 2019; Flamich *et al.*, 2020)) will be discussed in Section 3.4 and Section 8, and the asymptotic fixed-length approximate setting will be discussed in Sections 5, 6 and 7.

Fixed-length (FL) or Variable-length (VL) description. While Definition 2 allows a *variable-length* description $M \in \mathcal{C}_W \subseteq \{0,1\}^*$. One may also require a *fixed-length* description. If we require that the description M must fit within ℓ bits, then we would restrict $M \in \mathcal{C} = \{0,1\}^\ell$. For one-shot settings, we often allow a slightly finer granularlity and impose that the cardinality of the description does not exceed a fixed limit N, i.e., $M \in \{1, \ldots, \mathsf{N}\}$, where N does not need to be a power of 2. However, requiring a fixed-length description often necessitates relaxing the distribution constraint on Y to an approximate

constraint. The one-shot approximate fixed-length setting will be discussed in Section 8, and the asymptotic approximate fixed-length setting will be discussed in Sections 5, 7. It is possible to have fixed-length description with exact distribution constraint, which will be briefly discussed in Section 3.7, though it does not appear to yield elegant results, and may result in a significant penalty on the description length.

Known source distribution (KS) or Arbitrary source (AS). If the source distribution $X \sim P_X$ is known, then we can design the scheme according to P_X, and we are interested in the expected length $\mathbb{E}[|M|]$. If the source distribution is not known, then we cannot design the scheme according to P_X, and we are interested in the worst-case expected length $\mathbb{E}[|M|]$ among all input distribution P_X. It is straightforward to show that the worst-case expected length is $\sup_{x\in\mathcal{X}} \mathbb{E}[|M| \mid X = x]$. The arbitrary source case is perhaps closer to the usual interpretation of channels which can accept any input with or without known distribution. For results that apply to both known source distribution and arbitrary source, we denote them as "KAS".

No / Limited / Unlimited common randomness (NCR / LCR / UCR). Definition 2 allows the encoder and the decoder to share the common randomness W. We often do not mind the amount of common randomness, and allow P_W to be any (discrete/continuous) distribution. Unlimited common randomness is actually not an unreasonable assmption, which will be explained in Section 2.3. Nevertheless, if one cannot tolerate any common randomness, channel simulation is possible without any common randomness, but often at the expense of more complicated schemes that require more communication. Channel simulation without common randomness will be discussed in Sections 4, 6. One can also impose a limit on the common randomness and study the tradeoff between common randomness and communication, which will be discussed in Section 7. One can even limit the amount of *local* randomness available at the encoder and the decoder, which will be discussed in Section 9.4.

Privacy. As discussed in Section 1.6, if we intend to keep the data X private from the decoder, it is not sufficient to ensure that the channel $P_{Y|X}$ is differentially private since the decoder also knows (W, M). The actual "channel" mapping the data to the information at the decoder is $P_{W,M|X}$ instead of $P_{Y|X}$. We say that the channel simulation scheme is (ϵ, δ)-*locally differentially private* (Dwork *et al.*, 2006; Shah *et al.*, 2022; Shahmiri *et al.*, 2024) if the conditional distribution $P_{W,M|X}$ is (ϵ, δ)-locally differentially private (Definition 1). Given that the channel $P_{Y|X}$ to be simulated is differentially private, it is often of interest to determine whether $P_{W,M|X}$ is still differentially private.

Channel simulation results in this monograph will be labelled by the aforementioned abbreviations to indicate their assumptions. Refer to the Notations section for a lookup table. For example, the one-shot exact variable-length channel simulation setting for a general channel $P_{Y|X}$ with known source distribution and unlimited common randomness is abbreviated as "G/1/E/VL/KS/UCR". In this setting, the minimum expected length of the prefix-free description L^* is bounded in terms of the mutual information $I = I(X;Y)$ by (Harsha *et al.*, 2010; Li and El Gamal, 2018b; Li, 2024)

$$I \leq L^* \leq I + \log_2(I+2) + 3 \,\text{bits}.$$

This will be elaborated in Section 3.

2.2 Overview of Various Channel Simulation Approaches

We give an overview of various approaches to constructing channel simulation schemes discussed in this monograph.

Rejection sampling (G/1/E/VL/KAS/UCR). In Section 3.2, we study how the rejection sampling scheme for random number generation can be viewed as a channel simulation scheme, which is one of the earlier approaches to channel simulation (Steiner, 2000). The greedy version of the rejection sampling scheme (Harsha *et al.*, 2010) was the first one-shot channel simulation scheme attaining an expected length close to the mutual information.
Properties:

- One-shot, exact, variable-length, known or arbitrary source distribution, requires unlimited common randomness.
- General (discrete/continuous) sources and channels.
- Greedy rejection sampling (Harsha *et al.*, 2010; Flamich and Theis, 2023) attains an expected length of

  $$\mathbf{E}[|M|] \leq I(X;Y) + \log_2\left(I(X;Y) + \log_2(4e)\right) + \log_2(8e) \text{ bits}$$

  in one-shot, and hence is asymptotically optimal (attains the asymptotic rate $I(X;Y)$).
- (Non-exact) variants that ensure differential privacy exist (Bassily and Smith, 2015; Feldman and Talwar, 2021) (Sections 3.2.3, 3.5.3).
- Causal sampling scheme (Liu and Verdú, 2018) (see Section 3.5).

Exponential and Poisson functional representation (G/1/E/VL/KAS/UCR). In Section 3.3, we study how a Poisson process can be used as a codebook, giving a one-shot channel simulation scheme attaining an expected length close to the mutual information, with a proof of Theorem 4 with a small constant. This is the approach investigated in (Li and El Gamal, 2018b; Li and Anantharam, 2021).
Properties:

- One-shot, exact, variable-length, known or arbitrary source distribution, requires unlimited common randomness.
- General (discrete/continuous) sources and channels.
- Attains an expected length of

$$\mathbf{E}[|M|] \leq I(X;Y) + \log_2 \left(I(X;Y) + 2\right) + 3 \,\text{bits}$$

  in one-shot (the best known constant), and hence is asymptotically optimal (Li and El Gamal, 2018b; Li, 2024).
- Exact variants that ensure differential privacy exist (Liu *et al.*, 2024) (Section 3.3.5).
- Noncausal sampling scheme (Liu and Verdú, 2018) (see Section 3.5).

Likelihood encoder and minimal random coding (G/1/A/FL/KAS/UCR). In Section 3.4, we will return to a more conventional approach of generating a fixed-size codebook in an i.i.d. manner, with a likelihood encoder (Cuff, 2013; Watanabe *et al.*, 2015; Song *et al.*, 2016) for choosing a codeword randomly (for known source distribution). This is closely related to minimal random coding (Havasi *et al.*, 2019; Flamich *et al.*, 2020) (for arbitrary source). In Section 3.5.3, we also discuss ordered random coding (Theis and Yosri, 2022), a variant of minimal random coding that also combines ideas of Poisson functional representation. The analysis is deferred to Section 5.6.
Properties:

- One-shot, approximate (with vanishing total variation distance), fixed-length, known or arbitrary source distribution, requires unlimited common randomness.
- General (discrete/continuous) sources and channels.
- Asymptotically optimal.
- Ensures 2ϵ-differential privacy if $P_{Y|X}$ ensures ϵ-differential privacy (Shah *et al.*, 2022).
- Noncausal sampling scheme (see Section 3.5).

Subtractively dithered quantization and layered randomized quantizers (AC/1/E/(FL or VL)/KAS/UCR). In Section 3.6, we return to the earliest example of channel simulation for additive uniform noise channels via subtractive dithering (Roberts, 1962; Schuchman, 1964; Ziv, 1985; Gray and Stockham, 1993), and discuss how it can be generalized to nonuniform noises via layered randomized quantizers (Wilson, 2000; Hegazy and Li, 2022).
Properties:

- One-shot, exact, fixed or variable-length, known or arbitrary source distribution, requires unlimited common randomness.
- (Scalar/vector) additive noise channels with continuous noises only.
- One shot exactly optimal for uniform noise (in the sense of minimizing the conditional entropy).
- In the large-SNR asymptotic setting (X is uniform over a large set), direct layered randomized quantizer is optimal (Hegazy and Li, 2022).
- A variant that ensures differential privacy for additive Laplace noise exist (Shahmiri *et al.*, 2024) (Section 3.6.4).
- Not a sampling scheme.

Dyadic decomposition (C/1/E/VL/KAS/NCR). In Section 4.2, we discuss the dyadic decomposition scheme for simulating continuous channels (Li and El Gamal, 2017; Li and El Gamal, 2018a). The dyadic decomposition scheme has the advantage that it does not require any common randomness. Nevertheless, the results are often not stated in terms of the mutual information unless additional assumptions on X, Y are imposed.
Properties:

- One-shot, exact, variable-length, known or arbitrary source distribution, does not require any common randomness.
- Continuous channels only.
- Attains an expected length within a constant gap from $I(X;Y)$ in one-shot for jointly log-concave distributions over $\mathbb{R}^2$.
- Not differentially private.
- Not a sampling scheme.

Soft covering (G/(1 or F or ∞)/A/FL/KS/(N or L or UCR)). In Sections 5, 6 and 7, we discuss the *soft covering lemma* (Wyner, 1975a; Cuff, 2013) or *channel resolvability* (Han and Verdú, 1993), which is a popular theoretical tool for proving asymptotic channel simulation results, and various results that can be proved using this lemma. The one-shot soft-covering lemma will be discussed in Section 8.
Properties:

- Asymptotic (though one-shot bounds exist), approximate (with vanishing total variation distance), fixed-length, known source distribution, no/limited/unlimited common randomness.
- General (discrete/continuous) sources and channels (though most existing analyses focus on discrete).
- Asymptotically optimal.
- Not differentially private, but can be made secure against an eavesdropper that only observes the description but not the common randomness (Cuff, 2013) (Section 10.4).
- Noncausal sampling scheme.

Method of types (D/∞/E/VL/KAS/UCR). In Section 5.2, we will discuss the original asymptotic channel simulation scheme with unlimited common randomness in (Bennett *et al.*, 2002), which divide the input sequence into type classes, and transmit the index of the input-output pair within the joint type class.
Properties:

- Asymptotic, exact, variable-length, known or arbitrary source distribution, exponentially-large common randomness.
- Discrete sources and channels only.
- Asymptotically optimal.
- Not differentially private.
- Noncausal sampling scheme.

Table 2.1 compares these approaches and highlights their advantages and disadvantages.

Remark 3. The meaning of the columns of Table 2.1 are explained in Section 2.1. "1D additive unimodal noise channels" means "general 1-dimensional additive noise channel with continuous unimodal noise distribution". "n-D add. cont. noise channel" means "general

Approach	Scheme	Section	Type of channels simulated: Discrete channels?	1D additive unimodal noise channels?	n-D add. cont. noise channels?	General channels?	One-shot?	Exact? (otherwise approximate)	Fixed-length? (otherwise variable-length)	Arbitrary source? (otherwise known src. dist.)	Usable with no common randomness?	Optimal wrt mutual info. lower bound: Asymptotically optimal?	Logarithmic gap from optimum?	Constant gap from opt. for high SNR?	Differentially private?	Causal sampling scheme?
Rejection sampling	Classical rejection sampling	3.2.1	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✓
	Greedy rejection sampling	3.2.2	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✗	✗	✓
	Approximate rejection sampling	3.2.3	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓	✗	✗	✗	✓
	Sriramu-Wagner	5.1	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓	✗	✗	–
	Generic 1-bit protocol	3.2.3	✓	✓	✓	✓	✓	✗	✓	✓	✗	–	–	–	✓	✓
	Local pseudo-randomizer	3.5.3	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗	✓	✓
Poisson process	Poisson functional representation	3.3	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✗	✗	✗
	AD* coding	3.3.5	$Y \in \mathbb{R}$				✓	✓	✗	✓	✗	–	✓	✗	✗	✗
	Greedy Poisson rejection samp.	3.5.3	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✗	✗	✓
	Poisson private representation	3.3.5	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✗	✓	✗
Importance sampling	Minimal random coding	3.4	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓	✗	✗	✓	✗
	Ordered random coding	3.5.3	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓	✗	✗	✗
Simple quantization		3.6.1	✗	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✗	✓	–
Dithered quantization	Direct layered quantizer	3.6.2	✗	✓	✗	✗	✓	✓	✗	✓	✗	–	✗	✓	✗	–
	Shifted layered quantizer	3.6.2	✗	✓	✗	✗	✓	✓	✓	✓	✗	–	✗	✓	✗	–
	Shift-periodic quantizer	3.6.3	✗	✓	✓	✗	✓	✓	✗	✓	✗	✗	✗	✗	✗	–
	Rejection-sampled univ. quant.	3.6.3	✗	✓	✓	✗	✓	✓	✗	✓	✗	✗	✗	✓	✗	–
	Rotated dithered quantization	3.6.4	Gaussian only				✓	✗	✓	✓	✗	✓	✗	✓	✗	–
	Dyadic quantized Laplace	3.6.4	Laplace only				✓	✓	✗	✓	✗	–	✗	✓	✓	–
Nonnegative matrix factorization		4.1	✓	✗	✗	✗	✓	✓	✓	✓	✓	✗	✗	–	✗	–
Dyadic decomposition		4.2	✗	✓	✓	✗	✓	✓	✗	✓	✓	✗	✗	✗	✗	–
Soft covering		7.1	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✗	✗	✗	✗
Method of types		5.2	✓	✗	✗	✗	✗	✓	✗	✓	✗	✓	✗	–	✗	✗

Table 2.1: Properties of various channel simulation schemes. Refer to the "Section" column for details and references. There is no single scheme that is the best in all aspects, and the choice of scheme should depend on which properties are desired in the particular scenario. "–" means "not applicable". Refer to Remark 3 for explanations.

n-dimensional additive noise channel with continuous noise distribution". "Asymptotically optimal" means that when applied to simulate n copies of the channel $P_{Y|X}$ with known source distribution, i.e., $X^n \sim P_X^n$ and $Y^n|X^n \sim P_{Y|X}^n$ (perhaps approximately), the description rate approaches $I(X;Y)$ as $n \to \infty$.[2] "Logarithmic gap from optimum" means that for one-shot simulation with known source distribution, we have an expected description length $I(X;Y) + O(\log(I(X;Y)+1))$.[3] "Constant gap from opt. for high SNR" means that when applied on the additive noise channel $Y = X + Z$ with $X \sim \mathrm{Unif}(0,t)$ (or X uniform over a large ball for n-dimensional channels), the expected description length of the scheme is within a constant away from the optimum when $t \to \infty$ (see Sections 3.6.2 and 3.6.3).[4] The meaning of "causal sampling scheme" is explained in Section 3.5.[5]

2.3 Why Unlimited Common Randomness? Why Prefix-free? Why Expected Length?

One may have several questions regarding Definition 2: *Why can we assume unlimited common randomness? Why do we use variable-length prefix-free codes? Why is the expected length* $\mathbb{E}[|M|]$ *a reasonable performance measure?*

The answer to the first question is simple: *because common randomness is everywhere.* True common randomness can be shared between the encoder and the decoder in advance, for example, by storing a pre-generated common sequence of random numbers generated from a true random source. Alternatively, they can use publicly available randomness sources, sometimes known as *random beacons* (Rabin, 1983; Clark and Hengartner, 2010), such as a satellite broadcasting random numbers at regular time intervals (Rabin, 1983), stock market data (Clark and Hengartner, 2010), lotteries data (Bowe *et al.*, 2017), blockchain data (Bowe *et al.*, 2017), and a book containing lots of random digits (RAND Corporation, 2001).

Nevertheless, in practice, we do not usually require true randomness, and pseudo-randomness is often used instead. The encoder and the decoder can share a common random seed (e.g., 64-bit) beforehand using a small amount of communication, and use it to initialize two synchronized pseudorandom number generators (PRNGs), which will allow them to generate a practically unending shared stream of random numbers, providing the common randomness needed for all tasks that await them in the future. The amount of actual

[2] This is not applicable if the scheme only applies to 1D channels, or if the distribution of (X^n, Y^n) does not approach $P_X P_{Y|X}$ as $n \to \infty$.

[3] This is not applicable if the distribution of (X, Y) does not approach $P_X P_{Y|X}$ as the description length grows.

[4] This is not applicable if the scheme is not applicable to 1D additive noise channels with continuous noise, or if the distribution of (X, Y) does not approach $P_X P_{Y|X}$ as the description length grows.

[5] This is not applicable if the scheme is not a sampling scheme.

communication needed (the 64-bit random seed) is significantly smaller than the amount of (pseudo)randomness that can be generated by the PRNGs. Common randomness is so inexpensive that even an unlimited amount is realistically achievable.

The first usage of synchronized PRNGs to generate unlimited common randomness in channel simulation appeared in the pioneering work on subtractive dithering by Roberts (1962), which used synchronized PRNGs to generate the dither signals (see Section 3.6). The assumption that common randomness can be generated by synchronized PRNGs has become quite common in the study of subtractive dithering (Zamir and Feder, 1992), channel simulation (Havasi *et al.*, 2019; Agustsson and Theis, 2020; Flamich *et al.*, 2020) and differential privacy (Bassily and Smith, 2015), though it is not without practical concerns (Gray and Stockham, 1993; Wannamaker *et al.*, 2000). One concern is that the output of a PRNG is not truely random, and is merely a deterministic function of the seed, so it may violate the mathematical properties of ideal common randomness (Gray and Stockham, 1993). Whether PRNGs are acceptable substitutes for true randomness is an interesting discussion that is out of the scope of this monograph. Interested readers are referred to (Kneusel, 2018). This monograph adopts the common strategy of works on randomized algorithms—assume the availability of true randomness in theoretical analyses, and utilize PRNGs when discussing implementation aspects of the algorithms. If one prefers schemes that do not make the theoretically questionable assumption that unlimited common randomness can be generated from a finite seed, one may instead use the channel simulation schemes that require no common randomness in Section 4.

To answer the remaining questions, let us study the following practical scenario. We are compressing a heterogeneous data sequence $X_1, \dots, X_n$ and transmitting it through a bit stream (over a network or into a file), where each $X_i \sim P_{X_i}$ may follow a different distribution. For example, for an image or audio, different frequency components may have different distributions, and the entries in the metadata (e.g. resolution, date and location of creation, etc.) may even be lying in completely different spaces. We want the reconstructed sequence $Y_1, \dots, Y_n$ to follow some prescribed conditional distributions $Y_i|X_i \sim P_{Y_i|X_i}$. To achieve this,

- The encoder first initializes a PRNG using a random seed S and writes the seed to the bit stream.
- Then, for each $i = 1, \dots, n$, the encoder...
  - uses the PRNG to generate the common randomness W_i; and
  - applies a one-shot channel simulation scheme with common randomness W_i to encode X_i into $M_i \in \mathcal{C}_{i,W_i}$, and writes M_i to the bit stream.
- The decoder first reads S from the bit stream and initializes a PRNG.

- Then, for each $i = 1, \dots, n$, the decoder...
  - uses the PRNG to generate W_i;
  - uses the prefix-free codebook $\mathcal{C}_{i,W_i}$ to obtain M_i from the bit stream; and
  - uses one-shot channel simulation to decode M_i into Y_i.

The M_i's being prefix-free codewords guarantees that they can be concatenated even if they belong to different codebooks, and the encoder and the decoder will always be synchronized. Although unlimited common randomness might sound prohibitive, practically it only amounts to adding a small random seed at the beginning of the file.

Each channel simulation scheme that maps X_i to Y_i is a tiny component in the overall scheme, which must behave "responsibly". This means:

- They must keep the encoder and decoder synchronized, so the bits in the stream corresponding to different indices i's will not interfere each other. This is achieved by the use of prefix-free codes.
- They are allowed to use the PRNGs synchronized between the encoder and the decoder, which is an inexpensive "public resource" available to all components of the scheme. However, they must keep the PRNG synchronized after the scheme, so it can be reused for other tasks later. This means the encoder and the decoder must use their PRNGs the same number of times.
- The expected communication $\mathbb{E}[|M_i|]$ should be as little as possible.

Law of large numbers tells us that the total length is approximately $\sum_{i=1}^{n} \mathbb{E}[|M_i|]$. Therefore, as long as each component fulfil its responsibility to minimize $\mathbb{E}[|M_i|]$ separately, the total length will also be approximately minimized. Compared to conventional block codes in information theory where the whole sequence $X_1, \dots, X_n$ is encoded together, the one-shot variable-length approach, such as Huffman coding (Huffman, 1952) and the channel simulation schemes in this section, is more "modular" in the sense that the code for each X_i can be designed separately, allowing the code to be simpler and applicable to heterogeneous data. It also allows the decoder to obtain $Y_1, \dots, Y_n$ in a streaming manner, reducing the delay.

If the channel simulation scheme is a small component of a larger picture, it is reasonable to minimize the expected length $\mathbb{E}[|M|]$. On the other hand, if this one channel simulation scheme is the main part of the whole protocol (e.g., if we are compressing a large image or video by one use of a giant monolithic channel simulation scheme), and there is a strict limit on the number of bits used by the protocol, then we no longer have law of large numbers to help us, and we should instead give a small upper bound on $|M|$, i.e., we should minimize

ess sup $|M|$. This is the same as using *fixed-length schemes* where M is always a fixed-length bit sequence. Unfortunately, fixed-length exact channel simulation does not admit schemes and analyses as elegant as those for variable-length description. This will be briefly discussed in Section 3.7. It is also possible to relax the exact constraint on the distribution of Y to an approximate one, which will be discussed in Sections 5, 7, 8. Fixed-length schemes may require a significantly longer description than variable-length schemes, for the same reason that fixed-length lossless compression is less efficient than variable-length lossless compression (e.g., Huffman code)—the fixed-length encoding fails to adapt to the variability of the amount of information in the input X.

3 One-shot Channel Simulation with Unlimited Common Randomness

This section concerns one-shot channel simulation with unlimited common randomness (Definition 2). It is mostly devoted to the simulation of general channels in the one-shot exact variable-length case with unlimited common randomness, abbreviated as "G/1/E/VL/KAS/UCR", though we will also discuss schemes for restricted classes of channels and approximate schemes. The central result in this section is that the minimum expected length is approximately the mutual information $I(X;Y)$ for known source distribution, and approximately the channel capacity C for arbitrary source.

Theorem 4 (G/1/E/KAS/UCR (Harsha *et al.*, 2010; Li and El Gamal, 2018b))**.** *For the one-shot exact variable-length channel simulation setting for a general (discrete/continuous) channel $P_{Y|X}$ with unlimited common randomness:*

- *For known source distribution, the minimum expected length of the prefix-free description L^* is bounded in terms of the mutual information $I = I(X;Y)$ by*

$$I \le L^* \le I + \log_2(I+2) + 3 \text{ bits}.$$

- *For arbitrary source, the minimum worst-case expected length of the prefix-free description L^* is bounded in terms of the channel capacity $C := \sup_{P_X} I(X;Y)$ by*

$$C \le L^* \le C + \log_2(C+2) + 3 \text{ bits}.$$

Note that the lower bounds $L^* \ge I$ (for known source distribution) and $L^* \ge C$ (for arbitrary source) are immediate since $\mathbb{E}[|M|] \ge H(M) \ge I(X;M|W) \ge I(X;M,W) \ge I(X;Y)$ due to $I(X;W) = 0$ and the Markov chain $X \leftrightarrow (M,W) \leftrightarrow Y$.

The earliest form of this result was proved by Harsha *et al.* (2010), who showed that for discrete channels, $L^* \le I + (1+o(1))\log_2(I+1)$ for known source distribution, and $L^* \le C + (1+o(1))\log_2(C+1)$ for arbitrary source, using a *rejection sampling* scheme. This is improved by Braverman and Garg (2014) to $L^* \le I + \log_2(I+1) + O(1)$ for discrete channels and known source distribution. The first result for general (discrete/continuous) channels in the form of Theorem 4 was given by Li and El Gamal (2018b) using a construction called the *Poisson functional representation*. More recent analyses on the rejection sampling scheme can be found in (Liu and Verdú, 2018; Flamich and Theis, 2023). In Section 3.3, we will present the proof of Theorem 4 using the Poisson functional representation construction in (Li and El Gamal, 2018b; Li and Anantharam, 2021; Li, 2024), which gives the smallest constants for the bound in Theorem 4 to the best of the author's knowledge.[1]

[1] Li and El Gamal (2018b) gave the bound $I + \log_2(I+1) + 5$. This was improved by Li and Anantharam (2021) to $I + \log_2(I+1) + 4.732$. The bound $I + \log_2(I+2) + 3$ in Theorem 4, which is the best bound to the best of the author's knowledge, was given by Li (2024) using a slightly improved analysis.

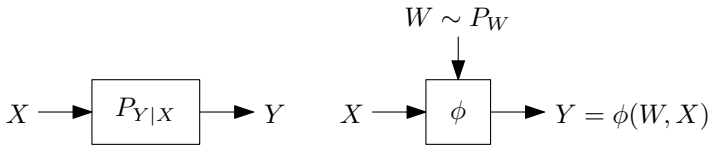


Figure 3.1: The conditional distribution formulation (left) and the functional formulation (Shannon, 1948) (right) of a channel.

Regarding the logarithmic term in Theorem 4, it was shown by Braverman and Garg (2014) that the logarithmic term is necessary for upper bounds that are stated only in terms of the mutual information $I = I(X;Y)$, in the sense that there is a sequence of source-channel pairs $(P_X, P_{Y|X})$ where $I \to \infty$ and $L^* \geq I + \log_2(I+1) + c$ where c is a constant. An explicit constant $c = -1$ is given by Li and El Gamal (2018b). Note that this does not mean $L^* \geq I + \log_2(I+1) + c$ in general (e.g., if $X = Y \sim \text{Unif}(\{0,1\}^\ell)$, then we have $L^* = I = \ell$).

This section is organized as follows. In Section 3.1, we present a useful view of the channel simulation setting with unlimited common randomness as a functional representation problem. In Sections 3.2, 3.3 and 3.4, we present three approaches to the simulation of general channels. These approaches are examples of sampling schmes, which are explained in Section 3.5. In Section 3.6, we discuss schemes based on dithering quantization for the simulation of additive noise channels. In Section 3.7, we discuss one-shot exact fixed-length channel simulation.

3.1 Functional Representation Lemma

Before we continue with the channel simulation discussion, let's take a huge step back and ask, *what is a channel?* We often think of a channel as a conditional distribution $P_{Y|X}$, which we call the *conditional distribution formulation*. However, in Shannon's original model of a communication channel (Shannon, 1948) (see Figure 3.1), a channel is instead given as a deterministic function $\phi : \mathcal{W} \times \mathcal{X} \to \mathcal{Y}$ mapping the input X and a random noise $W \sim P_W$ independent of X to the output Y, which we call the *functional formulation*. In the conditional distribution formulation, the channel is a "random mapping" from X to Y. In the functional formulation, the channel is a deterministic mapping ϕ, and all its randomness comes from the noise W.

For special cases such as additive noise channels where $\phi(w, x) = x + w$, it is apparent that it can be expressed using either formulation. It turns out that the two formulations of channels are equivalent in general. This fact is known as the *functional representation lemma* (El Gamal and Kim, 2011; Hajek and Pursley, 1979; Willems and Meulen, 1985; Kallenberg, 2002), which states that for any conditional distribution $P_{Y|X}$, there exists a

distribution P_W and a function $\phi : \mathcal{W} \times \mathcal{X} \to \mathcal{Y}$ such that if $X \sim P_X$ is independent of $W \sim P_W$, and $Y = \phi(W, X)$, then $(X, Y) \sim P_X P_{Y|X}$.[2]

Given a conditional distribution $P_{Y|X}$, its functional representation (P_W, ϕ) may not be unique. For instance, for the binary symmetric channel where $X, Y \in \{0, 1\}$, $P_{Y|X}(1|0) = P_{Y|X}(0|1) = \alpha \in [0, 1/2]$, we can take $W \sim \text{Bern}(\alpha)$ and $Y = X \oplus W$ (exclusive or), or take $P_W(0) = P_W(1) = \alpha$, $P_W(2) = 1 - 2\alpha$, and $Y = W$ if $W \neq 2$, and $Y = X$ if $W = 2$. There is no "canonical" way to find the functional representation (P_W, ϕ). However, some representations may be better under certain metrics. For example, we may choose "the noise W that is the most informative about the output Y", i.e., the representation that maximizes $I(W; Y)$, or equivalently, minimizes $H(Y|W)$. Loosely speaking, this noise W will decompose the output Y into "the part that comes from the noise" $I(W; Y)$, and "the part that comes from the input" $H(Y|W)$. Under this criterion, for the binary symmetric channel, the former representation gives $H(Y|W) = 1$, whereas the latter representation gives $H(Y|W) = 1 - 2\alpha$, and hence we would choose the latter representation over the former.

The functional representation of a channel is closely related to the channel simulation setting. A channel simulation code $(P_W, (\mathcal{C}_w)_{w \in \mathcal{W}}, f, g)$ with deterministic encoding function $f : \mathcal{W} \times \mathcal{X} \to \{0, 1\}^*$ and deterministic decoding function $g : \mathcal{W} \times \{0, 1\}^* \to \mathcal{Y}$ corresponds to the functional representation (P_W, ϕ) where $\phi(w, x) = g(w, f(w, x))$.[3] For the other direction, given a functional representation (P_W, ϕ) of $P_{Y|X}$, we can construct a code for channel simulation (with known source distribution) by taking the common randomness to be W, the encoding function $f(w, x)$ to be the encoding of $y = \phi(w, x)$ in the Huffman code constructed using the distribution $P_{Y|W}(\cdot|w)$ (we take the prefix-free codebook $\mathcal{C}_w$ to be the Huffman codebook constructed using the distribution $P_{Y|W}(\cdot|w)$), and the decoding function $g(w, m)$ to be the decoding of m into $y = \phi(w, x)$ using the Huffman code constructed using the distribution $P_{Y|W}(\cdot|w)$. The expected length of the encoding is bounded by

$$H(Y|W) \leq \mathbb{E}[|M|] \leq H(Y|W) + 1.$$

The optimality of Huffman code implies that this construction is optimal among channel simulation codes corresponding to the same functional representation (P_W, ϕ).

Therefore, the problem of finding the optimal channel simulation code with known source distribution is approximately equivalent to finding the functional representation (P_W, ϕ) with the smallest conditional entropy $H(Y|W)$. Define the *minimal conditional*

[2] Refer to (El Gamal and Kim, 2011) for a proof for the discrete case, and Kallenberg, 2002, Prop. 6.13 for the general case.

[3] If encoding and decoding are stochastic, we can always move the local randomness at the encoder and the decoder to the unlimited common randomness W. Therefore, without loss of generality, we can assume that encoding and decoding are deterministic, and all the randomness in the scheme is contained in W. This argument works only when privacy is not a concern, since moving the local randomness at the encoder to W may harm the privacy, considering that the decoder has access to W.

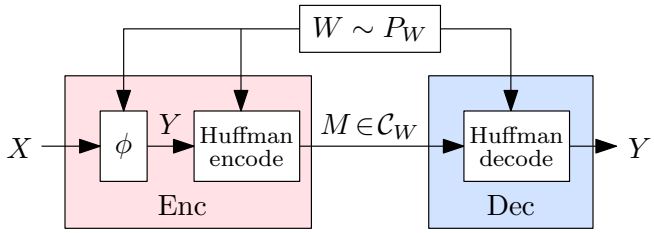


Figure 3.2: Functional representation and channel simulation.

entropy as (Li and El Gamal, 2018b)[4]

$$H^* := \inf_{P_{W|X,Y}: W \perp\!\!\!\perp X, H(Y|X,W)=0} H(Y|W). \tag{3.1}$$

Note that the condition $W \perp\!\!\!\perp X$ and $H(Y|X,W) = 0$ means that (P_W, ϕ) is a functional representation of the channel $P_{Y|X}$ for some ϕ. The answer to the channel simulation problem is within one bit from H^*:

$$H^* \le L^* \le H^* + 1.$$

Moreover, we have the following lower bound, which was observed in (Li and El Gamal, 2018b).

Proposition 5. *We have*

$$I(X;Y) \overset{(a)}{\le} H^* \le L^* \le H^* + 1, \tag{3.2}$$

and equality holds in the lower bound in (a) if there exists a functional representation (P_W, ϕ) with $X \leftrightarrow Y \leftrightarrow W$ forming a Markov chain (e.g., if W is a function of Y).

Proof. If $W \perp\!\!\!\perp X$ and $H(Y|X,W) = 0$, then

$$\begin{aligned} H(Y|W) &= I(X;Y|W) \\ &= I(X;Y|W) + I(X;W) \\ &= I(X;Y) + I(X;W|Y) \\ &\ge I(X;Y). \end{aligned}$$

Equality holds if $X \leftrightarrow Y \leftrightarrow W$ forms a Markov chain. □

Theorem 4 says that the mutual information lower bound in Proposition 5 is tight within a logarithmic gap. In the following sections, we will investigate various code constructions for the channel simulation problem, and how close they are to the mutual information lower bound.

[4] Li and El Gamal (2018b) studied the *excess functional information* instead, defined as $\Psi = H^* - I(X;Y)$.

3.2 Rejection Sampling

3.2.1 Classical Rejection Sampling

Rejection sampling (Devroye, 1986) is a popular method for random number generation, which, interestingly, can also serve as a channel simulation method. Early examples of applications of rejection sampling for channel simulation include the simulation of the Bell state via communication by Steiner (2000), and the direct sum result in communication complexity by Jain *et al.* (2003). Rejection sampling is also the basis of the schemes in (Harsha *et al.*, 2010; Braverman and Garg, 2014).

We review the rejection sampling construction. Suppose we can generate random variates from a *reference distribution* Q is a distribution over $\mathcal{Y}$, and we have generated $\bar{Y}_1, \bar{Y}_2, \ldots \overset{iid}{\sim} Q$. We have also generated $U_1, U_2, \ldots \overset{iid}{\sim} \mathrm{Unif}(0,1)$ independent of $(\bar{Y}_i)_{i\in\mathbb{N}^+}$. Now we want to obtain a random variate from another distribution P satisfying $P \ll Q$, i.e. P is absolutely continuous with respect to Q. To do so, we will read $(\bar{Y}_i, U_i)$ one by one, and accept the first $(\bar{Y}_i, U_i)$ where $U_i \le \gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_i)$, where $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Radon-Nikodym derivative,[5] and γ is a constant satisfying $\gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(y) \le 1$ for all $y \in \mathcal{Y}$. In other words, we output

$$Y = \bar{Y}_K, \text{ where } K := \min\left\{ i \in \mathbb{N}^+ : U_i \le \gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_i) \right\}.$$

It can be checked that $Y \sim P$ follows the desired distribution. Since the acceptance probability at each iteration is

$$\mathbb{P}\left(U_i \le \gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_i) \right) = \int_{\mathcal{Y}} \gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_i) Q(\mathrm{d}y) = \gamma,$$

the distribution of K is $\mathrm{Geom}(\gamma)$, i.e., the geometric distribution with support $\mathbb{N}^+$ and parameter γ.

It was observed by Steiner (2000) that the rejection sampling scheme can be treated as a remote generation scheme, and hence a channel simulation scheme. The encoder and the decoder agree on a reference distribution Q, and share the common randomness $W = (\bar{Y}_i)_{i\in\mathbb{N}^+}$. Suppose the encoder observes a distribution P, and wants the decoder to produce a sample from P. To accomplish this, the encoder applies rejection sampling to obtain the index $K := \min\{i \in \mathbb{N}^+ : U_i \le \gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_i)\}$ using the local randomness $U_1, U_2, \ldots \overset{iid}{\sim} \mathrm{Unif}(0,1)$, where γ is a constant (which may depend on P) satisfying $\gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(y) \le 1$ for all $y \in \mathcal{Y}$. The encoder encodes K into M and sends it to the decoder. The decoder recovers K from M and outputs $Y = \bar{Y}_K$. Refer to Figure 3.4 for an illustration.

To obtain a channel simulation scheme in the form of Definition 2 where the encoder observes X instead of P, we simply take $P = P_{Y|X}(\cdot|X)$, which guarantees that the decoder

[5] If P, Q are both discrete, then $\frac{\mathrm{d}P}{\mathrm{d}Q}(y) = P(y)/Q(y)$. If they are both continuous with probability density functions f_P, f_Q, then $\frac{\mathrm{d}P}{\mathrm{d}Q}(y) = f_P(y)/f_Q(y)$.

produces Y following the correct conditional distribution $P_{Y|X}(\cdot|X)$. Now the constant γ_X may depend on the choice of X, and must satisfy $\gamma_x \frac{\mathrm{d}P_{Y|X}(\cdot|x)}{\mathrm{d}Q}(y) \le 1$ for all x, y. The chosen index is

$$K = \min\left\{ i \in \mathbb{N}^+ : U_i \le \gamma_X \frac{\mathrm{d}P_{Y|X}(\cdot|X)}{\mathrm{d}Q}(\bar{Y}_i) \right\}. \tag{3.3}$$

In (Steiner, 2000), the values $\gamma_x = \gamma$ do not depend on x. In this case, we have $K \sim \mathrm{Geom}(\gamma)$ regardless of the value of X, giving the following result.

Theorem 6 (Classical rejection sampling (G/1/E/VL/KS/UCR) (Steiner, 2000))**.** *Let $\gamma > 0$ be a constant satisfying $\gamma \frac{\mathrm{d}P_{Y|X}(\cdot|x)}{\mathrm{d}Q}(y) \le 1$ for all x, y. Then the rejection sampling scheme with constant $\gamma_x = \gamma$ has a conditional entropy bounded by*

$$H(Y|W) \le -\log_2 \gamma - \frac{1-\gamma}{\gamma} \log_2(1-\gamma).$$

Proof. To bound the conditional entropy $H(Y|W)$, note that Y is a function of (K, W), and hence $H(Y|W) \le H(K|W) \le H(K)$ is upper-bounded by the entropy of $\mathrm{Geom}(\gamma)$. □

To implement this scheme, the encoder applies rejection sampling to obtain the index K, encodes it into $M \in \{0,1\}^*$ using Huffman coding (Huffman, 1952) for the distribution $K \sim \mathrm{Geom}(\gamma)$ with expected length $\mathbb{E}[|M|] \le H(K) + 1$, and sends it to the decoder. The decoder decodes K and outputs $\bar{Y}_K$ by looking up the common random sequence $(\bar{Y}_i)_i$.

The communication cost of classical rejection sampling in Theorem 6 is approximately $-\log_2 \gamma$, which is at least $\log_2 \sup_{x,y} \frac{\mathrm{d}P_{Y|X}(\cdot|x)}{\mathrm{d}Q}$ since $\gamma \frac{\mathrm{d}P_{Y|X}(\cdot|x)}{\mathrm{d}Q}(y) \le 1$ for all x, y. This is generally larger than $I(X;Y)$ which is approximately the optimal communication cost (Theorem 4). The problem is that the choice of γ must accommodate the largest values of $\frac{\mathrm{d}P_{Y|X}(\cdot|x)}{\mathrm{d}Q}(y)$, regardless of how unlikely those values are going to appear. In the next section, we will discuss a scheme based on rejection sampling with a smaller communication cost.

3.2.2 Greedy Rejection Sampling

We describe the strategy by Harsha *et al.* (2010), which modifies the acceptance rule of rejection sampling. In the original rejection sampling scheme for sampling from P using samples from Q, we accept the first $(\bar{Y}_i, U_i)$ where $U_i \le \gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_i)$. The acceptance rule does not depend on the iteration number. This may not be the best strategy if our goal is to accept as early as possible.

Assume we want to design the acceptance rule for the first iteration in a way that maximizes the acceptance probability at the first iteration. Assume P, Q are discrete for now.

We draw $\bar{Y}_1 \sim Q$ independent of $U_1 \sim \text{Unif}(0,1)$. Assume that we accept $\bar{Y}_1$ if $U_1 \le \rho_1(\bar{Y}_1)$, where $\rho_1(y) \in [0,1]$ is the acceptance probability at iteration 1 if the sample is y. Let K be the iteration number of the first acceptance. Since we want to obtain a sample from P at the end, we must have $P(y) \ge \mathbb{P}(Y = y,\, K = 1) = Q(y)\rho_1(y)$. Therefore, to maximize the acceptance probability at the first iteration, we should take $\rho_1(y) = \min\{P(y)/Q(y),\, 1\}$. We can write P as the following mixture

$$\begin{aligned} P(y) = \Big(\sum_{y'} Q(y')\rho_1(y')\Big)\frac{Q(y)\rho_1(y)}{\sum_{y'} Q(y')\rho_1(y')} \\ + \Big(1 - \sum_{y'} Q(y')\rho_1(y')\Big)\frac{P(y) - Q(y)\rho_1(y)}{1 - \sum_{y'} Q(y')\rho_1(y')}, \end{aligned}$$

where the first component $\frac{Q(y)\rho_1(y)}{\sum_{y'} Q(y')\rho_1(y')} = \mathbb{P}(Y = y \,|\, K = 1)$ is contributed by the first iteration. Therefore, the remaining iterations should yield a sample from the second component $\frac{P(y)-Q(y)\rho_1(y)}{1-\sum_{y'} Q(y')\rho_1(y')}$. We can then repeat this construction recursively to maximize the acceptance probability at the second iteration, third iteration, and so on.

We now describe the scheme in detail. Consider the general strategy where we accept the first $(\bar{Y}_i, U_i)$ where $U_i \le \rho_i(\bar{Y}_i)$, where $\rho_i(y) \in [0,1]$ is the acceptance probability at iteration i if the sample is y. In other words, we output

$$Y = \bar{Y}_K, \text{ where } K := \min\left\{i \in \mathbb{N}^+ :\, U_i \le \rho_i(\bar{Y}_i)\right\}. \tag{3.4}$$

We have

$$\begin{aligned} s_k &:= \mathbb{P}(K \ge k) \\ &= \mathbb{P}\left(\forall i \le k-1 :\, U_i > \rho_i(\bar{Y}_i)\right) \\ &= \prod_{i=1}^{k-1} \mathbb{P}\left(U_i > \rho_i(\bar{Y}_i)\right) \\ &= \prod_{i=1}^{k-1} \left(1 - \mathbb{E}[\rho_i(\bar{Y})]\right), \end{aligned}$$

where $\bar{Y} \sim Q$. For discrete P, Q, this scheme yields $Y \sim P$ where

$$\begin{aligned} P(y) &= \sum_{k=1}^{\infty} \mathbb{P}(K = k)\mathbb{P}(Y = y \,|\, K = k) \\ &= \sum_{k=1}^{\infty} \mathbb{P}(K = k)\frac{Q(y)\rho_k(y)}{\sum_{y'} Q(y')\rho_k(y')} \\ &= Q(y)\sum_{k=1}^{\infty} \mathbb{P}(K = k)\frac{\rho_k(y)}{\mathbb{E}[\rho_k(\bar{Y})]} \end{aligned}$$

$$
\begin{aligned}
&= Q(y) \sum_{k=1}^{\infty} \left(\prod_{i=1}^{k-1} \left(1 - \mathbb{E}[\rho_i(\bar{Y})]\right) - \prod_{i=1}^{k} \left(1 - \mathbb{E}[\rho_i(\bar{Y})]\right) \right) \frac{\rho_k(y)}{\mathbb{E}[\rho_k(\bar{Y})]} \\
&= Q(y) \sum_{k=1}^{\infty} \rho_k(y) \prod_{i=1}^{k-1} \left(1 - \mathbb{E}[\rho_i(\bar{Y})]\right) \\
&= Q(y) \sum_{k=1}^{\infty} \rho_k(y) s_k.
\end{aligned}
$$

For general P, Q, it can be shown that

$$
\frac{\mathrm{d}P}{\mathrm{d}Q}(y) = \sum_{k=1}^{\infty} \rho_k(y) s_k. \tag{3.5}
$$

Therefore, the goal is to design $\rho_i(y)$ satisfying (3.5) for given P, Q that results in a small K. The idea in (Harsha *et al.*, 2010) is to greedily maximize $\mathbb{P}(K = 1)$, and then maximize $\mathbb{P}(K = 2)$, and so on. To this end, we will maximize $\rho_1(y)$ subject to $\rho_1(y) \leq 1$ and $\rho_1(y)s_1 \leq \frac{\mathrm{d}P}{\mathrm{d}Q}(y)$ as required by (3.5), and then maximize $\rho_2(y)$ subject to $\rho_2(y) \leq 1$ and $\rho_1(y)s_1 + \rho_2(y)s_2 \leq \frac{\mathrm{d}P}{\mathrm{d}Q}(y)$, and so on. This yields the recursive formula

$$
\rho_k(y) = \min\left\{ \frac{1}{s_k} \left(\frac{\mathrm{d}P}{\mathrm{d}Q}(y) - \sum_{j=1}^{k-1} \rho_j(y) s_j \right), 1 \right\}, \tag{3.6}
$$

where s_k can also be recursively computed together with $\rho_k(y)$ as $s_1 = 1$ and

$$
s_k = s_{k-1}(1 - \mathbb{E}[\rho_{k-1}(\bar{Y})]) \tag{3.7}
$$

for $k \geq 2$. By construction, once $\rho_k(y) = 0$, we will always have $\rho_{k+1}(y) = \rho_{k+2}(y) = \cdots = 0$. Also, if $\rho_k(y) > 0$, $k \geq 2$, we must have $\rho_{k-1}(y) = 1$.[6] Hence, the sequence $(\rho_k(y))_{k \in \mathbb{N}^+}$ for any given y must be in the form $1, \ldots, 1, r, 0, 0, \ldots$, where $0 \leq r < 1$. Therefore, (3.6) can also be written as

$$
\rho_k(y) = \min\left\{ \max\left\{ \frac{1}{s_k} \left(\frac{\mathrm{d}P}{\mathrm{d}Q}(y) - \sum_{j=1}^{k-1} s_j \right), 0 \right\}, 1 \right\}. \tag{3.8}
$$

The recurrence relation on s_k in (3.7) can also be written as $s_1 = 1$, and

$$
s_k = s_{k-1} - \mathbb{E}\left[\min\left\{ \max\left\{ \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}) - \sum_{j=1}^{k-2} s_j, 0 \right\}, s_{k-1} \right\} \right] \tag{3.9}
$$

for $k \geq 2$, where $\bar{Y} \sim Q$. This is the *greedy rejection sampling scheme* in (Harsha *et al.*, 2010), and has been analyzed in (Liu and Verdú, 2018; Flamich and Theis, 2023). Refer

[6] If $\rho_k(y) > 0$, $k \geq 2$, then $\frac{\mathrm{d}P}{\mathrm{d}Q}(y) - \sum_{j=1}^{k-1} \rho_j(y) s_j > 0$, and $\rho_{k-1}(y) < s_{k-1}^{-1}(\frac{\mathrm{d}P}{\mathrm{d}Q}(y) - \sum_{j=1}^{k-2} \rho_j(y) s_j)$, so we have $\rho_{k-1}(y) = 1$.

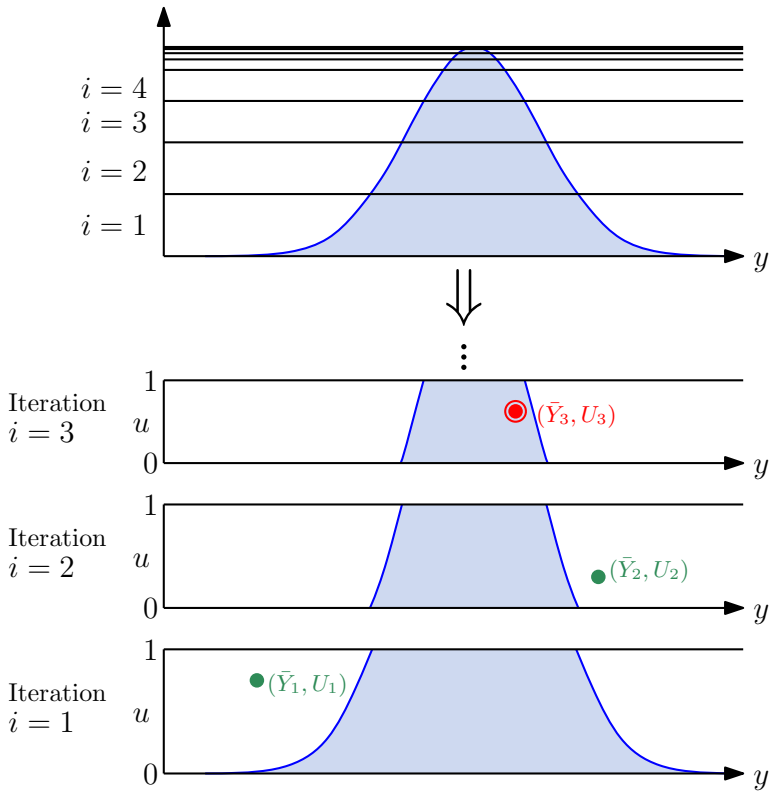


Figure 3.3: An illustration of the greedy rejection sampling scheme applied on Q being the uniform distribution over an interval, and P being a Gaussian distribution (truncated so it fits within the interval), shown as the blue shape in the figure. The distribution P is sliced horizontally. Iteration $i = 1$ corresponds to the lowest slice, and we accept $\bar{Y}_1$ if the point $(\bar{Y}_1, U_1)$ falls in the blue region. Iteration $i = 2$ corresponds to the second slice, and we accept $\bar{Y}_2$ if the point $(\bar{Y}_2, U_2)$ falls in the blue region, and so on until we accept a point (which is $\bar{Y}_3$ in the figure).

to Figure 3.3 for an illustration of greedy rejection sampling, and to Figure 3.4 for an illustration of the differences between classical rejection sampling and greedy rejection sampling.

The following result was proved in (Flamich and Theis, 2023), which refines upon the result in (Harsha *et al.*, 2010).[7]

Theorem 7 (Greedy rejection sampling (Harsha *et al.*, 2010; Flamich and Theis, 2023))**.** *The greedy rejection sampling scheme given by* (3.6) *achieves*

$$\mathbb{E}[\log_2 K] \leq D_{\mathrm{KL}}(P\|Q) + \log_2(2e).$$

[7] Harsha *et al.* (2010) showed $\mathbb{E}[\log_2 K] \leq D_{\mathrm{KL}}(P\|Q) + 2\log_2 e$ for discrete P, Q. Flamich and Theis (2023) proved $\mathbb{E}[\log_2 K] \leq D_{\mathrm{KL}}(P\|Q) + \log_2(2e)$ for general P, Q.

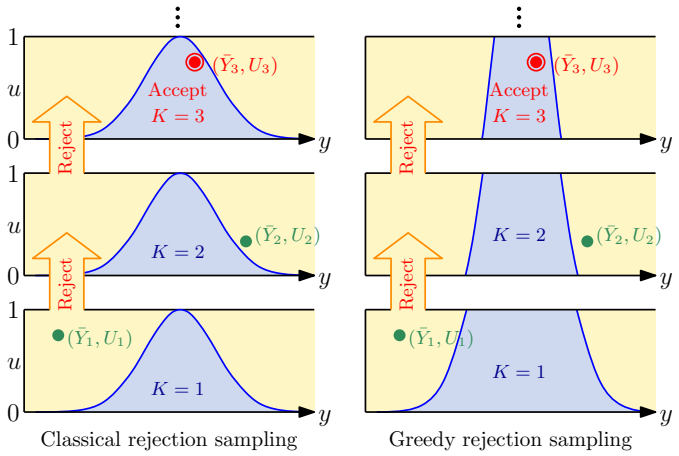


Figure 3.4: An illustration of classical rejection sampling and greedy rejection sampling, applied on Q being the uniform distribution over an interval, and P being a Gaussian distribution (truncated so it fits within the interval). While classical rejection sampling has the same acceptance probability at each iteration, greedy rejection sampling has a higher acceptance probability at the first iteration (indicated by the larger blue area).

Proof. The proof here is mostly based on (Flamich and Theis, 2023). Write $g(y) := (\mathrm{d}P/\mathrm{d}Q)(y)$. Define Y, K as in (3.4). Recall that $Y \sim P$ and $s_k = \mathbb{P}(K \geq k)$. We first show that the scheme terminates almost surely, or equivalently, $\lim_{k\to\infty} s_k = 0$. Assume the contrary that $\lim_{k\to\infty} s_k =: s_\infty > 0$. By (3.9), letting $\bar{Y} \sim Q$,

$$\begin{aligned}
1 - s_\infty &= \mathbb{E}\left[\sum_{k=2}^{\infty} \min\left\{ \max\left\{ \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}) - \sum_{j=1}^{k-2} s_j,\, 0 \right\},\, s_{k-1} \right\}\right] \\
&= \mathbb{E}\left[\sum_{k=2}^{\infty} \left(\min\left\{ \sum_{j=1}^{k-1} s_j,\, \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}) \right\} - \min\left\{ \sum_{j=1}^{k-2} s_j,\, \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}) \right\} \right)\right] \\
&= \mathbb{E}\left[\frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y})\right] \\
&= 1,
\end{aligned}$$

contradicting with $s_\infty > 0$. Hence, $\lim_{k\to\infty} s_k = 0$.

If $(K, Y) = (k, y)$ is a possible outcome, then $\rho_k(y) > 0$, and $g(y) > \sum_{j=1}^{k-1} s_j \geq (k-1)s_k$ by (3.8). Hence we have $K - 1 \leq g(Y)/s_K$ almost surely, and

$$\begin{aligned}
\mathbb{E}[\log_2 K] &\leq \mathbb{E}\left[\log_2\left(\frac{g(Y)}{s_K} + 1\right)\right] \\
&= \mathbb{E}\left[\log_2 \frac{g(Y)}{s_K}\right] + \mathbb{E}\left[\log_2\left(1 + \frac{s_K}{g(Y)}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\log_2 \frac{g(Y)}{s_K}\right] + \log_2\left(1 + \mathbb{E}\left[\frac{s_K}{g(Y)}\right]\right) \\
&\leq \mathbb{E}\left[\log_2 \frac{g(Y)}{s_K}\right] + \log_2\left(1 + \mathbb{E}\left[\frac{1}{g(Y)}\right]\right) \\
&\leq \mathbb{E}\left[\log_2 \frac{g(Y)}{s_K}\right] + \log_2 2,
\end{aligned}
$$

where the last line is because $\mathbb{E}[1/g(Y)] = \int 1/((\mathrm{d}P/\mathrm{d}Q)(y))P(\mathrm{d}y) \leq 1$. Note that $\mathbb{E}[\log_2 g(Y)] = D_{\mathrm{KL}}(P\|Q)$, and

$$
\begin{aligned}
\mathbb{E}[-\log_2 s_K] &= -\sum_{k=1}^{\infty} (\log_2 s_k)\mathbb{P}(K=k) \\
&= -\sum_{k=1}^{\infty} (\log_2 s_k)(s_k - s_{k+1}) \\
&= -\int_0^1 \log_2 s_{\max\{k:\, s_k \geq t\}}\, \mathrm{d}t \\
&\leq -\int_0^1 \log_2 t\, \mathrm{d}t \\
&= \log_2 e.
\end{aligned}
$$

Hence, $\mathbb{E}[\log_2 K] \leq D_{\mathrm{KL}}(P\|Q) + \log_2(2e)$. □

We can then use Theorem 7 to bound the performance of the greedy rejection sampling scheme applied to the channel simulation setting with known or arbitrary source distribution in Definition 2.

Corollary 8 (Greedy rejection sampling (G/1/E/VL/KAS/UCR) (Harsha *et al.*, 2010; Flamich and Theis, 2023))**.** *For the channel simulation setting, the greedy rejection sampling scheme given by* (3.6) *achieves:*

- *(Known source distribution)*

  $$\mathbb{E}[\log_2 K] \leq I(X;Y) + \log_2(2e),$$

  and a conditional entropy

  $$H(Y|W) \leq I(X;Y) + \log_2\left(I(X;Y) + \log_2(4e)\right) + \log_2(4e) \text{ bits}.$$

- *(Arbitrary source)*

  $$\sup_{x\in\mathcal{X}} \mathbb{E}[\log_2 K \,|\, X = x] \leq C + \log_2(2e),$$

  where $C := \sup_{P_X} I(X;Y)$*, and a worst-case expected length*

  $$\sup_{x\in\mathcal{X}} \mathbb{E}[|M| \,|\, X = x] \leq C + \log_2\left(C + \log_2(4e)\right) + \log_2(8e) \text{ bits}.$$

Proof. For the known source distribution case, applying the greedy rejection sampling scheme on $P = P_{Y|X}(\cdot|x)$, $Q = P_Y$, we have

$$\begin{aligned}\mathbb{E}[\log_2 K] &\le \mathbb{E}[D_{\mathrm{KL}}(P_{Y|X}(\cdot|X)\|P_Y)] + \log_2(2e)\\ &\le I(X;Y) + \log_2(2e).\end{aligned}$$

To bound the conditional entropy $H(Y|W)$, note that Y is a function of (K, W), and hence $H(Y|W) \le H(K|W) \le H(K)$. The bound on $H(K)$ can be obtained via the cross entropy between the distribution of K and the Zipf distribution $\mathrm{Zipf}(1 + 1/(I(X;Y) + \log_2(2e)))$, which is given in Appendix A.

For the arbitrary source case, we invoke the result by Kemperman (1974) (also see Polyanskiy and Wu, 2024, Theorem 5.9) about the saddle point characterization of channel capacity: if $C := \sup_{P_X} I(X;Y) < \infty$, then there exists a unique Q over $\mathcal{Y}$ such that

$$C = \sup_{P_X} \mathbb{E}_{X\sim P_X}\left[D_{\mathrm{KL}}(P_{Y|X}(\cdot|X)\,\|\,Q)\right]. \tag{3.10}$$

If the supremum in $C = \sup_{P_X} I(X;Y)$ is attained by P_X^*, then Q is the Y-marginal of $P_X^* P_{Y|X}$. Applying the greedy rejection sampling scheme on $P = P_{Y|X}(\cdot|x)$ and Q, we have

$$\begin{aligned}\mathbb{E}\left[\log_2 K \,\middle|\, X = x\right] &\le D_{\mathrm{KL}}(P_{Y|X}(\cdot|x)\,\|\,Q) + \log_2(2e)\\ &\le C + \log_2(2e).\end{aligned} \tag{3.11}$$

We encode K into M using a Shannon code (Shannon, 1948) constructed for the Zipf distribution $\mathrm{Zipf}(1 + 1/(C + \log_2(2e)))$. The expected length can be bounded by one plus the cross entropy between the distribution of K and the Zipf distribution, which is bounded in Appendix A. □

Implementation details. The pseudocode for the greedy rejection sampling algorithm (Harsha *et al.*, 2010; Liu and Verdú, 2018; Flamich and Theis, 2023) is given in Algorithm 1. The inputs to the algorithms are Q (the reference distribution, taken to be P_Y for the channel simulation setting), the Radon-Nikodym derivative $g(y) := (\mathrm{d}P/\mathrm{d}Q)(y)$ of the desired distribution P with respect to Q (for the channel simulation setting, take $P = P_{Y|X}(\cdot|X)$), an estimate of the channel capacity C (for the arbitrary source case; for known source distribution, take $C = I(X;Y)$; C is not needed if the Elias delta code (Elias, 1975) is used instead of the Shannon code), and a pseudorandom number generator (PRNG) $\mathfrak{G}$ synchronized between the encoder and the decoder (see Remark 10). The expected number of samples needed by the encoding and decoding algorithms is $O(2^{D_\infty(P\|Q)}) = O(\operatorname{ess\,sup}_{y\in\mathcal{Y}} \frac{\mathrm{d}P}{\mathrm{d}Q}(y))$ (Flamich and Theis, 2023).[8] Refer to Section 3.5 for more discussions on the sample complexity.

[8] Theorem 7 states that $\mathbb{E}[\log_2 K] \le D_{\mathrm{KL}}(P\|Q) + \log_2(2e)$. One might expect $\mathbb{E}[K] \lesssim 2^{D_{\mathrm{KL}}(P\|Q)}$ by taking exponential on both sides. This does not work since 2^t is convex instead of concave, and we instead

The pseudocode assumes that the Shannon code (Shannon, 1948) is used for the encoding of the index k. Other codes for positive integers such as the Elias delta code (Elias, 1975) can also be used. For more choices of codes, refer to Appendix A.

Remark 9. Algorithm 1 assumes that the expectation in Step 9 (corresponding to (3.9)) can be computed precisely. If $\mathcal{Y}$ is finite, this can be computed as a summation. However, if $\mathcal{Y}$ is continuous, this computation may or may not be feasible depending on P and Q. In Sections 3.3 and 3.4, we will study other channel simulation schemes that does not require computation of expectation or integral.

Remark 10. For the common randomness, the encoder should keep a PRNG $\mathfrak{G}_1$, and the decoder should keep a PRNG $\mathfrak{G}_2$. These two PRNGs should be synchronized at the beginning of the scheme (e.g., they are initialized by the same seed). The encoder and decoder must "behave responsibly" (see Section 2.3), meaning that $\mathfrak{G}_1$ and $\mathfrak{G}_2$ must also be synchronized when the scheme finishes, so the PRNGs can be reused for other tasks. In Algorithm 1, the encoder and the decoder use the PRNG the same number of times, and hence they are still synchronized after the algorithm. In practice, the encoder only needs to generate the random seed and send it to the decoder once, so that they can initialize their PRNGs in sync, and then use the PRNGs to generate common randomness for all the channel simulation tasks that follow.

Remark 11. Some PRNG algorithms allow fast "jumping ahead" to the state after n uses of the PRNG, i.e., advancing the state of the PRNG as if n random numbers are generated. For example, linear congruential generators and permuted congruential generators allow jumping ahead n steps in $O(\log n)$ time (O'Neill, 2014). For another class of PRNGs that allow jumping, counter-based pseudorandom number generators (Salmon *et al.*, 2011) are PRNGs where the internal state is an integer, and the state is incremented each time a random number is generated. A counter-based PRNG allows jumping to the state after n uses of the PRNG, simply by increasing the state by n. Therefore, in the decoding algorithm in Algorithm 1, if the number of calls to the PRNG per sample $y \sim Q$ is fixed, then we can jump to the state after the $k^* - 1$ rejected samples are generated in $O(1)$ time, without having to actually generate those $k^* - 1$ rejected samples, reducing the decoding time complexity from $O(k^*)$ to $O(1)$.[9] The use of counter-based PRNG in channel simulation

have $\mathbb{E}[K] = O(2^{D_\infty(P\|Q)})$ since the expectation is dominated by values of y with large $\frac{\mathrm{d}P}{\mathrm{d}Q}(y)$ that requires searching through a large number of samples to find.

[9] In the case where each sample $y \sim Q$ is generated using a variable number of calls to the PRNG (e.g., y is generated using rejection sampling), the decoder cannot jump to the state after $k^* - 1$ samples are generated. In this case, the encoder should instead send the cumulative number of calls to the PRNG when the k^*-th sample is generated (i.e., the state of the PRNG at that time, minus the initial state of the PRNG), so the decoder can increment the state of the PRNG by that number. As long as the expected number of calls per sample is finite, this only incurs a constant penalty on the encoding length.

Algorithm 1 Greedy rejection sampling (Harsha *et al.*, 2010) (also see (Flamich and Theis, 2023))

Procedure ENCODE$(Q, g, C, \mathfrak{G})$:

Input: distribution Q, density $g(y) := (\mathrm{d}P/\mathrm{d}Q)(y)$,
capacity C, PRNG $\mathfrak{G}$

Output: description $M \in \{0,1\}^*$

$s \leftarrow 1$ ▷ *the s_k in (3.9)*
$a \leftarrow 0$ ▷ *stores $\sum_{j=1}^{k-1} s_j$*
for $k = 1, 2, \ldots$ **do**
  Generate $y \sim Q$ using $\mathfrak{G}$
  $\rho \leftarrow \min\{\max\{\frac{1}{s}(g(y) - a), 0\}, 1\}$
  with probability ρ (using local randomness)
    return Shannon encoding of k for $\mathrm{Zipf}(1 + 1/(C + \log_2(2e)))$
    ▷ *see Appendix A for other codes*
  end with
  $s' \leftarrow s - \mathbb{E}_{\bar{Y}\sim Q}[\min\{\max\{g(\bar{Y}) - a, 0\}, s\}]$ ▷ *see Remark 9*
  $a \leftarrow a + s$
  $s \leftarrow s'$
end for

Procedure DECODE$(Q, C, M, \mathfrak{G})$:

Input: Q, C, $M \in \{0,1\}^*$, $\mathfrak{G}$

Output: sample Y

Decode M to k^* using Shannon code for $\mathrm{Zipf}(1 + 1/(C + \log_2(2e)))$
for $k = 1, 2, \ldots, k^*$ **do** ▷ *may jump the PRNG instead;*
  Generate $y \sim Q$ using $\mathfrak{G}$ ▷ *see Remark 11*
end for
return y

has been noted by Liu *et al.* (2024).

3.2.3 Other Variants of Rejection Sampling

Several other variants of the rejection sampling algorithm have been studied.

Braverman-Garg scheme for compressing the index (D/1/E/VL/KAS/UCR). The classical rejection sampling scheme uses a common randomness $W = (\bar{Y}_i)_{i\in\mathbb{N}^+}$, $\bar{Y}_1, \bar{Y}_2, \ldots \overset{iid}{\sim} Q$, and a local randomness $U_1, U_2, \ldots \overset{iid}{\sim} \mathrm{Unif}(0,1)$ at the encoder. Since common randomness is unlimited, we may move $(U_i)_i$ to the common randomness, and consider the common randomness to be $W = (\bar{Y}_i, U_i)_{i\in\mathbb{N}^+}$ instead. The advantage is that now the encoder can encode K conditional on $(\bar{Y}_i, U_i)_i$ since the decoder also knows $(\bar{Y}_i, U_i)_i$. In (Braverman and Garg, 2014), for a finite output space $\mathcal{Y}$, a rejection sampling scheme with a carefully designed encoding function of K given $(\bar{Y}_i, U_i)_i$ yields the following bound on the conditional entropy

$$H(Y|W) \le I(X;Y) + \log_2\left(I(X;Y)+1\right) + O(1),$$

where $O(1)$ denotes an absolute constant, similar to the bound for greedy rejection sampling (Corollary 8). Interested readers are referred to (Braverman and Garg, 2014) for details.

Generic 1-bit protocol (G/1/A/FL/KAS/UCR). The classical rejection sampling scheme, with the index K as the description, does not preserve privacy, in the sense that $P_{(\bar{Y}_i)_i,K|X}$ may not be differentially private (Definition 1) even when $P_{Y|X}$ is differentially private.[10] Intuitively, each iteration of rejection sampling is a measurement on the input X. Even if each measurement is noisy, taking a large number of independent noisy measurements may still reveal significant information.

To make the scheme differentially private, the idea of the *generic 1-bit protocol (1-bit-PROT)* by Bassily and Smith (2015) is to perform only *one* iteration of rejection sampling (3.3). If the encoder rejects the first sample, it immediately declares failure, i.e., we take $K = 1$ if $U_1 \le \gamma \frac{\mathrm{d}P_{Y|X}(\cdot|X)}{\mathrm{d}Q}(\bar{Y})$, and $K = 0$ (indicating failure) otherwise. The decoder receives K and outputs $\tilde{Y} = \bar{Y}$ if $K = 1$, and $\tilde{Y} = \mathrm{e}$ (a special erasure symbol) if $K = 0$. The erasure symbol $\tilde{Y} = \mathrm{e}$ indicates failure.

[10] Consider $X \sim \mathrm{Bern}(1/2)$, $Y|X \sim \mathrm{Bern}(a + (1-2a)X)$ for $0 < a < 1/2$ (i.e., $P_{Y|X}$ is a binary symmetric channel). This channel is $\ln((1-a)/a)$-locally differentially private. To apply classical rejection sampling, we take $Q = \mathrm{Bern}(1/2)$, and $0 < \gamma \le \frac{1}{2(1-a)}$ which ensures $\gamma \frac{\mathrm{d}P_{Y|X}(\cdot|x)}{\mathrm{d}Q}(y) \le 1$. However, $\mathbb{P}(\bar{Y}_1 = \cdots = \bar{Y}_n = 0, K > n \,|\, X = 0) = (1/2 - \gamma(1-a))^n$,which has an arbitrarily large ratio from $\mathbb{P}(\bar{Y}_1 = \cdots = \bar{Y}_n = 0, K > n \,|\, X = 1) = (1/2 - \gamma a)^n$ for large n, implying that rejection sampling is not ϵ-differentially private for any $\epsilon > 0$.

Technically, $\tilde{Y}$ does not follow the conditional distribution $P_{Y|X}$, but an "erased version" $P_{\tilde{Y}|X}$, where $\tilde{Y}$ is generated from $P_{Y|X}$ with probability γ, or $\tilde{Y} = \mathrm{e}$ with probability $1-\gamma$. This makes the scheme more suitable for the situation where there are a large number of n users, where user i has the data X_i and wants to convey a noisy version Y_i distributed according to a privacy-preserving randomizer $P_{Y|X}$ to the server (Bassily and Smith, 2015). The goal of the server is not to gain information about individual users, but rather to obtain an aggregate statistic using $Y_1, \dots, Y_n$ (e.g., the mean $n^{-1}\sum_{i=1}^{n} Y_n$). Suppose that instead of having Y_i following $P_{Y|X}$, the server now has the erased versions $\tilde{Y}_i$ following $P_{\tilde{Y}|X}$. The server will still be able to compute the statistic using the non-erased entries among $\tilde{Y}_1, \dots, \tilde{Y}_n$ (e.g., the mean can be estimated as $|\{i : \tilde{Y}_i \neq \mathrm{e}\}|^{-1}\sum_{i:\,\tilde{Y}_i \neq \mathrm{e}} \tilde{Y}_i$). The advantage of 1-bit-PROT is that only one bit of communication per user is needed to convey the K's, and the local differential privacy (Definition 1) of $P_{Y|X}$ is preserved (under certain conditions).

Approximate rejection sampling (G/1/A/FL/KAS/UCR). An approximate rejection sampling scheme where the distribution requirement on Y is approximate instead of exact were studied in (Block and Polyanskiy, 2023), where a fixed number of samples $\bar{Y}_1, \dots, \bar{Y}_n \overset{iid}{\sim} Q$ are used, and the encoder sends an arbitrary description if all n samples are rejected. To increase the acceptance probability and reduce the number of samples needed, instead of performing rejection sampling on the target distribution P, it is performed on the distribution of $Y \sim P$ conditional on the event that $\gamma_0 \frac{\mathrm{d}P}{\mathrm{d}Q}(Y) \leq 1$, where γ_0 is an appropriately chosen constant. Due to the two sources of error (having only a finite number of samples, and modifying the target distribution), this scheme can only simulate the target distribution approximately, though it can be carried out using a fixed-length description $K \in \{1, \dots, n\}$. Refer to (Block and Polyanskiy, 2023) for various upper and lower bounds on the number of samples needed. Also refer to (Flamich and Wells, 2024) for an improvement on the scheme in (Block and Polyanskiy, 2023).

Greedy rejection coding with partition process (G/1/E/VL/KAS/UCR). In (Flamich *et al.*, 2024), an additional partitioning step was introduced to the greedy rejection sampling algorithm, which can improve the running time. The idea (similar to Flamich *et al.* (2022)) is to partition the output space $\mathcal{Y}$ in a hierarchical manner, so as to allow faster searching of the accepted sample. It is suitable when $\mathrm{d}P/\mathrm{d}Q$ is unimodal.

More variants of the rejection sampling scheme will be discussed in Section 3.5.3. In the next section, we present a different but related scheme, which is used to prove Theorem 4.

3.3 Exponential and Poisson Functional Representation

3.3.1 Exponential Functional Representation for Discrete Output

In classical and greedy rejection sampling, we generate a sequence of samples $\bar{Y}_1, \bar{Y}_2, \ldots$, and select one sample as the output Y. The index i of $\bar{Y}_i$ can be regarded as the "time" of the sample. Samples with smaller time are more likely to be chosen. This property is useful for variable-length channel simulation schemes, since we can assign a shorter description to samples with a smaller time.

In this section, we will describe another channel simulation scheme for discrete $\mathcal{Y}$. Unlike the rejection sampling scheme, here all values in $\mathcal{Y}$ are regarded as "samples", and the time of a sample is drawn as an exponential random variable in an i.i.d. manner. The exponential distribution with rate r, denoted as $\mathrm{Exp}(r)$, is the continuous distribution with probability density function re^{-rz} for $z \geq 0$. If $Z \sim \mathrm{Exp}(r)$, then $Z/a \sim \mathrm{Exp}(ar)$ for $a > 0$. An elementary property of exponential random variables is that if we have two independent exponential random variables $Z_1 \sim \mathrm{Exp}(r_1)$, $Z_2 \sim \mathrm{Exp}(r_2)$, then $\min\{Z_1, Z_2\} \sim \mathrm{Exp}(r_1 + r_2)$, and $\mathbb{P}(Z_1 < Z_2) = r_1/(r_1 + r_2)$. Generalizing this to n exponential random variables, if we have a sequence $Z_1, \ldots, Z_n$ of independent random variables where $Z_i \sim \mathrm{Exp}(r_i)$, $r_i > 0$, then for any $a_1, \ldots, a_n \geq 0$ that are not all zeros,

$$\mathbb{P}\left(\operatorname*{argmin}_i \frac{Z_i}{a_i/r_i} = y\right) = \frac{a_y}{\sum_{i=1}^n a_i}, \tag{3.12}$$

where we assume $Z_i/(a_i/r_i) = \infty$ if $a_i = 0$. If $(a_1, \ldots, a_n)$ is a probability vector, then $\operatorname{argmin}_i Z_i/(a_i/r_i)$ gives a sample of the probability distribution. This provides a method to generate samples from a discrete probability distribution. This method (after taking logarithm) is known as the Gumbel-max trick (Huijben *et al.*, 2022), which is often used in machine learning.[11]

It was observed in (Li and El Gamal, 2018b) that (3.12) gives a scheme for the simulation of a channel with discrete output, which is called the *exponential functional representation* scheme. Suppose we want to simulate the discrete channel $P_{Y|X}$ with input distribution P_X, where $Y \in [n]$ is a finite discrete random variable. The common randomness is taken to be $W = (Z_1, \ldots, Z_n)$, which is a sequence of independent random variables where $Z_y \sim \mathrm{Exp}(Q(y))$, where Q is some distribution over $\mathcal{Y}$, called the *reference distribution*. Upon observing X, the encoder produces

$$Y = \operatorname*{argmin}_y \frac{Z_y}{P(y)/Q(y)}, \tag{3.13}$$

[11] One benefit of the Gumbel-max trick (Huijben *et al.*, 2022) is that the sequence $a_1, \ldots, a_n$ does not need to add up to 1. We can generate a sample from the normalized distribution by scanning through the sequence $a_1, \ldots, a_n$ only once, without the need to calculate $\sum_{i=1}^n a_i$ beforehand.

where $P(y) := P_{Y|X}(y|X)$, computes the rank K of Z_Y among $Z_1, \dots, Z_n$ (i.e., $K = 1$ if Z_Y is the smallest among $Z_1, \dots, Z_n$, $K = 2$ if Z_Y is the second smallest, etc.), and transmit K to the decoder. The decoder can then find the entry Z_Y which is ranked the K-th among $Z_1, \dots, Z_n$, and output its index Y. This scheme guarantees that Y follows the distribution $P(y) = P_{Y|X}(y|X)$, and hence the channel $P_{Y|X}$ is simulated.[12] The smaller Z_y's are more likely to be chosen in the argmin in (3.13). This makes K likely to be small, and require a short codeword if we encode K using a prefix-free code over positive integers (e.g., Elias delta code (Elias, 1975)). Here Z_y can be regarded as the "time" of the value y, and we are more likely to choose values that are seen "earlier". Refer to Figure 3.6 for an illustration. We defer the performance analysis to the next section.

3.3.2 Poisson Functional Representation for General Output

The exponential functional representation applies only to a discrete Y. In this subsection, we will present a generalization via Poisson processes which applies to general Y.

Informal limiting argument for the Poisson process. We first motivate the use of Poisson processes using an informal limiting argument. If $Y \in \mathbb{R}$ and Q is a continuous distribution, a natural attempt is to apply the exponential functional representation (3.13) on the quantized $\hat{Y}_\Delta = \Delta\lfloor Y/\Delta + 1/2\rfloor \in \Delta\mathbb{Z}$ and take the quantization step $\Delta \to 0$. Write $\hat{P}_\Delta$ for the distribution of $\hat{Y}_\Delta$ when $Y \sim P$, and $\hat{Q}_\Delta$ for the distribution of $\hat{Y}_\Delta$ when $Y \sim Q$, and let $\hat{Z}_{\hat{y}} \sim \mathrm{Exp}(\hat{Q}_\Delta(\hat{y}))$. The rule (3.13) becomes $\operatorname*{argmin}_{\hat{y}\in\Delta\mathbb{Z}} \frac{\hat{Z}_{\hat{y}}}{\hat{P}_\Delta(\hat{y})/\hat{Q}_\Delta(\hat{y})}$ which, when $\Delta \to 0$, intuitively "approaches" $\operatorname*{argmin}_{y\in\mathbb{R}} \frac{Z_y}{(\mathrm{d}P/\mathrm{d}Q)(y)}$ where $(\mathrm{d}P/\mathrm{d}Q)(y)$ is the Radon-Nikodym derivative (when P, Q are continuous, it is the ratio between their probability density functions). The remaining issue is the distribution of Z_y, which should be the "limit" of the stochastic process $\hat{Z}_{\hat{y}} \sim \mathrm{Exp}(\hat{Q}_\Delta(\hat{y}))$. If we plot the points $(\hat{y}, \hat{Z}_{\hat{y}}) \in \mathbb{R}^2$, these points lie on the lines $\hat{y} = \Delta k$ for $k \in \mathbb{Z}$, where the spacing Δ between the lines decreases to 0. Nevertheless, the "density of points" on each line $\hat{y} = \Delta k$ approaches 0 since each line only contains one point $\hat{Z}_{\hat{y}} \sim \mathrm{Exp}(\hat{Q}_\Delta(\hat{y}))$ with a mean $1/\hat{Q}_\Delta(\hat{y}) \to \infty$ as $\Delta \to 0$. Therefore, it is reasonable to expect the points $(\hat{y}, \hat{Z}_{\hat{y}})_{\hat{y}\in\Delta\mathbb{Z}}$ to have a limiting distribution.

To study this distribution, we order the points in ascending order of their $\hat{Z}$-coordinates. Let $(\bar{Y}_i, T_i)_{i\in\mathbb{N}^+}$ be the points $(\hat{y}, \hat{Z}_{\hat{y}})_{\hat{y}\in\Delta\mathbb{Z}}$ ordered according to $T_1 \le T_2 \le \cdots$.[13] Think

[12] Exponential functional representation can be regarded as a channel simulation scheme where the encoder observes X and wants the decoder to generate Y following $P = P_{Y|X}(\cdot|X)$, or equivalently, a remote generation scheme where the encoder observes a distribution P and wants the decoder to generate Y following P. Refer to the discussions after Definition 2.

[13] For a channel simulation scheme that utilizes this ordering idea without the limiting argument, refer to ordered random coding (Theis and Yosri, 2022) in Section 3.5.3.

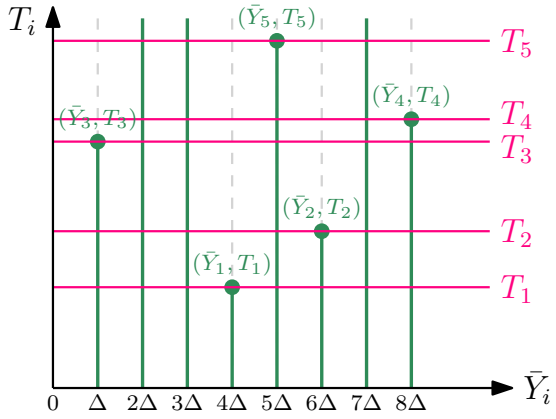


Figure 3.5: An illustration of the points $(\hat{y}, \hat{Z}_{\hat{y}})$ where $\hat{y} \in \Delta\mathbb{Z}$, $\hat{Z}_{\hat{y}} \sim \text{Exp}(\hat{Q}_\Delta(\hat{y}))$, sorted in ascending order of the second coordinate to become $(\bar{Y}_i, T_i)$ with $T_1 \le T_2 \le \cdots$. The arrival time along each vertical line at $\hat{y}$ is $\text{Exp}(\hat{Q}_\Delta(\hat{y}))$. At the beginning, we are waiting for any one of all the lines, and the waiting time is $T_1 \sim \text{Exp}(\sum_{\hat{y}} \hat{Q}_\Delta(\hat{y})) = \text{Exp}(1)$. After the first arrival, we are waiting for all except one line, and hence $T_2 - T_1$ should be approximately $\text{Exp}(1)$ as well.

of T_i as the arrival time of the point $(\bar{Y}_i, T_i)$. The first arrival time is $T_1 = \min_{\hat{y}} \hat{Z}_{\hat{y}} \sim \text{Exp}(\sum_{\hat{y}} \hat{Q}_\Delta(\hat{y})) = \text{Exp}(1)$, and we have $\bar{Y}_1 \sim \hat{Q}_\Delta$ due to the property of exponential random variables discussed in Section 3.3.1. We then study the next inter-arrival time $T_2 - T_1$. Condition on the event that the first point is $(\bar{Y}_1, T_1) = (\bar{y}_1, t_1)$. We know that we will not see another point with $\bar{Y}_i = \bar{y}_1$. Nevertheless, for another line $\hat{y} = \Delta k$ with $\Delta k \neq \bar{y}_1$, we know that the point on this line has not arrived yet by time t, and hence by the memoryless property of exponential distribution, the waiting time $\hat{Z}_{\hat{y}} - t \sim \text{Exp}(\hat{Q}_\Delta(\hat{y}))$ still follows the same exponential distribution. This gives $T_2 - T_1 \sim \text{Exp}(\sum_{\hat{y} \neq \bar{y}_1} \hat{Q}_\Delta(\hat{y})) = \text{Exp}(1 - \hat{Q}_\Delta(\bar{y}_1)) \approx \text{Exp}(1)$ conditioned on $(\bar{Y}_1, T_1) = (\bar{y}_1, t_1)$. We can see that the situation between time T_1 and time T_2 is the same as the situation between time 0 and time T_1, except that we no longer have points on the line $\hat{y} = \bar{y}_1$, which is inconsequential since this is merely one out of many lines. Hence, $T_2 - T_1$ should also approximately follow $\text{Exp}(1)$, and $\bar{Y}_2$ should approximately follow $\hat{Q}_\Delta \approx Q$, independent of $(\bar{Y}_1, T_1)$. Continuing this argument, we can see that T_1, $T_2 - T_1$, $T_3 - T_2, \ldots$ are approximately i.i.d. following $\text{Exp}(1)$, and $\bar{Y}_1, \bar{Y}_2, \ldots$ are approximately i.i.d. following Q. In other words, $(\bar{Y}_i, T_i)_i$ is approximately an i.i.d. process and a Poisson process put together. Refer to Figure 3.5 for an illustration.

Definition of the Poisson functional representation. We now formally define the Poisson process construction without the aforementioned informal limiting argument. We first discuss the case where Y is discrete. Recall that a Poisson process with rate r is a stochastic process $T_1 \le T_2 \le \cdots$ with i.i.d. inter-arrival times following $\text{Exp}(r)$, i.e.,

$T_1, T_2 - T_1, T_3 - T_2, \ldots \overset{iid}{\sim} \mathrm{Exp}(r)$. We write this as $(T_i)_{i\in\mathbb{N}^+} \sim \mathrm{PP}(r)$. Letting $(T_{y,i})_{i\in\mathbb{N}^+} \sim \mathrm{PP}(Q(y))$, independent across $y \in \mathcal{Y}$, and noting that $\min_i T_{y,i} = T_{y,1} \sim \mathrm{Exp}(Q(y))$, (3.13) can be equivalently written as

$$Y = \operatorname*{argmin}_y \min_i \frac{T_{y,i}}{P_{Y|X}(y|X)/Q(y)}. \tag{3.14}$$

We now merge the points $T_{y,i}$ into a single Poisson process. To this end, we utilize the splitting property of Poisson process (Last and Penrose, 2017), which states that if we consider the points in a Poisson process with rate r, where for each point, we randomly assign a class to it, with a probability $Q(y)$ of assigning it to class y (where Q is a discrete distribution), then the points in class y form a Poisson process with rate $rQ(y)$, and these Poisson processes for different y's are independent of each other. In other words, if $(T_i)_i \sim \mathrm{PP}(r)$, and $(\bar{Y}_i)_i \overset{iid}{\sim} Q$ is an i.i.d. sequence following a discrete distribution Q independent of $(T_i)_i$, then $(T_i)_{i:\,\bar{Y}_i=y} \sim \mathrm{PP}(rQ(y))$ (i.e., selecting only the points T_i where $\bar{Y}_i = y$) are independent across $y \in \mathcal{Y}$. Therefore, letting $(T_i)_i \sim \mathrm{PP}(1)$, and $(\bar{Y}_i)_i \overset{iid}{\sim} Q$ independent of $(T_i)_i$, (3.14) can be equivalently written as

$$Y = \bar{Y}_K, \text{ where } K := \mathrm{argmin}_i \frac{T_i}{P_{Y|X}(\bar{Y}_i|X)/Q(\bar{Y}_i)}. \tag{3.15}$$

Note that (3.15) does not only hold for discrete Q and $P_{Y|X}(\cdot|X)$, but also for general Q and $P_{Y|X}(\cdot|X)$. In this case, we have

$$Y = \bar{Y}_K, \text{ where } K := \mathrm{argmin}_i \frac{T_i}{g(\bar{Y}_i|X)}, \tag{3.16}$$

where

$$g(y|x) := \frac{\mathrm{d}P_{Y|X}(\cdot|x)}{\mathrm{d}Q}(y)$$

is the Radon-Nikodym derivative. This way, we can guarantee that $Y|X \sim P_{Y|X}$. Refer to Figure 3.6 for an illustration. The property that this construction yields $Y|X \sim P_{Y|X}$ has been observed by Maddison (2016) in a different context of simulating random samples in Monte Carlo simulations. Refer to the work on *A* sampling* by Maddison *et al.* (2014) for applications of this property to machine learning. This is also the key ingredient of the Poisson functional representation scheme for channel simulation by Li and El Gamal (2018b), which is utilized in the proof of Theorem 4.

Recall that in greedy rejection sampling (Section 3.2.2), each sample $\bar{Y}_i$ has a time $i \in \mathbb{N}^+$ which is a positive integer. In the Poisson functional representation, the time of the sample $\bar{Y}_i$ is T_i, where the times $(T_i)_i$ are generated according to a Poisson process. In greedy rejection sampling, the encoder does not only look at the samples $\bar{Y}_i$ and their times to determine the accepted sample; it also requires additional local randomness $U_i \sim \mathrm{Unif}(0,1)$ at the

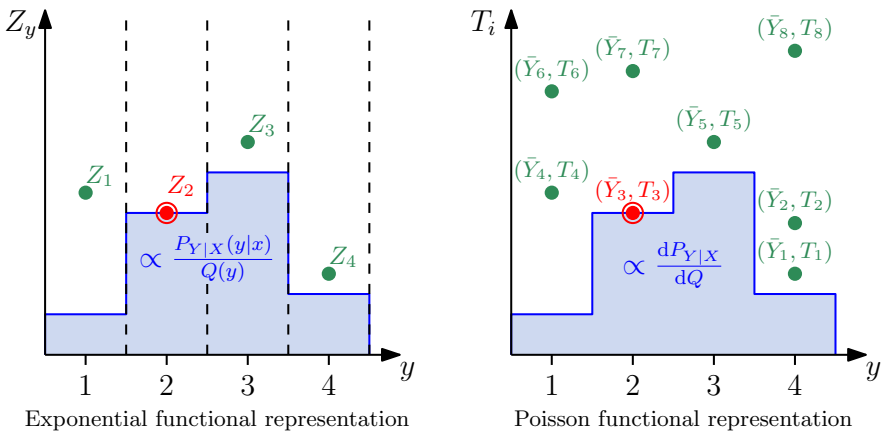


Figure 3.6: Left: an illustration of the exponential functional representation, where $\mathcal{Y} = \{1, 2, 3, 4\}$. For the argmin in (3.13), note that $\min_y Z_y/(P_{Y|X}(y|x)/Q(y)) = \min\{\gamma > 0 : \exists y : Z_y = \gamma P_{Y|X}(y|x)/Q(y)\}$, and hence the argmin in (3.13) can be regarded as having a shape $\gamma P_{Y|X}(y|x)/Q(y)$ (blue shape in the figure), which we keep "inflating" this shape by increasing γ until it hits the first point Z_y, and the point it hits corresponds to the Y selected in (3.13) ($Y = 2$ in the figure).
Right: an illustration of the Poisson functional representation. We can convert the exponential functional representation to the Poisson functional representation by having a Poisson process of points for each value of y (instead of only one exponential random variable Z_y), and then merging the points into a single Poisson point process over the 2D plane. We again have a shape $\gamma \frac{\mathrm{d}P_{Y|X}(\cdot|x)}{\mathrm{d}Q}(y)$ (blue shape in the figure), which we keep "inflating" by increasing γ until it hits the first point $(\bar{Y}_i, T_i)$, and the point it hits corresponds to the Y selected in (3.16) ($\bar{Y}_K = 2$ in the figure).

encoder. In the Poisson functional representation, the selection rule (3.16) depends only on the samples $\bar{Y}_i$ and their times T_i, which are parts of the common randomness. Therefore, a major difference between rejection sampling and Poisson functional representation is that Poisson functional representation moves the randomness needed for the selection of sample to the times T_i, which is part of the common randomness.

A point process of pairs $(\bar{Y}_i, T_i)$. Another way to view the Poisson functional representation construction is to treat $(\bar{Y}_i, T_i)_i$ as a point process over the space $\mathcal{Y} \times [0, \infty)$. To this end, we utilize a more general definition of Poisson process over a general measure space (Last and Penrose, 2017). A (proper) Poisson process with intensity measure μ (an s-finite measure) over a measurable space is a random collection of points $(X_i)_{i \in [N]}$ (where the number of points $N \in \mathbb{N}_0 \cup \{\infty\}$ is random and can be countably infinite) such that for any measurable set B in the measurable space, the number of points in B follows a Poisson distribution with rate $\mu(B)$, i.e.,

$$|\{i : X_i \in B\}| \sim \mathrm{Poi}\left(\mu(B)\right),$$

and also for disjoint $B_1, \ldots, B_n$, the numbers of points $|\{i : X_i \in B_j\}|$ are independent across $j \in [n]$, i.e., the points at different locations are independent of each other. We write this as $(X_i)_i \sim \mathrm{PP}(\mu)$.

Note that the Poisson process over $[0, \infty)$ with rate 1 is a Poisson process with intensity measure $\lambda_{[0,\infty)}$, which denotes the Lebesgue measure over $[0, \infty)$. Since $(\bar{Y}_i)_i$ is generated independently from $(T_i)_i$, it follows from the marking theorem (Last and Penrose, 2017) that $(\bar{Y}_i, T_i)_i \sim \mathrm{PP}(Q \times \lambda_{[0,\infty)})$ is a Poisson point process with intensity measure $Q \times \lambda_{[0,\infty)}$ (the product measure between Q and $\lambda_{[0,\infty)}$). We can think of $\bar{Y}_i$ as an i.i.d. generated label attached to the point T_i.

If $(X_i)_i$ is a Poisson process, then for any measurable function f, the mapped points $(f(X_i))_i$ is a Poisson process as well.[14] Since we are interested in generating Y following $P_{Y|X}(\cdot|x)$ instead of Q, our goal is to apply a suitable mapping on $(\bar{Y}_i, T_i)_i$ so that the new intensity measure is $P_{Y|X}(\cdot|x) \times \lambda_{[0,\infty)}$, which means that the first coordinates of the points in the new process will be i.i.d. following $P_{Y|X}(\cdot|x)$. Note that if $(T_i)_i$ is a Poisson process over $[0, \infty)$ with rate r, then $(aT_i)_i$ is a Poisson process over $[0, \infty)$ with rate r/a for $a > 0$. Therefore, we can apply a larger scaling factor on T_i if we want to make the value of $\bar{Y}_i$ more rare. Loosely speaking, if we apply a scaling factor $1/g(y|x) = 1/(\mathrm{d}P_{Y|X}(\cdot|x)/\mathrm{d}Q)(y)$ to the points with $\bar{Y}_i = y$, then we can change the "intensity" of those points from $Q(y)$ to $P_{Y|X}(y|x)$. Indeed, the points

$$(\bar{Y}_i, \tilde{T}_i)_{i:\, g(\bar{Y}_i|x) \neq 0}, \text{ where } \tilde{T}_i := \frac{T_i}{g(\bar{Y}_i|x)}, \tag{3.17}$$

[14] This is a consequence of the mapping theorem (Last and Penrose, 2017).

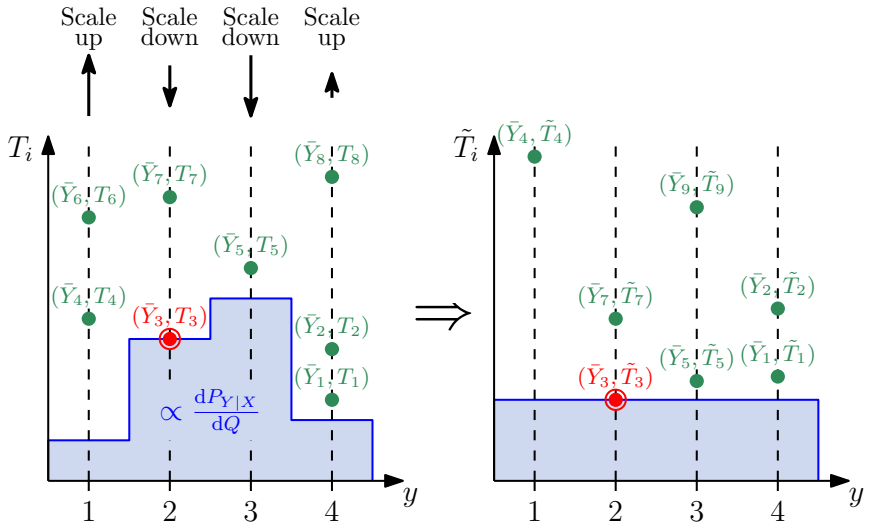


Figure 3.7: Left: a Poisson process $(\bar{Y}_i, T_i)_i \sim \mathrm{PP}(Q \times \lambda_{[0,\infty)})$. Right: a Poisson process $(\bar{Y}_i, \tilde{T}_i)_{i:\, g(\bar{Y}_i)\neq 0} \sim \mathrm{PP}(P \times \lambda_{[0,\infty)})$, obtained by scaling the time coordinate of the points in the original Poisson process by $\tilde{T}_i := T_i / g(\bar{Y}_i)$, $g(y) := (\mathrm{d}P/\mathrm{d}Q)(y)$. Basically, we look at all points in the original process on the vertical line $\bar{Y}_i = y$, and scale the points up or down by a factor $1/g(y)$. Such a scaling will affect the rate of the points on the line $\bar{Y}_i = y$ by a factor of $g(y)$, and hence the new process will have an intensity measure $Qg \times \lambda_{[0,\infty)} = P \times \lambda_{[0,\infty)}$.

form a Poisson process with intensity measure $P_{Y|X}(\cdot|x) \times \lambda_{[0,\infty)}$, which follows from the mapping theorem (Last and Penrose, 2017). The process $(\bar{Y}_i, \tilde{T}_i)_i$ can be formed by attaching i.i.d. labels following $P_{Y|X}(\cdot|x)$ to a Poisson process over $[0, \infty)$ with rate 1 (up to reordering of the points). Therefore, for the point $(\bar{Y}_i, \tilde{T}_i)$ with the smallest $\tilde{T}_i$, its value of $\bar{Y}_i$ will follow $P_{Y|X}(\cdot|x)$. Refer to Figure 3.7 for an illustration.

3.3.3 Analyses of the Communication Cost

To prove Theorem 4, we utilize the following result in Li and Anantharam, 2021, Eqn. (29). The bound (3.19) is referred to as the *Poisson matching lemma* in (Li and Anantharam, 2021).

Lemma 12 (Poisson functional representation (Li and Anantharam, 2021))**.** *Fix two distributions $P \ll Q$,*[15] *and let $g(y) := (\mathrm{d}P/\mathrm{d}Q)(y)$. Let $(T_i)_{i\in\mathbb{N}^+}$, $0 \le T_1 \le T_2 \le \cdots$ be a Poisson point process with rate 1, $(\bar{Y}_i)_{i\in\mathbb{N}^+} \overset{iid}{\sim} Q$ independent of $(T_i)_{i\in\mathbb{N}^+}$, and*

$$Y = \bar{Y}_K, \text{ where } K := \mathrm{argmin}_i \frac{T_i}{g(\bar{Y}_i)},$$

[15] $P \ll Q$ means P is absolutely continuous with respect to Q, which is necessary for $\mathrm{d}P/\mathrm{d}Q$ to be defined.

where $T_i/g(\bar{Y}_i) = \infty$ *if* $g(\bar{Y}_i) = 0$. *Then* $Y \sim P$, *and* $K \in \mathbb{N}^+$ *is conditionally a geometric random variable given* Y, *with*

$$K|Y \sim \text{Geom}\left(\left(\mathbb{E}_{Y' \sim Q}[\max\{g(Y),\, g(Y')\} \,|\, Y] \right)^{-1} \right).$$

As a result,[16]

$$\mathbb{E}[\log_2 K] \le D_{\text{KL}}(P\|Q) + 1 \text{ bit}, \tag{3.18}$$

and

$$\mathbb{P}(K > 1 \,|\, Y) \le 1 - \frac{1}{g(Y)+1} \le g(Y) \tag{3.19}$$

almost surely.

Proof. The following arguments are based on (Li and Anantharam, 2021).[17] We now condition on the event $(\bar{Y}_K, T_K) = (y_0, t_0)$, i.e., the point $(\bar{Y}_i, T_i)_i$ with the smallest $T_i/g(\bar{Y}_i)$ is $(\bar{Y}_K, T_K) = (y_0, t_0)$. Since a point $(\bar{Y}_i, T_i)$ with $g(\bar{Y}_i) = 0$ will never be selected as $(\bar{Y}_K, T_K)$ (it gives $T_i/g(\bar{Y}_i) = \infty$), we can assume $g(y_0) \neq 0$. The event $(\bar{Y}_K, T_K) = (y_0, t_0)$ means two things: 1) there is no point in the region $\{(y,t) : t/g(y) < t_0/g(y_0)\}$, and 2) there exists a point (y_0, t_0). Since the points in a Poisson process at different locations are independent of each other, our knowledge about the two regions $\{(y,t) : t/g(y) < t_0/g(y_0)\}$ and $\{(y_0,t_0)\}$ does not affect the distribution of points outside of the two regions, which are $(\bar{Y}_i, T_i)_{i \neq K}$. Let $\mathcal{S} := \{(y,t) : t/g(y) \ge t_0/g(y_0)\}$. Conditional on $(\bar{Y}_K, T_K) = (y_0, t_0)$, we know that the remaining points $(\bar{Y}_i, T_i)_{i \neq K} \sim \text{PP}((Q \times \lambda_{[0,\infty)})|_{\mathcal{S}})$ follows a Poisson process with intensity measure $(Q \times \lambda_{[0,\infty)})|_{\mathcal{S}}$, which denotes the measure $Q \times \lambda_{[0,\infty)}$ restricted to the set $\mathcal{S}$, i.e., $(Q \times \lambda_{[0,\infty)})|_{\mathcal{S}}(\mathcal{A}) = (Q \times \lambda_{[0,\infty)})(\mathcal{A} \cap \mathcal{S})$.

Recall that K is the rank of the point $(\bar{Y}_K, T_K)$ among $(\bar{Y}_i, T_i)_i$ when ordered in increasing order of T_i. Therefore

$$K = |\{i \neq K : T_i < T_K\}| + 1.$$

Conditional on $(\bar{Y}_K, T_K) = (y_0, t_0)$, we have $(\bar{Y}_i, T_i)_{i \neq K} \sim \text{PP}((Q \times \lambda_{[0,\infty)})|_{\mathcal{S}})$, and the number of points with $T_i < T_K = t_0$ follows a Poisson distribution with rate given by the $(Q \times \lambda_{[0,\infty)})|_{\mathcal{S}}$-measure of the set $\{(y,t) : t < t_0\}$, which is

$$\int_{\mathcal{Y}} \lambda_{[0,\infty)} \left(t : t < t_0,\, t/g(y) \ge t_0/g(y_0) \right) Q(\mathrm{d}y)$$

[16] It was proved in Li and Anantharam, 2021, Prop. 4 that $\mathbb{E}[\log_2 K] \le D_{\text{KL}}(P\|Q) + \log_2 e$. This was slightly improved in (Li, 2024) to $\mathbb{E}[\log_2 K] \le D_{\text{KL}}(P\|Q) + 1$.

[17] A more general result where K is the index of the j-th smallest $T_i/g(\bar{Y}_i)$ was proved in Li and Anantharam, 2021, Appendix A. The proof here is based on specializing (Li and Anantharam, 2021) to the case $j = 1$. Readers may also refer to (Khisti *et al.*, 2024) for another proof of (3.19).

$$
\begin{aligned}
&= \int_{\mathcal{Y}} \max\left\{t_0 - \frac{t_0 g(y)}{g(y_0)},\, 0\right\} Q(\mathrm{d}y) \\
&= \frac{t_0}{g(y_0)} \int_{\mathcal{Y}} \max\left\{g(y_0) - g(y),\, 0\right\} Q(\mathrm{d}y) \\
&= \frac{t_0}{g(y_0)} \mathbb{E}_{Y' \sim Q}\left[\max\left\{g(y_0) - g(Y'),\, 0\right\}\right] \\
&= \alpha t_0, \qquad (3.20)
\end{aligned}
$$

where $\alpha := \mathbb{E}_{Y' \sim Q}[\max\{g(y_0) - g(Y'), 0\}]/g(y_0)$. Hence,

$$
(K-1) \,|\, \{(\bar{Y}_K, T_K) = (y_0, t_0)\} \;\sim\; \mathrm{Poi}(\alpha t_0). \qquad (3.21)
$$

We now consider the distribution of T_K. Let $\tilde{T}_i := T_i / g(\bar{Y}_i)$. We have seen that $(\bar{Y}_i, \tilde{T}_i)_{i:\, g(\bar{Y}_i) \neq 0} \sim \mathrm{PP}(P \times \lambda_{[0,\infty)})$, which can be formed by attaching i.i.d. labels following P to a Poisson process over $[0, \infty)$ with rate 1. Therefore, $\tilde{T}_K = \min_{i:\, g(\bar{Y}_i) \neq 0} \tilde{T}_i \sim \mathrm{Exp}(1)$ is an exponential random variable, independent of $\bar{Y}_K$ which is an independent label attached to $\tilde{T}_K$. Conditional on $\bar{Y}_K = y_0$, we still have $\tilde{T}_K \sim \mathrm{Exp}(1)$, and hence $T_K = \tilde{T}_K g(y_0) \sim \mathrm{Exp}(1/g(y_0))$. Therefore, the conditional distribution of $K - 1$ conditional on $\bar{Y}_K = y_0$ is an exponential mixture of the Poisson distributions in (3.21), i.e., we have $T_K \sim \mathrm{Exp}(1/g(y_0))$ and $(K-1)|T_K \sim \mathrm{Poi}(\alpha T_K)$. The resultant distribution of K is the geometric distribution with parameter[18]

$$
\begin{aligned}
&\frac{1}{1 + \alpha g(y_0)} \\
&= \frac{1}{1 + \mathbb{E}_{Y' \sim Q}\left[\max\left\{g(y_0) - g(Y'),\, 0\right\}\right]} \\
&= \frac{1}{\mathbb{E}_{Y' \sim Q}\left[\max\left\{g(y_0),\, g(Y')\right\}\right]},
\end{aligned}
$$

where the last line is because $\mathbb{E}_{Y' \sim Q}[g(Y')] = \int (\mathrm{d}P/\mathrm{d}Q)(y) Q(\mathrm{d}y) = \int P(\mathrm{d}y) = 1$. Therefore, by Jensen's inequality,[19]

$$
\begin{aligned}
\mathbb{E}[\log_2 K] &\leq \mathbb{E}\left[\log_2 \mathbb{E}[K|Y]\right] \\
&= \mathbb{E}\left[\log_2 \mathbb{E}_{Y' \sim Q}\big[\max\{g(Y),\, g(Y')\} \,\big|\, Y\big]\right] \\
&\leq \mathbb{E}\left[\log_2 \mathbb{E}_{Y' \sim Q}\big[g(Y) + g(Y') \,\big|\, Y\big]\right]
\end{aligned}
$$

[18] There is an intuitive reason why an exponential mixture of Poisson distributions is a geometric distribution with support $\mathbb{N}_0$. Fix $\lambda, \alpha > 0$. Let $(T_i)_i \sim \mathrm{PP}(\lambda+\alpha)$ labeled with $(W_i)_i \overset{iid}{\sim} \mathrm{Bern}(\alpha/(\lambda+\alpha))$, and $K := \min\{i : W_i = 0\}$. Then $K \sim \mathrm{Geom}(\lambda/(\lambda+\alpha))$. Also, since $(T_i)_{i:\, W_i = 0} \sim \mathrm{PP}(\lambda)$ and $(T_i)_{i:\, W_i = 1} \sim \mathrm{PP}(\alpha)$ are two independent Poisson processes, we have $T_K = \min_{i:\, W_i = 0} T_i \sim \mathrm{Exp}(\lambda)$, and $K - 1 = |\{i : W_i = 1,\, T_i < T_K\}| \sim \mathrm{Poi}(\alpha t)$ conditional on $T_K = t$. Hence if $T \sim \mathrm{Exp}(\lambda)$ and $(K-1)|T \sim \mathrm{Poi}(\alpha T)$, then $K \sim \mathrm{Geom}(\lambda/(\lambda+\alpha)) = \mathrm{Geom}(1/(1+\alpha/\lambda))$.

[19] This step appeared in (Li, 2024).

$$
\begin{aligned}
&= \mathbb{E}\left[\log_2\left(g(Y)+1\right)\right] \\
&= \mathbb{E}\left[\log_2 g(Y)\right] + \mathbb{E}\left[\log_2\left(1+\frac{1}{g(Y)}\right)\right] \\
&\le \mathbb{E}\left[\log_2 g(Y)\right] + \log_2\left(1+\mathbb{E}\left[\frac{1}{g(Y)}\right]\right) \\
&\le \mathbb{E}\left[\log_2 g(Y)\right] + \log_2 2,
\end{aligned}
$$

where the last line is because $\mathbb{E}[1/g(Y)] = \int 1/((\mathrm{d}P/\mathrm{d}Q)(y))P(\mathrm{d}y) \le 1$. Hence, (3.18) follows from $\mathbb{E}\left[\log_2 g(Y)\right] = D_{\mathrm{KL}}(P\|Q)$. For (3.19),

$$
\begin{aligned}
\mathbb{P}(K>1\,|\,Y) &= 1 - \left(\mathbb{E}_{Y'\sim Q}\left[\max\{g(Y),\, g(Y')\}\,|\,Y\right]\right)^{-1} \\
&\le 1 - \left(\mathbb{E}_{Y'\sim Q}\left[g(Y)+g(Y')\,|\,Y\right]\right)^{-1} \\
&= 1 - \frac{1}{g(Y)+1}.
\end{aligned}
$$

□

We remark that Lemma 12 can also be used to prove several other results, including various multi-terminal source and channel coding results in (Li and Anantharam, 2021), and the pairwise optimal coupling results in (Angel and Spinka, 2019; Li and Anantharam, 2019), though we will not discuss these applications here.

We now prove the "known source distribution" case in Theorem 4, which is the consequence of Proposition 5 and the following result called the *strong functional representation lemma* in (Li and El Gamal, 2018b). We present the version in (Li, 2024) with a slightly better constant than (Li and El Gamal, 2018b).

Theorem 13 (Strong functional representation lemma (G/1/E/VL/KAS/UCR) (Li and El Gamal, 2018b; Li, 2024))**.** *For any P_X and $P_{Y|X}$, there exists a functional representation (P_W, ϕ) (i.e., $\phi : \mathcal{W}\times\mathcal{X}\to\mathcal{Y}$ satisfying that if $X\sim P_X$ is independent of $W\sim P_W$, and $Y=\phi(W,X)$, then $(X,Y)\sim P_X P_{Y|X}$) such that*

$$
H(Y|W) \le I(X;Y) + \log_2\left(I(X;Y)+2\right) + 2 \text{ bits}.
$$

Moreover, if $\mathcal{X},\mathcal{Y}$ are finite, then we can have $|\mathcal{W}| \le |\mathcal{X}|(|\mathcal{Y}|-1)+2$.

Proof. Let $Q=P_Y$ be the Y-marginal of $P_X P_{Y|X}$. Let $(T_i)_i \sim \mathrm{PP}(1)$, $(\bar{Y}_i)_i \overset{iid}{\sim} P_Y$ independent of $(T_i)_i$, and $Y=\bar{Y}_K$ where $K := \mathrm{argmin}_i T_i/g(\bar{Y}_i|X)$, and

$$
g(y|x) := \frac{\mathrm{d}P_{Y|X}(\cdot|x)}{\mathrm{d}P_Y}(y).
$$

By Lemma 12,

$$\mathbb{E}\left[\log_2 K \,|\, X = x\right] \le D_{\mathrm{KL}}(P_{Y|X}(\cdot|x) \,\|\, P_Y) + 1. \tag{3.22}$$

Taking expectation over X gives

$$\mathbb{E}[\log_2 K] \le I(X;Y) + 1. \tag{3.23}$$

Take the common randomness to be $W = (\bar{Y}_i, T_i)_i$. To bound the conditional entropy $H(Y|W)$, note that Y is a function of (K, W), and hence $H(Y|W) \le H(K|W) \le H(K)$. The bound on $H(K)$ can be bounded via the cross entropy between the distribution of K and the Zipf distribution $\mathrm{Zipf}(1 + 1/(I(X;Y)+1))$, which is given in Appendix A. Please refer to (Li and El Gamal, 2018b) for the cardinality bound $|\mathcal{Z}| \le |\mathcal{X}|(|\mathcal{Y}| - 1) + 2$. □

To prove the "arbitrary source" case in Theorem 4, we use the same arguments as in the proof of Corollary 8. The proof is similar and is omitted.

Although the Poisson functional representation shares some similarities with rejection sampling, there is one fundamental difference. Rejection sampling is *causal* (Liu and Verdú, 2018), in the sense that one can read the samples $\tilde{Y}_i$ one by one, and decide when to stop and output the current $\tilde{Y}_i$ as Y without looking at future $\tilde{Y}_i$'s. Poisson functional representation is *noncausal* in the sense that in order to decide whether to output the current $\tilde{Y}_i$, we may need to look at the $\tilde{Y}_i$'s with a larger time. Refer to Section 3.5 for more discussions and bounds on causal and noncausal sampling schemes.

3.3.4 Implementation Considerations

The pseudocode for the Poisson functional representation (Li and El Gamal, 2018b) (also see (Maddison, 2016; Theis and Yosri, 2022)) is given in Algorithm 2. The inputs to the algorithms are Q (the reference distribution, taken to be P_Y for the channel simulation setting), the Radon-Nikodym derivative $g(y) := (\mathrm{d}P/\mathrm{d}Q)(y)$ of the desired distribution P with respect to Q (for the channel simulation setting, take $P = P_{Y|X}(\cdot|X)$), an upper-bound $g^* \ge \sup_y g(y)$,[20] an estimate of the channel capacity C (for the arbitrary source case; for known source distribution, take $C = I(X;Y)$; C is not needed if the Elias delta code (Elias, 1975) is used instead of the Shannon code), and a pseudorandom number generator (PRNG) $\mathfrak{G}$. The expected number of samples needed by the encoding and decoding algorithms is $O(2^{D_\infty(P\|Q)}) = O(\operatorname{ess\,sup}_{y\in\mathcal{Y}} \frac{\mathrm{d}P}{\mathrm{d}Q}(y))$ (Maddison, 2016). Refer to Section 3.5 for more discussions on the sample complexity, and to Appendix A for various methods for the encoding of the index k^*.

[20] g^* is used for detecting when we can stop generating the Poisson process in Algorithm 2. This technique has appeared in (Maddison, 2016).

Algorithm 2 Poisson functional representation (Li and El Gamal, 2018b; Maddison, 2016) (also see (Theis and Yosri, 2022))

Procedure ENCODE$(Q, g, g^*, C, \mathfrak{G})$:

Input: distribution Q, density $g(y) := (\mathrm{d}P/\mathrm{d}Q)(y)$, bound $g^* \geq \sup_y g(y)$, capacity C, PRNG $\mathfrak{G}$

Output: description $M \in \{0,1\}^*$

Initialize PRNG $\mathfrak{G}'$ using a seed generated by $\mathfrak{G}$
▷ *it is preferrable to jump the PRNG instead; see Remark 14*
$t \leftarrow 0$, $\tilde{t}^* \leftarrow \infty$, $k^* \leftarrow 0$
for $k = 1, 2, \ldots$ **do**
  $t \leftarrow t + \mathrm{Exp}(1)$ ▷ $\mathrm{Exp}(1)$ *is a new exponential random variate generated using local randomness (not* $\mathfrak{G}, \mathfrak{G}'$*)*
  if $t/g^* \geq \tilde{t}^*$ **then** ▷ *no new points can have* $t/g(y) < \tilde{t}^*$
    return Shannon encoding of k^* for $\mathrm{Zipf}(1 + 1/(C+1))$
    ▷ *see Appendix A for other codes*
  end if
  Generate $y \sim Q$ using $\mathfrak{G}'$
  $\tilde{t} \leftarrow t/g(y)$
  if $\tilde{t} < \tilde{t}^*$ **then**
    $\tilde{t}^* \leftarrow \tilde{t}$ ▷ *update minimizer of* $t/g(y)$
    $k^* \leftarrow k$
  end if
end for

Procedure DECODE$(Q, C, M, \mathfrak{G})$:

Input: Q, C, $M \in \{0,1\}^*$, $\mathfrak{G}$

Output: sample Y

Initialize PRNG $\mathfrak{G}'$ using a seed generated by $\mathfrak{G}$
▷ *it is preferrable to jump the PRNG instead; see Remark 14*
Decode M to k^* using Shannon code for $\mathrm{Zipf}(1 + 1/(C+1))$
for $k = 1, 2, \ldots, k^*$ **do** ▷ *may jump the PRNG instead;*
  Generate $y \sim Q$ using $\mathfrak{G}'$ ▷ *see Remark 11*
end for
return y

Remark 14. For the common randomness in Algorithm 2, the encoder should keep a PRNG $\mathfrak{G}_1$, and the decoder should keep a PRNG $\mathfrak{G}_2$, where the two PRNGs are assumed to be synchronized at the beginning of the scheme, and must still be synchronized after the scheme so the PRNGs can be reused for other tasks. The need of synchronization is the reason why the encoder and the decoder need to initialize another PRNG $\mathfrak{G}'$ with a seed generated using $\mathfrak{G}_1, \mathfrak{G}_2$, and then use $\mathfrak{G}'$ instead of $\mathfrak{G}_1, \mathfrak{G}_2$ in the algorithm. This way, $\mathfrak{G}_1, \mathfrak{G}_2$ can be synchronized since they are used once in both the encoding and the decoding function. If we ignore this step and use $\mathfrak{G}_1, \mathfrak{G}_2$ directly, then $\mathfrak{G}_1$ may be invoked a larger number of times than $\mathfrak{G}_2$ since k^* may be less than the number of samples $y \sim Q$ generated by the encoder. This is a minor nuisance of noncausal sampling schemes, which is not needed for causal sampling schemes (e.g., Algorithm 1). See Section 2.3 for the importance of synchronization, and Section 3.5 for discussions on causal and noncausal sampling schemes.

The aforementioned approach to use the existing PRNG to initialize a new PRNG relies on the assumption that the random numbers generated by the new PRNG are independent of the future random numbers generated by the existing PRNG. This may or may not be a reasonable assumption depending on the PRNG used. A better approach is to instead perform the following at the beginning of the encoding and decoding algorithms: 1) create $\mathfrak{G}'$, the PRNG to be used in the algorithm, as a (deep) copy of the existing PRNG $\mathfrak{G}$ (i.e., $\mathfrak{G}'$ has the same state as $\mathfrak{G}$); and 2) jump $\mathfrak{G}$ ahead a large number of steps, say 2^{50} (i.e., update the state of $\mathfrak{G}$ to be as if 2^{50} random numbers are generated). This way, we can ensure that the number generated by $\mathfrak{G}'$ will not overlap the future numbers generated by $\mathfrak{G}$, unless we call $\mathfrak{G}'$ more than 2^{50} times which takes an extremely long time. There are fast methods to jump a PRNG (Haramoto *et al.*, 2008a; Haramoto *et al.*, 2008b; O'Neill, 2014), which are implemented in popular libraries (e.g., (NumPy Developers, 2024)). Also, it is straightforward to jump counter-based pseudorandom number generators (Salmon *et al.*, 2011) (see Remark 11).

3.3.5 Variants of Poisson Functional Representation

We discuss several variants of the Poisson functional representation and the A* sampling (Maddison *et al.*, 2014; Maddison, 2016) below.

Space partitioning (G/1/E/VL/KAS/UCR). A shortcoming of Algorithm 2 is that it has an exponential encoding time complexity $O(2^{D_\infty(P\|Q)}) = O(\sup_{y\in\mathcal{Y}} \frac{\mathrm{d}P}{\mathrm{d}Q}(y))$ (Maddison, 2016), which can be prohibitive for large $D_\infty(P\|Q)$. More efficient algorithms can be designed if we impose additional assumptions on the distributions. For example, Flamich *et al.* (2022) proposed several methods based on A* sampling (Maddison *et al.*, 2014; Maddison, 2016), namely AS*, AD* and DAD* coding, for the situation where $Y \in \mathbb{R}$ is a scalar.

The real line $\mathbb{R}$ is hierarchically partitioned as a binary tree. The algorithm then searches the tree to locate the sample Y to be communicated. Under the assumption that $\frac{\mathrm{d}P}{\mathrm{d}Q}(y)$ is unimodal, AS* coding achieves a linear time complexity $O(D_\infty(P\|Q))$. Nevertheless, the expected description length of AS* coding has a multiplicative gap from the optimum. On the other hand, the expected description length of AD* coding has a logarithmic gap from the optimum, but it is not proved (only conjectured) that AD* coding has a time complexity $O(D_\infty(P\|Q))$ (Flamich *et al.*, 2022). Also refer to (Flamich, 2023) for a scheme that also utilizes a partition of the real line, and to (He *et al.*, 2024b) for a modification of the Poisson functional representation, which utilizes a partition of the space to quickly search for samples lying in the support of the target distribution.

Poisson private representation (G/1/E/VL/KAS/UCR). The exponential and Poisson functional representations do not preserve differential privacy.[21] The reason is that Y is a deterministic function of X and the common randomness. To ensure privacy, the encoder should be stochastic. In the *Poisson private representation* by (Liu *et al.*, 2024), the encoder also applies (3.17) to transform the points points $(\bar{Y}_i, T_i)$ (a Poisson process with intensity measure $Q \times \lambda_{[0,\infty)}$) by scaling the time

$$\tilde{T}_i := \frac{T_i}{g(\bar{Y}_i|x)},$$

where $g(y|x) = (\mathrm{d}P_{Y|X}(\cdot|x)/\mathrm{d}Q)(y)$, to form a Poisson process $(\bar{Y}_i, \tilde{T}_i)$ with intensity measure $P_{Y|X}(\cdot|x) \times \lambda_{[0,\infty)}$. Now, instead of choosing the point with the smallest $\tilde{T}_i$, the encoder chooses the point $(\bar{Y}_K, \tilde{T}_K)$ randomly with

$$\mathbb{P}(K = k) = \frac{\tilde{T}_k^{-\alpha}}{\sum_{i=1}^\infty \tilde{T}_i^{-\alpha}},$$

where $\alpha > 1$ is a parameter. This ensures $\bar{Y}_K \sim P_{Y|X}(\cdot|x)$ exactly. Note that only $(\bar{Y}_i)_i$ is shared in the common randomness. The arrival times $(T_i)_i$, as well as the randomness in the generation of K above, are local randomness of the encoder. It was shown by Liu *et al.* (2024) that

$$\mathbb{E}[\log_2 K] \le D(P_{Y|X}(\cdot|x)\|Q) + \frac{\log_2 3.56}{\min\{(\alpha-1)/2, 1\}},$$

and if $P_{Z|X}$ is (ϵ, δ)-locally differentially private (Definition 1), then $P_{(\bar{Y}_i)_i, K|X}$ (where $(\bar{Y}_i)_i, K$ are the decoder's observations) is $(2\alpha\epsilon, 2\delta)$-locally differentially private. Refer to (Liu *et al.*, 2024) for other results on the differential privacy of Poisson private representation.

[21] Consider $X \sim \mathrm{Bern}(1/2)$ and $Y|X \sim \mathrm{Bern}(a + (1-2a)X)$ (i.e., a binary symmetric channel) with $0 < a < 1/2$. This channel is $\ln((1-a)/a)$-locally differentially private. Take $Q = \mathrm{Bern}(1/2)$ in the exponential functional representation. If $Z_0 = Z_1$, then $Y = \mathrm{argmin}_y Z_y/(P_{Y|X}(y|X)/Q(y))$ will be equal to X. If the decoder observes $Z_0 = Z_1$, it can know the precise value of X.

More variants of the Poisson functional representation scheme will be discussed in Section 3.5.3.

3.4 Likelihood Encoder and Minimal Random Coding

Importance sampling (Kloek and Van Dijk, 1978; Liu, 2004) is a technique that allows us to estimate the mean of a function of a distribution when we can only access samples from another distribution. Suppose we are given i.i.d. samples $\bar{Y}_1, \dots, \bar{Y}_{\mathsf{N}}$ following the *reference distribution* Q. If we are interested in estimating $\mathbb{E}_{Y\sim P}[f(Y)]$ for a different *target distribution* P, we can give an estimate

$$\sum_{i=1}^{\mathsf{N}} f(\bar{Y}_i) \frac{\frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_i)}{\sum_{j=1}^{\mathsf{N}} \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_j)},$$

where $\mathrm{d}P/\mathrm{d}Q$ is the Radon-Nikodym derivative. The idea is that, while $\bar{Y}_1, \dots, \bar{Y}_{\mathsf{N}}$ are samples of Q, we can transform it into "weighted samples" of P by attaching an importance weight $\frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_i)$ to $\bar{Y}_i$.

This is the basis of the *minimal random coding* scheme (Havasi *et al.*, 2019) for channel simulation with arbitrary source. Suppose a reference distribution Q over $\mathcal{Y}$ is known to both the encoder and the decoder. The encoder and decoder have common randomness $\bar{Y}_1, \dots, \bar{Y}_{\mathsf{N}} \overset{iid}{\sim} Q$. Now the encoder observes the input X, and wants to convey a sample $\bar{Y}$ with distribution $P = P_{Y|X}(\cdot|X)$ to the decoder.[22] To this end, the encoder perform the following steps:

1. Compute the importance weights $\alpha_i := \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_i)$ of each sample.
2. Generate a random number $K \in [\mathsf{N}]$ with $\mathbb{P}(K = k) = \frac{\alpha_k}{\sum_{i=1}^{\mathsf{N}} \alpha_i}$.
3. Encode K into $\lceil \log_2 \mathsf{N} \rceil$ bits and transmit it.

The decoder simply decodes K and outputs $\tilde{Y} = \bar{Y}_K$. For the implementation, we can employ the Gumbel-max trick or the exponential-min trick in Section 3.3.1 to draw the random number K without having to store the whole sequences of $\bar{Y}_i$'s and α_i's, which has been noted by Theis and Yosri (2022). Refer to Algorithm 3 for the pseudocode, and refer to Figure 3.8 for an illustration.

[22] Minimal random coding can be regarded as a channel simulation scheme where the encoder observes X and wants the decoder to generate Y following $P = P_{Y|X}(\cdot|X)$, or equivalently, a remote generation scheme where the encoder observes a distribution P and wants the decoder to generate Y following P. Refer to the discussions after Definition 2.

Algorithm 3 Minimal random coding and likelihood encoder (Havasi *et al.*, 2019; Cuff, 2013) (also see (Theis and Yosri, 2022))

Procedure ENCODE$(Q, g, \mathsf{N}, \mathfrak{G})$:

Input: distribution Q, density $g(y) := (\mathrm{d}P/\mathrm{d}Q)(y)$, sample size N, PRNG $\mathfrak{G}$

Output: index $k \in [\mathsf{N}]$

$t^* \leftarrow \infty$
for $i = 1, \ldots, \mathsf{N}$ **do**
  Generate $y \sim Q$ using $\mathfrak{G}$
  $t \leftarrow \mathrm{Exp}(1)/g(y)$ ▷ Exp(1) *is a new exponential random variate*
    ▷ *generated using local randomness (not* $\mathfrak{G}$*)*
  if $t < t^*$ **then**
    $t^* \leftarrow t$
    $y^* \leftarrow y$
    $k \leftarrow i$
  end if
end for
return k

Procedure DECODE$(Q, \mathsf{N}, k, \mathfrak{G})$:

Input: Q, N, $k \in [\mathsf{N}]$, $\mathfrak{G}$

Output: sample Y

for $i = 1, \ldots, \mathsf{N}$ **do** ▷ *may jump the PRNG instead;*
  Generate $y \sim Q$ using $\mathfrak{G}$ ▷ *see Remark 11*
  if $i = k$ **then**
    $y^* \leftarrow y$
  end if
end for
return y^*

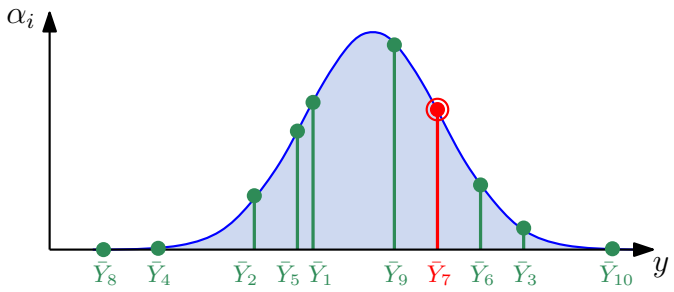


Figure 3.8: An illustration of the minimal random coding scheme, applied on Q being the uniform distribution over an interval, and $P = P_{Y|X}(\cdot|x)$ being a Gaussian distribution (truncated so it fits within the interval), shown as the blue shape in the figure. Each sample point in $\bar{Y}_1, \dots, \bar{Y}_{\mathsf{N}} \overset{iid}{\sim} Q$ is selected with a probability proportional to $\alpha_i = \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_i)$, indicated by the length of the stem at $\bar{Y}_i$ in the figure. Even though the points $\bar{Y}_i$ are generated from Q, since we are more likely to select a point $\bar{Y}_i$ with a large $\frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_i)$, the selected point can have a distribution close to P.

If the source distribution P_X is known, then we can take $Q = P_Y$ to be the Y-marginal of $P_X P_{Y|X}$. If X, Y are discrete, then

$$\alpha_i = \frac{P_{Y|X}(\bar{Y}_i|X)}{P_Y(\bar{Y}_i)} = \frac{P_{X|Y}(X|\bar{Y}_i)}{P_X(X)}$$

is proportional to the likelihood $P_{X|Y}(X|\bar{Y}_i)$. In this case, the above procedure is known as the *likelihood encoder* (Cuff, 2013; Watanabe *et al.*, 2015; Song *et al.*, 2016).

Note the similarity between this scheme and the random coding scheme for lossy compression (Cover and Thomas, 2006), which also involves a random codebook $\bar{Y}_1, \dots, \bar{Y}_{\mathsf{N}} \overset{iid}{\sim} Q$. The difference is that, for lossy compression, we select the $\bar{Y}_i$ that is closest to the input X. For likelihood encoder/minimal random coding, we instead use a stochastic encoder to select $\bar{Y}_i$ in a random manner. Such randomness is unnecessary for lossy compression since the goal is only to minimize the distortion between the input and the output, but can be helpful for channel simulation since it helps shaping the output into the desired distribution.

Approximation guarantees of minimal random coding. This scheme is approximate in the sense that the distribution of $\tilde{Y}$ is not exactly $P = P_{Y|X}(\cdot|X)$. This is the reason we denote the output as $\tilde{Y}$ instead of Y since $P_{\tilde{Y}|X}$ (the actual conditional distribution of the output) is not exactly $P_{Y|X}$. Nevertheless, it is straightforward to check that the distribution of $\tilde{Y}$ approaches P as $\mathsf{N} \to \infty$. In (Havasi *et al.*, 2019), it was shown that $\mathsf{N} \approx 2^{D_{\mathrm{KL}}(P\|Q)}$ suffices for a good approximation for a fixed $P = P_{Y|X}(\cdot|x)$ (the situation where X is not fixed will be discussed later in this section), where D_{KL} denotes the Kullback-Leibler divergence. Therefore, the encoding length of the index $K \in [\mathsf{N}]$ is approximately

$\log_2 \mathsf{N} \approx D_{\mathrm{KL}}(P\|Q)$. We state Lemma D.1 of Theis and Yosri (2022), which was proved using Theorem 3.2 of Havasi *et al.* (2019) (which in turn was proved by invoking Theorem 1.2 of Chatterjee and Diaconis (2018)).

Theorem 15 (Minimal random coding (G/1/A/FL/KAS/UCR) (Theis and Yosri, 2022))**.** *Consider any fixed reference distribution Q and target distribution $P = P_{Y|X}(\cdot|x) \ll Q$. Assume $\mathsf{N} = 2^{D_{\mathrm{KL}}(P\|Q)+t}$ for some $t \geq 0$. The distribution $\tilde{P}$ of the output $\tilde{Y}$ of minimal random coding satisfies*

$$\delta_{\mathrm{TV}}(\tilde{P}, P) \leq 4\sqrt{2^{-t/4} + 2\sqrt{\mathbb{P}\left(\iota_{P\|Q}(Y) > D_{\mathrm{KL}}(P\|Q) + \frac{t}{2}\right)}},$$

where $Y \sim P$, $\iota_{P\|Q}(Y) := \log_2 \frac{\mathrm{d}P}{\mathrm{d}Q}(Y)$, and $\delta_{\mathrm{TV}}(P_{\tilde{Y}}, P) := \sup_{E\subseteq\mathcal{Y}} |\tilde{P}(E) - P(E)|$ is the total variation distance (Section 5.3).

A small $\delta_{\mathrm{TV}}(\tilde{P}, P)$ means that $\tilde{Y}$ is approximately distributed as P. Readers are referred to Section 5.3 for discussions on δ_{TV}. Note that $\mathbb{E}[\iota_{P\|Q}(Y)] = D_{\mathrm{KL}}(P\|Q)$. In order to guarantee a small $\delta_{\mathrm{TV}}(\tilde{P}, P)$ when $\mathsf{N} \approx 2^{D_{\mathrm{KL}}(P\|Q)}$, we need $\iota_{P\|Q}(Y)$ to be concentrated around its mean $D_{\mathrm{KL}}(P\|Q)$, so the probability that $\iota_{P\|Q}(Y) > D_{\mathrm{KL}}(P\|Q) + t/2$ is small. We will also include an analysis of the likelihood encoder using the techniques in (Cuff, 2013; Yassaee, 2015) later in Sections 5.6 and 8.2, where we discuss fixed-length approximate schemes in detail.

A downside of minimal random coding is that the sample size N must be exponential in $D_{\mathrm{KL}}(P\|Q)$ similar to greedy rejection sampling and Poisson functional representation, leading to an exponential time complexity. Refer to Section 3.5 for more discussions on the sample complexity. Techniques for alleviating this downside were investigated in (Flamich *et al.*, 2020; Flamich *et al.*, 2022).

Also, note that if our goal is to simulate a channel $P_{Y|X}$ where the input X is not fixed, then we apply minimal random coding with the target distribution $P = P_{Y|X}(\cdot|X)$ that depends on X. In order to guarantee that the output $\tilde{Y}$ approximately follows $P_{Y|X}$ for most input X, we require a description length $\log_2 \mathsf{N} \geq D_{\mathrm{KL}}(P_{Y|X}(\cdot|X)\|Q)$ with high probability for $X \sim P_X$ (in the known source distribution case in Definition 2). Assuming that we take $Q = P_Y$ to be the Y-marginal of $P_X P_{Y|X}$, we have to select a description length $\log_2 \mathsf{N}$ large enough to accommodate the largest possible values of $D_{\mathrm{KL}}(P_{Y|X}(\cdot|x)\|P_Y)$ (and also the variability of $\iota_{P_{Y|X}(\cdot|x)\|P_Y}(Y)$ as in Theorem 15), and hence $\log_2 \mathsf{N}$ will generally have to be larger than $I(X;Y) = \mathbb{E}_{X\sim P_X}[D_{\mathrm{KL}}(P_{Y|X}(\cdot|X)\|P_Y)]$. This is a downside of using a fixed-length code to encode K, instead of a variable-length code as in greedy rejection sampling and Poisson functional representation that can adapt to the value of X. In Section 3.5.3, we will discuss ordered random coding (Theis and Yosri, 2022), which uses a variable-length

code to compress K in the minimal random coding in order to reduce the description length to close to $I(X;Y)$.

Differential privacy of minimal random coding. A nice property of minimal random coding is that it guarantees differential privacy (Definition 1) as long as the original channel $P_{Y|X}$ is differentially private, albeit with a two-fold increase of the privacy budget (Shah *et al.*, 2022).

Theorem 16 (ϵ-DP of minimal random coding (Shah *et al.*, 2022))**.** *If $P_{Y|X}$ is ϵ-locally differentially private, then minimal random coding $P_{(\bar{Y}_i)_i,K|X}$ (with any reference distribution Q) is 2ϵ-locally differentially private.*

Proof. We use similar arguments as Shah *et al.* (2022), restricted to the case where Y is discrete for the sake of simplicity. Consider any $x_1, x_2 \in \mathcal{X}$. By Definition 1, we have $e^{-\epsilon} P_{Y|X}(y|x_2) \le P_{Y|X}(y|x_1) \le e^{\epsilon} P_{Y|X}(y|x_2)$ for all $y \in \mathcal{Y}$. For every $(\bar{y}_i)_{i\in[\mathsf{N}]} \in \mathcal{Y}^{\mathsf{N}}$ and $k \in [\mathsf{N}]$,

$$
\begin{aligned}
&P_{(\bar{Y}_i)_i,K|X}((\bar{y}_i)_i, k|x_1) \\
&= P_{(\bar{Y}_i)_i}((\bar{y}_i)_i) P_{K|(\bar{Y}_i)_i,X}(k|(\bar{y}_i)_i, x_1) \\
&= P_{(\bar{Y}_i)_i}((\bar{y}_i)_i) \frac{P_{Y|X}(\bar{y}_k|x_1)/Q(\bar{y}_k)}{\sum_{i=1}^{\mathsf{N}} P_{Y|X}(\bar{y}_i|x_1)/Q(\bar{y}_i)} \\
&\le P_{(\bar{Y}_i)_i}((\bar{y}_i)_i) \frac{e^{\epsilon} P_{Y|X}(\bar{y}_k|x_2)/Q(\bar{y}_k)}{\sum_{i=1}^{\mathsf{N}} e^{-\epsilon} P_{Y|X}(\bar{y}_i|x_2)/Q(\bar{y}_i)} \\
&= e^{2\epsilon} P_{(\bar{Y}_i)_i,K|X}((\bar{y}_i)_i, k|x_2),
\end{aligned}
$$

where the first equality is because $(\bar{Y}_i)_i$ is independent of X. Therefore, $P_{(\bar{Y}_i)_i,K|X}$ is 2ϵ-locally differentially private. □

The penalty factor 2 can be reduced, but at the expense of having a small privacy leakage δ as in (ϵ, δ)-differential privacy. Interested readers are referred to (Shah *et al.*, 2022) for details. Another scheme for federated learning with differential privacy based on minimal random coding has been studied in (Triastcyn *et al.*, 2021).

3.5 Discussions on Sampling Schemes

3.5.1 Properties of Sampling Schemes

In this section, we discuss greedy rejection sampling in Section 3.2.2, exponential and Poisson functional representation in Section 3.3, and minimal random coding in Section

3.4. We first start with their similarities. All of these three schemes are *sampling schemes* (Liu and Verdú, 2018) for remote generation, which utilize an i.i.d. sequence $\bar{Y}_1, \bar{Y}_2, \ldots \overset{iid}{\sim} Q$ generated from a reference distribution Q as common randomness, with an encoder that selects an entry $\bar{Y}_K$ from the sequence and transmits the index K to the decoder, and a decoder that outputs $Y = \bar{Y}_K$ which is required to follow the target distribution P (taken to be $P_{Y|X}(\cdot|X)$ for the channel simulation setting in Definition 2) exactly or approximately.

A subclass of sampling schemes is *causal sampling schemes*, where at iteration k, the encoder is not allowed to look at future samples $\bar{Y}_{k+1}, \bar{Y}_{k+2}, \ldots$ in order to decide whether to select $K = k$, i.e., the event $K = k$ is independent of $(\bar{Y}_{k+1}, \bar{Y}_{k+2}, \ldots)$ (Liu and Verdú, 2018). Greedy rejection sampling is causal, whereas Poisson functional representation and minimal random coding are noncausal. Noncausal schemes may require extra care for the synchronization between the encoder and the decoder (see Remark 14).

All these schemes satisfy $\mathbb{E}[\log_2 K] \lesssim D_{\mathrm{KL}}(P\|Q)$ (possibly under some additional assumptions),[23] and hence requires a communication cost approximately $D_{\mathrm{KL}}(P\|Q)$ (possibly with a logarithmic gap). Schemes that achieves an approximate $D_{\mathrm{KL}}(P\|Q)$ communication cost are termed *relative entropy coding* in (Flamich *et al.*, 2020).

Sample complexity. Readers may notice in the previous sections that all these schemes have sample complexity (number of samples $\bar{Y}_1, \bar{Y}_2, \ldots$ that the encoder reads) exponential in $D_{\mathrm{KL}}(P\|Q)$ (or $D_\infty(P\|Q)$). This requirement is actually fundamental, as shown by Liu and Verdú (2018).

Theorem 17 (Liu and Verdú 2018)**.** *For any sampling scheme on $\bar{Y}_1, \bar{Y}_2, \ldots \overset{iid}{\sim} Q$ with an output distribution $\bar{Y}_K \sim P$, we have*

$$\mathbb{E}[K] \geq 2^{D_2(P\|Q)-1} \geq 2^{D_{\mathrm{KL}}(P\|Q)-1},$$

where $D_2(P\|Q) := \log_2 \mathbb{E}_{Y\sim P}[(\mathrm{d}P/\mathrm{d}Q)(Y)]$ is the order-2 Rényi divergence.

Proof. We repeat the arguments by Liu and Verdú (2018) here. First assume $\bar{Y}_i$ is discrete. We have

$$\mathbb{P}(K = k|\bar{Y}_K = y) = \frac{\mathbb{P}(K = k, \bar{Y}_k = y)}{P(y)} \leq \frac{Q(y)}{P(y)}.$$

Hence

$$\mathbb{E}[K|\bar{Y}_K = y] = \sum_{k=0}^{\infty} \mathbb{P}(K > k|\bar{Y}_K = y)$$

[23] Greedy rejection sampling has $\mathbb{E}[\log_2 K] \leq D_{\mathrm{KL}}(P\|Q) + \log_2(2e)$ (Theorem 7). Poisson functional representation has $\mathbb{E}[\log_2 K] \leq D_{\mathrm{KL}}(P\|Q) + 1$ (Theorem 12). For minimal random coding, we have $\mathbb{E}[\log_2 K] \leq \mathbb{E}[\log_2 \mathsf{N}] \lesssim D_{\mathrm{KL}}(P\|Q)$ if $\log_2 \frac{\mathrm{d}P}{\mathrm{d}Q}(Y)$ is concentrated around $D_{\mathrm{KL}}(P\|Q)$ (Theorem 15).

$$
\begin{aligned}
&\geq \sum_{k=0}^{\infty} \max\left\{1 - k\frac{Q(y)}{P(y)}, 0\right\} \\
&\geq \int_0^{\infty} \max\left\{1 - t\frac{Q(y)}{P(y)}, 0\right\} \mathrm{d}t \\
&= \frac{P(y)}{2Q(y)},
\end{aligned}
$$

and $\mathbb{E}[K] \geq 2^{-1}\mathbb{E}_{Y\sim P}[P(Y)/Q(Y)]$. If $\bar{Y}_i$ is not discrete, we discretize it by considering $\bar{Y}_i' = f(\bar{Y}_i)$ for some function f with discrete codomain, which gives $\mathbb{E}[K] \geq 2^{D_2(P'\|Q')-1}$ where P', Q' are discretized versions of P, Q, respectively. The proof is completed by taking the supremum over all such functions f. □

This lower bound can be improved to $\mathbb{E}[K] \geq 2^{D_\infty(P\|Q)}$ for causal sampling schemes (Goc and Flamich, 2024; Liu and Verdú, 2018). Another hardness result was given by Agustsson and Theis (2020), who showed that there is no polynomial-time (with respect to $D_{\mathrm{KL}}(P\|Q)$) sampling algorithm with an output distribution within a total variation distance at most $1/12$ from the target distribution P, assuming RP $\neq$ NP where RP is the class of problems with randomized polynomial-time algorithms. Also refer to (Block and Polyanskiy, 2023; Goc and Flamich, 2024; Flamich and Wells, 2024) for other lower bounds on the sample complexity of sampling schemes. This limitation makes these sampling schemes suitable only for simulating channels with small capacities (e.g., privacy-preserving channels (Bassily and Smith, 2015; Liu *et al.*, 2024), or additive noise channels with a low dimension). To improve the running time, the approach by Flamich *et al.* (2022), Flamich *et al.* (2024), and Flamich (2023) is to partition the space $\mathcal{Y}$ so as to allow more efficient searching for the $\bar{Y}_K$ that is more likely to be chosen. This approach often requires additional assumptions on P, Q (e.g., $\mathrm{d}P/\mathrm{d}Q$ is unimodal).

3.5.2 Comparison between Different Approaches to Sampling Schemes

We highlight the key differences between the three approaches:

Rejection sampling is a causal sampling scheme, where the encoder scans $\bar{Y}_1, \bar{Y}_2, \ldots$ one by one until it stops and chooses the current $\bar{Y}_K$, and sends the variable-length encoding of K (see Algorithm 1). It does not require looking into samples of larger time in order to determine whether to accept the current sample. Rejection sampling is exact in the sense that $Y \sim P$ exactly. Greedy rejection sampling achieves an expected length upper-bounded by $I(X;Y) + \log_2(I(X;Y) + \log_2(4e)) + \log_2(8e)$ in one-shot (Corollary 8). The algorithm for greedy rejection sampling requires computation of expectations (see (3.9) and Step 9 in Algorithm 1), which may or may not be feasible depending on P and Q.

Poisson functional representation is a noncausal sampling scheme, where the encoder scans $\bar{Y}_1, \bar{Y}_2, \ldots$ one by one until it stops and chooses a current or past sample $\bar{Y}_K$, and sends the variable-length encoding of K (see Algorithm 2). Poisson functional representation is exact, and achieves an expected length upper-bounded by $I(X;Y) + \log_2(I(X;Y)+2) + 3$ in one-shot (Theorem 13, the best known constant). The algorithm for Poisson functional representation (Algorithm 2) requires only a way to generate the samples $\bar{Y}_i$, the Radon-Nikodym derivative $g(y) = (\mathrm{d}P/\mathrm{d}Q)(y)$, and an upper bound $g^* \geq \sup_y g(y)$.

Minimal random coding and likelihood encoder are noncausal sampling schemes, where the encoder scans all samples $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_{\mathsf{N}}$, selects a sample $\bar{Y}_K$ (non-uniformly) at random, and sends the fixed-length encoding of K. Unlike rejection sampling and Poisson functional representation, there is no "time" attached to a sample, and all samples are treated equally. Since there are no "samples with smaller times" that are more likely to be chosen, there is no reason to use a variable-length code, and a fixed-length code should be used to encode K. Minimal random coding is approximate, in the sense that the conditional distribution of Y given X approaches $P_{Y|X}$ as $\mathsf{N} \to \infty$. Minimal random coding may or may not achieve a length close to $I(X;Y)$ in one-shot, though it achieves a rate $I(X;Y)$ in the asymptotic setting (see Sections 5.6 and 8.2). The algorithm for minimal random coding (Algorithm 3) requires only a way to generate the samples $\bar{Y}_i$ and the Radon-Nikodym derivative $g(y) = (\mathrm{d}P/\mathrm{d}Q)(y)$.

In sum, the advantage of greedy rejection sampling is its exactness and causality; the advantage of Poisson functional representation is its exactness, requiring only the computation of simple quantities ($g(y)$ and g^*) in the algorithm, and a smaller bound on the encoding length; and the advantage of minimal random coding is requiring only the computation of simple quantities ($g(y)$) in the algorithm, and a fixed number of samples, which makes fixed-length encoding applicable.

3.5.3 Other Sampling Schemes

We review some other sampling schemes studied in the literature. Some of them are combinations of the three approaches discussed in the previous sections.

Ordered random coding (G/1/E/VL/KAS/UCR). A method which is based on the Poisson functional representation and minimal random coding, called *ordered random coding*, was proposed by Theis and Yosri (2022). Recall that we may utilize the Gumbel-max trick or the exponential-min trick to select the sample $\bar{Y}_K$ in minimal random coding—we generate $\bar{Y}_1, \ldots, \bar{Y}_{\mathsf{N}} \overset{iid}{\sim} Q$, $Z_1, \ldots, Z_{\mathsf{N}} \overset{iid}{\sim} \mathrm{Exp}(1)$, and take $K = \mathrm{argmin}_i Z_i / \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_i)$ (see Algorithm 3). In ordered random coding, $Z_1, \ldots, Z_{\mathsf{N}}$ are generated using the common randomness (instead

of the local randomness at the encoder), and the pairs $(\bar{Y}_i, Z_i)$ are sorted in ascending order of Z_i, to form $(\bar{Y}_{(i)}, Z_{(i)})_{i\in[\mathsf{N}]}$ where $Z_{(1)} \leq \cdots \leq Z_{(\mathsf{N})}$.[24] After sorting, the samples $\bar{Y}_{(i)}$ with smaller indices i become more likely to be chosen by $K = \operatorname{argmin}_i Z_{(i)}/\frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_{(i)})$, and hence we can use a variable-length code on K that assigns a shorter length to smaller values in order to reduce the communication cost (compared to minimal random coding).[25]

Ordered random coding has the same conditional distribution of the output $\tilde{Y} = \bar{Y}_{(K)}$ given the input X as minimal random coding, and hence the bound in Theorem 15 also applies. Like minimal random coding, ordered random coding is approximate, and is subject to the same requirements on N as minimal random coding (see Section 3.4). Nevertheless, unlike minimal random coding where the description length $\log_2 \mathsf{N}$ increases unboundedly with N, the description length of ordered random coding does not grow unboundedly as N increases. In fact, ordered random coding approaches the Poisson functional representation (Section 3.3.2) when $\mathsf{N} \to \infty$. Similar to Poisson functional representation, ordered random coding can achieve a conditional entropy upper-bounded by $I(X;Y) + \log_2(I(X;Y)+1) + 4$ for channel simulation in the known source distribution case (Definition 2) (Theis and Yosri, 2022).

Compared to Poisson functional representation, ordered random coding ensures that $K \leq \mathsf{N}$, and at most N samples $\bar{Y}_1, \ldots, \bar{Y}_{\mathsf{N}}$ are needed, whereas Poisson functional representation (Algorithm 2) requires a number of samples that is unbounded but has a finite expectation. Also note that ordered random coding (like minimal random coding) is approximate, whereas Poisson functional representation is exact.

Readers are also referred to (Phan *et al.*, 2024) for a related importance sampling based compression scheme.

Greedy Poisson rejection sampling (G/1/E/VL/KAS/UCR). The greedy Poisson rejection sampling scheme, which is a causal sampling scheme (like greedy rejection sampling) that utilizes a Poisson process (like Poisson functional representation), was investigated by Flamich (2023). As in Poisson functional representation, we let $(T_i)_i \sim \mathrm{PP}(1)$, $(\bar{Y}_i)_i \overset{iid}{\sim} Q$. Unlike Poisson functional representation which selects $\bar{Y}_K$ with $K = \operatorname{argmin}_i T_i/g(\bar{Y}_i|X)$ for $g(y|x) := \frac{\mathrm{d}P_{Y|X}(\cdot|x)}{\mathrm{d}Q}(y)$ (3.16), here we select $\bar{Y}_K$ with

$$K = \min\left\{k : T_i \leq \sigma(g(\bar{Y}_i|X))\right\},$$

[24] Theis and Yosri (2022) considers the Gumbel random variables $G_i = -\log_2 Z_i$ instead, which is an equivalent formulation after taking logarithm.

[25] Refer to (Theis and Yosri, 2022) for a more efficient algorithm that does not require generating all N samples. Note that ordered random coding does not retain the differential privacy property of minimal random coding, since the sequence $Z_1, \ldots, Z_{\mathsf{N}}$ becomes a part of the common randomness instead of the local randomness at the encoder. Without this local randomness, the encoder in ordered random coding becomes deterministic, and cannot preserve privacy.

for a suitable function $\sigma : [0,\infty) \to [0,\infty)$ defined in (Flamich, 2023). This allows us to scan the points in increasing order of T_i and accept the first point satisfying $T_i \le \sigma(g(\bar{Y}_i|X))$, without the need of looking at future points. Greedy Poisson rejection sampling achieves a conditional entropy $H(Y|W) \le I(X;Y) + \log_2(I(X;Y)+1) + 6$ (Flamich, 2023).

Local pseudo-randomizer (G/1/A/FL/KAS/NCR). In the classical rejection sampling scheme (Section 3.2.1), we generate $\bar{Y}_1, \bar{Y}_2, \ldots \overset{iid}{\sim} Q$ over $\mathcal{Y}$ and $U_1, U_2, \ldots \overset{iid}{\sim} \mathrm{Unif}(0,1)$, and accept the first $(\bar{Y}_K, U_K)$ where $U_K \le \gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_K)$. In the sampling schemes discussed in the previous sections, the encoder transmits the index K of the selected sample $\bar{Y}_K$. In case $\mathcal{Y}$ is a finite set, the encoder may simply transmit $Y = \bar{Y}_K$. Unfortunately, in practice, we often desire an output Y that lies in a large set, or is even continuous. But can we *really* output a continuous Y in practice? If we are using a pseudorandom number generator (PRNG) that has a fixed-size internal state to produce $\bar{Y}_1, \bar{Y}_2, \ldots$, then the number of choices of Y will also be limited to the number of possible states.

This is the idea utilized in the *local pseudo-randomizer* by Feldman and Talwar (2021). Assume $\mathfrak{G} : \{0,1\}^\ell \to \mathcal{Y}$ is a PRNG, satisfying that when $S \sim \mathrm{Unif}(\{0,1\}^\ell)$ is a random seed, then $\mathfrak{G}(S)$ approximately follows Q. The local pseudo-randomizer algorithm generates $S_1, S_2, \ldots \overset{iid}{\sim} \mathrm{Unif}(\{0,1\}^\ell)$ and $U_1, U_2, \ldots \overset{iid}{\sim} \mathrm{Unif}(0,1)$, computes $\bar{Y}_i = \mathfrak{G}(S_i)$, accepts the first $(\bar{Y}_K, U_K)$ where $U_K \le \gamma \frac{\mathrm{d}P}{\mathrm{d}Q}(\bar{Y}_K)$, and transmits S_K (instead of K) to the decoder using ℓ bits. The decoder simply outputs $\tilde{Y} = \mathfrak{G}(S_K)$. No common randomness between the encoder and the decoder is needed. This algorithm only produces $\tilde{Y}$ that approximately follows the desired conditional distribution $P_{Y|X}$, and the quality of $\tilde{Y}$ depends on the quality of the PRNG. An advantage of local pseudo-randomizer is its privacy property. The decoder only observes S_K, which has the same information as $\tilde{Y} = \mathfrak{G}(S_K)$ about X. Therefore, as long as $\tilde{Y}$ does not leak too much information about X, S_K will not leak too much information as well.

To compare local pseudo-randomizer with minimal random coding, note that local pseudo-randomizer also uses a fixed set of 2^ℓ candidates $\{\mathfrak{G}(s)\}_{s\in\{0,1\}^\ell}$. The rejection sampling procedure eventually selects one of the candidates $\tilde{Y}$ with a probability proportional to $\frac{\mathrm{d}P}{\mathrm{d}Q}(\tilde{Y})$. The difference between minimal random coding and local pseudo-randomizer is that the latter fixes the PRNG $\mathfrak{G}$ and hence the set of candidate, whereas minimal random coding uses the common randomness to randomize the set of candidates. Also, local pseudo-randomizer utilizes a rejection sampling procedure to select $\tilde{Y}$ without scanning through all candidates, making the scheme efficient even when ℓ is large.

GenProt (G/1/A/FL/KAS/UCR). A scheme proposed by Bun *et al.* (2019), called GenProt, also uses a fixed number of samples $\bar{Y}_1, \ldots, \bar{Y}_{\mathsf{N}} \overset{iid}{\sim} Q$ like minimal random coding. It then utilizes a different method to select the sample $\bar{Y}_K$ by first randomly rejecting a subset

of the samples, and then selecting a non-rejected sample uniformly at random. GenProt is an approximate channel simulation scheme (the channel is only simulated within a small total variation distance). An advantage of GenProt is that it can convert an approximate differential privacy mechanism into a pure differential privacy mechanism.

3.6 Subtractively Dithered Quantization and Universal Quantization

3.6.1 Subtractively Dithered Scalar Quantization

In this section, we discuss the notion of subtractively dithered quantization (Roberts, 1962; Schuchman, 1964; Ziv, 1985; Gray and Stockham, 1993) mentioned in the introduction, which is perhaps the earliest example of channel simulation. We first review the original non-dithered quantization, where $X \in \mathbb{R}$ is mapped to the closest reconstruction level in $\ldots, -2\Delta, -\Delta, 0, \Delta, \ldots$, which is given by

$$Y = \Delta \left\lfloor \frac{X}{\Delta} + \frac{1}{2} \right\rfloor.$$

For the encoding operation, the encoder maps X to $\lfloor X/\Delta + 1/2 \rfloor \in \mathbb{Z}$, which is then encoded into a sequence of bits using either a fixed-length code (if X has a known bound, so $\lfloor X/\Delta + 1/2 \rfloor$ has finitely many possible values), or a variable-length code (e.g. the Huffman code (Huffman, 1952) or the signed Elias delta code over the integers (Elias, 1975)). The decoder decodes $\lfloor X/\Delta + 1/2 \rfloor$ and outputs $Y = \Delta \lfloor X/\Delta + 1/2 \rfloor$.

Since quantization maps a continuous value to a discrete value that can be communicated using finitely many bits, one straightforward "channel simulation" scheme for a channel $P_{Y|X}$ with $Y \in \mathbb{R}$ is to have the encoder generate Y following $P_{Y|X}$ locally, quantize Y into $\tilde{Y} = \Delta \lfloor Y/\Delta + 1/2 \rfloor$, and transmit $M = \lfloor Y/\Delta + 1/2 \rfloor \in \mathbb{Z}$ so the decoder can recover $\tilde{Y}$. We call this *simple quantization*. This approach is popular for the compression of privacy mechanisms (e.g., Andrés *et al.* (2013)), since the privacy properties of $P_{Y|X}$ are preserved by $P_{\tilde{Y}|X}$ as $\tilde{Y}$ is merely a function of Y. Nevertheless, this scheme is not exact since $P_{\tilde{Y}|X}$ deviates from the original $P_{Y|X}$, and does not utilize the noise in the channel $P_{Y|X}$ to reduce the description length.

In order to allow exact simulation, the quantization will be carried out in a random manner. The idea of subtractive dithering (Roberts, 1962; Schuchman, 1964; Ziv, 1985; Gray and Stockham, 1993) is to introduce a random shift to the reconstruction levels by an amount $-W\Delta$, where $W \sim \mathrm{Unif}(-1/2, 1/2)$ is the common randomness. Now the reconstruction levels are $\ldots, (-2 - W)\Delta$, $(-1 - W)\Delta$, $-W\Delta$, $(1 - W)\Delta, \ldots$, and the reconstruction level closest to X is

$$Y = \Delta \left(\left\lfloor \frac{X}{\Delta} + W + \frac{1}{2} \right\rfloor - W \right). \tag{3.24}$$

The encoder maps X to $K := \lfloor X/\Delta + W + 1/2 \rfloor \in \mathbb{Z}$, which is then encoded into a sequence of bits. The decoder decodes K and outputs $Y = \Delta(K - W)$. Refer to Figure 1.2 for an illustration.

The quantization noise $Y - X \sim \mathrm{Unif}(-\Delta/2, \Delta/2)$ is independent of X, and hence this is a channel simulation scheme for the additive noise channel with noise distribution $\mathrm{Unif}(-\Delta/2, \Delta/2)$. We now show this claim. Condition on $X = x$. Note that changing W into $W + a$ for any $a \in \mathbb{Z}$ will not affect the value of Y in 3.24, and hence the value of Y depends only on X and $W - \lfloor W \rfloor$ (the fractional part of W). We have $W - \lfloor W \rfloor \sim \mathrm{Unif}(0, 1)$ since $W \sim \mathrm{Unif}(-1/2, 1/2)$ is uniformly distributed over an interval of length 1. We still have $W - \lfloor W \rfloor \sim \mathrm{Unif}(0, 1)$ if $W \sim \mathrm{Unif}(-x/\Delta - 1/2,\ -x/\Delta + 1/2)$ instead. In this case, we have $\lfloor x/\Delta + W + 1/2 \rfloor = 0$, and $Y \sim \mathrm{Unif}(x - \Delta/2,\ x + \Delta/2)$. The result follows.

If $\mathbf{X} = (X_1, \dots, X_n) \in \mathbb{R}^n$ is a (not necessarily i.i.d.) random vector, and we want to simulate an additive noise channel $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$ where $Z_i \overset{iid}{\sim} \mathrm{Unif}(-\Delta/2, \Delta/2)$, i.e., $\mathbf{Z}$ is uniformly distributed over the hypercube $[-\Delta/2, \Delta/2]^n$, we can simply apply dithered quantization on each entry X_i separately, using the common randomness $\mathbf{W} = (W_1, \dots, W_n)$, $W_1, \dots, W_n \overset{iid}{\sim} \mathrm{Unif}(-1/2, 1/2)$. This is the optimal way to simulate the additive noise channel with uniform noise over the hypercube, in the sense that its conditional entropy (defined in Section 3.1) attains the mutual information lower bound in Proposition 5. This has been observed in (Zamir and Feder, 1992), and follows directly from Proposition 5 since $W_i \equiv -Y_i/\Delta \pmod 1$ is a function of Y_i.

Proposition 18 (Optimality of dithered quantization (nDAC/1/E/VL/KS/UCR) (Zamir and Feder, 1992))**.** *Fix $\Delta > 0$. Consider any joint distribution $P_{\mathbf{X}}$ over $\mathbb{R}^n$. Let $\mathbf{W}, \mathbf{Y}$ be the common randomness and output of the dithered quantization scheme (3.24) applied on each entry X_i separately. Then the conditional entropy of this scheme is*

$$H(\mathbf{Y}|\mathbf{W}) = I(\mathbf{X}; \mathbf{Y}) = h(\mathbf{Y}) - n \log_2 \Delta.$$

Hence, the dithered quantization scheme attains the minimum conditional entropy (3.1) for simulating the additive noise channel channel $P_{\mathbf{Y}|\mathbf{X}}$ with noise $Z_i \overset{iid}{\sim} \mathrm{Unif}(-\Delta/2, \Delta/2)$. If we encode $\mathbf{Y}$ using the Huffman code conditional on $\mathbf{W}$, then the expected length is at most 1 bit away from the minimal expected length L^*.

Universal quantization. Moreover, this scheme is *universally almost optimal*, not only among channel simulation schemes, but also among lossy compression schemes under mean squared error, in the sense that its conditional entropy is within a constant bit per dimension n from the rate-distortion function, regardless of the distribution of $\mathbf{X}$. This is established in the work on universal quantization by Ziv (1985) and Zamir and Feder (1992). Also refer to the work by Gish and Pierce (1968).

Theorem 19 (Universal quantization (Ziv, 1985; Zamir and Feder, 1992))**.** *Fix $D > 0$. Consider any joint distribution $P_{\mathbf{X}}$. Let $\mathbf{W}, \mathbf{Y}$ be the common randomness and output of the dithered quantization scheme (3.24) applied on each X_i separately, with $\Delta = 2\sqrt{3D}$ (so $n^{-1}\mathbb{E}[\|\mathbf{X} - \mathbf{Y}\|^2] = D$). Then the conditional entropy of this scheme satisfies*

$$\begin{aligned} H(\mathbf{Y}|\mathbf{W}) &\le R(D) + \frac{n}{2}\log_2\frac{\pi e}{3} \\ &\le R(D) + 0.755n \text{ bits}, \end{aligned}$$

where

$$R(D) := \inf_{P_{\hat{\mathbf{X}}|\mathbf{X}}: n^{-1}\mathbb{E}[\|\mathbf{X}-\hat{\mathbf{X}}\|^2]\le D} I(\mathbf{X};\hat{\mathbf{X}})$$

is the rate-distortion function. If we encode $\mathbf{Y}$ using the Huffman code conditional on $\mathbf{W}$, then the expected length is at most $0.755n + 1$ bits away from $R(D)$.[26]

Proof. We present the arguments in (Zamir and Feder, 1992) here. Let $\mathbf{Z} \sim \mathrm{Unif}([-\Delta/2, \Delta/2]^n)$ be independent of $\mathbf{X}$, and $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$. Consider any $P_{\hat{\mathbf{X}}|\mathbf{X}}$ with $n^{-1}\mathbb{E}[\sum_{i=1}^n (X_i - \hat{X}_i)^2] \le D$. Let $\hat{\mathbf{X}}|\mathbf{X} \sim P_{\hat{\mathbf{X}}|\mathbf{X}}$, with $(\mathbf{X}, \hat{\mathbf{X}})$ being independent of $\mathbf{Z}$. By Proposition 18,

$$\begin{aligned} H(\mathbf{Y}|\mathbf{W}) &= I(\mathbf{X};\mathbf{Y}) \\ &\le I(\mathbf{X};\mathbf{Y},\hat{\mathbf{X}}) \\ &= I(\mathbf{X};\hat{\mathbf{X}}) + I(\mathbf{X};\mathbf{Y}|\hat{\mathbf{X}}) \\ &= I(\mathbf{X};\hat{\mathbf{X}}) + h(\mathbf{Y}|\hat{\mathbf{X}}) - h(\mathbf{Y}|\mathbf{X},\hat{\mathbf{X}}) \\ &= I(\mathbf{X};\hat{\mathbf{X}}) + h(\mathbf{Y}-\hat{\mathbf{X}}|\hat{\mathbf{X}}) - h(\mathbf{Z}) \\ &\le I(\mathbf{X};\hat{\mathbf{X}}) + h(\mathbf{Y}-\hat{\mathbf{X}}) - n\log_2\Delta \\ &= I(\mathbf{X};\hat{\mathbf{X}}) + h(\mathbf{Y}-\hat{\mathbf{X}}) - \frac{n}{2}\log_2(12D), \end{aligned}$$

where

$$\begin{aligned} h(\mathbf{Y}-\hat{\mathbf{X}}) &\le \sum_{i=1}^n h(Y_i - \hat{X}_i) \\ &\overset{(a)}{\le} \sum_{i=1}^n \frac{1}{2}\log_2\left(2\pi e \mathrm{Var}\left[Y_i - \hat{X}_i\right]\right) \\ &= \sum_{i=1}^n \frac{1}{2}\log_2\left(2\pi e\left(\mathrm{Var}\left[X_i - \hat{X}_i\right] + \mathrm{Var}\left[Z_i\right]\right)\right) \end{aligned}$$

[26] It was shown by Gish and Pierce (1968) and Ziv (1985) that if $\mathbf{X}$ contains i.i.d. entries with a smooth probability density function, in the "high resolution limit" where $D \to 0$, we have $\limsup_{D\to 0}(H(\mathbf{Y}|\mathbf{W}) - R(D)) \le (n/2)\log_2(2\pi e/12) \le 0.255n$ bits.

$$
\begin{aligned}
&= \sum_{i=1}^{n} \frac{1}{2} \log_2 \left(2\pi e \left(\mathrm{Var}\left[X_i - \hat{X}_i \right] + \frac{\Delta^2}{12} \right) \right) \\
&\overset{(b)}{\leq} \frac{n}{2} \log_2 \left(2\pi e \left(\frac{1}{n} \sum_{i=1}^{n} \mathrm{Var}\left[X_i - \hat{X}_i \right] + \frac{\Delta^2}{12} \right) \right) \\
&\leq \frac{n}{2} \log_2 \left(2\pi e \left(D + D \right) \right) \\
&= \frac{n}{2} \log_2 \left(4\pi e D \right),
\end{aligned}
$$

where (a) is because the Gaussian distribution maximizes differential entropy for a fixed variance, and (b) is by Jensen's inequality. Hence,

$$
\begin{aligned}
H(\mathbf{Y}|\mathbf{W}) &\leq I(\mathbf{X}; \hat{\mathbf{X}}) + \frac{n}{2} \log_2 \frac{4\pi e}{12} \\
&= I(\mathbf{X}; \hat{\mathbf{X}}) + \frac{n}{2} \log_2 \frac{\pi e}{3}.
\end{aligned}
$$

□

Theorem 19 provides a method for designing lossy compression schemes, which has an expected length per sample at most a constant away from the optimum. To compress $\mathbf{X}$, we simply apply dithered quantization to each entry, and encode $\mathbf{Y}$ conditional on $\mathbf{W}$ using Huffman code or another entropy encoding. We do not have to tailor the quantizer for the distribution of $\mathbf{X}$. Nevertheless, the Huffman code still needs to take the distribution of $\mathbf{X}$ into account.

3.6.2 Layered Randomized Quantizers for Additive Noise Channel Simulation

Subtractive dithering produces an additive noise with distribution $\mathrm{Unif}(-\Delta/2, \Delta/2)$. If the decoder adds a shift $B \in \mathbb{R}$ to the output Y, the noise distribution would be $\mathrm{Unif}(B - \Delta/2,\ B + \Delta/2)$, which can be uniform over any given interval. Using this, we can construct a channel simulation scheme for any unimodal noise distribution over $\mathbb{R}$.[27] The idea, which appeared (in a different form) in the *layered multishift coupler* in (Wilson, 2000), noted in (Agustsson and Theis, 2020) for the Gaussian case, and investigated in (Hegazy and Li, 2022) for the general case, is to express the noise distribution with probability density function f as a mixture of uniform distributions over intervals. Let

$$
L_s^+(f) := \{ x \in \mathbb{R} :\ f(x) \geq s \}
$$

[27] A probability density function $f : \mathbb{R} \to \mathbb{R}$ is unimodal if there exists $c \in \mathbb{R}$ such that $f(x)$ is nondecreasing over $x \in (-\infty, c]$, and nonincreasing over $x \in [c, \infty)$.

be the superlevel set of f. For a unimodal f, the superlevel set is an interval (or empty). Moreover, f can be expressed as a mixture over uniform distributions over $L_s^+(f)$, as

$$\begin{aligned} f(x) &= \int_0^\infty \mathbf{1}\{x \in L_s^+(f)\}\mathrm{d}s \\ &= \int_0^\infty \mathrm{Vol}(L_s^+(f)) \cdot \mathrm{Unif}(x; L_s^+(f))\mathrm{d}s, \end{aligned}$$

where $\mathrm{Unif}(x; L_s^+(f)) := \mathbf{1}\{x \in L_s^+(f)\}/\mathrm{Vol}(L_s^+(f))$ is the probability density function of $\mathrm{Unif}(L_s^+(f))$. Therefore, we can simulate the noise distribution f, by randomly selecting S and applying subtractive dithering to produce a noise distribution $\mathrm{Unif}(L_s^+(f))$.

The channel simulation scheme for an additive noise channel with a unimodal noise distribution f in (Hegazy and Li, 2022), called *(direct) layered randomized quantizer*, is constructed as follows. First, generate a common randomness (S, W), where $S \in [0, \infty)$ has a probability density function $f_S(s) := \mathrm{Vol}(L_s^+(f)) = \sup L_s^+(f) - \inf L_s^+(f)$ and $W \sim \mathrm{Unif}(-1/2, 1/2)$. The encoder observes S, W, X, computes $\Delta = f_S(S)$,

$$K := \left\lfloor \frac{X}{\Delta} + W + \frac{1}{2} \right\rfloor \in \mathbb{Z},$$

and encodes and transmit K. The decoder observes S, W, decodes K, and computes $\Delta = f_S(S)$,

$$B := \frac{\sup L_S^+(f) + \inf L_S^+(f)}{2},$$

$$Y = \Delta \cdot (K - W) + B. \tag{3.25}$$

Conditional on $S = s$, the layered randomized quantizer would become a subtractive dithering scheme with a noise distribution

$$\mathrm{Unif}\big(\inf L_s^+(f), \sup L_s^+(f)\big) = \mathrm{Unif}(L_s^+(f)). \tag{3.26}$$

Therefore, randomizing over S, the noise distribution becomes f. The algorithm, based on (Hegazy and Li, 2022; Hegazy *et al.*, 2024), is given in Algorithm 4. In particular, if the desired noise distribution is $N(0, \sigma^2)$, then Δ can equivalently be generated as $G \sim \mathrm{Gamma}(3/2, 1/2)$ and $\Delta = 2\sigma\sqrt{G}$ (Walker, 1999; Agustsson and Theis, 2020; Hasırcıoğlu and Gündüz, 2024). Refer to Figure 3.9 for an illustration.

Asymptotic optimality. Although the layered randomized quantizer does not enjoy the universal optimality property in Proposition 18, it is asymptotically optimal in the *high signal-to-noise-ratio (SNR) limit*, where we consider the input distribution $X \sim \mathrm{Unif}(0, t)$, and take $t \to \infty$. In this case, the mutual information $I(X; X+Z)$ (where $Z \sim f$) grows like $\log_2 t + O(1)$, and hence we expect the optimal conditional entropy for channel simulation to

grow like this as well. It remains to characterize the "$O(1)$" term. Theorem 4 can only give a scheme with conditional entropy $\log_2 t + \log_2 \log_2 t + O(1)$, which is not strong enough to give us a "$\log_2 t + O(1)$" result, and hence a different analysis is needed. Note that this is different from the *large blocklength limit* (to be discussed in Section 5), which is what "asymptotic" usually mean in this monograph, where we simulate a memoryless channel with input sequence $X_1, \dots, X_n$ and take $n \to \infty$. In the large blocklength limit, the communication needed for channel simulation grows like $\Theta(n)$.

The following result in (Hegazy and Li, 2022) characterizes the precise high SNR limit of additive noise channel simulation with a unimodal noise, and showed that the layered randomized quantizer is optimal in the high SNR limit. Readers are referred to (Hegazy and Li, 2022) for the proof.

Theorem 20 (High SNR limit of additive noise channel simulation (1DAC/1/E/VL/KS/UCR) (Hegazy and Li, 2022))**.** *Fix a unimodal probability density function $f : \mathbb{R} \to \mathbb{R}$ with a finite mean. Let $H^*_{f,t}$ be the minimum conditional entropy (3.1) among channel simulation schemes for the additive noise channel $X \to X + Z$, where $X \sim \mathrm{Unif}(0, t)$, $Z \sim f$. Then we have*

$$H^*_{f,t} = \log_2 t - h_L(f) + O\Big(\frac{1}{t}\Big) \tag{3.27}$$

as $t \to \infty$, where

$$h_L(f) := \int_0^\infty \mathrm{Vol}(L_s^+(f)) \log_2 \mathrm{Vol}(L_s^+(f))\, \mathrm{d}s \tag{3.28}$$

is called the layered entropy *of f (Hegazy and Li, 2022). More precisely, we have the following for any $t > 0$:*

$$0 \le H^*_{f,t} - (\log_2 t - h_L(f)) \le \frac{8\delta_Z \log_2 e}{t}, \tag{3.29}$$

where $\delta_Z := \mathbb{E}[|Z - \mathrm{median}(Z)|]$ is the mean absolute deviation of Z around the median (which is not greater than the standard deviation). The conditional entropy of the direct layered randomized quantizer satisfies (3.27) and the bounds in (3.29).

Nonasymptotic bound. Moreover, we have the following nonasymptotic bound on the conditional entropy of layered randomized quantizer, which is a slight generalization of the result by Kobus *et al.* (2024a) on the layered randomized quantizer for Gaussian channel simulation.

Theorem 21 (1DAC/1/E/VL/KS/UCR (Kobus *et al.*, 2024a))**.** *Fix a unimodal probability density function $f : \mathbb{R} \to \mathbb{R}$ with a finite mean, and any known source distribution P_X. The conditional entropy of the direct layered randomized quantizer satisfies*

$$H(Y|S, W) \le I(X; Y) + h(f) - h_L(f),$$

where $h(f)$ is the differential entropy, and $h_L(f)$ is the layered entropy (3.28).[28] *Hence, the expected description length can be upper-bounded by $I(X;Y) + h(f) - h_L(f) + 1$.*

Proof. We have

$$\begin{aligned} H(Y|S,W) - I(X;Y) &\overset{(a)}{=} I(X;S,W|Y) \\ &= I(X;S|Y) + I(X;W|S,Y) \\ &\overset{(b)}{=} I(X;S|Y) \\ &\le I(X,Y;S) \\ &\overset{(c)}{=} I(Y;S|X) \\ &= h(Y|X) - h(Y|X,S) \\ &\overset{(d)}{=} h(f) - \mathbb{E}\left[\log_2 f_S(S)\right] \\ &= h(f) - h_L(f), \end{aligned}$$

where (a) was shown in the proof of Proposition 5, (b) is because $W \equiv -Y/f_S(S) \pmod 1$ is a function of (S,Y), (c) is due to $I(X;S) = 0$, and (d) is because for a fixed S, the channel from X to Y is an additive noise channel with noise distribution $\mathrm{Unif}(L_s^+(f))$, with $\mathrm{Vol}(L_s^+(f)) = f_S(s)$ (3.26). □

The bound in Theorem 21 was evaluated for scalar additive Gaussian noise channels by Kobus *et al.* (2024a), which gives $H(Y|S,W) \le I(X;Y) + 0.521$ bits.

Shifted layered randomized quantizer. One modification of the direct layered randomized quantizer is to flip the left half of the graph of f upside down. This is the construction in the layered multishift coupler in (Wilson, 2000), and is referred to as the *shifted layered randomized quantizer* in (Hegazy *et al.*, 2024). The scheme for an additive noise channel with a unimodal noise distribution f works as follows (Hegazy *et al.*, 2024). Let $f^* := \max_x f(x)$. First, generate a common randomness (S,W), where $S \in [0, f^*]$ has a probability density function

$$f_S(s) := \sup L_s^+(f) - \inf L_{f^*-s}^+(f),$$

and $W \sim \mathrm{Unif}(-1/2, 1/2)$. The encoder observes S, W, X, computes $\Delta = f_S(S)$,

$$K := \left\lfloor \frac{X}{\Delta} + W + \frac{1}{2} \right\rfloor \in \mathbb{Z}$$

[28] Note that $h_L(f) \le h(f)$. Refer to (Hegazy and Li, 2022; Ling and Li, 2024) for the proof.

and encodes and transmit K. The decoder observes S, W, decodes K, and computes $\Delta = f_S(S)$,

$$B := \frac{\sup L_S^+(f) + \inf L_{f^*-S}^+(f)}{2}$$
$$Y = \Delta \cdot (K - W) + B.$$

Refer to Figure 3.9 for an illustration. Although the shifted layered randomized quantizer has a larger conditional entropy compared to the direct construction (3.25), and cannot achieve the high SNR limit in Theorem 20, the advantage is that the quantization step Δ can be lower-bounded, and hence K can be bounded if X is bounded. If $|X| \le \gamma$, then we have $|K| \le \gamma / \inf_s f_S(s) + 1/2$, and hence K can be encoded into a fixed-length codeword. This is desirable if a fixed-length code is needed, for example, if we have a strict communication budget for every component in the protocol. Readers are referred to (Hegazy *et al.*, 2024) for bounds on the performance of the shifted layered randomized quantizer. The algorithm, based on (Wilson, 2000; Hegazy *et al.*, 2024), is given in Algorithm 4.

Layered randomized quantizers has been applied to simulate additive noise mechanisms for differential privacy settings in (Hasırcıoğlu and Gündüz, 2024; Shahmiri *et al.*, 2024; Hegazy *et al.*, 2024; Yan *et al.*, 2023) (see Section 1.6).

3.6.3 Subtractively Dithered Lattice Quantization

Recall that the conditional entropy of the direct layered randomized quantizer applied to a scalar additive Gaussian noise channel is upper-bounded as $H(Y|S,W) \le I(X;Y) + 0.521$ in Theorem 21 (Kobus *et al.*, 2024a). If our goal is to simulate n copies of this scalar channel, we require a description length $\approx nI(X;Y) + 0.521n$, significantly longer than the $nI(X;Y) + \log_2(nI(X;Y)+1) + O(1)$ description length given by greedy rejection sampling and Poisson functional representation. In order to reduce the description length, we should treat these n copies of scalar channel as a single vector channel with input and output in $\mathbb{R}^n$, and perform quantization over these n dimensions together.

We first review some basic concepts in lattice quantization (Conway and Sloane, 2013; Zamir, 2014). A *lattice* in $\mathbb{R}^n$ with a generator matrix $\mathbf{G} \in \mathbb{R}^{n \times n}$ (a full-rank matrix) is the set $\mathbf{G}\mathbb{Z}^n = \{\mathbf{Gi} : \mathbf{i} \in \mathbb{Z}^n\}$. A *fundamental cell* of the lattice is a bounded subset $\mathcal{P}_\mathbf{0} \subseteq \mathbb{R}^n$ such that $\{\mathcal{P}_\mathbf{0} + \mathbf{y} : \mathbf{y} \in \mathbf{G}\mathbb{Z}^n\}$ (the set of translations $\mathcal{P}_\mathbf{0} + \mathbf{y} = \{\mathbf{x} + \mathbf{y} : \mathbf{x} \in \mathcal{P}_\mathbf{0}\}$ of $\mathcal{P}_\mathbf{0}$ by lattice points) forms a partition of $\mathbb{R}^n$. Given a fundamental cell, we can define a lattice quantization function $Q : \mathbb{R}^n \to \mathbb{R}^n$, where $Q(\mathbf{x})$ is taken to be the vector $\mathbf{y} \in \mathbf{G}\mathbb{Z}^n$ such that $\mathbf{x} \in \mathcal{P}_\mathbf{0} + \mathbf{y}$. In particular, if $Q(\mathbf{x}) = \operatorname{argmin}_{\mathbf{y} \in \mathbf{G}\mathbb{Z}^n} \|\mathbf{x} - \mathbf{y}\|$ is the closest lattice point to $\mathbf{x}$, then the corresponding fundamental cell $Q^{-1}(\{\mathbf{0}\}) = \{\mathbf{x} \in \mathbb{R}^n : \operatorname{argmin}_{\mathbf{y} \in \mathbf{G}\mathbb{Z}^n} \|\mathbf{x} - \mathbf{y}\| = \mathbf{0}\}$ is the *Voronoi cell* of the lattice.

Subtractive dithering can also be applied on lattice quantization (Kirac and Vaidyanathan, 1996; Zamir, 2014). Consider a generator matrix $\mathbf{G}$ and a fundamental cell $\mathcal{P}_\mathbf{0}$ with a corre-

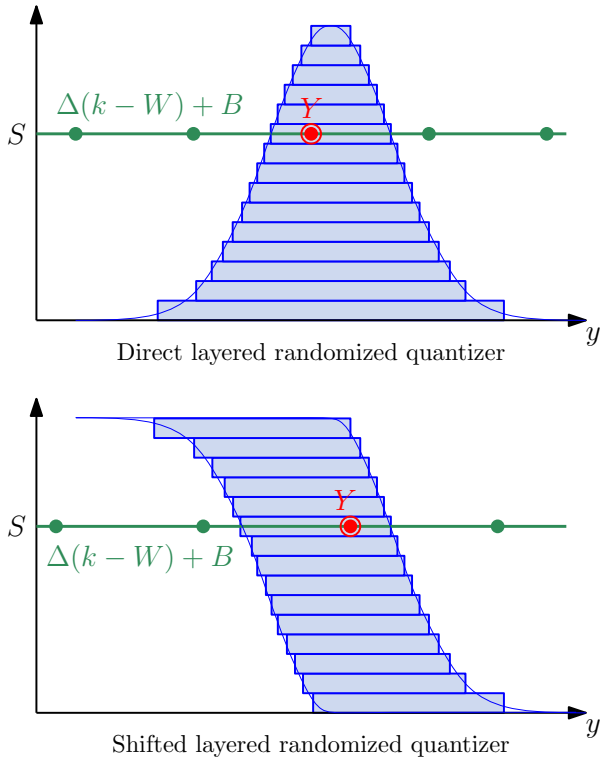


Figure 3.9: Top: an illustration of the direct layered randomized quantizer applied to simulate a Gaussian additive noise, where we first select the vertical coordinate S at random (as a part of the common randomness), and then apply subtractive dithering to create a noise uniform over superlevel set $L_S^+(f)$ (which is the blue interval with vertical coordinate S). Among the reconstruction levels $(\Delta(k-W)+B)_{k\in\mathbb{Z}}$ shown as the green points in the figure (refer to (3.25)), we select the point that lies in the superlevel set $L_S^+(f)$ (blue interval) as the reconstruction Y (red point). There is exactly one reconstruction level in $L_S^+(f)$ since Δ is the length of $L_S^+(f)$.
Bottom: an illustration of the shifted layered randomized quantizer applied to simulate a Gaussian additive noise. Compared to the direct layered randomized quantizer, the left half of the graph of the Gaussian density function is flipped upside down. This guarantees that the lengths of the blue intervals are bounded away from 0.

Algorithm 4 Direct/shifted layered randomized quantizer (Hegazy and Li, 2022; Wilson, 2000; Hegazy *et al.*, 2024)

Procedure ENCODE$(f, x, \text{mode}, \mathfrak{G})$:

Input: density function f, input $x \in \mathbb{R}$,
mode $\in \{\text{direct}, \text{shifted}\}$, PRNG $\mathfrak{G}$

Output: $k \in \mathbb{Z}$

$f^* \leftarrow \max_x f(x)$
if mode = direct **then**
  Generate $s \in [0, f^*]$ with density $f_S(s) = \sup L_s^+(f) - \inf L_s^+(f)$ using $\mathfrak{G}$
  $\Delta \leftarrow \sup L_s^+(f) - \inf L_s^+(f)$
else if mode = shifted **then**
  Generate $s \in [0, f^*]$ with density $f_S(s) = \sup L_s^+(f) - \inf L_{f^*-s}^+(f)$ using $\mathfrak{G}$
  $\Delta \leftarrow \sup L_s^+(f) - \inf L_{f^*-s}^+(f)$
end if
Generate $w \sim \text{Unif}(-1/2, 1/2)$ using $\mathfrak{G}$
return $\lfloor x/\Delta + w + 1/2 \rfloor$

Procedure DECODE$(f, k, \text{mode}, \mathfrak{G})$:

Input: f, k, mode, $\mathfrak{G}$

Output: sample Y

$f^* \leftarrow \max_x f(x)$
if mode = direct **then**
  Generate $s \in [0, f^*]$, $f_S(s) = \sup L_s^+(f) - \inf L_s^+(f)$ using $\mathfrak{G}$
  $\Delta \leftarrow \sup L_s^+(f) - \inf L_s^+(f)$
  $b \leftarrow \frac{\sup L_s^+(f) + \inf L_s^+(f)}{2}$
else if mode = shifted **then**
  Generate $s \in [0, f^*]$, $f_S(s) = \sup L_s^+(f) - \inf L_{f^*-s}^+(f)$ using $\mathfrak{G}$
  $\Delta \leftarrow \sup L_s^+(f) - \inf L_{f^*-s}^+(f)$
  $b \leftarrow \frac{\sup L_s^+(f) + \inf L_{f^*-s}^+(f)}{2}$
end if
Generate $w \sim \text{Unif}(-1/2, 1/2)$ using $\mathfrak{G}$
return $\Delta \cdot (k - w) + b$

sponding quantization function Q. The common randomness is $\mathbf{W} \sim \text{Unif}(\mathcal{P}_\mathbf{0})$. The encoder maps $\mathbf{X} \in \mathbb{R}^n$ to

$$\mathbf{K} := \mathbf{G}^{-1} Q(\mathbf{X} + \mathbf{W}) \in \mathbb{Z}^n, \tag{3.30}$$

which is then encoded into a sequence of bits. The decoder decodes $\mathbf{K}$ and outputs $\mathbf{Y} = \mathbf{G}\mathbf{K} - \mathbf{W}$.

The quantization noise has a distribution

$$\mathbf{Y} - \mathbf{X} = Q(\mathbf{X} + \mathbf{W}) - \mathbf{W} - \mathbf{X} \sim \text{Unif}(-\mathcal{P}_\mathbf{0}),$$

independent of the input $\mathbf{X}$. Refer to (Kirac and Vaidyanathan, 1996) for the proof. Note that $-\mathcal{P}_\mathbf{0} = \{-\mathbf{v} : \mathbf{v} \in \mathcal{P}_\mathbf{0}\}$ is a fundamental cell as well. Hence, using subtractive quantization, we can perform channel simulation on the additive noise channel $\mathbf{X} \to \mathbf{X} + \mathbf{Z}$ where $\mathbf{Z}$ is uniform over an arbitrary fundamental cell. This includes the entrywise dithered quantization in Section 3.6.1 as a special case, by taking $\mathbf{G} = \Delta\mathbf{I}$ and $\mathcal{P}_\mathbf{0} = [-\Delta/2, \Delta/2)^n$.

Similar to Proposition 18, the subtractively dithered lattice quantization scheme is also optimal for simulating the corresponding additive noise channel. This has been observed in (Zamir and Feder, 1992), and follows directly from Proposition 5 since $\mathbf{W}$ is the unique point in $\mathcal{P}_\mathbf{0}$ such that there exists $\mathbf{t} \in \mathbf{G}\mathbb{Z}^n$ with $\mathbf{W} = \mathbf{t} - \mathbf{Y}$, and hence $\mathbf{W}$ is a function of $\mathbf{Y}$.

Proposition 22 (Optimality of dithered lattice quantization (*n*DAC/1/E/VL/KS/UCR) (Zamir and Feder, 1992))**.** *Consider a generator matrix* $\mathbf{G}$ *and a fundamental cell* $\mathcal{P}_\mathbf{0}$ *with a corresponding quantization function* Q*. Consider any distribution* $P_\mathbf{X}$ *over* $\mathbb{R}^n$*. Let* $\mathbf{W} \sim \text{Unif}(\mathcal{P}_\mathbf{0})$ *and* $\mathbf{Y} = Q(\mathbf{X} + \mathbf{W}) - \mathbf{W}$ *be the common randomness and output of the dithered lattice quantization scheme* (3.30)*. Then the conditional entropy of this scheme is*

$$H(\mathbf{Y}|\mathbf{W}) = I(\mathbf{X}; \mathbf{Y}) = h(\mathbf{Y}) - \log_2 \text{Vol}(\mathcal{P}_\mathbf{0}).$$

Hence, the dithered vector quantization scheme attains the minimum conditional entropy (3.1) *for simulating the additive noise channel* $\mathbf{X} \to \mathbf{X} + \mathbf{Z}$ *where* $\mathbf{Z} \sim \text{Unif}(-\mathcal{P}_\mathbf{0})$*. If we encode* $\mathbf{Y}$ *using the Huffman code conditional on* $\mathbf{W}$*, then the expected length is at most* 1 *bit away from the minimal expected length* L^**.*

A universal quantization result similar to Theorem 19 for vector quantization is also proved in (Zamir and Feder, 1992). Readers are referred to (Zamir and Feder, 1992) for the result and the proof.

Rejection-sampled universal quantizer. We have seen how dithered lattice quantization can simulate an additive noise channel with noise uniform over a fundamental cell. For example, in $\mathbb{R}^2$, dithered lattice quantization can create a noise uniform over a parallelogram

or a regular hexagon. However, it cannot create a noise uniform over a circular disk, which is not a fundamental cell. In order to shape the error distribution into a uniform distribution over an arbitrary set $\mathcal{A} \subseteq \mathbb{R}^n$, we may combine dithered lattice quantization with rejection sampling (Section 3.2), which is the approach considered by Ling and Li (2024), called *rejection-sampled universal quantizer (RSUQ)* (also refer to (Theis and Yosri, 2022) for a closely related construction called hybrid coding, which combines dithered quantization with ordered random coding discussed in Section 3.5.3). Suppose $\mathcal{P}_{\mathbf{0}} \supseteq -\mathcal{A}$ is a fundamental cell. We repeatedly apply subtractive dithering (3.30) with the fundamental cell $\mathcal{P}_{\mathbf{0}}$ using different dither signals, until the quantization noise $\mathbf{Y} - \mathbf{X}$ falls in $\mathcal{A}$. More precisely, letting the common randomness be a sequence of dither signals $\mathbf{W}_1, \mathbf{W}_2, \ldots \overset{iid}{\sim} \mathrm{Unif}(\mathcal{P}_{\mathbf{0}})$, the encoder maps $\mathbf{X} \in \mathbb{R}^n$ to $(J, \mathbf{K})$ where $J := \min\{i : Q(\mathbf{X} + \mathbf{W}_i) - \mathbf{W}_i - \mathbf{X} \in \mathcal{A}\}$ is the index of the first dither signal $\mathbf{W}_J$ that makes the quantization noise fall in $\mathcal{A}$, and $\mathbf{K} := \mathbf{G}^{-1} Q(\mathbf{X} + \mathbf{W}_J)$. The decoder maps $(J, \mathbf{K})$ to $\mathbf{Y} = \mathbf{G}\mathbf{K} - \mathbf{W}_J$. This way, we can ensure that $\mathbf{Y} - \mathbf{X} \sim \mathrm{Unif}(\mathcal{A})$. It has been shown by Ling and Li (2024) that RSUQ can outperform conventional lattice quantizers in certain aspects, e.g., attaining a smaller maximum error and/or mean squared error for certain dimensions in the high resolution limit.

If the desired noise distribution is nonuniform with a probability density function $f : \mathbb{R}^n \to \mathbb{R}$, the method by Ling and Li (2024), called *layered RSUQ (LRSUQ)*, is to adopt the same strategy as the layered randomized quantizer in Section 3.6.2, where f is decomposed into a mixture of uniform distributions over the superlevel sets $L_S^+(f) := \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \geq S\}$ for a randomly chosen $S \geq 0$ with probability density function $f_S(s) := \mathrm{Vol}(L_s^+(f))$, and then RSUQ is applied on $L_S^+(f)$ for a random S shared as a part of the common randomness. LRSUQ can be used to show a high SNR result for simulating a vector additive noise channel with general continuous noise (Ling and Li, 2024). Nevertheless, unlike the 1D layered randomized quantizer that is asymptotically optimal in the high SNR limit (Theorem 3.6.2), LRSUQ is only asymptotically within $\log_2 e$ bits from the optimum.

Theorem 23 (High SNR limit of vector additive noise channel simulation (*n*DAC/1/E/VL/KS/UCR) (Ling and Li, 2024))**.** *Fix a probability density function $f : \mathbb{R}^n \to \mathbb{R}$ satisfying the regularity condition*

$$\int_0^\infty \Big(\sup_{\mathbf{z}:\|\mathbf{z}\|_2 \geq \gamma} f(\mathbf{z}) \Big) \gamma^{n-1} \log_2(1+\gamma) \mathrm{d}\gamma < \infty.$$

Let $H_{f,t}^*$ be the minimum conditional entropy (3.1) among channel simulation schemes for the additive noise channel $\mathbf{X} \to \mathbf{X} + \mathbf{Z}$, where $\mathbf{X} \sim \mathrm{Unif}(tB^n)$ is uniform over the n-ball $tB^n = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 \leq t\}$, and $\mathbf{Z} \sim f$. Then we have

$$-h_L(f) \leq H_{f,t}^* - \log_2 \mathrm{Vol}(tB^n) \leq -h_L(f) + \log_2 e + o(1) \tag{3.31}$$

as $t \to \infty$, where $h_L(f)$ is the layered entropy (3.28). The bounds (3.31) are satisfied by LRSUQ.

Another construction that achieves an error uniform over an arbitrary set $\mathcal{A} \subseteq \mathbb{R}^n$ is *shift-periodic quantization* (Ling and Li, 2023), which aims at dissecting $\mathcal{A}$ into (possibly infinitely many) pieces that can be reconstructed to form a fundamental cell $-\mathcal{P}_\mathbf{0}$. Readers are referred to (Ling and Li, 2023) for details.

3.6.4 Dithering-based Schemes for Specific Noise Distributions

We discuss several other channel simulation schemes based on dithered quantization.

Rotated dithered quantization (*n*DAC/1/A/(FL or VL)/KAS/UCR). Consider the simulation of a vector additive noise channel with Gaussian noise $\mathbf{Z} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, which is a rotationally invariant distribution, i.e., $\mathbf{R}Z$ has the same distribution as $\mathbf{Z}$ for any orthogonal matrix $\mathbf{R} \in \mathbb{R}^{n\times n}$. Dithered lattice quantization produces an error uniform over a fundamental cell, which cannot be rotationally invariant. To ensure a rotationally invariant error, the idea by Kobus *et al.* (2024a) is to apply a rotation by a uniformly randomly generated orthogonal matrix $\mathbf{R}$ (included in the common randomness). The encoding and decoding are given by

$$\mathbf{K} = \mathbf{G}^{-1}Q(\mathbf{R}^T\mathbf{X} + \mathbf{W}) \in \mathbb{Z}^n,$$

$$\begin{aligned}\mathbf{Y} &= \mathbf{R}(\mathbf{G}\mathbf{K} - \mathbf{W} + \tilde{\mathbf{Z}}) \\ &= \mathbf{R}(Q(\mathbf{R}^T\mathbf{X} + \mathbf{W}) - \mathbf{W} + \tilde{\mathbf{Z}}),\end{aligned}$$

where Q is the lattice quantization function (Section 3.6.2) for some fundamental cell $\mathcal{P}_\mathbf{0}$ of a lattice $\mathbf{G}\mathbb{Z}^n$, and $\tilde{\mathbf{Z}}$ is an additional noise added to make the overall noise distribution closer to Gaussian. The case where $\mathbf{G}\mathbb{Z}^n$ is a scaled integer lattice ($\mathbf{G}$ is a scalar multiple of $\mathbf{I}$), and $\tilde{\mathbf{Z}}$ contains i.i.d. Weibull-distributed entries, was analyzed by Kobus *et al.* (2024a). While this scheme does not achieve an exact Gaussian noise distribution, the KL divergence of the overall noise distribution from the Gaussian distribution is bounded by $O(n^{-1})$ (Kobus *et al.*, 2024a).

Dyadic quantized Laplace mechanism (1DAC/1/E/VL/KAS/UCR). The layered randomized quantizers (Section 3.6.2) does not satisfy differential privacy or metric privacy even when Z is a privacy-perserving noise. The decoder, observing W, Δ and $K = \lfloor X/\Delta + W + 1/2 \rfloor$, would know precisely which quantization cell X lies in, and a slight

change to X can result in a deterministic change to K (when W is fixed) if X is close to the boundary of the quantization cell. An observation by Shahmiri *et al.* (2024) is that, if the encoder adds a local noise with a piecewise constant probability density function to the input before applying dithered quantization, then differential privacy can be achieved. To ensure that the error follows the Laplace distribution, the construction in (Shahmiri *et al.*, 2024) decomposes the Laplace distribution into a mixture of piecewise linear probability density functions. This allows the Laplace mechanism (Dwork *et al.*, 2006) to be simulated exactly with a finite amount of communication, while retaining a (weaker) differential privacy guarantee.

Interested readers are also referred to (Shlezinger *et al.*, 2020; Amiri *et al.*, 2021; Lang *et al.*, 2023) for the use of dithered vector quantization in various learning and privacy tasks.

3.7 Exact Fixed-Length Channel Simulation

In this section, we briefly discuss exact channel simulation with a strict constraint on the description length. We require that the description M has a fixed length. For the fixed-length setting, we would modify Definition 2 to have $M \in [\mathsf{N}]$ instead of $M \in \mathcal{C}_W$, where $\mathsf{N} \in \mathbb{N}^+$ is the description size. The problem becomes minimizing N such that there exists a scheme with $Y|X \sim P_{Y|X}$ exactly. Note that among the schemes for one-shot exact channel simulation with unlimited common randomness discussed in this section, only subtractive dithering and the shifted layered randomized quantizer in Section 3.6 can be made into a fixed-length scheme (for a bounded input signal). Although minimal random coding (Section 3.4) is a fixed-length scheme, it only simulates the channel approximately.

Consider the conditional probability matrix $\mathbf{P}_{Y|X} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ with entries $(\mathbf{P}_{Y|X})_{x,y} = P_{Y|X}(y|x)$ (here we assume $X \in \mathcal{X} = [|\mathcal{X}|]$, and $Y \in \mathcal{Y} = [|\mathcal{Y}|]$ are integers). Fix a channel simulation scheme. For each value w of the common randomness W, consider the conditional probability matrix $\mathbf{P}_{Y|X,W=w} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ induced by the scheme, with entries $(\mathbf{P}_{Y|X,W=w})_{x,y} = P_{Y|X,W}(y|x,w)$. Since $M \in [\mathsf{N}]$ and $Y = g(W,M)$, there are at most N possible values of Y for each fixed value w of the common randomness W. Therefore, there are at most N columns in $\mathbf{P}_{Y|X,W=w}$ that are not all zeros, or equivalently, $\|\mathbf{1}^T\mathbf{P}_{Y|X,W=w}\|_0 \le \mathsf{N}$, where $\|\mathbf{a}\|_0 = |\{i : a_i \neq 0\}|$ is the sparsity of $\mathbf{a}$. Therefore, $\mathbf{P}_{Y|X} = \sum_w P_W(w)\mathbf{P}_{Y|X,W=w}$ is a convex combination of conditional probability matrices with at most N nonzero columns. Therefore, we obtain the following formula, which has been first observed by Cubitt *et al.* (2011).

Theorem 24 (D/1/E/FL/KAS/UCR (Cubitt *et al.*, 2011))**.** *For the one-shot exact fixed-length channel simulation setting for finite discrete $\mathcal{X}$, $P_{Y|X}$, with unlimited common*

randomness, and known[29] *or arbitrary source distribution, the optimal description size* N^* *is given by*

$$\mathsf{N}^* = \min\left\{k : \mathbf{P}_{Y|X} \in \operatorname{conv}(\{\mathbf{Q}_{Y|X} : \|\mathbf{1}^T\mathbf{Q}_{Y|X}\|_0 \le k\})\right\}, \tag{3.32}$$

where $\mathbf{Q}_{Y|X}$ *ranges over conditional probability matrices from* $\mathcal{X}$ *to* $\mathcal{Y}$*, where* $\|\mathbf{1}^T\mathbf{Q}_{Y|X}\|_0 = |\{y : \sum_x (\mathbf{Q}_{Y|X})_{x,y} > 0\}| \le k$*, and* $\operatorname{conv}(\mathcal{M})$ *denotes the convex hull of a set of conditional probability matrices* $\mathcal{M}$*.*

Unfortunately, the expression (3.32) is difficult to evaluate, and does not relate to simpler quantities like mutual information as in Theorem 4. Furthermore, the penalty for requiring fixed length description can be significant. For example, consider the channel $Y \in [a]$, $X \in \binom{[a]}{b}$ (where $1 \le b \le a$, and $\binom{[a]}{b}$ denotes the set of subsets of $[a]$ with size b), with $Y|X \sim \operatorname{Unif}(X)$, i.e., the channel takes a size-b subset of $[a]$ as input, and outputs a random element in the subset. Variable-length channel simulation would require an expected length within a logarithmic gap from $\max_{P_X} I(X;Y) = \log_2(a/b)$. On the other hand, for fixed-length channel simulation, we can show that it is impossible to have a scheme with $\mathsf{N} \le a - b$. This is because for a scheme with $\mathsf{N} = a - b$, for each fixed value w of the common randomness W, there are at most $a - b$ possible values of Y, and hence the scheme will fail when the input X is a set that does not contain any of those possible values. We can construct a scheme with $\mathsf{N} = a - b + 1$ by having $W \sim \operatorname{Unif}(\binom{[a]}{a-b+1})$, the encoder chooses a an element in $X \cap W$ randomly and transmit its index in W, and the decoder recovers and outputs that element. This means that $\log_2 \mathsf{N}^* = \log_2(a - b + 1)$, which can be significantly larger than the $L^* \le \log_2(a/b) + \log_2(\log_2(a/b) + 2) + 3$ for variable-length channel simulation in Theorem 4 (e.g., when $a = 2b$, $\log_2 \mathsf{N}^* = \log_2(b + 1)$, whereas $L^* \le \log_2 3 + 4$).

The approximate setting where the output is only required to follow $P_{Y|X}$ approximately is deferred to Section 8.

[29] For known source distribution, we assume $P_X(x) > 0$ for all $x \in \mathcal{X}$.

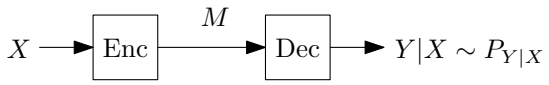


Figure 4.1: One-shot channel simulation without common randomness.

4 One-shot Channel Simulation without Common Randomness

In this section, we study the one-shot channel simulation without common randomness between the encoder and the decoder. It is applicable to scenarios where true common randomness is too expensive, the initial step for establishing synchronized pseudorandom number generators in Section 2.3 is unavailable, or if pseudorandomness is considered unacceptable. Although the setting is similar to Definition 2 where unlimited common randomness is available, the techniques for the case without common randomness are vastly different, and the results are often not stated in terms of the mutual information $I(X;Y)$ (except in Theorem 31 where additional constraints are needed). We state the definition of the setting (including the fixed and variable-length cases, and the known source distribution and arbitrary source cases) here.

Definition 25 (One-shot variable-length exact channel simulation without common randomness)**.** Consider a general (discrete/continuous) channel $P_{Y|X}$ from $\mathcal{X}$ to $\mathcal{Y}$, and a source (discrete/continuous) distribution P_X (for the known source distribution case). A one-shot variable-length channel simulation scheme without common randomness is characterized by a tuple $(\mathcal{C}, P_{M|X}, P_{Y|M})$ described below:

- **Codebook.**
  - For the variable-length setting, the set of possible descriptions $\mathcal{C} \subseteq \{0,1\}^*$ is a prefix-free codebook, which we can design as a part of the coding scheme.
  - For the fixed-length setting, the set of possible descriptions must be $\mathcal{C} = [\mathsf{N}]$, where N is the description size.

- **Encoder.** The encoder observes a source symbol X (with $X \sim P_X$ for the known source distribution case), and sends a description $M \in \mathcal{C}$, $M|X \sim P_{M|X}$ produced by passing X through a conditional distribution $P_{M|X}$ from $\mathcal{X}$ to $\mathcal{C}$ (called the *encoding Markov kernel*).

- **Decoder.** The decoder then outputs $Y|M \sim P_{Y|M}$ produced by passing M through a conditional distribution $P_{Y|M}$ from $\mathcal{C}$ to $\mathcal{Y}$ (called the *decoding Markov kernel*).

- **Requirement.** We require $Y|X \sim P_{Y|X}$ exactly.

- **Performance metric.**
  - For the variable-length setting, we are interested in the smallest expected length $\mathbb{E}[|M|]$ of the prefix-free description M. Let

$$L^* := \begin{cases} \inf \mathbb{E}[|M|] & \text{(known source dist.)} \\ \inf \sup_{x \in \mathcal{X}} \mathbb{E}\big[|M| \,\big|\, X = x\big] & \text{(arbitrary source)} \end{cases}$$

    be the *minimum expected length* (known source distribution) or *minimum worst-case expected length* (arbitrary source), where the infimum is over schemes $(\mathcal{C}, P_{M|X}, P_{Y|X})$ satisfying the requirement.
  - For the fixed-length setting, we are interested in the smallest description size N. Let N^* be the minimum of the set of achievable description sizes among all schemes.

In the following two subsections, we will investigate this setting for discrete channels and continuous channels, respectively.

4.1 Discrete Channels

4.1.1 Fixed-Length Setting

We have

$$(X, M, Y) \sim P_X P_{M|X} P_{Y|M},$$

i.e., the distribution of (X, M, Y) is obtained by first generating $X \sim P_X$, and passing it through the encoding Markov kernel to obtain $M|X \sim P_{M|X}$, and then passing it through the decoding Markov kernel to obtain $Y|M \sim P_{Y|M}$. The only condition this setting imposes on (X, M, Y) is that X, Y are conditionally independent given M, i.e., $X \leftrightarrow M \leftrightarrow Y$ forms a Markov chain. Therefore, for the fixed-length setting, to find the minimum N given $P_{X,Y}$ is equivalent to find the conditional distribution $P_{M|X,Y}$ where $X \leftrightarrow M \leftrightarrow Y$ forms a Markov chain, and M has the smallest cardinality.

Considering that $X \leftrightarrow M \leftrightarrow Y$ forms a Markov chain, the data processing inequality implies that $\log_2 \mathsf{N} \geq H(M) \geq I(X;Y)$ must hold. One channel simulation scheme is to have the encoder generate $Y|X \sim P_{Y|X}$ and encode Y into M (stochastic encoder), and to have the decoder decode M into Y (deterministic decoder). Another scheme is to have the encoder encode X into M (deterministic encoder), and to have the decoder decode M

into X and generate $Y|X \sim P_{Y|X}$ (stochastic decoder). Choosing the better one of the two scheme, we know that $\mathsf{N} = \min\{|\mathcal{X}|, |\mathcal{Y}|\}$ is achievable. Hence, we have the bound

$$2^{I(X;Y)} \le \mathsf{N}^* \le \min\{|\mathcal{X}|, |\mathcal{Y}|\}.$$

Consider the conditional probability matrix $\mathbf{P}_{Y|X} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ with entries $(\mathbf{P}_{Y|X})_{x,y} = P_{Y|X}(y|x)$ (here we assume $X \in \mathcal{X} = [|\mathcal{X}|]$, and $Y \in \mathcal{Y} = [|\mathcal{Y}|]$ are integers). Define $\mathbf{P}_{M|X}$ and $\mathbf{P}_{Y|M}$ similarly. We then have

$$\mathbf{P}_{Y|X} = \mathbf{P}_{M|X}\mathbf{P}_{Y|M}.$$

As shown in (Cubitt *et al.*, 2011; Zhang, 2012; Jain *et al.*, 2013), the minimum cardinality N of the description M is given by the *nonnegative rank* (Berman and Plemmons, 1994; Lee and Seung, 1999) $\mathrm{rank}_+(\mathbf{P}_{Y|X})$, defined as

$$\mathrm{rank}_+(\mathbf{A}) := \min\left\{k \in \mathbb{N}_0 : \exists \mathbf{B} \in \mathbb{R}_{\ge 0}^{m\times k}, \mathbf{C} \in \mathbb{R}_{\ge 0}^{k\times n} : \mathbf{A} = \mathbf{BC}\right\}$$

for $\mathbf{A} \in \mathbb{R}_{\ge 0}^{m\times n}$, where $\mathbb{R}_{\ge 0}^{m\times n}$ is the set of $m \times n$ matrices of nonnegative entries. The factorization $\mathbf{A} = \mathbf{BC}$ is called *nonnegative matrix factorization* (Berman and Plemmons, 1994; Lee and Seung, 1999). The nonnegative rank is lower-bounded by the rank, i.e., $\mathrm{rank}_+(\mathbf{A}) \ge \mathrm{rank}(\mathbf{A})$. Unlike the rank which can be computed efficiently, the computation of $\mathrm{rank}_+(\mathbf{A})$ is NP-hard (Vavasis, 2010). We also refer interested readers to (Cohen and Rothblum, 1993) which discusses properties of the nonnegative rank. The following result holds regardless of whether the source distribution P_X is known, or arbitrary (i.e., the scheme must guarantee $Y|\{X = x\} \sim P_{Y|X}(\cdot|x)$ for every value of $x \in \mathcal{X}$).

Proposition 26 (D/1/E/FL/KAS/NCR (Cubitt *et al.*, 2011; Zhang, 2012))**.** *For the discrete one-shot exact fixed-length channel simulation setting with no common randomness and known*[1] *or arbitrary source distribution (Definition 25), the minimum cardinality of the description is given by the nonnegative rank:*

$$\mathsf{N}^* = \mathrm{rank}_+(\mathbf{P}_{Y|X}).$$

4.1.2 Variable-Length Setting

We now study the variable-length setting where M is no longer an integer in $[\mathsf{N}]$, but a codeword in a prefix-free codebook $\mathcal{C} \subseteq \{0,1\}^*$. The encoding Markov kernel would be $P_{M|X}$ from $\mathcal{X}$ to $\mathcal{C}$, and the decoding Markov kernel would be $P_{Y|M}$ from $\mathcal{C}$ to $\mathcal{Y}$. Our goal is to minimize the expected description length $\mathbb{E}[|M|]$. Utilizing the Huffman code

[1] For known source distribution, we assume $P_X(x) > 0$ for all $x \in \mathcal{X}$.

(Huffman, 1952), we know that the minimum expected length $\mathbb{E}[|M|]$ is bounded between $H(M)$ and $H(M)+1$, and hence the minimum of $H(M)$ would provide a good approximate of the minimum of $\mathbb{E}[|M|]$. This gives the following approximate characterization shown in (Kumar *et al.*, 2014).

Proposition 27 (D/1/E/VL/KS/NCR (Kumar *et al.*, 2014))**.** *For the discrete one-shot exact variable-length channel simulation setting with no common randomness and known source distribution (Definition 25), the minimum entropy $H(M)$ of the description is given by the* common entropy

$$G(X;Y) := \min_{P_{U|X,Y}\colon X \leftrightarrow U \leftrightarrow Y} H(U). \tag{4.1}$$

Hence, the minimum expected length is bounded by $G(X;Y) \le L^* \le G(X;Y)+1$.

The optimization problem in $G(X;Y)$ is non-convex (Kumar *et al.*, 2014), and it is unclear whether there is an efficient algorithm for computing $G(X;Y)$. For the values of $G(X;Y)$ for specific distributions, refer to (6.5) for the case where $P_{Y|X}$ is a binary erasure channel, and to Theorem 31 for a bound when $(X,Y) \in \mathbb{R}^2$ is Gaussian.

The asymptotic version of this setting will be discussed in Section 6. This setting is closely related to the one-shot distributed source simulation problem (Wyner, 1975a; Kumar *et al.*, 2014), where two terminals want to simulate a pair of correlated random variables (X,Y) using common randomness. This will be discussed in Section 9.2.

Yu and Tan (Yu and Tan, 2020) studied a quantity which generalizes both $\log_2 \operatorname{rank}_+(\mathbf{P}_{Y|X})$ and $G(X;Y)$, called *common Rényi entropy* of order $\alpha \in [0,\infty]$, defined as

$$G_\alpha(X;Y) := \min_{P_{U|X,Y}\colon X \leftrightarrow U \leftrightarrow Y} H_\alpha(U),$$

where

$$H_\alpha(U) := \begin{cases} \frac{1}{1-\alpha}\log_2\left(\sum_u (P_U(u))^\alpha\right) & \text{if } \alpha \notin \{0,1,\infty\} \\ \log_2 |\{u:\ P_U(u)>0\}| & \text{if } \alpha = 0 \\ H(U) & \text{if } \alpha = 1 \\ -\log_2 \max_u P_U(u) & \text{if } \alpha = \infty \end{cases}$$

is the Rényi entropy (Rényi, 1961). We have $G_0(X;Y) = \log_2 \operatorname{rank}_+(\mathbf{P}_{Y|X})$ and $G_1(X;Y) = G(X;Y)$.

4.2 Continuous Channels

At first glance, the one-shot channel simulation setting without common randomness appears to be reasonable only for discrete X and Y. For continuous X, Y, it seems that the entropy

of the description must be infinite, for the same reason that the discrete Shannon entropy of a continuous random variable is infinite. This is not true in general, and we can often find a discrete random variable M such that X and Y are conditionally independent given M, making $G(X;Y)$ (4.1) finite.

In fact, we can have an even stronger *universal remote generation* setting (Li and El Gamal, 2018a), where the stochastic encoder observes an arbitrary probability density function $f : \mathbb{R}^n \to \mathbb{R}$ that is continuous almost everywhere,[2] and sends a codeword $M \in \mathcal{C}$ in a prefix-free codebook $\mathcal{C}$, so that the stochastic decoder can output a sample Y following the distribution f using M. To apply this to the channel simulation setting in Definition 25, we can take $f(\mathbf{y}) = f_{\mathbf{Y}|X}(\mathbf{y}|x)$. The existence of a universal remote generation scheme is perhaps counter-intuitive, considering that there are significantly more continuous distributions over $\mathbb{R}^n$ than codewords in a prefix-free codebook, so such "transmission of continuous distributions" should not be possible.

We now describe the constructions by Li and El Gamal (2017) and Li and El Gamal (2018a). In the channel simulation setting without common randomness, if the encoder sends a description m, the decoder will output $Y \sim P_{Y|M}(\cdot|m)$. Therefore, the task is to design the countable collection of distributions $(P_{Y|M}(\cdot|m))_m$ such that any continuous distribution f (that is continuous almost everywhere) can be expressed as a mixture of distributions in that collection. For the one-dimensional case $n = 1$, we take $(P_{Y|M}(\cdot|m))_m$ to be the collection of uniform distributions over *dyadic intervals*, which are intervals in the form $[2^{-k}v,\, 2^{-k}(v+1))$ for $k, v \in \mathbb{Z}$. Any interval can be expressed as a disjoint union of dyadic intervals (possibly differing at the boundary points), by repeatedly including the longest dyadic interval that can fit within the uncovered part. For example, $[2/5, 7/3]$ can be expressed as $[1,2) \cup [1/2,1) \cup [2,9/4) \cup [7/16,1/2) \cup [9/4,37/16) \cup \cdots$. Any continuous distribution can then be expressed as a mixture of uniform distributions over dyadic intervals, by first expressing it as a mixture of uniform distributions over arbitrary intervals (see Section 3.6.2). This is referred to as *dyadic decomposition* by Li and El Gamal (2017) and Li and El Gamal (2018a).

For the general n-dimensional case, the dyadic decomposition (Li and El Gamal, 2017; Li and El Gamal, 2018a) is the decomposition of a distribution into mixture of uniform distributions in the form $\mathrm{Unif}(C_{k,\mathbf{v}})$, where $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{Z}^n$ is an integer vector, $k \in \mathbb{Z}$, and

$$\begin{aligned} C_{k,\mathbf{v}} &:= 2^{-k}([0,1)^n + \mathbf{v}) \\ &= [2^{-k}v_1,\, 2^{-k}(v_1+1)) \times \cdots \times [2^{-k}v_n,\, 2^{-k}(v_n+1)) \end{aligned}$$

is a hypercube, which is referred to as a *dyadic hypercube*. The dyadic hypercubes with side length 2^{-k} partitions $\mathbb{R}^n$, and the partition by dyadic hypercubes with side length $2^{-(k+1)}$

[2] This means that f has a set of discontinuities with measure 0 with respect to the Lebesgue measure over $\mathbb{R}^n$.

is a refinement of the partition by dyadic hypercubes with side length 2^{-k}. The idea is to "discretize" the probability density function $f : \mathbb{R}^n \to \mathbb{R}$ so that the discretized function is constant over each dyadic hypercube with side length 2^{-k}. Let

$$\begin{aligned}\tilde{f}_k(\mathbf{y}) &:= \inf f(C_{k,\lfloor 2^k \mathbf{y} \rfloor}) \\ &= \sum_{\mathbf{v} \in \mathbb{Z}^n} (\inf f(C_{k,\mathbf{v}})) \mathbf{1}_{C_{k,\mathbf{v}}}(\mathbf{y}),\end{aligned}$$

where $\lfloor \mathbf{z} \rfloor := (\lfloor z_1 \rfloor, \ldots, \lfloor z_n \rfloor)$ for $\mathbf{z} \in \mathbb{R}^n$, and $\inf f(S) = \inf\{f(\mathbf{y}) : \mathbf{y} \in S\}$. We can see that $\tilde{f}_k$ is the largest function upper-bounded by f, that is constant over each dyadic hypercube with side length 2^{-k}, or equivalently, can be decomposed into a weighted sum of indicator functions $\mathbf{1}_{C_{k,\mathbf{v}}}(\mathbf{y})$ of such dyadic hypercubes.

Let $\tilde{f}_\infty(\mathbf{y}) := \lim_{k \to \infty} \tilde{f}_k(\mathbf{y})$. We have $\int \tilde{f}_\infty(\mathbf{y}) \mathrm{d}\mathbf{y} = 1$ if f is continuous almost everywhere,[3] and hence $\tilde{f}_\infty$ is the probability density function of the same distribution as f. Hence, we can decompose this distribution as

$$\tilde{f}_\infty(\mathbf{y}) = \sum_{k=-\infty}^{\infty} (\tilde{f}_k(\mathbf{y}) - \tilde{f}_{k-1}(\mathbf{y})),$$

where $\tilde{f}_k(\mathbf{y}) - \tilde{f}_{k-1}(\mathbf{y})$ is constant over each dyadic hypercube with side length 2^{-k}. This way, $\tilde{f}_\infty$ can be decomposed into a weighted sum of indicator functions of dyadic hypercubes. More explicitly, the distribution f can be expressed as the following mixture of uniform distributions over dyadic hypercubes:

$$\sum_{k \in \mathbb{Z}, \mathbf{v} \in \mathbb{Z}^n} 2^{-nk} (\inf f(C_{k,\mathbf{v}}) - \inf f(C_{k-1,\lfloor \mathbf{v}/2 \rfloor})) \mathrm{Unif}(C_{k,\mathbf{v}}).$$

Given such a decomposition of the distribution f into a mixture of uniform distributions over dyadic hypercubes, the universal remote generation scheme (Li and El Gamal, 2018a) operates as follows:

- The encoder observes f, and generate a dyadic hypercube $C_{k,\mathbf{v}}$, $k \in \mathbb{Z}$, $\mathbf{v} \in \mathbb{Z}^n$ by sampling from the probability mass function

$$p(k, \mathbf{v}) = 2^{-nk} (\inf f(C_{k,\mathbf{v}}) - \inf f(C_{k-1,\lfloor \mathbf{v}/2 \rfloor})). \tag{4.2}$$

  The encoder encodes $k, \mathbf{v}_1, \ldots, \mathbf{v}_n$ into M using any code over integers, for example, by concatenating their signed Elias delta encodings (Elias, 1975) of $k, \mathbf{v}_1, \ldots, \mathbf{v}_n$.

[3] $\tilde{f}_\infty(\mathbf{y}) < f(\mathbf{y})$ only when $\mathbf{y}$ is a point of discontinuity, since for any k, we can find a dyadic hypercube with side length 2^{-k} containing both $\mathbf{y}$ and a point $\mathbf{y}'$ with $f(\mathbf{y}') < (f(\mathbf{y}) + \tilde{f}_\infty(\mathbf{y}))/2$, and hence we can find points $\mathbf{y}'$ arbitrarily close to $\mathbf{y}$ with $f(\mathbf{y}')$ bounded away from $f(\mathbf{y})$. Therefore, $\int (f(\mathbf{y}) - \tilde{f}_\infty(\mathbf{y})) \mathrm{d}y = 0$ since the set of points of discontinuity has measure 0.

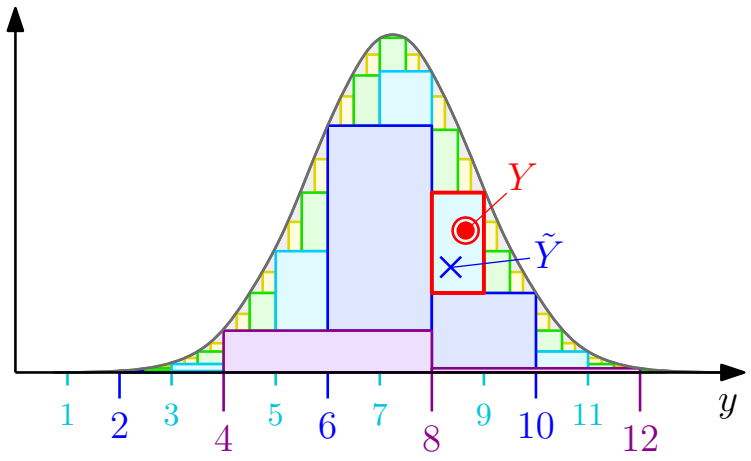


Figure 4.2: The dyadic decomposition scheme, applied on $P_{Y|X}(\cdot|x)$ being a Gaussian distribution f. The purple rectangles are the largest rectangles with width 4 and horizontal coordinates of corners being multiples of 4 (i.e., the horizontal range is a dyadic interval with length 4) that fit below f. The blue rectangles are the largest rectangles staking on top of the purple rectangles with horizontal ranges being dyadic intervals of length 2 that fit below f, and so on.
According to (4.3), the encoder generate $\tilde{Y} \sim f$, $U|\tilde{Y} \sim \text{Unif}(0, f(\tilde{Y}))$, and take $K = \min\{k \in \mathbb{Z} : \inf f(2^{-k}([0,1) + \lfloor 2^k \tilde{Y} \rfloor)) \geq U\}$ and $V = \lfloor 2^K \tilde{Y} \rfloor$. This is equivalent to first generating a point $(\tilde{Y}, Uf(\tilde{Y}))$ uniformly over the region between the graph of f and the horizontal axis (the blue cross in the figure), selecting the rectangle containing that point, and transmitting the $K, V \in \mathbb{Z}$ corresponding to the horizontal range $2^{-K}([0,1) + V)$ of that rectangle ($(K, V) = (0, 8)$ in the figure). The decoder recovers the horizontal range $2^{-K}([0,1) + V)$, and outputs a point uniformly distributed over this range (the red point in the figure), which may not be the same as the point generated by the encoder (the blue cross).

- The decoder decodes $k, \mathbf{v}$ from M, and generates Y uniformly over $C_{k,\mathbf{v}}$.

One method to sample from the distribution (4.2) is to generate $\tilde{\mathbf{Y}} \sim f, U|\tilde{\mathbf{Y}} \sim \text{Unif}(0, f(\tilde{\mathbf{Y}}))$, and take

$$K = \min\left\{k \in \mathbb{Z} : \tilde{f}_k(\tilde{\mathbf{Y}}) \geq U\right\}, \ \mathbf{V} = \lfloor 2^K \tilde{\mathbf{Y}} \rfloor. \tag{4.3}$$

Refer to Figure 4.2 for an illustration. The algorithm in (Li and El Gamal, 2018a) is given in Algorithm 5.

4.2.1 The Communication Costs of Dyadic Decomposition Schemes

If $\mathcal{Y} = [-b/2, b/2]^n$ is bounded, then for a fixed k, the coordinates $\mathbf{v}_1, \ldots, \mathbf{v}_n$ in the universal remote generation scheme are bounded, and hence we can use a fixed-length code to encode them. Applying the universal remote generation scheme on $f = f_{\mathbf{Y}|X}(\cdot|x)$ gives the following result on channel simulation with arbitrary source (Li and El Gamal, 2018a), which applies

Algorithm 5 Dyadic decomposition (Li and El Gamal, 2017; Li and El Gamal, 2018a)

Procedure ENCODE(f) :

Input: almost everywhere continuous density $f : \mathbb{R}^n \to \mathbb{R}$

Output: $k \in \mathbb{Z}$, $\mathbf{v} \in \mathbb{Z}^n$

1: Generate $\tilde{\mathbf{y}} \sim f$

2: Generate $u \sim \text{Unif}(0, f(\tilde{\mathbf{y}}))$

3: $k \leftarrow \min\left\{k \in \mathbb{Z} : \inf f(C_{k,\lfloor 2^k \tilde{\mathbf{y}} \rfloor}) \geq u\right\}$ ▷ $C_{k,\mathbf{v}} := 2^{-k}([0,1)^n + \mathbf{v})$

4: $\mathbf{v} \leftarrow \lfloor 2^k \tilde{\mathbf{y}} \rfloor$

5: **return** $k, \mathbf{v}$

Procedure DECODE($k, \mathbf{v}$) :

Input: $k \in \mathbb{Z}$, $\mathbf{v} \in \mathbb{Z}^n$

Output: sample $\mathbf{Y}$

1: **return** $\mathbf{y} \sim \text{Unif}(C_{k,\mathbf{v}})$

to bounded orthogonally concave distributions.[4] Readers are referred to (Li and El Gamal, 2018a) for the proof.

Theorem 28 (C/1/E/VL/AS/NCR (Li and El Gamal, 2018a))**.** *For the one-shot exact variable-length channel simulation setting with no common randomness and arbitrary source (Definition 25), where $\mathcal{Y} = [-b/2, b/2]^n$ and $\mathcal{X}$ is arbitrary, if $P_{\mathbf{Y}|X}(\cdot|x)$ is continuous and always has an orthogonally concave probability density function $f_{\mathbf{Y}|X}(\cdot|x)$ satisfying $\sup_{x,\mathbf{y}} f_{\mathbf{Y}|X}(\mathbf{y}|x) \leq c$, then the minimum worst-case expected length is bounded by*

$$L^* \leq n \log_2(4enb^n c) + 2 \log_2 \log_2(8enb^n c) + 1.$$

The bound on the communication cost when f is orthogonally concave but not necessarily bounded, which is considerably more complicated, is given below (Li and El Gamal, 2018a).

Theorem 29 (Universal remote generation (Li and El Gamal, 2018a))**.** *There is a universal remote generation scheme for generating arbitrary continuous distributions f over $\mathbb{R}^n$ that is continuous almost everywhere, satisfying that when f is orthogonally concave, we have*

[4] A set $S \subseteq \mathbb{R}^n$ is *orthogonally convex* if for any line L parallel to one of the n axes, $L \cap A$ is a connected set, i.e., empty, a point or an interval. A functional $f : \mathbb{R}^n \to \mathbb{R}$ is *orthogonally concave* if the hypograph $\text{hyp}(f) := \{(\mathbf{y}, z) : \mathbf{y} \in \mathbb{R}^n, z \leq f(y)\} \subseteq \mathbb{R}^{n+1}$ is orthogonally convex. A distribution is orthogonally concave if its probability density function is orthogonally concave. Any log-concave or quasiconcave function is orthogonally concave.

an expected description length $\mathbb{E}[|M|] \le \mathrm{UR}(f)$*, where*

$$\mathrm{UR}(f) := \inf_{\hat{\mathbf{y}} \in \mathbb{R}^n} \Big\{ n\ell_\delta((n-1)\log_2 r_{\hat{\mathbf{y}}} + \log_2(\|\hat{\mathbf{y}}\|_\infty + r_{\hat{\mathbf{y}}}) + \log_2 c + 4n + 8)$$
$$+ \ell_\delta(\log_2\left((n-1)\log_2 r_{\hat{\mathbf{y}}} + 2\max\{\log_2 r_{\hat{\mathbf{y}}}, 0\} + \log_2 c + 4n + 10\right) + 2)\Big\},$$

where $c := \sup_{\mathbf{y}} f(\mathbf{y})$*,* $r_{\hat{\mathbf{y}}} := \mathbb{E}_{\mathbf{Y}\sim P}[\|\mathbf{Y} - \hat{\mathbf{y}}\|_\infty]$*, and* $\ell_\delta(t) := t + 2\log_2 t$*.*

Applying the universal remote generation scheme on the distribution $P_{\mathbf{Y}|X}(\cdot|x)$, we have the following result for the channel simulation setting in Definition 25.

Corollary 30 (C/1/E/VL/AS/NCR (Li and El Gamal, 2018a))**.** *For the one-shot exact variable-length channel simulation setting with no common randomness and arbitrary source (Definition 25), if* $P_{\mathbf{Y}|X}(\cdot|x)$ *always has an orthogonally concave probability density function* $f_{\mathbf{Y}|X}(\cdot|x)$ *over* $\mathbb{R}^n$*, then the minimum worst-case expected length is bounded by*

$$L^* \le \sup_{x\in\mathcal{X}} \mathrm{UR}(f_{\mathbf{Y}|X}(\cdot|x)),$$

where UR *is defined in Theorem 29.*

We now study the known source distribution setting. For the special case where $X, Y \in \mathbb{R}$, and the joint probability density function $f_{X,Y}$ is log-concave, i.e., $\log_2 f_{X,Y}(x,y)$ is a concave function (which holds when X, Y are jointly Gaussian, or when (X,Y) is uniform over a convex subset of $\mathbb{R}^2$), then it was shown by Li and El Gamal (2017) that the entropy and the expected length of the description M can be upper-bounded by $I(X;Y)$ plus a constant.[5] The construction in (Li and El Gamal, 2017) is to apply dyadic decomposition not on $f_{Y|X}$ or $f_{X,Y}$, but on the uniform distribution $(X,Y,Z) \sim \mathrm{Unif}(\mathcal{S})$ over the positive part of the hypograph $\mathcal{S} := \{(x,y,z) : 0 \le z \le f_{X,Y}(x,y)\} \subseteq \mathbb{R}^3$. After decomposing $\mathcal{S}$ into dyadic cubes, the encoder can generate $(Y,Z)|X \sim P_{Y,Z|X}$ based on the observed X, and send the size and location of the dyadic cube containing (X,Y,Z) to the decoder. The decoder can generate $(\tilde{X},\tilde{Y},\tilde{Z})$ uniformly over the dyadic cube, and output $\tilde{Y}$. We also need to apply scaling on $\mathcal{S}$ in order to obtain the desired bound. Readers are referred to (Li and El Gamal, 2017) for details.

Theorem 31 (C/1/E/VL/KS/NCR (Li and El Gamal, 2017))**.** *For jointly continuous random variables* $X, Y \in \mathbb{R}$*, if the joint probability density function* $f_{X,Y}$ *is log-concave, then*

$$G(X;Y) \le I(X;Y) + 24 \text{ bits}.$$

[5] Note that here we allow M to be any discrete (finite or countably infinite) random variable, as long as $H(M)$ is finite. It is uncertain whether the minimum $\min_{P_{W|X,Y}: X\leftrightarrow W\leftrightarrow Y} H(W)$ exists for continuous (X,Y), so we instead take the infimum $G(X;Y) = \inf_{P_{W|X,Y}: X\leftrightarrow W\leftrightarrow Y} H(W)$ here.

Therefore, for the one-shot exact variable-length channel simulation setting with no common randomness and known source distribution (Definition 25), the minimum expected length satisfies $L^* \leq I(X;Y) + 25$.

5 Asymptotic Channel Simulation with Unlimited Common Randomness

5.1 Exact Variable-Length Channel Simulation

In this section, we discuss the asymptotic exact channel simulation setting with variable-length description and unlimited common randomness. Consider the variable-length setting in Definition 2, where the source is an i.i.d. sequence $X^n = (X_1, \dots, X_n) \sim P_X^n$ (where P_X^n denotes the i.i.d. distribution of length n where each entry follows P_X; the n here is called the *blocklength*), the output $Y^n = (Y_1, \dots, Y_n)$ is a sequence as well, and the channel to be simulated is a memoryless channel $P_{Y^n|X^n} = P_{Y|X}^n$ with $P_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^n P_{Y|X}(y_i|x_i)$. Let L_n^* be the infimum of the set of achievable expected lengths among all codes with blocklength n. The goal is to characterize the optimal asymptotic rate

$$R^* = \limsup_{n\to\infty} \frac{L_n^*}{n}, \tag{5.1}$$

i.e., the optimal asymptotic number of bits used per symbol simulated. For the arbitrary source case, we require the code to work for every sequence $x^n \in \mathcal{X}^n$, L_n^* is defined as (2.1), and R^* is defined in the same way as (5.1).

One can always simply substitute a one-shot result into (5.1) to give a characterization or a bound for R^*. For example, using Theorem 4, and observing that $\log_2(I(X^n;Y^n)+2)+3 = o(n)$ as $n \to \infty$, we have

$$R^* = \limsup_{n\to\infty} \frac{I(X^n;Y^n)}{n}. \tag{5.2}$$

This is called an *n-letter characterization* since it involves a sequence of random variables with length n that tends to infinity. In asymptotic settings, we are more interested in *single-letter characterizations*, i.e., expressions where the number of variables is fixed. Fortunately, we have $I(X^n;Y^n) = nI(X;Y)$ since $(X_i, Y_i) \overset{iid}{\sim} P_X P_{Y|X}$, and hence (5.2) is simply $R^* = I(X;Y)$, which is a single-letter characterization.

The following result, called the reverse Shannon theorem, was first proved in (Bennett *et al.*, 2002) for the discrete case using the method of types. Here we prove it as a direct corollary of Theorem 4. We will also briefly discuss the proof by the method of types in Section 5.2.

Theorem 32 (G/∞/E/VL/KAS/UCR (Bennett *et al.*, 2002))**.** *For the asymptotic exact variable-length channel simulation setting for a general (discrete/continuous) channel $P_{Y|X}$ with unlimited common randomness*:

- *For known source distribution, the optimal rate is $R^* = I(X;Y)$.*

- *For arbitrary source, the optimal rate is the channel capacity* $R^* = C := \sup_{P_X} I(X;Y)$.

Proof. For known source distribution, applying Theorem 4 on (X^n, Y^n), we have

$$nI(X;Y) = I(X^n;Y^n) \le L_n^* \le nI(X;Y) + \log_2(nI(X;Y)+2) + 3.$$

Dividing by n and taking $n \to \infty$ give the desired result. The same holds for the arbitrary source case. □

Let $R_n^* := L_n^*/n$ be the optimal rate when the blocklength is n. The proof of Theorem 32 shows that

$$I(X;Y) \le R_n^* \le I(X;Y) + \frac{\log_2 n}{n} + O\left(\frac{1}{n}\right)$$

as $n \to \infty$. It is natural to ask whether R_n^* is closer to the upper bound or the lower bound. It turns out that R_n^* is usually in the middle. In the remainder of this subsection, assume X and Y are discrete with finite supports. It was shown by Sriramu and Wagner (2024) that

$$R_n^* \le I(X;Y) + \frac{\log_2 n}{2n} + o\left(\frac{\log_2 n}{n}\right)$$

if the channel $P_{Y|X}$ is non-singular, i.e., $\frac{\mathrm{d}P_{Y|X}(\cdot|X)}{\mathrm{d}P_Y}(Y)$ is not a deterministic function of Y.[1] Moreover,

$$R_n^* = I(X;Y) + \frac{\log_2 n}{2n} + o\left(\frac{\log_2 n}{n}\right)$$

if $P_{Y|X}$ is non-singular and $P_{X,Y}(x,y) > 0$ for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$. In case if $P_{Y|X}$ is singular,

$$R_n^* = I(X;Y) + o\left(\frac{\log_2 n}{n}\right).$$

Interested readers are referred to (Sriramu and Wagner, 2024) for the proofs of the above results, where a rejection sampling technique is employed.

The fixed-length approximate version of Theorem 32 will be discussed in Section 5.3 onward.

5.2 Method of Types

We briefly discuss the proof in (Bennett *et al.*, 2002) for Theorem 32 for discrete X, Y based on the method of types. Given a sequence $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$, its *type* (Csiszár, 1998; Cover, 1999) refers the empirical distribution induced by x^n, denoted as

$$\mathbf{P}_{x^n} : \mathcal{X} \to [0,1],\ \mathbf{P}_{x^n}(x) := \frac{|\{i \in [n] : x_i = x\}|}{n}.$$

[1] This means there is a positive probability that $\frac{\mathrm{d}P_{Y|X}(\cdot|X)}{\mathrm{d}P_Y}(Y) \neq \mathbb{E}\left[\frac{\mathrm{d}P_{Y|X}(\cdot|X)}{\mathrm{d}P_Y}(Y) \middle| Y\right]$ when $(X,Y) \sim P_{X,Y}$.

Let

$$\mathcal{P}_n(\mathcal{X}) := \{\mathbf{P}_{x^n} : x^n \in \mathcal{X}^n\}$$

be the *set of types with denominator* n (Csiszár, 1998; Cover, 1999), which contains probability mass functions over $\mathcal{X}$ where each probability is a multiple of $1/n$. We can divide the set of sequences $\mathcal{X}^n$ into *type classes* according to their types, where sequences in each type class have the same type. While there are $|\mathcal{X}|^n$ sequences in $\mathcal{X}^n$, which is exponential in n, there are only $|\mathcal{P}_n(\mathcal{X})| = \binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1}$ type classes by the standard stars and bars argument, which is polynomial in n. Similarly, for $x^n \in \mathcal{X}^n$, $y^n \in \mathcal{Y}^n$, their *joint type* is

$$\mathbf{P}_{x^n,y^n} : \mathcal{X} \times \mathcal{Y} \to [0,1],\ \mathbf{P}_{x^n,y^n}(x,y) := \frac{|\{i \in [n] : (x_i, y_i) = (x,y)\}|}{n},$$

which is simply the type of the sequence of pairs $(x_i, y_i)_{i\in[n]}$.

To generate $Y^n|X^n \sim P_{Y|X}^n$ given X^n, we can first generate the joint type $\mathbf{P}_{X^n,Y^n}$ conditional on X^n, and then generate Y^n conditional on $(X^n, \mathbf{P}_{X^n,Y^n})$. The idea is to offload the randomness needed in the second step to the common randomness. Fix $\epsilon > 0$. The scheme in (Bennett *et al.*, 2002) operates as follows:

- Generate the common randomness $W := (\bar{Y}_{p,i}^n)_{p,i}$ as a collection of sequences $\bar{Y}_{p,i}^n \in \mathcal{Y}^n$, where $p \in \mathcal{P}_n(\mathcal{X})$ ranges over the set of types over $\mathcal{X}$, $i \in [2^{n(I(p,P_{Y|X})+\epsilon)}]$ (where $I(p, P_{Y|X})$ denotes $I(\tilde{X};\tilde{Y})$ when $(\tilde{X},\tilde{Y}) \sim pP_{Y|X}$), and the entries of $\bar{Y}_{p,i}^n$ are generated from the Y-marginal distribution of $pP_{Y|X}$, independently across p, i.

- Given X^n, the encoder generates the joint type $Q \in \mathcal{P}_n(\mathcal{X},\mathcal{Y})$ following the conditional distribution of $\mathbf{P}_{X^n,Y^n}$ given X^n. To do so, the encoder can simply generate $\tilde{Y}^n|X^n \sim P_{Y|X}^n$, and then take $Q = \mathbf{P}_{X^n,\tilde{Y}^n}$.

- The encoder then takes $p = \mathbf{P}_{X^n}$, and finds i such that the joint type $\mathbf{P}_{X^n,\bar{Y}_{p,i}^n} = Q$, and encodes $(0,p,i)$ into the description M, which takes $\approx n(I(p,P_{Y|X}) + \epsilon)$ bits since $p \in \mathcal{P}_n(\mathcal{X})$ only has a polynomial number of choices. If there are multiple i's satisfying this requirement, select the smallest i. If there is no such i, the encoder encodes $(1,\tilde{Y}^n)$ into the description M, which takes $\approx n\log_2|\mathcal{Y}|$ bits.

- The decoder decodes M into either $(0,p,i)$ or $(1,\tilde{Y}^n)$, and outputs either $Y^n = \bar{Y}_{p,i}^n$ or $Y^n = \tilde{Y}^n$ respectively.

To check the correctness of the scheme, note that the conditional distribution of Y^n given $(X^n, \mathbf{P}_{X^n,Y^n})$ is uniform over the set $\{y^n \in \mathcal{Y}^n : \mathbf{P}_{X^n,y^n} = \mathbf{P}_{X^n,Y^n}\}$. Since $P_{Y^n|X^n}(y^n|x^n)$ depends only on the joint type $\mathbf{P}_{x^n,y^n}$, if we fix the joint type, then every y^n with that joint

type with X^n has the same probability. By symmetry, we can see that Y^n output by the scheme follows the correct conditional distribution given $(X^n, \mathbf{P}_{X^n,Y^n})$, and hence follows the correct conditional distribution given X^n.

If we can show that the encoder can find i such that $\mathbf{P}_{X^n,\bar{Y}^n_{p,i}} = Q$ with probability approaching 1 as $n \to \infty$, and ϵ is small, then the expected length would be $\approx n\mathbb{E}[I(\mathbf{P}_{X^n}, P_{Y|X})] \approx nI(X;Y)$ for known source distribution P_X (since $\mathbf{P}_{X^n} \approx P_X$ by law of large numbers), or upper-bounded by $\approx n \max_{p\in\mathcal{P}_n(\mathcal{X})} I(p, P_{Y|X}) \le nC$ for arbitrary source, giving the desired result in Theorem 32.

Therefore, it is left to show that there exists $i \in [2^{n(I(p,P_{Y|X})+\epsilon)}]$ with $\mathbf{P}_{X^n,\bar{Y}^n_{p,i}} = Q$ with probability approaching 1. The proof requires several standard techniques in the method of types, which will not be included in this monograph since we will not be using the method of types in any other part of this monograph. Interested readers are referred to (Bennett *et al.*, 2002) for the proof, and (Csiszár, 1998; Cover, 1999) for techniques in the method of types. Very loosely speaking, we have to choose $2^{n(I(p,P_{Y|X})+\epsilon)}$ sequences in $\mathcal{Y}^n$ to "cover" the type class p, so that for most sequences $x^n \in \mathcal{X}^n$ with $\mathbf{P}_{x^n} = p$, we can find a chosen sequence that looks like it cound be coming from the conditional distribution $P^n_{Y|X}(\cdot|x^n)$. The general theme "we need $\approx 2^{nI(X;Y)}$ sequences in $\mathcal{Y}^n$ to cover the sequence $X^n \sim P^n_X$ with high probability" is the same as that of the covering lemma for lossy compression (El Gamal and Kim, 2011), and also for the soft covering lemma, a powerful technique for proving channel simulation results that will be discussed in Section 5.5.

5.3 Total Variation Distance

Most schemes discussed in this monograph so far are variable-length schemes. Nevertheless, for asymptotic settings, we are often more interested in fixed-length schemes, i.e., we require the description $M \in \{0,1\}^{\lfloor nR \rfloor}$ (or $M \in [\lfloor 2^{nR} \rfloor]$) to fit within $\approx nR$ bits, where R is the rate. The reason is that taking $n \to \infty$ allows us to utilize the law of large numbers to argue that the length of M should concentrate around its mean, and hence there is no longer a strong reason to allow the flexibility of variable-length schemes. Indeed, one way to prove an asymptotic fixed-length result is to simply concatenate many copies of the variable-length code given by Theorem 4, and then either truncate M if it is too long, or pad M with zeros if it is too short, in order to fix its length.

One caveat of fixed-length settings is that it may no longer be possible to simulate the desired channel exactly, i.e., having Y^n follow the conditional distribution $P_{Y^n|X^n}$ exactly. If we only allow $\lfloor nR \rfloor$ bits for the description, it may be possible that some particular sequences X^n's require more than $\lfloor nR \rfloor$ bits, resulting in a distortion in the distribution of Y^n. In the aforementioned truncation approach, if M is too long and has to be truncated, it will affect the distribution of Y^n. In channel coding, this corresponds to an error event

where the decoder outputs the wrong message. In channel simulation, there is no "wrong" output Y^n. Instead of the error probability, we will control the distance between the ideal joint distribution of X^n, Y^n and the actual distribution.

The *total variation distance* (Csiszár and Körner, 2011) (also called *variational distance*, *statistical distance* and *statistical difference*) is a distance between probability distributions defined as follows.

Definition 33 (Total variation distance (Csiszár and Körner, 2011))**.** For two probability distributions P, Q over the same measurable space $(\Omega, \mathcal{F})$, the *total variation (TV) distance* is given as the largest difference between the probabilities they can assign to the same event, i.e.,

$$\delta_{\mathrm{TV}}(P, Q) := \sup_{E \in \mathcal{F}} |P(E) - Q(E)|. \tag{5.3}$$

Note that $\delta_{\mathrm{TV}}(P, Q) = 0$ if and only if $P = Q$. We have the triangle inequality

$$\delta_{\mathrm{TV}}(P_1, P_3) \le \delta_{\mathrm{TV}}(P_1, P_2) + \delta_{\mathrm{TV}}(P_2, P_3),$$

which follows directly from definition, and hence δ_{TV} is a metric.

A small $\delta_{\mathrm{TV}}(P, Q)$ implies that P and Q assign similar probabilities to the same event, and hence they are difficult to be distinguished. Suppose we perform an action on X (which may follow P or Q) with an outcome $f(X)$ (e.g., f can be a statistical test), then having a small TV distance $\delta_{\mathrm{TV}}(P, Q) \le \epsilon$ guarantees that $|\mathbb{P}_{X\sim P}(f(X) = y) - \mathbb{P}_{X\sim Q}(f(X) = y)| \le \epsilon$ for every y, meaning that it is difficult to distinguish P and Q by looking at the outcome. This makes the TV distance a popular metric in security and cryptography applications (Oded, 2004), for example, to guarantee that the ciphertext cannot be distinguished from pure noise unless one has the key.

In the context of channel simulation, having a small TV distance $\delta_{\mathrm{TV}}(P_{X,Y}, P_{X,\tilde{Y}}) \le \epsilon$, where X is the input, Y is the ideal output following $P_{Y|X}$, and $\tilde{Y}$ is the actual output of the approximate scheme, guarantees that the probability of any event that depends on (X, Y) will not be greatly affected by the inexactness of the scheme. For example, the probability of excess distortion $\mathbb{P}(d(X, Y) > D)$ (Section 1.5) will be increased by at most ϵ if we use $\tilde{Y}$ instead of Y. If (X, Y) are further processed (e.g., through the neural networks discussed in Section 1.4), the probability of any failure event in the downstream will also be increased by at most ϵ if we use $\tilde{Y}$ instead of Y.[2] Readers are referred to the discussions in (Flamich and Wells, 2024) for the use of TV distance in channel simulation.

[2]This argument holds if the channel simulation scheme is used once. However, if we apply a scheme with $\delta_{\mathrm{TV}}(P_{X,Y}, P_{X,\tilde{Y}}) \le \epsilon$ on each entry in a sequence $X_1, \ldots, X_n$ to give $\tilde{Y}_1, \ldots, \tilde{Y}_n$, then we only have the bound $\delta_{\mathrm{TV}}(P_{X^n,Y^n}, P_{X^n,\tilde{Y}^n}) \le n\epsilon$, and the TV distance is increased by at most n fold. This highlights the advantage of exact schemes, where the TV distance is 0 and will not accumulate when the scheme is used more than once.

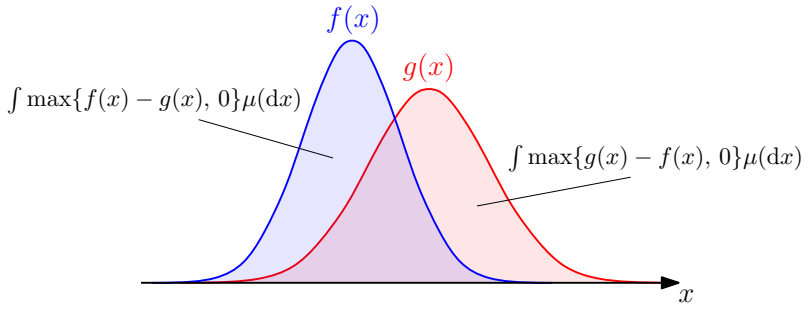


Figure 5.1: For two probability density functions $f(x)$, $g(x)$, the region under $f(x)$ but not under $g(x)$ has area $\int \max\{f(x) - g(x), 0\}\mu(\mathrm{d}x)$, whereas the region under $g(x)$ but not under $f(x)$ has area $\int \max\{g(x) - f(x), 0\}\mu(\mathrm{d}x)$. Both areas are equal to the total variation distance $\delta_{\mathrm{TV}}(f, g)$.

There are several equivalent characterizations of the total variation distance given below (e.g., see (Han and Verdú, 1993; Cuff, 2013)). Refer to Figure 5.1 for an illustration.

Proposition 34 (Equivalent characterizations of total variation distance)**.** *We have*

- $$\delta_{\mathrm{TV}}(P, Q) = \sup_{E \in \mathcal{F}} \left(P(E) - Q(E)\right). \tag{5.4}$$

- $$\delta_{\mathrm{TV}}(P, Q) = \sup_{\psi:\Omega \to [0,1]} \left|\mathbb{E}_{X\sim P}[\psi(X)] - \mathbb{E}_{X\sim Q}[\psi(X)]\right|, \tag{5.5}$$

  where the supremum is over measurable functions $\psi : \Omega \to [0, 1]$.

- *If* $P \ll \mu$ *and* $Q \ll \mu$ *for a sigma-finite measure* μ *over* $(\Omega, \mathcal{F})$,[3] *with density functions* $f(x) := \frac{\mathrm{d}P}{\mathrm{d}\mu}(x)$, $g(x) := \frac{\mathrm{d}Q}{\mathrm{d}\mu}(x)$, *then*

$$\delta_{\mathrm{TV}}(P, Q) = \frac{1}{2}\int |f(x) - g(x)|\mu(\mathrm{d}x) \tag{5.6}$$

$$= \int \max\{f(x) - g(x), 0\}\mu(\mathrm{d}x) \tag{5.7}$$

$$= \int \max\{g(x) - f(x), 0\}\mu(\mathrm{d}x). \tag{5.8}$$

  In particular, if P, Q *are discrete distributions, then*

$$\delta_{\mathrm{TV}}(P, Q) = \frac{1}{2}\sum_x |P(x) - Q(x)|$$

[3] We can always find such a μ, for example, $\mu = (P + Q)/2$.

$$= \sum_x \max\{P(x) - Q(x),\, 0\}$$
$$= \sum_x \max\{Q(x) - P(x),\, 0\},$$

where x takes values over the union of the supports of P and Q, i.e., the TV distance is simply half of the ℓ_1 distance between the probability vectors.

- *Write $\Gamma(P, Q)$ for the set of couplings of P and Q, where a probability measure γ over the product measurable space $(\Omega^2, \mathcal{F} \otimes \mathcal{F})$ [4] is called a coupling of P and Q if $(X, Y) \sim \gamma$ implies $X \sim P$ and $Y \sim Q$. Then*

$$\delta_{\mathrm{TV}}(P, Q) = \min_{\gamma \in \Gamma(P,Q)} \mathbb{P}_{(X,Y)\sim\gamma}(X \neq Y). \tag{5.9}$$

This is known as the coupling lemma *(see (Lindvall, 2002) and Mitzenmacher and Upfal, 2017, Lemma 51).*

Proof. These equalities can be considered folklore, though we briefly prove them for the sake of completeness.

- (5.3)=(5.4): For every $E \in \mathcal{F}$, we can find $E' \in \mathcal{F}$ such that $|P(E) - Q(E)| = P(E') - Q(E')$ by taking $E' = E$ or $E' = E^c$.
- (5.3)≤(5.5): For every $E \in \mathcal{F}$, take $\psi(x) = \mathbf{1}_E(x)$ to be the indicator function of E. We have $|\mathbb{E}_{X\sim P}[\psi(X)] - \mathbb{E}_{X\sim Q}[\psi(X)]| = |P(E) - Q(E)|$.
- (5.3)≥(5.5): Let $L_t^+(\psi) := \{x \in \mathcal{X} : \psi(x) \geq t\}$ be the superlevel set of ψ. We have

$$\mathbb{E}_{X\sim P}[\psi(X)] = \mathbb{E}_{X\sim P}\left[\int_0^1 \mathbf{1}\{t \in L_t^+(\psi)\}\mathrm{d}t\right]$$
$$= \int_0^1 P(L_t^+(\psi))\mathrm{d}t.$$

Hence,

$$|\mathbb{E}_{X\sim P}[\psi(X)] - \mathbb{E}_{X\sim Q}[\psi(X)]|$$
$$= \left|\int_0^1 \left(P(L_t^+(\psi)) - Q(L_t^+(\psi))\right)\mathrm{d}t\right|$$
$$\leq \left|\int_0^1 \delta_{\mathrm{TV}}(P, Q)\mathrm{d}t\right|$$
$$= \delta_{\mathrm{TV}}(P, Q).$$

[4] $\mathcal{F} \otimes \mathcal{F}$ is the smallest sigma-algebra containing $\mathcal{F} \times \mathcal{F}$.

- (5.7)=(5.8):

$$\begin{aligned}&\int \left(\max\{f(x)-g(x),\,0\}-\max\{g(x)-f(x),\,0\}\right)\mu(\mathrm{d}x)\\&=\int \left(f(x)-g(x)\right)\mu(\mathrm{d}x)\\&=0.\end{aligned}$$

- (5.7)=(5.6):

$$\begin{aligned}&\int \max\{f(x)-g(x),\,0\}\mu(\mathrm{d}x)\\&=\frac{1}{2}\int \left(\max\{f(x)-g(x),\,0\}+\max\{g(x)-f(x),\,0\}\right)\mu(\mathrm{d}x)\\&=\frac{1}{2}\int |f(x)-g(x)|\mu(\mathrm{d}x).\end{aligned}$$

- (5.9)≤(5.7): Write $\delta := \delta_{\mathrm{TV}}(P,Q)$. Define the probability density functions (with respect to μ)

$$\begin{aligned}h(x)&:=\frac{\min\{f(x),g(x)\}}{1-\delta},\\\tilde{f}(x)&:=\frac{\max\{f(x)-g(x),\,0\}}{\delta},\\\tilde{g}(x)&:=\frac{\max\{g(x)-f(x),\,0\}}{\delta}.\end{aligned}$$

  If $\delta=1$, take an arbitrary h. If $\delta=0$, take arbitrary $\tilde{f},\tilde{g}$. We can see that $f(x)=(1-\delta)h(x)+\delta\tilde{f}(x)$ and $g(x)=(1-\delta)h(x)+\delta\tilde{g}(x)$ are mixtures of $h,\tilde{f},\tilde{g}$. Define random variables $Z\sim h$, $\tilde{X}\sim\tilde{f}$, $\tilde{Y}\sim\tilde{g}$ all independent, and let γ be the distribution of (X,Y) where

$$(X,Y)=\begin{cases}(Z,Z) & \text{with prob. } 1-\delta,\\(\tilde{X},\tilde{Y}) & \text{with prob. } \delta.\end{cases}$$

  We have $\gamma\in\Gamma(P,Q)$ and $\mathbb{P}_{(X,Y)\sim\gamma}(X\neq Y)\le\delta$.

- (5.7)≤(5.3): Let $E=\{x\in\mathcal{X}:\,f(x)\ge g(x)\}$. We have

$$\begin{aligned}P(E)-Q(E)&=\int_E (f(x)-g(x))\mu(\mathrm{d}x)\\&=\int \max\{f(x)-g(x),\,0\}\mu(\mathrm{d}x).\end{aligned}$$

- (5.3)≤(5.9): For every event E and coupling $\gamma\in\Gamma(P,Q)$,

$$P(E)-Q(E)=\mathbb{P}_{(X,Y)\sim\gamma}(X\in E)-\mathbb{P}_{(X,Y)\sim\gamma}(Y\in E)$$

$$\leq \mathbb{P}_{(X,Y)\sim\gamma}(X \in E \text{ and } Y \notin E)$$
$$\leq \mathbb{P}_{(X,Y)\sim\gamma}(X \neq Y).$$

□

Another property of TV distance is that passing two random variables through the same Markov kernel will not increase their TV distance.

Lemma 35 (Data processing inequality)**.** *For $X \sim P_X$, $Y|X \sim P_{Y|X}$, $\tilde{X} \sim P_{\tilde{X}}$, $\tilde{Y}|\tilde{X} \sim P_{Y|X}$,*

$$\delta_{TV}(P_Y, P_{\tilde{Y}}) \leq \delta_{TV}(P_X, P_{\tilde{X}}).$$

Proof. This is a direct consequence of the data processing inequality for f-divergence (e.g., see (Polyanskiy and Wu, 2024)). We include a proof for the sake of completeness. Consider any measurable set $E \subseteq \mathcal{Y}$. We have

$$\begin{aligned}
&\left|\mathbb{P}(Y \in E) - \mathbb{P}(\tilde{Y} \in E)\right| \\
&= \left|\mathbb{E}\left[\mathbb{P}(Y \in E \,|\, X)\right] - \mathbb{E}\left[\mathbb{P}(\tilde{Y} \in E \,|\, \tilde{X})\right]\right| \\
&= \left|\mathbb{E}_{X\sim P_X}\left[\mathbb{P}(Y \in E \,|\, X)\right] - \mathbb{E}_{X\sim P_{\tilde{X}}}\left[\mathbb{P}(Y \in E \,|\, X)\right]\right| \\
&\leq \delta_{TV}(P_X, P_{\tilde{X}}),
\end{aligned}$$

where the last inequality is by applying (5.5) on the function $x \mapsto \mathbb{P}(Y \in E \,|\, X = x)$. □

A consequence is that for two joint distributions $P_{X,Y}, Q_{X,Y}$ with X-marginals P_X, Q_X, we have $\delta_{TV}(P_X, Q_X) \leq \delta_{TV}(P_{X,Y}, Q_{X,Y})$. This follows from applying Lemma 35 on the Markov kernel that maps (X, Y) to X.

We then prove a useful result for bounding the TV distance. This result can be considered folklore, though we include the proof for the sake of completeness.

Lemma 36 (Chain rule bound for TV distance)**.** *For two distributions P_X, Q_X over the same space $\mathcal{X}$, and two conditional distributions $P_{Y|X}, Q_{Y|X}$ from $\mathcal{X}$ to $\mathcal{Y}$, we have*

$$\begin{aligned}
&\delta_{TV}(P_X P_{Y|X},\, Q_X Q_{Y|X}) \\
&\leq \delta_{TV}(P_X, Q_X) + \mathbb{E}_{X\sim P_X}\left[\delta_{TV}(P_{Y|X}(\cdot|X),\, Q_{Y|X}(\cdot|X))\right].
\end{aligned}$$

Equality holds if $P_X = Q_X$ or $P_{Y|X} = Q_{Y|X}$.

Proof. Consider any measurable set $E \subseteq \mathcal{X} \times \mathcal{Y}$. Write $E_x = \{y : (x, y) \in E\}$ for the section of E. We have

$$\begin{aligned}
&\left|P_X P_{Y|X}(E) - Q_X Q_{Y|X}(E)\right| \\
&= \left|\mathbb{E}_{X\sim P_X}\left[P_{Y|X}(E_X|X)\right] - \mathbb{E}_{X\sim Q_X}\left[Q_{Y|X}(E_X|X)\right]\right| \\
&\le \left|\mathbb{E}_{X\sim P_X}\left[Q_{Y|X}(E_X|X)\right] - \mathbb{E}_{X\sim Q_X}\left[Q_{Y|X}(E_X|X)\right]\right| \\
&\quad + \left|\mathbb{E}_{X\sim P_X}\left[P_{Y|X}(E_X|X)\right] - \mathbb{E}_{X\sim P_X}\left[Q_{Y|X}(E_X|X)\right]\right| \\
&\overset{(a)}{\le} \delta_{TV}(P_X, Q_X) + \mathbb{E}_{X\sim P_X}\left[\left|P_{Y|X}(E_X|X) - Q_{Y|X}(E_X|X)\right|\right] \\
&\le \delta_{TV}(P_X, Q_X) + \mathbb{E}_{X\sim P_X}\left[\delta_{TV}(P_{Y|X}(\cdot|X),\, Q_{Y|X}(\cdot|X))\right],
\end{aligned}$$

where (a) is by applying (5.5) on the function $x \mapsto Q_{Y|X}(E_x|x)$.

For the equality case $P_X = Q_X$, assume $P_X = Q_X$. Let $X \sim P_X$ and $R_{Y|X} := (P_{Y|X} + Q_{Y|X})/2$. By (5.7),

$$\begin{aligned}
&\mathbb{E}\left[\delta_{TV}(P_{Y|X}(\cdot|X),\, Q_{Y|X}(\cdot|X))\right] \\
&= \mathbb{E}\left[\int \max\left\{\frac{\mathrm{d}P_{Y|X}(\cdot|X)}{\mathrm{d}R_{Y|X}(\cdot|X)}(y) - \frac{\mathrm{d}Q_{Y|X}(\cdot|X)}{\mathrm{d}R_{Y|X}(\cdot|X)}(y), 0\right\} R_{Y|X}(\mathrm{d}y|X)\right] \\
&= \mathbb{E}\left[\int \max\left\{\frac{\mathrm{d}P_X P_{Y|X}}{\mathrm{d}P_X R_{Y|X}}(X, y) - \frac{\mathrm{d}P_X Q_{Y|X}}{\mathrm{d}P_X R_{Y|X}}(X, y), 0\right\} R_{Y|X}(\mathrm{d}y|X)\right] \\
&= \delta_{TV}(P_X P_{Y|X},\, Q_X Q_{Y|X}).
\end{aligned}$$

For the other equality case, we assume $P_{Y|X} = Q_{Y|X}$. We have

$$\delta_{TV}(P_X P_{Y|X},\, Q_X P_{Y|X}) \ge \delta_{TV}(P_X, Q_X)$$

by applying Lemma 35 on the Markov kernel that maps (X, Y) to X, and hence

$$\delta_{TV}(P_X P_{Y|X},\, Q_X P_{Y|X}) = \delta_{TV}(P_X, Q_X).$$

□

For a random variable $X \sim P_X$, we often write $\delta_{\mathrm{TV}}(X, Q) = \delta_{\mathrm{TV}}(P_X, Q)$ for brevity. Similar to the conditional expectation notation $\mathbb{E}[X|Z]$, for a conditional distribution $Q_{X|Z}$, we write

$$\delta_{\mathrm{TV}}(X, Q_{X|Z} \,|\, Z) = \delta_{\mathrm{TV}}(P_{X|Z}(\cdot|Z),\, Q_{X|Z}(\cdot|Z)), \tag{5.10}$$

where $P_{X|Z}$ is the conditional distribution of X given Z. Note that $\delta_{\mathrm{TV}}(X, Q_{X|Z}|Z)$ is a random variable and is a function of Z. Similarly, for random variables X, Y, Z, we write

$$\delta_{\mathrm{TV}}(X, Y|Z) = \delta_{\mathrm{TV}}(P_{X|Z}(\cdot|Z),\, P_{Y|Z}(\cdot|Z)).$$

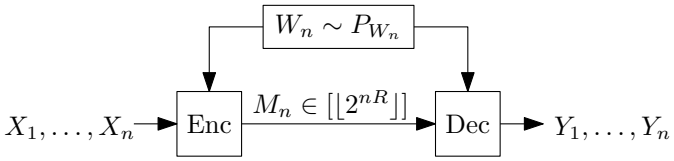


Figure 5.2: Asymptotic approximate fixed-length channel simulation with common randomness.

5.4 Approximate Fixed-Length Channel Simulation

The asymptotic approximate fixed-length channel simulation setting is defined as follows.

Definition 37 (Asymptotic approximate fixed-length channel simulation)**.** Consider a general channel $P_{Y|X}$ and a source distribution P_X. An asymptotic approximate fixed-length channel simulation scheme with rate $R \geq 0$ and common randomness rate $R_0 \in [0, \infty]$ is characterized by a sequence of tuples $(P_{W_n}, P_{M_n|W_n,X^n}, P_{\tilde{Y}^n|W_n,M_n})_{n\in\mathbb{N}^+}$ described below:

- **Common randomness.** There is a common random source $W_n \in \mathcal{W}_n$, $W_n \sim P_{W_n}$ available to the encoder and the decoder. If $R_0 = \infty$ (unlimited common randomness) we can choose an arbitrary distribution P_{W_n} as a part of the coding scheme. If $R_0 \neq \infty$, P_{W_n} is fixed to be $\mathrm{Unif}([\lfloor 2^{nR_0} \rfloor])$. Note that $R_0 = 0$ is the no common randomness case.[5]

- **Encoder.** The encoder observes W_n and a source sequence $X^n \sim P_X^n$ (for the arbitrary source case, we can have any $X^n \in \mathcal{X}^n$), and sends $M_n|(W_n, X^n) \sim P_{M_n|W_n,X^n}$ produced by passing W_n, X^n through an encoding Markov kernel $P_{M_n|W_n,X^n}$ from $\mathcal{W}_n \times \mathcal{X}^n$ to $[\lfloor 2^{nR} \rfloor]$.

- **Decoder.** The decoder then outputs $\tilde{Y}^n|(W_n, M_n) \sim P_{\tilde{Y}^n|W_n,M_n}$ produced by passing W_n, M_n through a decoding Markov kernel $P_{\tilde{Y}^n|W_n,M_n}$ from $\mathcal{W}_n \times [\lfloor 2^{nR} \rfloor]$ to $\mathcal{Y}^n$.

- **Requirement.**

[5] Loosely speaking, one can think of the common randomness as a sequence of $\approx nR_0$ i.i.d. coin flips. If $R_0 = \infty$, this means an infinite sequence of coin flips is available. One can convert the sequence into a uniform real number over $[0, 1]$, which can be used to simulate any distribution over a Polish space (see Itô, 1984, Theorem 2.4.1). Therefore, when $R_0 = \infty$, we can select an arbitrary common randomness distribution.

– For the known source distribution case, we require that $\tilde{Y}^n$ follows the conditional distribution $P_{Y|X}^n$ approximately, in the sense that

$$\delta_{\mathrm{TV}}\big((X^n, \tilde{Y}^n),\, P_X^n P_{Y|X}^n\big) \to 0$$

as $n \to \infty$. Or equivalently, $\delta_{\mathrm{TV}}(\tilde{Y}^n,\, P_{Y|X}^n \,|\, X^n) \to 0$ in probability as $n \to \infty$ by Lemma 36.

– For the arbitrary source case, we require

$$\sup_{x^n \in \mathcal{X}^n} \delta_{\mathrm{TV}}(\tilde{Y}^n,\, P_{Y|X}^n \,|\, X^n = x^n) \to 0$$

as $n \to \infty$.

- **Performance metrics.** We say that the rate pair (R, R_0) is *achievable* if there exists a scheme with rate R and common randomness rate R_0 satisfying the above requirement. The *optimal rate region* is the closure of the set of achievable pairs. For $R_0 \in [0,\, \infty]$, write $R^*(R_0)$ for the infimum of R such that (R, R_0) is achievable.

Although Definition 37 covers the no common randomness, limited common randomness and unlimited common randomness cases, this section focuses on the unlimited common randomness case. Refer to Section 6 for the no common randomness case, and to Section 7 for the limited common randomness case.

Unlike Definitions 2 and 25 where the output is written as Y, here we write $\tilde{Y}^n$ to emphasize that $\tilde{Y}^n$ does not necessarily follow the conditional distribution $P_{Y|X}^n$ given X^n. The output $\tilde{Y}^n$ is only an "approximate" version of the ideal output Y^n.

The following result was proved in (Bennett *et al.*, 2002) and (Winter, 2002), showing that the "approximate fixed-length" setting and the "exact variable-length" setting in Theorem 32 share the same optimal rate when unlimited common randomness is available.

Theorem 38 (D/∞/A/FL/KAS/UCR (Bennett *et al.*, 2002; Winter, 2002))**.** *Assume X, Y are discrete and finite. For the asymptotic approximate fixed-length channel simulation setting (Definition 37) with unlimited common randomness ($R_0 = \infty$):*

- *For known source distribution, the optimal rate is $R^*(\infty) = I(X;Y)$.*
- *For arbitrary source, the optimal rate is the channel capacity $R^*(\infty) = C = \max_{P_X} I(X;Y)$.*

We will present two proofs of Theorem 38: as a corollary of Theorem 4 given below; and via the soft covering lemma and the likelihood encoder in Section 5.6.

Proof. Consider the known source distribution case. Write $I = I(X;Y)$. We first show that there exists a fixed-length scheme with rate at most $\tilde{R} := I + \log_2(I+2) + 4$. By Theorem 4, we have a variable-length scheme with expected description length $\mathbb{E}[|M|] \le \tilde{R} - 1 < \tilde{R}$. We now construct a fixed-length scheme. For a blocklength n, we apply the above variable-length scheme n times, and concatenate these n descriptions. Let $\bar{M}_n \in \{0,1\}^*$ be this concatenation, and Y^n be the output of the variable-length scheme (which satisfies $Y^n|X^n \sim P_{Y|X}^n$ exactly). By law of large numbers,

$$\mathbb{P}\left(|\bar{M}_n| > \lfloor n\tilde{R} \rfloor\right) \to 0$$

as $n \to \infty$. We produce the description $M_n \in \{0,1\}^{\lfloor n\tilde{R} \rfloor}$ by taking the first $\lfloor n\tilde{R} \rfloor$ bits of $\bar{M}_n$, and padding with zeros if necessary. The decoder is the same as the variable-length scheme. Let its output be $\tilde{Y}^n$. The decoder uses a prefix-free code so padding with zeros does not affect the output. The only situation where $\tilde{Y}^n \ne Y^n$ is when $\bar{M}_n$ is truncated, i.e., $|\bar{M}_n| > \lfloor n\tilde{R} \rfloor$, which happens with vanishing probability. Therefore, by the coupling lemma (Proposition (34)),

$$\begin{aligned}
\delta_{\mathrm{TV}}\big((X^n, \tilde{Y}^n),\, P_X^n P_{Y|X}^n\big) &\le \mathbb{P}\big((X^n, \tilde{Y}^n) \ne (X^n, Y^n)\big) \\
&= \mathbb{P}(\tilde{Y}^n \ne Y^n) \\
&\le \mathbb{P}(|\bar{M}_n| > \lfloor n\tilde{R} \rfloor) \\
&\to 0
\end{aligned}$$

as $n \to \infty$.

We then show that there exists a fixed-length scheme with rate R whenever $R > I$. Apply the above result on X^k (i.i.d. sequence with length k) and $Y^k|X^k \sim P_{Y|X}^k$ for some $k \in \mathbb{N}^+$, we have a fixed-length scheme for simulating Y^k with rate at most

$$\begin{aligned}
&I(X^k;Y^k) + \log_2(I(X^k;Y^k)+2) + 4 \\
&= kI + \log_2(kI+2) + 4
\end{aligned}$$

bits per Y^k simulated. Since each Y^k consists of k symbols, the above scheme can be regarded as a scheme for simulating Y with rate at most[6]

$$\frac{1}{k}\left(kI + \log_2(kI+2) + 5\right) = I + \frac{\log_2(kI+2) + 5}{k}.$$

Therefore, taking k large enough, the rate is upper-bounded by R.

[6] To convert a scheme for simulating Y^k with rate $\tilde{R}$ to a scheme for simulating Y with rate $(\tilde{R}+1)/k$, for a blocklength n, we apply the scheme $\lceil n/k \rceil$ times, which requires $\lceil \lceil n/k \rceil \tilde{R} \rceil \le n(\tilde{R}+1)/k$ bits for large enough n.

Consider the arbitrary source case. We first show that there exists a fixed-length scheme with rate at most $\tilde{R} := C + \log_2(C+2) + 5$. By Theorem 4, we have a variable-length scheme with

$$\max_{x \in \mathcal{X}} \mathbb{E}\big[|\tilde{M}(x)|\big] < \tilde{R} - 1,$$

where $\tilde{M}(x) \in \{0,1\}^*$ denotes the description given by the scheme when the input is x. Fix $\epsilon > 0$. Applying the law of large numbers on each of the finitely many $x \in \mathcal{X}$, we have $\mathbb{P}(|\tilde{M}^\ell(x)| > \ell(\tilde{R}-1)) \to 0$ as $\ell \to 0$ for every $x \in \mathcal{X}$, where $\tilde{M}^\ell(x) \in \{0,1\}^*$ denotes the concatenation of ℓ i.i.d. copies of $\tilde{M}(x)$. Hence, there exists $L_\epsilon \in \mathbb{N}^+$ such that

$$\mathbb{P}(|\tilde{M}^\ell(x)| > \ell(\tilde{R}-1)) \le \epsilon$$

for every $x \in \mathcal{X}$ as long as $\ell \ge L_\epsilon$. For a blocklength n, we apply the above variable-length scheme n times, and concatenate these n descriptions. Let $\bar{M}(x^n) \in \{0,1\}^*$ be this concatenation when the input is $x^n = (x_1, \dots, x_n)$. Let $n_x := |\{i : x_i = x\}|$. We have, for n large enough such that $\lfloor n\tilde{R} \rfloor / (n + |\mathcal{X}|L_\epsilon) \ge \tilde{R} - 1$,

$$\begin{aligned}
\mathbb{P}(|\bar{M}(x^n)| > \lfloor n\tilde{R} \rfloor) &= \mathbb{P}\Big(\sum_{x \in \mathcal{X}} |\tilde{M}^{n_x}(x)| > \lfloor n\tilde{R} \rfloor \Big) \\
&\le \mathbb{P}\Big(\sum_{x \in \mathcal{X}} |\tilde{M}^{n_x + L_\epsilon}(x)| > \lfloor n\tilde{R} \rfloor \Big) \\
&\le \sum_{x \in \mathcal{X}} \mathbb{P}\Big(|\tilde{M}^{n_x + L_\epsilon}(x)| > \frac{(n_x + L_\epsilon)\lfloor n\tilde{R} \rfloor}{n + |\mathcal{X}|L_\epsilon} \Big) \\
&\le \sum_{x \in \mathcal{X}} \mathbb{P}\Big(|\tilde{M}^{n_x + L_\epsilon}(x)| > (n_x + L_\epsilon)(\tilde{R} - 1) \Big) \\
&\le |\mathcal{X}|\epsilon.
\end{aligned}$$

Taking $\epsilon \to 0$, we have $\max_{x^n \in \mathcal{X}^n} \mathbb{P}(|\bar{M}(x^n)| > \lfloor n\tilde{R} \rfloor) \to 0$ as $n \to \infty$. We produce the description $M_n \in \{0,1\}^{\lfloor n\tilde{R} \rfloor}$ by taking the first $\lfloor n\tilde{R} \rfloor$ bits of $\bar{M}_n$, and padding with zeros if necessary. Let its output of the decoder be $\tilde{Y}^n$. By the same arguments as the known source distribution case,

$$\begin{aligned}
\max_{x^n \in \mathcal{X}^n} \delta_{\mathrm{TV}}(\tilde{Y}^n,\, P_{Y|X}^n \,|\, X^n = x^n) &\le \max_{x^n \in \mathcal{X}^n} \mathbb{P}(|\bar{M}(x^n)| > \lfloor n\tilde{R} \rfloor) \\
&\to 0
\end{aligned}$$

as $n \to \infty$. The rest of the proof for the arbitrary source case is the same as the proof for the known source distribution case.

Refer to the proof of Theorem 45 for the converse $R^*(\infty) \ge I(X;Y)$ for the known source distribution case. For the converse $R^*(\infty) \ge \max_{P_X} I(X;Y)$ for the arbitrary source case, it follows from the converse for the known source distribution case by considering the P_X that maximizes $I(X;Y)$. □

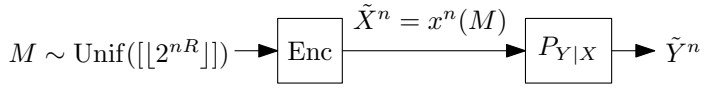


Figure 5.3: Asymptotic soft covering lemma or channel resolvability.

5.5 Soft Covering Lemma

We discuss an important tool for deriving asymptotic channel simulation results, called the *soft covering lemma* (Wyner, 1975a; Cuff, 2013) or *channel resolvability* (Han and Verdú, 1993).[7] If we are allowed unlimited description length and common randomness, then we can perform channel simulation perfectly. However, if we limit the description length and common randomness, we may distort the distribution of the output. Intuitively, the purpose of the soft covering lemma is to characterize how many different values of the description and common randomness we need in order to "cover" the distribution, so limiting to those values will not distort the distribution too much.

Consider the channel coding setting with a memoryless channel $P_{Y|X}$ and an input distribution P_X. If an encoder generates an input sequence $\tilde{X}^n$ following the i.i.d. distribution P_X^n (the distribution of an i.i.d. sequence of length n following P_X), then the output $\tilde{Y}^n$ will also follow an i.i.d. distribution P_Y^n, where P_Y is the Y-marginal of the joint distribution $P_X P_{Y|X}$ induced by the input distribution P_X and the channel $P_{Y|X}$.

The question is: how much randomness does the encoder need in order to ensure that $\tilde{Y}^n$ (approximately) follows P_Y^n? We assume the encoder encodes a uniform random message $M \sim \text{Unif}([\lfloor 2^{nR} \rfloor])$ into $\tilde{X}^n = x^n(M)$ using a codebook $\mathfrak{X} := (x^n(1), \ldots, x^n(\lfloor 2^{nR} \rfloor)) \in (\mathcal{X}^n)^{\lfloor 2^{nR} \rfloor}$, and passes it through the memoryless channel $P_{Y|X}$. How large does the rate R need to be so that $\tilde{Y}^n$ approximately follows P_Y^n? Would an i.i.d. random codebook construction, where $x_i(m) \sim P_X$ i.i.d. for $i = 1, \ldots, n$, $m = 1, \ldots, \lfloor 2^{nR} \rfloor$, suffice?[8]

For the special case where the channel is completely noisy, with an output that is independent of the input, no randomness is needed at the encoder since we always have $\tilde{Y}^n \sim P_Y^n$ regardless of the input, and the smallest R is 0. For the other extreme where the channel is completely noiseless and $Y = X$, we will require $R \geq H(X)$ since the only randomness in $X = Y$ comes from M. We can see that the minimum R is larger when Y is more dependent on X.

The soft covering lemma (Wyner, 1975a; Han and Verdú, 1993) states that $\tilde{Y}^n$ approximately follows P_Y^n with a vanishing total variation distance as long as $R > I(X;Y)$.

[7] The soft covering lemma is also useful for proving results in information-theoretic secrecy (Wyner, 1975b), though this is out of the scope of this monograph.

[8] The channel resolvability setting (Han and Verdú, 1993) allows the encoder to choose any codebook, though we do not consider it here for the sake of simplicity.

Lemma 39 (Soft covering lemma (Wyner, 1975a; Han and Verdú, 1993))**.** *Consider a finite discrete memoryless channel $P_{Y|X}$ and a finite discrete input distribution P_X. Fix $R > I(X;Y)$. Let $\mathfrak{X} = (x^n(m))_{m\in[\lfloor 2^{nR}\rfloor]}$ be a random codebook with i.i.d. entries following P_X, $M \sim \mathrm{Unif}([\lfloor 2^{nR}\rfloor])$ independent of $\mathfrak{X}$, $\tilde{X}^n := x^n(M)$, and $\tilde{Y}^n|\tilde{X}^n \sim P_{Y|X}^n$. Then*

$$\delta_{\mathrm{TV}}(\tilde{Y}^n, P_Y^n \,|\, \mathfrak{X}) \to 0$$

in probability (and in expectation) as $n \to \infty$.

The proof of the soft covering lemma is deferred to Section 8.1. Recall that $\delta_{\mathrm{TV}}(\tilde{Y}^n, P_Y^n \,|\, \mathfrak{X}) = \delta_{\mathrm{TV}}(P_{\tilde{Y}^n|\mathfrak{X}}(\cdot|\mathfrak{X}), P_Y^n)$ is a random variable as defined in (5.10). We now explain the (succinct yet slightly confusing) statement of Lemma 39. There are three sources of randomness in this setting: the random codebook $\mathfrak{X}$, the random message M, and the noisy channel. The only sources of randomness that contribute to the goal (making $\tilde{Y}^n$ approximately follows P_Y^n) are M and the channel. The randomness in $\mathfrak{X}$ is just a fictitious construction in our scheme. In reality, we always have to fix a codebook, just like the standard random coding arguments for source and channel coding. It would be cheating if we include the randomness of $\mathfrak{X}$ and only argue that the unconditional distribution $P_{\tilde{Y}^n}$ is close to P_Y^n (which is trivially true for any R). Since $\mathfrak{X}$ will eventually be fixed, for a fixed codebook $\mathfrak{X} = \mathfrak{x}$ to be considered satisfactory, we require the conditional distribution $P_{\tilde{Y}^n|\mathfrak{X}}(\cdot|\mathfrak{x})$ to be close to P_Y^n. A fixed codebook $\mathfrak{x}$ is ϵ-good ($\epsilon > 0$) if $P_{\tilde{Y}^n|\mathfrak{X}}(\cdot|\mathfrak{x})$ is ϵ-close to P_Y^n in total variation distance, i.e., $\delta_{\mathrm{TV}}(P_{\tilde{Y}^n|\mathfrak{X}}(\cdot|\mathfrak{x}), P_Y^n) \le \epsilon$. The soft covering lemma says that as long as $R > I(X;Y)$, for every $\epsilon > 0$, if we generate the codebook at random, then the probability of getting an ϵ-good codebook approaches 1 as $n \to \infty$. Refer to Figure 5.4 for an illustration.

5.6 Likelihood Encoder for Asymptotic Coding

In Section 3.4, we had briefly discussed the likelihood encoder (Cuff, 2013; Watanabe *et al.*, 2015; Song *et al.*, 2016) and minimal random coding (Havasi *et al.*, 2019; Flamich *et al.*, 2020). In this section, we will describe the likelihood encoder applied to the asymptotic fixed-length channel simulation setting with unlimited common randomness.

Take the common randomness to be $W = (\bar{Y}^n(i))_{i\in\lfloor 2^{nR}\rfloor}$, where $\bar{Y}^n(i) \overset{iid}{\sim} P_Y^n$ for $i \in [\lfloor 2^{nR}\rfloor]$. Here P_Y is the Y- marginal of the joint distribution $P_X P_{Y|X}$. The encoder observes X^n, computes the likelihood

$$\alpha_i(X^n) := \frac{\mathrm{d}P_{Y|X}^n(\cdot|X^n)}{\mathrm{d}P_Y^n}(\bar{Y}^n(i)),$$

and generates the description $M \in [\lfloor 2^{nR}\rfloor]$ with

$$\mathbb{P}(M = m \,|\, W, X^n) = \frac{\alpha_m(X^n)}{\sum_{i=1}^{\lfloor 2^{nR}\rfloor} \alpha_i(X^n)}.$$

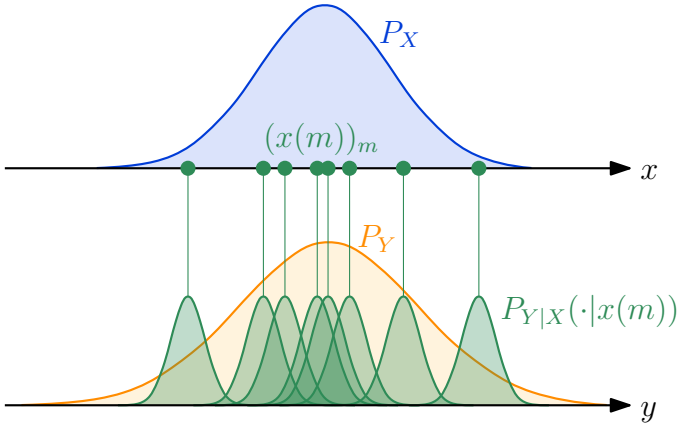


Figure 5.4: An illustration of the soft covering lemma, where we take $n = 1$ for simplicity. We draw an i.i.d. codebook $(x(m))_{m\in[\mathsf{N}]}$ following P_X, and note the mixture distribution $\mathsf{N}^{-1}\sum_m P_{Y|X}(\cdot|x(m))$, which is the distribution of $\tilde{Y}$ if $M \sim \mathrm{Unif}([\mathsf{N}])$, $\tilde{X} = x(M)$, and $\tilde{Y}|\tilde{X} \sim P_{Y|X}$. If the number of samples N is large enough, then the mixture distribution should be close to P_Y (the Y-marginal of $P_X P_{Y|X}$) with high probability.

The decoder simply outputs $\tilde{Y}^n = \bar{Y}^n(M)$.

We now prove the achievability part of the known source distribution case of Theorem 38 using the soft covering lemma and the techniques in (Cuff, 2013).

Proof of Theorem 38 using likelihood encoder. Assume X, Y are finite and discrete. Fix $R > I(X;Y)$. Consider the "reverse channel" $P_{X|Y}$ computed using the joint distribution $P_X P_{Y|X}$. Let $\hat{M} \sim \mathrm{Unif}([\lfloor 2^{nR} \rfloor])$, $\hat{Y}^n = \bar{Y}^n(\hat{M})$, and $\hat{X}^n|\hat{Y}^n \sim P_{X|Y}^n$. Note that $(\hat{X}^n, \hat{Y}^n) \sim P_X^n P_{Y|X}^n$, and the conditional distribution of $\hat{M}$ given $W, \hat{X}^n$ is

$$\begin{aligned} P_{\hat{M}|W,\hat{X}^n}(m \,|\, (\bar{Y}^n(i))_i, \hat{X}^n) &= \frac{P_{X|Y}^n(\hat{X}^n|\bar{Y}^n(m))}{\sum_{i=1}^{\lfloor 2^{nR} \rfloor} P_{X|Y}^n(\hat{X}^n|\bar{Y}^n(i))} \\ &= \frac{\alpha_m(\hat{X}^n)}{\sum_{i=1}^{\lfloor 2^{nR} \rfloor} \alpha_i(\hat{X}^n)}, \end{aligned}$$

which is the same as the likelihood encoder. Therefore, the proof would be completed if we can simply swap $\hat{X}^n$ with X^n. Nevertheless, the obstacle is that $\hat{X}^n$ is not independent of W, but X^n (the input given to the encoder) must be independent of W (the pre-shared common randomness generated before X^n is observed).

This is where we utilize the soft covering lemma. Since $R > I(X;Y)$, applying the soft covering lemma on $P_{X|Y}$, we know that

$$\delta_{\mathrm{TV}}(\hat{X}^n,\, P_X^n \,|\, W) \to 0 \tag{5.11}$$

in probability as $n \to \infty$. Therefore, although $P_{\hat{X}^n|W}$ is not P_X^n, they are quite close.

We now swap $\hat{X}^n$ with X^n and redefine the other random variables accordingly. Let $M|(W, X^n) \sim P_{\hat{M}|W,\hat{X}^n}(\cdot|W, X^n)$ be the description given by the likelihood encoder when the input is X^n instead of $\hat{X}^n$, and $\tilde{Y}^n = \bar{Y}^n(M)$. Since $P_{\tilde{Y}^n|W,X^n} = P_{\hat{Y}^n|W,\hat{X}^n}$ by construction, applying Lemma 36, we have

$$\begin{aligned}
&\delta_{\mathrm{TV}}\big((X^n, \tilde{Y}^n),\, P_X^n P_{Y|X}^n\big) \\
&= \delta_{\mathrm{TV}}\big((X^n, \tilde{Y}^n),\, (\hat{X}^n, \hat{Y}^n)\big) \\
&\overset{(a)}{\le} \delta_{\mathrm{TV}}\big((W, X^n),\, (W, \hat{X}^n)\big) \\
&\overset{(b)}{=} \mathbb{E}\left[\delta_{\mathrm{TV}}\big(X^n,\, \hat{X}^n \,|\, W\big)\right] \\
&\overset{(c)}{\to} 0
\end{aligned} \tag{5.12}$$

as $n \to \infty$, where (a) is due to Lemma 36 since $P_{\tilde{Y}^n|W,X^n} = P_{\hat{Y}^n|W,\hat{X}^n}$, (b) is also due to Lemma 36, and (c) is due to (5.11). □

The analysis of the likelihood encoder for the one-shot setting is deferred to Section 8.2.

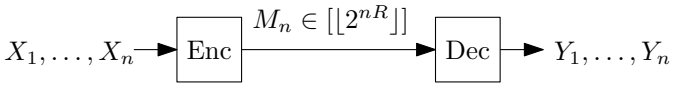


Figure 6.1: Asymptotic channel simulation without common randomness.

6 Asymptotic Channel Simulation without Common Randomness

In this section, we study the asymptotic channel simulation setting without common randomness, which is the asymptotic version of the one-shot setting in Section 4. We will study two cases: the approximate fixed-length case, and the exact fixed-length/variable-length case.

6.1 Approximate Case—Wyner's Common Information

In this section, we study the asymptotic fixed-length channel simulation setting without common randomness in Figure 6.1, i.e., Definition 37 with $R_0 = 0$. We present a result due to (Cuff, 2008; Cuff, 2013), showing that the optimal rate is given by the following single-letter expression called Wyner's common information (Wyner, 1975a).

Theorem 40 (D/∞/A/FL/KS/NCR (Wyner, 1975a; Cuff, 2008; Cuff, 2013))**.** *For the asymptotic approximate fixed-length channel simulation setting (Definition 37) with no common randomness, known source distribution, and finite discrete X, Y, the optimal rate $R^*(0)$ is given by* Wyner's common information

$$J(X;Y) := \min_{P_{U|X,Y}:\, X \leftrightarrow U \leftrightarrow Y} I(X,Y;U).$$

Note that $I(X;Y) \le J(X;Y) \le G(X;Y)$ (where $G(X;Y) = \min_{X \leftrightarrow U \leftrightarrow Y} H(U)$ is the common entropy (4.1)) since $I(X;Y) \le I(X,Y;U) \le H(U)$ whenever $X \leftrightarrow U \leftrightarrow Y$ forms a Markov chain.

Proof. We adopt the proof strategy in (Cuff, 2008; Cuff, 2013). First prove the achievability part. We omit the subscript in M_n and simply write M. Fix $P_{U|X,Y}$ satisfying $X \leftrightarrow U \leftrightarrow Y$. Fix $R > I(X,Y;U)$. Generate a random codebook $\mathfrak{U} := (u^n(m))_{m \in [\lfloor 2^{nR} \rfloor]}$, where $u_i(m) \sim P_U$ i.i.d. across $i \in [n]$, $m \in [\lfloor 2^{nR} \rfloor]$. Let $\bar{M} \sim \mathrm{Unif}([\lfloor 2^{nR} \rfloor])$ and $\bar{U}^n := u^n(\bar{M})$. Define random variables $(\bar{X}^n, \bar{Y}^n)$ with conditional distribution $(\bar{X}^n, \bar{Y}^n)|(\bar{U}^n, \bar{M}, \mathfrak{U}) \sim P^n_{X,Y|U}$. Since $X \leftrightarrow U \leftrightarrow Y$ forms a Markov chain, $P_{X,Y|U} = P_{X|U} P_{Y|U}$, and hence $\bar{X}^n \leftrightarrow \bar{U}^n \leftrightarrow \bar{Y}^n$ forms a Markov chain. Since $\bar{U}^n = u^n(\bar{M})$ is a function of $(\bar{M}, \mathfrak{U})$, $\bar{X}^n \leftrightarrow \bar{M} \leftrightarrow \bar{Y}^n$ forms a Markov chain conditional on $\mathfrak{U}$.

We take the encoding Markov kernel to be $P_{\bar{M}|\bar{X}^n,\mathfrak{U}}(\cdot|\cdot,\mathfrak{U})$, and the decoding Markov kernel to be $P_{\bar{Y}^n|\bar{M},\mathfrak{U}}(\cdot|\cdot,\mathfrak{U})$ (these kernels are "random" and depend on $\mathfrak{U}$, though we will later argue that there exists a fixed value of $\mathfrak{U}$ that gives good kernels). This ensures that, if the input to the encoder is $\bar{X}^n$, then the joint distribution of $\bar{X}^n$ and the decoder's output will be the same as the joint distribution of $(\bar{X}^n, \bar{Y}^n)$. Applying the soft covering lemma on $P_{X,Y|U}$, we know that

$$\delta_{\mathrm{TV}}\big((\bar{X}^n, \bar{Y}^n),\, P_{X,Y}^n \,\big|\, \mathfrak{U}\big) \to 0 \tag{6.1}$$

in probability as $n \to \infty$.

This means the joint distribution of the input and output will be close to $P_{X,Y}^n$, which would be the desired result, except that the input is actually not $\bar{X}^n$, but $X^n \sim P_X^n$. We have to swap $\bar{X}^n$ with X^n and redefine the other random variables accordingly. Let $M|(X^n, \mathfrak{U}) \sim P_{\bar{M}|\bar{X}^n,\mathfrak{U}}(\cdot|X^n, \mathfrak{U})$ be the description given by the encoding Markov kernel when the input is X^n instead of $\bar{X}^n$, and $\tilde{Y}^n|(M, X^n, \mathfrak{U}) \sim P_{\bar{Y}^n|\bar{M},\mathfrak{U}}(\cdot|M, \mathfrak{U})$ be the output given by the decoding Markov kernel when the description is M. Since $P_{\tilde{Y}^n|X^n,\mathfrak{U}} = P_{\bar{Y}^n|\bar{X}^n,\mathfrak{U}}$ by construction, applying Lemma 36, we have

$$\begin{aligned}
&\delta_{\mathrm{TV}}\big((X^n, \tilde{Y}^n),\, (\bar{X}^n, \bar{Y}^n) \,\big|\, \mathfrak{U}\big) \\
&= \delta_{\mathrm{TV}}\big(X^n,\, \bar{X}^n \,\big|\, \mathfrak{U}\big) \\
&\to 0
\end{aligned} \tag{6.2}$$

in probability as $n \to \infty$ due to (6.1). Combining (6.1) and (6.2) using the triangle inequality,

$$\delta_{\mathrm{TV}}\big((X^n, \tilde{Y}^n),\, P_{X,Y}^n \,\big|\, \mathfrak{U}\big) \to 0$$

in probability as $n \to \infty$. This means there is a fixed choice $\mathfrak{u}$ of $\mathfrak{U}$ (which depends on n) that gives $\delta_{\mathrm{TV}}((X^n, \tilde{Y}^n),\, P_{X,Y}^n \,|\, \mathfrak{U} = \mathfrak{u}) \to 0$, which is the desired result. Refer to the proof of Theorem 45 for the converse. □

Wyner's common information was originally studied in the asymptotic distributed source simulation problem (Wyner, 1975a) where two terminals want to simulate a pair of correlated random sequences $\tilde{X}^n, \tilde{Y}^n$ approximately following the i.i.d. distribution $P_{X,Y}^n$, using the smallest amount of common randomness. This setting will be discussed in Section 9.2.

For the case with arbitrary source, similar to how the optimal rate $\max_{P_X} I(X;Y)$ for arbitrary source and unlimited common randomness can be obtained by considering the worst-case source distribution P_X in the optimal rate $I(X;Y)$ for known source distribution and unlimited common randomness, the optimal rate for arbitrary source and no common randomness can also be obtained by considering the worst-case source distribution of the corresponding known source distribution setting. The following result is proved in (Bennett *et al.*, 2014) using the method of types. Refer to (Bennett *et al.*, 2014) for the proof.

Theorem 41 (D/∞/A/FL/AS/NCR (Bennett *et al.*, 2014))**.** *For the asymptotic approximate fixed-length channel simulation setting (Definition 37) with no common randomness, arbitrary source, and finite discrete X, Y, the optimal rate of the description is given by*

$$\max_{P_X} J(X;Y).$$

6.2 Exact Case—Exact Common Information Rate

In the previous section, we have studied channel simulation with vanishing TV distance. In this section, we strengthen the constraint and require the output $(X^n, \tilde{Y}^n)$ to follow $(P_X P_{Y|X})^n$ *exactly*. One might expect that, since $(X^n, \tilde{Y}^n)$ "approaches" $(P_X P_{Y|X})^n$ in TV distance as $n \to \infty$, by continuity, $(X^n, \tilde{Y}^n)$ should exactly follow $(P_X P_{Y|X})^n$ at the "limit". This argument does not work since $(P_X P_{Y|X})^n$ changes as n increases, so there is never a fixed target that $(X^n, \tilde{Y}^n)$ can approach. It turns out that the exact case is more complicated than the approximate case, and a simple single-letter expression for the solution (as in Theorem 40) has not been found.

For the sake of clarity, we give the definition of the setting below (which is basically an asymptotic version of Definition 25). As in Section 4, there are two ways to impose the communication constraint. First, we can restrict $M \in [\lfloor 2^{nR} \rfloor]$ to be a fixed-length description with $\approx nR$ bits. Second, we can allow $M \in \{0,1\}^*$ to be a variable-length codeword, and restrict the average length $\mathbb{E}[|M|]$ instead.

Definition 42 (Asymptotic exact channel simulation without common randomness)**.** Consider a general channel $P_{Y|X}$ from $\mathcal{X}$ to $\mathcal{Y}$, and a general source distribution P_X. An asymptotic variable-length channel simulation scheme without common randomness is characterized by a tuple $(\mathcal{C}_n, P_{M_n|X^n}, P_{Y^n|M_n})_{n \in \mathbb{N}^+}$ described below:

- **Codebook.**
  - For the variable-length setting, the set of possible descriptions $\mathcal{C}_n \subseteq \{0,1\}^*$ is a prefix-free codebook, which we can design as a part of the coding scheme.
  - For the fixed-length setting, the set of possible descriptions must be $\mathcal{C}_n = [\lfloor 2^{nR} \rfloor]$, where $R \geq 0$ is the description rate.

- **Encoder.** The encoder observes an i.i.d. source sequence $X^n \sim P_X^n$, and sends a description $M_n \in \mathcal{C}_n$, $M_n | X^n \sim P_{M_n|X^n}$.

- **Decoder.** The decoder then outputs $Y^n|M_n \sim P_{Y^n|M_n}$.

- **Requirement.** We require $Y^n|X^n \sim P_{Y|X}^n$ exactly.

- **Performance metric.**
  - For the variable-length setting, we are interested in the smallest rate of increase of the expected length $\mathbb{E}[|M_n|]$. Let
  $$R^* := \inf \limsup_{n\to\infty} \frac{1}{n}\mathbb{E}[|M_n|]$$
  be the *optimal rate*, where the infimum is over schemes $(\mathcal{C}_n, P_{M_n|X^n}, P_{Y^n|M_n})_{n\in\mathbb{N}^+}$ satisfying the requirement.
  - For the fixed-length setting, we are interested in the smallest description rate R. Let R^* be the infimum of R over schemes $(\mathcal{C}_n, P_{M_n|X^n}, P_{Y^n|M_n})_{n\in\mathbb{N}^+}$ satisfying the requirement.

From the operational meaning of Wyner's common information in Theorem 40, we can see that Wyner's common information $J(X^2;Y^2)$ between $X^2 = (X_1, X_2)$ and $Y^2 = (Y_1, Y_2)$, where $(X_1, Y_1), (X_2, Y_2) \overset{iid}{\sim} P_{X,Y}$, is simply given by $2J(X_1;Y_1)$, since we can use a channel simulation scheme for the channel $(X_1, X_2) \to (Y_1, Y_2)$ to simulate the channel $X_1 \to Y_1$ at twice the rate. More generally, as long as (X_1, Y_1) is independent of (X_2, Y_2), we have

$$J(X^2;Y^2) = J(X_1;Y_1) + J(X_2;Y_2). \tag{6.3}$$

This is known as the *tensorization property* of Wyner's common information. Note that the mutual information also satisfies the tensorization property.

This argument fails for one-shot channel simulation in Propositions 26 and 27, since there is no such thing as "simulate at twice the rate" in a one-shot setting. While we generally have

$$\log_2 \operatorname{rank}_+(\mathbf{P}_{Y^2|X^2}) \le \log_2 \operatorname{rank}_+(\mathbf{P}_{Y_1|X_1}) + \log_2 \operatorname{rank}_+(\mathbf{P}_{Y_2|X_2}),$$

and

$$G(X^2;Y^2) \le G(X_1;Y_1) + G(X_2;Y_2), \tag{6.4}$$

as long as (X_1, Y_1) is independent of (X_2, Y_2),[1] the other direction does not hold in general. Refer to (Vandaele *et al.*, 2016) for an example of matrix $\mathbf{A}$ satisfying that $\log_2 \operatorname{rank}_+(\mathbf{A} \otimes$

[1] If $X_1 \leftrightarrow U_1 \leftrightarrow Y_1$ and $X_2 \leftrightarrow U_2 \leftrightarrow Y_2$, then $(X_1, X_2) \leftrightarrow (U_1, U_2) \leftrightarrow (Y_1, Y_2)$ and $H(U_1, U_2) \le H(U_1) + H(U_2)$.

$\mathbf{A}) < 2\log_2 \operatorname{rank}_+(\mathbf{A})$ (where $\mathbf{A} \otimes \mathbf{A}$ denotes the Kronecker product), and (Kumar *et al.*, 2014) for an example of $P_{X,Y}$ where $G(X^2;Y^2) < 2G(X_1;Y_1)$.

This creates an obstacle for the characterization of the optimal rate for the exact setting. Currently, there is no known single-letter expression for the asymptotic exact fixed-length or variable-length channel simulation settings with no common randomness. The result for the fixed-length setting can merely be stated as a limit of the nonnegative rank, as given in (Yu and Tan, 2020; Yu and Tan, 2022).

Proposition 43 (D/∞/E/FL/KAS/NCR (Yu and Tan, 2020; Yu and Tan, 2022))**.** *For the asymptotic exact fixed-length channel simulation setting with no common randomness, known*[2] *or arbitrary source distribution, and finite discrete X, Y, the optimal rate is given by*

$$R^* = \lim_{n\to\infty} \frac{1}{n} \log_2 \operatorname{rank}_+(\mathbf{P}_{Y|X}^{\otimes n}),$$

where $\mathbf{P}_{Y|X}^{\otimes n}$ is the n-fold Kronecker power of $\mathbf{P}_{Y|X}$, which is the same as the conditional probability matrix $\mathbf{P}_{Y^n|X^n}$.

Readers are also referred to (Braun *et al.*, 2017) for the relation between the nonnegative rank and Wyner's common information.

For the variable-length setting, the limit is called the exact common information rate (Kumar *et al.*, 2014), as given below.

Proposition 44 (D/∞/E/VL/KS/NCR (Kumar *et al.*, 2014))**.** *For the asymptotic exact variable-length channel simulation setting with no common randomness, known source distribution, and finite discrete X, Y, the optimal rate is given by the* exact common information rate

$$R^* = \overline{G}(X;Y) = \lim_{n\to\infty} \frac{1}{n} G(X^n;Y^n),$$

where $(X_i, Y_i) \overset{iid}{\sim} P_X P_{Y|X}$, and $G(A;B) := \min_{P_{U|A,B}:\, A\leftrightarrow U\leftrightarrow B} H(U)$ is the common entropy (4.1).

We have the following relations:

$$I(X;Y) \le J(X;Y) \le \overline{G}(X;Y) \le G(X;Y) \le \log_2 \operatorname{rank}_+(\mathbf{P}_{Y|X}).$$

For each of the above inequalities, there are examples where the inequality is strict. For example, consider $X \sim \text{Bern}(1/2)$ and $P_{Y|X}$ is a binary erasure channel with erasure

[2] For known source distribution, we assume $P_X(x) > 0$ for all $x \in \mathcal{X}$.

probability $0 < p < 1$, i.e., $Y = \{0, 1, e\}$, $P_{Y|X}(0|0) = P_{Y|X}(1|1) = 1 - p$, $P_{Y|X}(e|0) = P_{Y|X}(e|1) = p$. We have

$$\begin{aligned} I(X;Y) &= 1 - p, \\ J(X;Y) &= H_b(\max\{p, 1/2\}), \\ \overline{G}(X;Y) &= H_b(\max\{p, 1/2\}), \\ G(X;Y) &= \min\{H_b(p) + 1 - p,\, 1\}, \\ \log_2 \operatorname{rank}_+(\mathbf{P}_{Y|X}) &= 1, \end{aligned} \tag{6.5}$$

where $H_b(a)$ is the binary entropy function (the entropy of $\mathrm{Bern}(a)$). The formula for $J(X;Y)$ was proved in (Cuff, 2013),[3] whereas the formulae for $\overline{G}(X;Y)$ and $G(X;Y)$ were proved in (Kumar *et al.*, 2014).[4]

The problem whether there exists an example where $J(X;Y) < \overline{G}(X;Y)$, which is not yet resolved by the binary erasure channel example, was stated as a conjecture in (Kumar *et al.*, 2014). There are several cases where the equality $J(X;Y) = \overline{G}(X;Y)$ holds. Generalizing the binary erasure channel example, it was shown in (Vellambi and Kliewer, 2018; Yu and Tan, 2020) that the equality holds if $P_{X,Y}$ is a *pseudo-product distribution*, i.e., there exists distributions Q_X, Q_Y over $\mathcal{X}, \mathcal{Y}$ respectively and a set $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{Y}$ such that $P_{X,Y}$ is the conditional distribution of $Q_X Q_Y$ condititional on $\mathcal{A}$, that is, $P_{X,Y}(x, y) \propto \mathbf{1}\{(x, y) \in \mathcal{A}\} Q_X(x) Q_Y(y)$. This is generalized to the case where $P_{X,Y}$ is a *Wyner-product distribution* in (Yu and Tan, 2020) (refer to (Yu and Tan, 2020) for the definition).

The conjecture on whether it is possible to have $J(X;Y) < \overline{G}(X;Y)$ was resolved in (Yu and Tan, 2020), which showed that $J(X;Y) < \overline{G}(X;Y)$ when $X \sim \mathrm{Bern}(1/2)$ and $Y|X$ is a binary symmetric channel (i.e., $Y = \{0, 1\}$, $P_{Y|X}(0|1) = P_{Y|X}(1|0) = p$), or equivalently, (X, Y) is a doubly symmetric binary source. Despite being one of the simpliest joint distribution, the computation of $\overline{G}(X;Y)$ for the doubly symmetric binary source is highly nontrivial. Interested readers are referred to (Yu and Tan, 2020; Yu and Tan, 2022).

[3] The minimum in $J(X;Y) = \min_{P_{U|X,Y}: X \leftrightarrow U \leftrightarrow Y} I(X, Y; U)$ is attained by $U = X$ when $p \le 1/2$, giving $I(X, Y; U) = H(X) = 1$. When $p > 1/2$, it is attained by taking $P_{U|X}$ to be a binary erasure channel with erasure probability $2p - 1$, and $P_{Y|U}$ to be a binary erasure channel with erasure probability $1/2$, giving $I(X, Y; U) = H(X, Y) - H(X, Y|U) = 1 + H_b(p) - 1 = H_b(p)$ (Cuff, 2013).

[4] The minimum in $G(X;Y) = \min_{P_{U|X,Y}: X \leftrightarrow U \leftrightarrow Y} H(U)$ is attained by either $U = X$ (giving $H(U) = 1$), or $U = Y$ (giving $H(U) = H_b(p) + 1 - p$) (Kumar *et al.*, 2014).

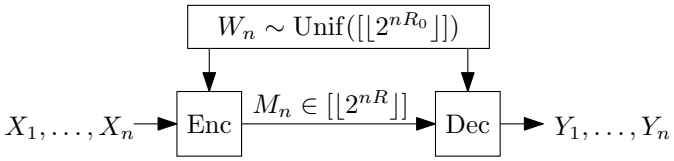


Figure 7.1: Asymptotic approximate fixed-length channel simulation with limited common randomness.

7 Asymptotic Channel Simulation with Limited Common Randomness

7.1 Approximate Channel Simulation

In the previous sections, we have studied asymptotic approximate fixed-length channel simulation without common randomness and with unlimited common randomness. Here we study the setting for a general common randomness rate R_0 in Figure 7.1 (see Definition 37). The precise single-letter characterization of the optimal rate region has been given by (Cuff, 2008; Cuff, 2013) as follows.

Theorem 45 (D/∞/A/FL/KS/LCR (Cuff, 2008; Cuff, 2013))**.** *For the asymptotic approximate fixed-length channel simulation setting (Definition 37) with known source distribution, limited common randomness and finite discrete X, Y, the optimal rate region is given by*

$$\bigcup_{P_{U|X,Y}:\, X \leftrightarrow U \leftrightarrow Y} \left\{ (R, R_0) \in \mathbb{R}^2 : \begin{array}{l} R \geq I(X;U), \\ R_0 + R \geq I(X,Y;U) \end{array} \right\}. \tag{7.1}$$

Moreover, it suffices to consider U with cardinality $|\mathcal{U}| \leq |\mathcal{X}||\mathcal{Y}| + 1$.

Refer to Figure 7.2 for an illustration of the optimal rate region in (7.1).

Proof. We adopt the strategy in (Cuff, 2008). First prove the achievability part. Assume $P_{U|X,Y}$ satisfies that $X \leftrightarrow U \leftrightarrow Y$ forms a Markov chain. Assume $R > I(X;U)$ and $R_0 + R > I(X,Y;U)$. Let $\mathfrak{U} = (u^n(m,w))_{m \in [\lfloor 2^{nR} \rfloor], w \in [\lfloor 2^{nR_0} \rfloor]}$ be a random codebook with i.i.d. entries following P_U. Let $\bar{M} \sim \mathrm{Unif}([\lfloor 2^{nR} \rfloor])$ be independent of $W \sim \mathrm{Unif}([\lfloor 2^{nR_0} \rfloor])$. Let $\bar{U}^n := u^n(\bar{M}, W)$, and $(\bar{X}^n, \bar{Y}^n)|\bar{U}^n \sim P_{X|U}^n P_{Y|U}^n$. Applying the soft covering lemma on the channel $P_{X,Y|U}$, since $R + R_0 > I(X,Y;U)$, we have

$$\mathbb{E}\big[\delta_{\mathrm{TV}}((\bar{X}^n, \bar{Y}^n),\, P_{X,Y}^n \,|\, \mathfrak{U})\big] \to 0 \tag{7.2}$$

as $n \to \infty$.

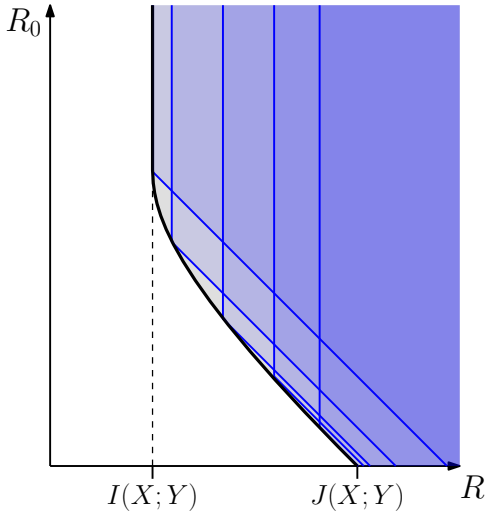


Figure 7.2: An illustration of the optimal rate region in (7.1), similar to the figure in (Cuff, 2013). It is the union of regions in the form $\{(R, R_0) : R \geq I(X;U), R_0 + R \geq I(X,Y;U)\}$ for different $P_{U|X,Y}$'s satisfying $X \leftrightarrow U \leftrightarrow Y$ (the blue polygons in the figure). As noted in (Cuff, 2013), there are two interesting extreme points. If R_0 is unlimited, the smallest possible R is $I(X;Y)$, reducing to the unlimited common randomness case in Theorem 38. If $R_0 = 0$, the smallest possible R is $J(X;Y) := \min_{P_{U|X,Y}: X \leftrightarrow U \leftrightarrow Y} I(X,Y;U)$, reducing to the no common randomness case in Theorem 40.

Now we study the distribution of $\bar{X}^n$ conditional on $(W, \mathfrak{U})$. For a fixed W, $\bar{U}^n$ is randomly picked from the codebook $(u^n(m, W))_{m \in [\lfloor 2^{nR} \rfloor]}$. Hence, the soft covering lemma gives us

$$\mathbb{E}\big[\delta_{\mathrm{TV}}(\bar{X}^n,\, P_X^n \,|\, W, \mathfrak{U})\big] \to 0 \tag{7.3}$$

as $n \to \infty$, since $R > I(X;U)$. We take the encoding Markov kernel to be $P_{\bar{M}|\bar{X}^n,W,\mathfrak{U}}(\cdot|\cdot,\cdot,\mathfrak{U})$, and the decoding Markov kernel to be $P_{\bar{Y}^n|\bar{M},W,\mathfrak{U}}(\cdot|\cdot,\cdot,\mathfrak{U})$ (these kernels are "random" and depend on $\mathfrak{U}$, though we will later argue that there exists a fixed value of $\mathfrak{U}$ that gives good kernels).

Similar to the proof of Theorem 40, since the actual input is $X^n \sim P_X^n$ (independent of the common randomness $W \sim \mathrm{Unif}([\lfloor 2^{nR_0} \rfloor]))$ instead of $\bar{X}^n$, we have to swap $\bar{X}^n$ with X^n and redefine the other random variables accordingly. Let $M|(X^n, W, \mathfrak{U}) \sim P_{\bar{M}|\bar{X}^n,W,\mathfrak{U}}(\cdot|X^n, W, \mathfrak{U})$ be the description given by the encoding Markov kernel when the input is X^n instead of $\bar{X}^n$, and $\tilde{Y}^n|(M, W, \mathfrak{U}) \sim P_{\bar{Y}^n|\bar{M},W,\mathfrak{U}}(\cdot|M, W, \mathfrak{U})$ be the output given by the decoding Markov kernel when the description is M. Since $P_{\tilde{Y}^n|X,W,\mathfrak{U}} = P_{\bar{Y}^n|X,W,\mathfrak{U}}$ by construction, using Lemma 36, we have

$$\begin{aligned}
&\delta_{\mathrm{TV}}\big((X^n, \tilde{Y}^n, W),\, (\bar{X}^n, \bar{Y}^n, W) \,\big|\, \mathfrak{U}\big) \\
&= \delta_{\mathrm{TV}}\big((X^n, W),\, (\bar{X}^n, W) \,\big|\, \mathfrak{U}\big)
\end{aligned}$$

$$= \mathbb{E}\left[\delta_{\mathrm{TV}}(X^n,\, \bar{X}^n \,|\, W, \mathfrak{U}) \,\middle|\, \mathfrak{U}\right]$$
$$\to 0 \tag{7.4}$$

in expectation (and hence in probability) as $n \to \infty$ due to (7.3), where both equalities are due to Lemma 36. Combining (7.2) and (7.4) using the triangle inequality,

$$\delta_{\mathrm{TV}}((X^n, \tilde{Y}^n),\, P^n_{X,Y} \,|\, \mathfrak{U}) \to 0$$

in probability as $n \to \infty$. This means there is a fixed choice $\mathfrak{u}$ of $\mathfrak{U}$ (which depends on n) that gives $\delta_{\mathrm{TV}}((X^n, \tilde{Y}^n),\, P^n_{X,Y} \,|\, \mathfrak{U} = \mathfrak{u}) \to 0$, which is the desired result.

We now prove the converse using a modified version of the arguments in (Cuff, 2013). Assume (R, R_0) is achievable. Consider any scheme with

$$\epsilon := \delta_{\mathrm{TV}}((X^n, \tilde{Y}^n),\, P^n_{X,Y}) = \mathbb{E}\left[\delta_{\mathrm{TV}}(\tilde{Y}^n,\, P^n_{Y|X} \,|\, X^n)\right].$$

Our goal is to show that there exist random variables X, Y, U satisfying $(X, Y) \sim P_X P_{Y|X}$, $X \leftrightarrow U \leftrightarrow Y$, and approximately satisfying (7.1), with a gap that tends to 0 as $\epsilon \to 0$. Let $Q \sim \mathrm{Unif}([n])$. While we can take $X = X_Q$ (which has the correct distribution $X_Q \sim P_X$ since $X_i \overset{iid}{\sim} P_X$), we cannot find Y with $(X, Y) \sim P_X P_{Y|X}$ among existing random variables, since $\tilde{Y}^n$ does not actually follow the ideal conditional distribution $P^n_{Y|X}$. Hence, we have to artificially construct the ideal Y^n. Applying the coupling lemma (Proposition 34) on $P_{\tilde{Y}^n|X^n}(\cdot|X^n)$ and $P^n_{Y|X}(\cdot|X^n)$, we can have a random sequence Y^n with $Y^n|X^n \sim P^n_{Y|X}$ and $\mathbb{P}(\tilde{Y}^n \neq Y^n) \leq \epsilon$. We have

$$\begin{aligned}
nR &\overset{(a)}{\geq} I(X^n; M|W) \\
&\overset{(b)}{=} I(X^n; M, W) \\
&= \sum_{i=1}^n I(X_i; M, W|X^{i-1}) \\
&\overset{(c)}{=} \sum_{i=1}^n I(X_i; M, W, X^{i-1}) \\
&\geq \sum_{i=1}^n I(X_i; M, W) \\
&= nI(X_Q; M, W|Q) \\
&\overset{(d)}{=} nI(X_Q; M, W, Q),
\end{aligned} \tag{7.5}$$

where (a) is because $M \in [\lfloor 2^{nR} \rfloor]$, (b) is because $I(W; X^n) = 0$, (c) is because $I(X_i; X^{i-1}) = 0$ since X_i's are i.i.d., and (d) is because $I(X_Q; Q) = 0$. By the exact same arguments

applied on (X^n, Y^n) instead of X^n,

$$\begin{aligned} n(R_0 + R) &\geq I(X^n, Y^n; M, W) \\ &\geq nI(X_Q, Y_Q; M, W, Q). \end{aligned}$$

If we take $U := (M, W, Q)$, we would have $R \geq I(X_Q; U)$ and $R_0 + R \geq I(X_Q, Y_Q; U)$, though we unfortunately have $X_Q \leftrightarrow U \leftrightarrow \tilde{Y}_Q$ (since the decoder outputs $\tilde{Y}_Q$ based on M, W, Q) instead of $X_Q \leftrightarrow U \leftrightarrow Y_Q$. Since $\mathbb{P}(\tilde{Y}^n \neq Y^n) \leq \epsilon$ is small, the Markov chain "$X_Q \leftrightarrow U \leftrightarrow Y_Q$" approximately holds, and we can indeed make it exactly holds by introducing a random variable V with a small entropy. By Lemma 76 (proved in Appendix B), there exists V such that $X_Q \leftrightarrow (U, V) \leftrightarrow Y_Q$ holds and $H(V) \leq \delta_{|\mathcal{X}|,|\mathcal{Y}|}(\epsilon)$, where $\delta_{|\mathcal{X}|,|\mathcal{Y}|}(\epsilon)$ is a function that tends to 0 as $\epsilon \to 0$ for any fixed $|\mathcal{X}|, |\mathcal{Y}|$. We have $R \geq I(X_Q; U) \geq I(X_Q; U, V) - H(V)$ and $R_0 + R \geq I(X_Q, Y_Q; U) \geq I(X_Q, Y_Q; U, V) - H(V)$. Letting $\epsilon \to 0$, we have $H(V) \to 0$, and (R, R_0) lies in the region in (7.1). Therefore, (7.1) is the capacity region. Readers are referred to (Cuff, 2013) for the cardinality bound $|\mathcal{U}| \leq |\mathcal{X}||\mathcal{Y}| + 1$. □

Similar to Theorem 38 and Theorem 41, the optimal rate region for arbitrary source can be obtained by considering the worst-case source distribution P_X. The following result is proved in (Bennett *et al.*, 2014) using the method of types. Refer to (Bennett *et al.*, 2014) for the proof.

Theorem 46 (D/∞/A/FL/AS/LCR (Bennett *et al.*, 2014))**.** *For the asymptotic approximate fixed-length channel simulation setting (Definition 37) with limited common randomness, arbitrary source, and finite discrete X, Y, the optimal rate region is given by*

$$\bigcap_{P_X} \bigcup_{P_{U|X,Y}: X \leftrightarrow U \leftrightarrow Y} \left\{ \begin{array}{l} (R, R_0) \in \mathbb{R}^2 : \\ R \geq I(X; U), \\ R_0 + R \geq I(X, Y; U) \end{array} \right\}.$$

It is also possible to study the local randomness needed in this setting. Refer to Section 9.4.

7.2 Exact Channel Simulation

Although this section is focused on the asymptotic approximate fixed-length setting with limited common randomness, the asymptotic exact variable-length setting with limited common randomness, i.e., the scenario where the distribution requirement on $\tilde{Y}^n$ is exact, and the description M is variable-length (i.e., the G/∞/E/VL/KS/LCR setting), has also

been studied in (Yu and Tan, 2019). More precisely, we make two modifications to Definition 37. First, we require $\tilde{Y}^n|X^n \sim P_{Y|X}^n$ exactly. Second, instead of $M_n \in [\lfloor 2^{nR} \rfloor]$, we have $M_n \in \{0,1\}^n$, with the requirement that $M_n \in \mathcal{C}_{n,W_n}$ almost surely, where $(\mathcal{C}_{n,w})_{n\in\mathbb{N}^+,\, w\in\mathcal{W}_n}$ is a collection of prefix-free codebooks that we can design as a part of the coding scheme (similar to Definition 2), and the requirement that $\limsup_{n\to\infty} n^{-1}\mathbb{E}[|M_n|] \le R$. Note that when the common randomness rate R_0 is finite, the common randomness W_n still follows the uniform distribution $\mathrm{Unif}([\lfloor 2^{nR_0} \rfloor])$.

It was proved in (Yu and Tan, 2019) that the rate pair $(R, R_0) = (I(X;Y), H(Y|X))$ is achievable for the exact variable-length setting. More generally, (Yu and Tan, 2019) showed the following inner bound of the optimal rate region. Interested readers are referred to (Yu and Tan, 2019; Yu and Tan, 2022) for the proof, and for more inner and outer bounds.

Theorem 47 (G/∞/E/VL/KS/LCR (Yu and Tan, 2019))**.** *For the asymptotic exact variable-length channel simulation setting with known source distribution, limited common randomness and finite discrete X, Y, the optimal rate region is a superset of the following set:*

$$\bigcup_{P_{U|X,Y}:\, X\leftrightarrow U\leftrightarrow Y} \left\{ (R, R_0) \in \mathbb{R}^2 : \begin{array}{l} R \ge I(X;U), \\ R_0 + R \ge H(U) \end{array} \right\}.$$

8 One-shot Bounds for Fixed-Length Channel Simulation

In previous sections, we have studied one-shot settings with variable-length descriptions, where results are stated in terms of simple quantities like the mutual information (e.g. Theorem 4). The simplicity of these results is because we only care about the expected length of the description, which naturally corresponds to first-order quantities like mutual information which is the "expected amount of shared information". We have also studied asymptotic fixed-length settings, where results are again stated in terms of first-order quantities, due to the law of large numbers.

In this section, we study one-shot settings with fixed-length descriptions. The result will no longer be in terms of the familiar first-order quantities. Instead of the entropy

$$H(X) = \mathbb{E}\left[\log_2 \frac{1}{P_X(X)}\right],$$

we will use the *self-information*

$$\iota_X(x) := \log_2 \frac{1}{P_X(x)},$$

which captures the amount of information in the particular value x of the random variable X. Note that $H(X) = \mathbb{E}[\iota_X(X)]$, i.e., the entropy is the average amount of information in X. For the lossless compression of the source X, unlike the one-shot variable-length setting (where Huffman coding (Huffman, 1952) gives us an expected length between $H(X)$ and $H(X)+1$) and the asymptotic fixed-length setting (where the source coding theorem gives us the optimal compression rate of $H(X)$ bits per sample X_i), the optimal error probability of compressing one sample of X into a fixed-size description $M \in [\mathsf{N}]$ is upper-bounded by

$$\mathbb{P}\left(\iota_X(X) \geq \log_2 \mathsf{N}\right). \tag{8.1}$$

Refer to Theorem 11.4 in (Polyanskiy and Wu, 2024) for a proof. Intuitively, the amount of information in X is $\iota_X(X)$ which is random, and we have an error if $\iota_X(X)$ cannot fit within the $\log_2 \mathsf{N}$ number of bits in M. This highlights the limitation of one-shot fixed-length results. Since the amount of information is random, but the number of bits $\log_2 \mathsf{N}$ is fixed, we have to make $\log_2 \mathsf{N}$ large enough to accomodate most values of $\iota_X(X)$. The number of bits needed is no longer given by the expectation of $\iota_X(X)$, but the tail of $\iota_X(X)$.

By the same logic, for channel simulation results, instead of Theorems 40, 4 and 38 which are in terms of the mutual information, the results will be in terms of the *information density*

$$\iota_{X;Y}(x;y) := \log_2 \frac{\mathrm{d}P_{X,Y}}{\mathrm{d}P_X P_Y}(x,y),$$

where $\mathrm{d}P_{X,Y}/\mathrm{d}P_X P_Y$ is the Radon-Nikodym derivative between the joint distribution $P_{X,Y}$ and the product distribution $P_X P_Y$ of the marginals. We often write $\iota(x;y) = \iota_{X;Y}(x;y)$ if

the random variables are clear from the context. For discrete X, Y, we have

$$\iota_{X;Y}(x;y) = \log_2 \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}.$$

Note that $I(X;Y) = \mathbb{E}[\iota_{X;Y}(X;Y)]$. For example, one would expect that, for one-shot fixed-length channel simulation with unlimited common randomness, instead of having a result in terms of $I(X;Y)$ as in Theorems 4 and 38, the result will be in terms of the distribution of $\iota_{X;Y}(X;Y)$.

What makes a good one-shot fixed-length result? Since there is only one X, we can simply write the operational setting as an optimization problem, and "characterize" the optimal scheme. For example, the optimal error probability for the one-shot lossless source coding setting is simply

$$\min_{f:\mathcal{X}\to[k],\, g:[\mathsf{N}]\to\mathcal{X}} \mathbb{P}(X \neq g(f(X))) \tag{8.2}$$

This is not that different from asymptotic results like Theorem 40 which also involves an optimization problem. The reason why we prefer (8.1) over (8.2) is that (8.1) implies the asymptotic result as a corollary. If we compress an i.i.d. source $X^n = (X_1, \dots, X_n)$ into $M \in [\lfloor 2^{nR} \rfloor]$, we have $\iota_{X^n}(x^n) = -\log_2 \prod_{i=1}^n P_X(x_i) = \sum_{i=1}^n \iota_X(x_i)$, and hence $\iota_{X^n}(X^n)/n \to \mathbb{E}[\iota_X(X)] = H(X)$ by law of large numbers, and $\mathbb{P}(\iota_{X^n}(X^n) \geq \log_2 \lfloor 2^{nR} \rfloor) \to 0$ as long as $R > H(X)$. The basic expectation of a one-shot result is that it should imply the asymptotic result directly. Sometimes one-shot results can give us more refined estimates on the rate and the error probability, for example, in second-order results where we use the central limit theorem instead of the law of large numbers (Hayashi, 2008; Hayashi, 2009; Polyanskiy *et al.*, 2010; Watanabe *et al.*, 2015).

8.1 One-shot Soft Covering Lemma

Recall that in Section 5.5, we discussed the soft covering lemma where the input sequence $\tilde{X}^n$ is picked randomly from a codebook, and sent through a memoryless channel $P_{Y|X}$ to give $\tilde{Y}^n$. In this section, we will study the case $n = 1$. There are various one-shot and nonasymptotic versions of the soft covering lemma, e.g., (Han and Verdú, 1993; Hayashi, 2006; Cuff, 2013; Watanabe and Hayashi, 2014; Watanabe *et al.*, 2015; Cuff, 2016; Yagli and Cuff, 2019; Liu *et al.*, 2019; Li and Anantharam, 2021). Here we present the version by Yassaee (Yassaee, 2015), which admits a relatively simple statement and proof.

Lemma 48 (One-shot soft covering lemma (Yassaee, 2015))**.** *Consider a discrete channel $P_{Y|X}$ and a discrete input distribution P_X. Fix a positive integer N. Let $\mathfrak{X} = (x(m))_{m\in[\mathsf{N}]}$ be a random codebook with i.i.d. entries following P_X, $M \sim \mathrm{Unif}([\mathsf{N}])$, $\tilde{X} := x(M)$, and*

$\tilde{Y}|\tilde{X} \sim P_{Y|X}$. *Then*

$$\mathbb{E}\Big[\delta_{\mathrm{TV}}(\tilde{Y},\, P_Y \,|\, \mathfrak{X})\Big] \le \sqrt{1 - \mathbb{E}^2\left[\left(1 + \mathsf{N}^{-1}2^{\iota(X;Y)}\right)^{-1/2}\right]}. \tag{8.3}$$

The asymptotic soft covering lemma (Lemma 39) follows as a corollary of Lemma 48. To see this, consider the channel $P_{Y|X}^n$, input distribution P_X^n, and $\mathsf{N} = \lfloor 2^{nR} \rfloor$. Note that

$$\begin{aligned}
\iota(X^n;Y^n) &= \log_2 \frac{P_{X,Y}^n(X^n,Y^n)}{P_X^n(X^n)P_Y^n(Y^n)} \\
&= \log_2 \prod_{i=1}^n \frac{P_{X,Y}(X_i,Y_i)}{P_X(X_i)P_Y(Y_i)} \\
&= \sum_{i=1}^n \log_2 \frac{P_{X,Y}(X_i,Y_i)}{P_X(X_i)P_Y(Y_i)} \\
&= \sum_{i=1}^n \iota(X_i;Y_i).
\end{aligned} \tag{8.4}$$

Hence, $n^{-1}\iota(X^n;Y^n) \to \mathbb{E}[\iota(X_1;Y_1)] = I(X;Y)$ in probability by law of large numbers. If $R > I(X;Y)$, then $\lfloor 2^{nR} \rfloor^{-1} 2^{\iota(X^n;Y^n)} \to 0$ in probability, and hence the right hand side of (8.3) tends to 0, and $\delta_{\mathrm{TV}}(\tilde{Y},\, P_Y \,|\, \mathfrak{X}) \to 0$ in mean.

We now present the proof of Lemma 48 using the arguments in (Yassaee, 2015).

Proof. The strategy in (Yassaee, 2015) is that, instead of directly bounding the TV distance, we bound it via the *fidelity* (also known as the *Bhattacharyya coefficient*) (Kailath, 1967) defined as

$$F(P,Q) := \mathbb{E}_{X\sim Q}\left[\sqrt{\frac{\mathrm{d}P}{\mathrm{d}Q}(X)}\right]$$

for two distributions P, Q, where $\mathrm{d}P/\mathrm{d}Q(x)$ is the Radon-Nikodym derivative. For discrete P, Q, we have

$$F(P,Q) = \sum_x \sqrt{P(x)Q(x)}.$$

We have $0 \le F(P,Q) \le 1$, and $F(P,Q) = 1$ if and only if $P = Q$. The TV distance can be bounded in terms of the fidelity as (Yassaee *et al.*, 2013)

$$\delta_{\mathrm{TV}}(P,Q) \le \sqrt{1 - F^2(P,Q)}. \tag{8.5}$$

This can be shown for the discrete case by

$$\sqrt{\delta_{\mathrm{TV}}^2(P,Q) + F^2(P,Q)}$$

$$
\begin{aligned}
&= \sqrt{\Big(\sum_x \frac{1}{2}\,|P(x)-Q(x)|\Big)^2 + \Big(\sum_x \sqrt{P(x)Q(x)}\Big)^2} \\
&\overset{(a)}{\le} \sum_x \sqrt{\Big(\frac{1}{2}\,|P(x)-Q(x)|\Big)^2 + \Big(\sqrt{P(x)Q(x)}\Big)^2} \\
&= \sum_x \sqrt{\Big(\frac{1}{2}(P(x)+Q(x))\Big)^2} \\
&= 1,
\end{aligned}
$$

where (a) is by the triangle inequality on the vectors

$$
\Big(\frac{1}{2}|P(x)-Q(x)|,\, \sqrt{P(x)Q(x)}\Big) \in \mathbb{R}^2
$$

for $x \in \mathcal{X}$.

To prove the lemma, for a fixed $\mathfrak{X} = (x(m))_{m\in[\mathsf{N}]}$,

$$
\begin{aligned}
F(P_Y, P_{\tilde{Y}|\mathfrak{X}}) &= \sum_y \sqrt{P_Y(y)\Big(\frac{1}{\mathsf{N}}\sum_{m=1}^{\mathsf{N}} P_{Y|X}(y|x(m))\Big)} \\
&= \frac{1}{\sqrt{\mathsf{N}}} \sum_{y,m} P_{Y|X}(y|x(m)) \sqrt{\frac{P_Y(y)}{\sum_{m'} P_{Y|X}(y|x(m'))}} \\
&= \frac{1}{\sqrt{\mathsf{N}}} \sum_{y,m} P_{Y|X}(y|x(m)) \Big(\sum_{m'} 2^{\iota_{X;Y}(x(m');y)}\Big)^{-1/2} \\
&= \frac{1}{\sqrt{\mathsf{N}}} \sum_{y,m} P_{Y|X}(y|x(m)) \Big(2^{\iota_{X;Y}(x(m);y)} + \sum_{m'\neq m} 2^{\iota_{X;Y}(x(m');y)}\Big)^{-1/2} \\
&= \sqrt{\mathsf{N}}\mathbb{E}\left[\Big(2^{\iota_{X;Y}(\tilde{X};\tilde{Y})} + \sum_{m'\neq M} 2^{\iota_{X;Y}(x(m');\tilde{Y})}\Big)^{-1/2} \,\middle|\, \mathfrak{X}\right],
\end{aligned}
$$

where for the last equality, recall that $M \sim \mathrm{Unif}([\mathsf{N}])$, $\tilde{X} := x(M)$, and $\tilde{Y}|\tilde{X} \sim P_{Y|X}$. Taking expectation over $\mathfrak{X}$,

$$
\begin{aligned}
&\mathbb{E}[F(P_Y, P_{\tilde{Y}|\mathfrak{X}})] \\
&= \sqrt{\mathsf{N}}\mathbb{E}\left[\Big(2^{\iota_{X;Y}(\tilde{X};\tilde{Y})} + \sum_{m'\neq M} 2^{\iota_{X;Y}(x(m');\tilde{Y})}\Big)^{-1/2}\right] \\
&= \sqrt{\mathsf{N}}\mathbb{E}\left[\mathbb{E}\left[\Big(2^{\iota_{X;Y}(\tilde{X};\tilde{Y})} + \sum_{m'\neq M} 2^{\iota_{X;Y}(x(m');\tilde{Y})}\Big)^{-1/2} \,\middle|\, \tilde{X},\tilde{Y}\right]\right] \\
&\overset{(b)}{\ge} \sqrt{\mathsf{N}}\mathbb{E}\left[\mathbb{E}^{-1/2}\left[2^{\iota_{X;Y}(\tilde{X};\tilde{Y})} + \sum_{m'\neq M} 2^{\iota_{X;Y}(x(m');\tilde{Y})} \,\middle|\, \tilde{X},\tilde{Y}\right]\right]
\end{aligned}
$$

$$\overset{(c)}{=} \sqrt{\mathsf{N}}\mathbb{E}\left[\left(2^{\iota_{X;Y}(\tilde{X};\tilde{Y})}+\mathsf{N}-1\right)^{-1/2}\right]$$
$$\geq \mathbb{E}\left[\left(1+\mathsf{N}^{-1}2^{\iota_{X;Y}(\tilde{X};\tilde{Y})}\right)^{-1/2}\right],$$

where (b) is by Jensen's inequality, and (c) is because $\mathbb{E}[2^{\iota_{X;Y}(X';Y')}] = \mathbb{E}[P_{X,Y}(X',Y')/(P_X(X')P_Y(Y'))] = 1$ when $X' \sim P_X$ is independent of $Y' \sim P_Y$. The result follows from (8.5). □

8.2 One-shot Fixed-length Channel Simulation

A one-shot fixed-length channel simulation setting is defined in a similar manner as Definition 37, with the blocklength set to $n = 1$. We include the definition of the one-shot fixed-length setting here for the sake of clarity.

Definition 49 (One-shot approximate fixed-length channel simulation)**.** Consider a general channel $P_{Y|X}$ and a general input distribution P_X. A one-shot approximate fixed-length channel simulation scheme with description size $\mathsf{N} \in \mathbb{N}^+$ and common randomness size $\mathsf{N}_0 \in \mathbb{N}^+ \cup \{\infty\}$ is characterized by a tuple $(P_W, P_{M|W,X}, P_{\tilde{Y}|W,M})$ described below:

- **Common randomness.** There is a common random source $W \in \mathcal{W}$, $W \sim P_W$ available to the encoder and the decoder. If $\mathsf{N}_0 = \infty$ (unlimited common randomness), we are allowed to choose an arbitrary P_W. If $\mathsf{N}_0 \neq \infty$, P_W is fixed to be $\mathrm{Unif}([\mathsf{N}_0])$. Note that $\mathsf{N}_0 = 1$ is the no common randomness case.
- **Encoder.** The encoder observes W and a source symbol $X \sim P_X$, and sends $M|(W,X) \sim P_{M|W,X}$ produced by passing W, X through an encoding Markov kernel $P_{M|W,X}$ from $\mathcal{W} \times \mathcal{X}$ to $[\mathsf{N}]$.
- **Decoder.** The decoder then outputs $\tilde{Y}|(W,M) \sim P_{\tilde{Y}|W,M}$ produced by passing W, M through a decoding Markov kernel $P_{\tilde{Y}|W,M}$ from $\mathcal{W} \times [\mathsf{N}]$ to $\mathcal{Y}$.
- **Performance metric.** We say that the code achieves a TV distance ϵ if
$$\mathbb{E}[\delta_{\mathrm{TV}}(\tilde{Y},\, P_{Y|X} \,|\, X)] \leq \epsilon.$$
The goal is to study the trade-off between N, N_0 and ϵ.

Now that we have a one-shot version of the soft covering lemma, we can simply use the same arguments as in Section 5.6, with the asymptotic soft covering lemma replaced by the one-shot one, to prove a one-shot bound for the likelihood encoder for the setting with unlimited common randomness. Let us first restate the likelihood encoder in Sections 3.4

and 5.6 applied to the one-shot setting. Take the common randomness to be $W = (\bar{Y}(i))_{i\in[\mathsf{N}]}$, where $\bar{Y}(i) \overset{iid}{\sim} P_Y$ for $i \in [\mathsf{N}]$. The encoder observes X, computes the likelihood

$$\alpha_i(X) := \frac{\mathrm{d}P_{Y|X}(\cdot|X)}{\mathrm{d}P_Y}(\bar{Y}(i)),$$

and generates the description $M \in [\mathsf{N}]$ with

$$\mathbb{P}(M = m \,|\, W, X) = \frac{\alpha_m(X)}{\sum_{i=1}^{\mathsf{N}} \alpha_i(X)}.$$

The decoder simply outputs $\tilde{Y} = \bar{Y}(M)$. We have the following bound.

Theorem 50 (D/1/A/FL/KS/UCR)**.** *For the one-shot approximate fixed-length channel simulation setting with known source distribution, unlimited common randomness and finite discrete X, Y, the likelihood encoder achieves a TV distance*

$$\begin{aligned}&\delta_{\mathrm{TV}}\big((X,\tilde{Y}),\, P_{X,Y}\big)\\ &\le \sqrt{1 - \mathbb{E}^2\left[\left(1 + \mathsf{N}^{-1}2^{\iota(X;Y)}\right)^{-1/2}\right]}.\end{aligned}$$

Proof. The proof is the same as that of Theorem 38 in Section 5.6 with $n = 1$, except that we have

$$\begin{aligned}&\mathbb{E}\left[\delta_{\mathrm{TV}}\big(\tilde{Y},\, P_{Y|X} \,|\, X\big)\right]\\ &\le \mathbb{E}\left[\delta_{\mathrm{TV}}\big(X,\, \hat{X} \,|\, W\big)\right]\\ &\le \sqrt{1 - \mathbb{E}^2\left[\left(1 + \mathsf{N}^{-1}2^{\iota(X;Y)}\right)^{-1/2}\right]}\end{aligned}$$

by Lemma 48. □

Note that this implies Theorem 38 (known source distribution, achievability part) by substituting $X = X^n$ and $Y = Y^n$ to be i.i.d. sequences and $\mathsf{N} = \lfloor 2^{nR} \rfloor$, since $n^{-1}\iota(X^n; Y^n) \to I(X;Y)$ in probability due to law of large numbers (see (8.4)).

Next, we use the same arguments as the proof of Theorem 40, with the asymptotic soft covering lemma replaced by the one-shot version, to prove the following one-shot bound for the setting with no common randomness.

Theorem 51 (D/1/A/FL/KS/NCR)**.** *For the one-shot approximate fixed-length channel simulation setting with known source distribution, no common randomness and finite discrete X, Y, there exists a scheme with a TV distance*

$$\delta_{\mathrm{TV}}\big((X,\tilde{Y}),\, P_{X,Y}\big) \le 2\sqrt{1 - \mathbb{E}^2\left[\left(1 + \mathsf{N}^{-1}2^{\iota(X,Y;U)}\right)^{-1/2}\right]}$$

for any $P_{U|X,Y}$ satisfying that $X \leftrightarrow U \leftrightarrow Y$ forms a Markov chain.

Proof. The proof is the same as that of Theorem 40 with $n = 1$, except that instead of (6.1), we have the following bound by Lemma 48:

$$\begin{aligned}&\mathbb{E}\big[\delta_{\mathrm{TV}}\big((\bar{X},\bar{Y}),\, P_{X,Y}\,\big|\,\mathfrak{U}\big)\big]\\&\le\sqrt{1-\mathbb{E}^2\left[\left(1+\mathsf{N}^{-1}2^{\iota(X,Y;U)}\right)^{-1/2}\right]}.\end{aligned}$$

The coefficient 2 is because the above bound is use twice in (6.1) and (6.2), and combined using triangle inequality. □

We then use the same arguments as the proof of Theorem 45, with the asymptotic soft covering lemma replaced by the one-shot one, to prove the following one-shot bound for the setting with limited common randomness.

Theorem 52 (D/1/A/FL/KS/LCR)**.** *For the one-shot approximate fixed-length channel simulation setting with known source distribution, limited common randomness and finite discrete X, Y, there exists a scheme with a TV distance*

$$\begin{aligned}&\delta_{\mathrm{TV}}\big((X,\tilde{Y}),\, P_{X,Y}\big)\\&\le\sqrt{1-\mathbb{E}^2\left[\left(1+(\mathsf{N}\mathsf{N}_0)^{-1}2^{\iota(X,Y;U)}\right)^{-1/2}\right]}\\&\quad+\sqrt{1-\mathbb{E}^2\left[\left(1+\mathsf{N}^{-1}2^{\iota(X;U)}\right)^{-1/2}\right]}\end{aligned}$$

for any $P_{U|X,Y}$ satisfying that $X \leftrightarrow U \leftrightarrow Y$ forms a Markov chain.

Proof. The proof is the same as that of Theorem 45 with $n = 1$, except that instead of (7.2), we have the following bound by Lemma 48:

$$\begin{aligned}&\mathbb{E}\big[\delta_{\mathrm{TV}}\big((\bar{X},\bar{Y}),\, P_{X,Y}\,\big|\,\mathfrak{U}\big)\big]\\&\le\sqrt{1-\mathbb{E}^2\left[\left(1+(\mathsf{N}\mathsf{N}_0)^{-1}2^{\iota(X,Y;U)}\right)^{-1/2}\right]},\end{aligned}$$

and instead of (7.3), we have the following bound by Lemma 48:

$$\begin{aligned}&\mathbb{E}\big[\delta_{\mathrm{TV}}\big(\bar{X},\, P_X\,\big|\,W,\mathfrak{U}\big)\big]\\&\le\sqrt{1-\mathbb{E}^2\left[\left(1+\mathsf{N}^{-1}2^{\iota(X;U)}\right)^{-1/2}\right]}.\end{aligned}$$

□

There are tighter one-shot bounds than Theorem 52. For example, the following bound was proved in (Yassaee, 2015).

Theorem 53 (D/1/A/FL/KS/LCR (Yassaee, 2015)). *For the one-shot approximate fixed-length channel simulation setting with known source distribution, limited common randomness and finite discrete X, Y, there exists a scheme with a TV distance*

$$\begin{aligned}&\delta_{\mathrm{TV}}\left((X,\tilde{Y}),\, P_{X,Y}\right)\\ &\leq \sqrt{1-\mathbb{E}^2\left[\left(1+(\mathsf{N}\mathsf{N}_0)^{-1}2^{\iota(X,Y;U)}\right)^{-1/2}\left(1+\mathsf{N}^{-1}2^{\iota(X;U)}\right)^{-1/2}\right]}.\end{aligned}$$

for any $P_{U|X,Y}$ satisfying that $X \leftrightarrow U \leftrightarrow Y$ forms a Markov chain.

The aforementioned results are for the case where the source distribution P_X is known. For the case with arbitrary source, interested readers are referred to (Cao *et al.*, 2022b; Cao *et al.*, 2022a) which give upper and lower bounds on the optimal description size in terms of the smooth max-divergence.

One-shot fixed-length results are useful for deriving finite-blocklength results, where we substitute $X = X^n$ and $Y = Y^n$ to be sequences. They are also useful when the input distribution P_{X^n} is not i.i.d., and when the channel $P_{Y^n|X^n}$ is not memoryless. Nevertheless, one-shot fixed-length results still require the information density terms (such as $\iota(X;Y)$, or $\iota(X^n;Y^n)$ if applied on sequences) to concentrate around its mean in order to give any meaningful bound. Take Theorem 50 as an example. If $\iota(X;Y)$ has a large variance, the fixed description size N has to be large enough so that $\iota(X;Y) \leq \log_2 \mathsf{N}$ with high probability, and hence $\log_2 \mathsf{N}$ may have to be significantly larger than the mean $I(X;Y)$. Therefore, despite being "one-shot", these results are still better suited for the situation where X, Y are "uniform enough", or "large enough" for concentration to occur. In contrast, one-shot variable-length results (such as Theorem 4) are suitable for general X, Y, including the situation where X, Y are small, and the information density $\iota(X;Y)$ is spread out. Therefore, if one has to perform channel simulation on the sequences X^n, Y^n, one can either treat X^n as a whole block (or divide X^n into large groups) and apply a one-shot/finite-blocklength fixed-length or variable-length scheme, or encode each symbol X_i (or each small group of symbols) separately using a one-shot variable-length scheme. Although technically one can apply a one-shot fixed-length scheme on each symbol X_i separately, this would not be a good idea.

There are other lines of research on finite-blocklength results and refined asymptotics for soft covering and channel simulation that are not covered in this monograph. Interested readers are referred to (Watanabe and Hayashi, 2014; Cuff, 2016; Cao *et al.*, 2022a) for second-order asymptotics of soft covering and channel simulation, which describe how the optimal rate for a finite blocklength n approaches the optimal asymptotic rate as $n \to \infty$ for a fixed total variation distance. Readers are also referred to (Yagli and Cuff, 2019; Yassaee, 2019) for the exponential rate of decay of the total variation distance in the soft covering lemma for a fixed rate R as the blocklength $n \to \infty$, called the soft covering exponent.

Another popular technique for proving channel simulation results is output statistics of random binning (Yassaee *et al.*, 2014), which can also apply to finite blocklength settings (Yassaee *et al.*, 2013).

9 Source and Channel Simulation with Limited Local Randomness

In the channel simulation settings studied in this monograph, the quantities of interest are usually the amount of communication and the amount of common randomness. We usually do not consider the amount of local randomness since local randomness, especially pseudo-randomness, is often considered inexpensive. Nevertheless, the resources required to generate true random numbers may still be nontrivial. In this section, we will consider the local randomness needed in various source and channel simulation tasks, namely source simulation, distributed source simulation,[1] local channel simulation, and channel simulation with limited common and local randomness.

9.1 Source Simulation

9.1.1 One-shot Source Simulation

Consider the following fundamental problem: how many fair coin flips are needed to generate a random variate $X \sim P_X$? This is commonly referred to as *random number generation* or *source simulation* (Knuth and Yao, 1976; Han and Verdú, 1993; Altuğ and Wagner, 2012). Similar to source coding and channel simulation, there are two main flavors of this setting: the *one-shot variable-length* setting where we simulate one variate $X \sim P_X$ using a variable number of coin flips, and the *asymptotic fixed-length* setting where we simulate an approximately i.i.d. sequence $\tilde{X}^n$ using a fixed number of coin flips. We start with the one-shot variable-length setting (Knuth and Yao, 1976).

Definition 54 (One-shot variable-length source simulation)**.** Consider a discrete distribution P_X over $\mathcal{X}$. A one-shot variable-length source simulation scheme is characterized by a pair $(\mathcal{C}, f)$ described below:

- **Codebook.** $\mathcal{C} \subseteq \{0,1\}^*$ is a full prefix-free codebook (which may be infinite), i.e., $\mathcal{C}$ is a prefix-free codebook where the equality in Kraft's inequality (Kraft, 1949) holds: $\sum_{c \in \mathcal{C}} 2^{-|c|} = 1$.

- **Simulator.** Given a sequence of coin flips $W_1, W_2, \ldots \overset{iid}{\sim} \mathrm{Bern}(1/2)$, the simulator reads the coin flips one by one, until the sequence of observed coin flips is found in

[1] Technically, the quantity of interest in distributed source simulation is the amount of common randomness instead of the amount of local randomness, though it is included in this section since it is a natural generalization of the source simulation setting.

the codebook, i.e., the simulator stops at

$$N := \min\{n \in \mathbb{N}_0 : W^n \in \mathcal{C}\}.$$

The simulator then outputs $X := f(W^N)$, where $f : \mathcal{C} \to \mathcal{X}$ is the sampling function.[2]

- **Requirement.** We require $X \sim P_X$ exactly.

- **Performance metric.** We are interested in the smallest expected number of coin flips needed $\mathbb{E}[N]$. Let $L^* := \inf \mathbb{E}[N]$ be the optimal expected number of coin flips, where the infimum is over all schemes satisfying the requirement.

This is a variable-length scheme since the number of coin flips N is not fixed. There are several different ways one can understand $\mathcal{C}$ and the sampling process. First, one may regard N as a stopping time of $W_1, W_2, \ldots$, where the encoder's decision on whether to stop at time $N = n$ can only depend on the currently observed coin flips W^n. Alternatively, one can also imagine the simulator to be traversing a complete binary tree, where each non-leaf node has two children: left (connected by an edge labeled 0) and right (connected by an edge labeled 1). The simulator starts at the root. When the simulator reads a coin flip that is 0, the simulator moves to the left child of the current node. When the simulator reads a 1, the simulator moves to the right child. The process stops when the simulator reaches a leaf node, which is a node where the path $c \in \{0,1\}^*$ from the root to that node corresponds to a codeword in $\mathcal{C}$. Finally, the simulator outputs the label of the leaf node, which is given by $f(c)$. This is called a *discrete distribution-generating* (DDG) tree in (Knuth and Yao, 1976), which is depicted in Figure 9.1.

Operationally, the simulator either possesses a long sequence of i.i.d. fair random bits (e.g., from some true random source, or from a random number book (RAND Corporation, 2001) in an old-fashioned manner), or has access to a (true or pseudo) random number generator (RNG). The simulator reads the bits one by one (or invokes the RNG interactively) until it decides to stop at time N. Since N is a stopping time that does not depend on future bits, the remaining unread bits in the long sequence are not "tainted", and are still i.i.d. fair random bits that can be reused for other tasks. Refer to Figure 9.2.

The optimal expected number of coin flips is within 2 bits from the entropy $H(X)$, as proved in (Knuth and Yao, 1976).

[2] This process is guaranteed to terminate almost surely. Let $N = \infty$ if there is no n such that $W^n \in \mathcal{C}$. Then we have $\mathbb{P}(N < \infty) = \sum_{c \in \mathcal{C}} \mathbb{P}(W^N = c) = \sum_{c \in \mathcal{C}} 2^{-|c|} = 1$.

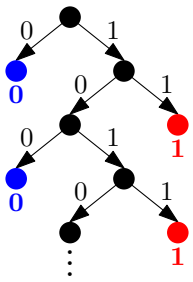


Figure 9.1: A discrete distribution-generating tree for the distribution Bern(1/3). Note that this tree has infinitely many nodes. The probability of reaching a leaf node with depth ℓ is $2^{-\ell}$. The probabilities of the leaf nodes labelled 0 (the blue nodes) sum up to 2/3, whereas the probabilities of the leaf nodes labelled 1 (the red nodes) sum up to 1/3. This tree corresponds to $\mathcal{C} = \{0, 11, 100, 1011, \ldots\}$, $f(0) = f(100) = \cdots = 0$, $f(11) = f(1011) = \cdots = 1$.

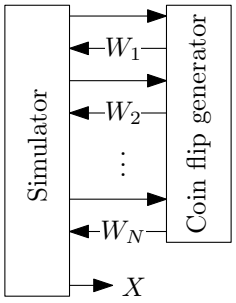


Figure 9.2: One-shot variable-length source simulation. The simulator interactively invokes the random number generator (i.e., calls the generator, obtains a bit W_1, calls the generator again, obtains a bit W_2, etc.) until the stopping time N, and then outputs X using $W_1, \ldots, W_N$.

Theorem 55 (Knuth and Yao 1976)**.** *For the one-shot variable-length source simulation setting, the optimal expected number of coin flips is bounded by*

$$H(X) \le L^* \le H(X) + 2.$$

Proof. The proof of the upper bound requires an infinite version of Kraft's inequality (Kraft, 1949): for any finite or countably infinite collection of nonnegative integers $(\ell_i)_{i\in\mathcal{I}}$ with $\sum_{i\in\mathcal{I}} 2^{-\ell_i} \le 1$, there exists a prefix-free codebook $\mathcal{C} \subseteq \{0,1\}^*$ and a bijective function $g : \mathcal{C} \to \mathcal{I}$ such that $|c| = \ell_{g(c)}$ for all $c \in \mathcal{C}$. We briefly describe a standard construction. Without loss of generality, assume $\mathcal{I} = [|\mathcal{I}|]$ or $\mathbb{N}^+$, and $\ell_1 \le \ell_2 \le \cdots$ are ordered in ascending order (this is possible since $|\{i \in \mathcal{I} : \ell_i = \ell\}|$ is always finite for every $\ell \in \mathbb{N}_0$ due to $\sum_i 2^{-\ell_i} \le 1$). Let $c_i \in \{0,1\}^*$ be the first ℓ_i binary digits of $\sum_{j=1}^{i-1} 2^{-\ell_j}$ after the decimal point. For $i' > i$, the first ℓ_i binary digits of $\sum_{j=1}^{i'-1} 2^{-\ell_j} \ge \sum_{j=1}^{i-1} 2^{-\ell_j} + 2^{-\ell_i}$ must be different from the first ℓ_i binary digits of $\sum_{j=1}^{i-1} 2^{-\ell_j}$, and hence c_i cannot be a prefix of $c_{i'}$. Hence, we can take $\mathcal{C} = \{c_i : i \in \mathcal{I}\}$ and $g(c_i) = i$.

We prove the upper bound $L^* \le H(X) + 2$. For $x \in \mathcal{X}$, let $P_X(x) = \sum_{i=1}^{\infty} b_{x,i} 2^{-i}$, $b_{x,i} \in \{0,1\}$ be the binary representation of $P_X(x)$. Since $\sum_x \sum_{i=1}^{\infty} b_{x,i} 2^{-i} = 1$, we invoke Kraft's inequality to construct a prefix-free codebook $\mathcal{C} \subseteq \{0,1\}^*$ and a bijective function $g : \mathcal{C} \to \{(x,i) : b_{x,i} = 1\}$ (write $g(c) = (g_1(c), g_2(c))$) such that $|c| = g_2(c)$. Take $f(c) := g_1(c)$. For $W_1, W_2, \ldots \overset{iid}{\sim} \mathrm{Bern}(1/2)$, $N = \min\{n : W^n \in \mathcal{C}\}$, we have

$$\begin{aligned}
\mathbb{P}(f(W^N) = x) &= \sum_{c\in\mathcal{C}:\, f(c)=x} \mathbb{P}(W^N = c) \\
&= \sum_{c\in\mathcal{C}:\, f(c)=x} 2^{-g_2(c)} \\
&= \sum_{i=1}^{\infty} b_{x,i} 2^{-i} \\
&= P_X(x).
\end{aligned}$$

Also,

$$\begin{aligned}
\mathbb{E}[N] &= \sum_{c\in\mathcal{C}} \mathbb{P}(W^N = c) \cdot |c| \\
&= \sum_x \sum_{i=1}^{\infty} b_{x,i} \cdot 2^{-i} \cdot i \\
&= \sum_x \sum_{j=0}^{\infty} \sum_{i=j+1}^{\infty} b_{x,i} 2^{-i} \\
&= \sum_x \sum_{j=0}^{\infty} \left(P_X(x) - 2^{-j} \left\lfloor 2^j P_X(x) \right\rfloor \right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_x \sum_{j=0}^{\infty} \min\left\{P_X(x),\, 2^{-j}\right\} \\
&= \sum_x \left(\left(\lfloor -\log_2 P_X(x) \rfloor + 1\right) P_X(x) + \sum_{j=\lfloor -\log_2 P_X(x)\rfloor + 1}^{\infty} 2^{-j} \right) \\
&= \sum_x \left(\left(\lfloor -\log_2 P_X(x) \rfloor + 1\right) P_X(x) + 2^{-\lfloor -\log_2 P_X(x)\rfloor} \right) \\
&\overset{(a)}{\leq} \sum_x \max\left\{ \left(-\log_2 P_X(x)\right) P_X(x) + 2^{\log_2 P_X(x)+1}, \right. \\
&\qquad \left. \left(-\log_2 P_X(x) + 1\right) P_X(x) + 2^{\log_2 P_X(x)} \right\} \\
&= \sum_x \left(-P_X(x) \log_2 P_X(x) + 2P_X(x)\right) \\
&= H(X) + 2,
\end{aligned}
$$

where (a) is because $t \mapsto (t+1)P_X(x) + 2^{-t}$ is convex, and hence its maximum is attained at the lower or upper bound of t.

For the lower bound, for any scheme, we have

$$
\begin{aligned}
H(X) &\leq H(W^N) \\
&= \mathbb{E}[-\log_2 P_{W^N}(W^N)] \\
&= \mathbb{E}[-\log_2(2^{-N})] \\
&= \mathbb{E}[N].
\end{aligned}
$$

□

For a computationally more efficient construction, Han and Hoshi (Han and Hoshi, 1997) proposed the *interval algorithm*, which can achieve an expected length of $\mathbb{E}[L] \leq H(X) + 3$. Refer to (Oohama, 2011; Watanabe and Han, 2020) for more detailed analyses on the interval algorithm. There are also works on the generation of random variates using biased coins instead of fair coins. Interested readers are referred to (Von Neumann, 1963; Hoeffding and Simons, 1970; Elias, 1972; Roche, 1991; Peres, 1992).

9.1.2 Asymptotic Source Simulation

We can also study an asymptotic source simulation setting where we want to generate an i.i.d. sequence $X^n \sim P_X^n$. If we are allowed to read a variable number of coin flips, we can simply apply Theorem 55 on X^n to obtain a scheme with expected number of coin flips $\leq nH(X) + 2$. Nevertheless, the asymptotic setting allows us to invoke the law of

large numbers to argue that the number of coin flips is concentrated, and hence we can use a fixed number of coin flips to generate X^n. Since there is a small probability that the amount of randomness needed is more than the number of coin flips available (if we are allowed ℓ coin flips, and $-\log_2 P_{X^n}(x^n) > \ell$, i.e., $P_{X^n}(x^n) < 2^{-\ell}$, then it is impossible to obtain a probability $P_{X^n}(x^n)$ using ℓ coin flips), we are only able to generate $\tilde{X}^n$ that is approximately i.i.d. following P_X. The setting is defined below.

Definition 56 (Asymptotic fixed-length source simulation)**.** Consider a discrete distribution P_X over $\mathcal{X}$. An asymptotic fixed-length source simulation scheme is characterized by a sequence $(f_n)_{n\in\mathbb{N}^+}$ described below:

- **Simulator.** Given $W_n \sim \mathrm{Unif}([\lfloor 2^{nR} \rfloor])$ (where $R \geq 0$ is the randomness rate), the simulator outputs $\tilde{X}^n := f_n(W_n)$, where $f_n : [\lfloor 2^{nR} \rfloor] \to \mathcal{X}^n$ is the sampling function.

- **Requirement.** We require

$$\delta_{\mathrm{TV}}(\tilde{X}^n,\, P_X^n) \to 0 \;\text{ as } n \to \infty.$$

- **Performance metric.** We are interested in the smallest randomness rate. Let R^* be the infimum of R over all schemes satisfying the requirement.

As one would naturally expect, the optimal randomness rate is given by the entropy $H(X)$. This follows from Theorem 55 (Knuth and Yao, 1976) as a direct corollary, and has been studied in (Han and Verdú, 1993; Steinberg and Verdú, 1996) for the more general case where the desired distribution of $\tilde{X}^n$ is a general stochastic process.

Theorem 57 (Knuth and Yao 1976; Han and Verdú 1993)**.** *For the asymptotic fixed-length source simulation setting, the optimal randomness rate is*

$$R^* = H(X).$$

Proof. We present two proofs of the achievability. For the first proof, we invoke the one-shot variable-length scheme in Theorem 55 n times to obtain a variable-length scheme that uses $\sum_{i=1}^n N_i$ coin flips, where N_i is the number of coin flips needed to simulate X_i, satisfying $\mathbb{E}[N_i] \leq H(X)+2$. By law of large numbers, $\mathbb{P}(\sum_{i=1}^n N_i > n(H(X)+3)) \to 0$ as $n \to \infty$. We then construct a fixed-length scheme by capping the number of coin flips to $n(H(X)+3)$, where the simulator experiences a failure and outputs any $\tilde{X}^n$ if the number of coin flips

is insufficient.[3] The probability of failure approaches 0. Hence, we can achieve a rate $R = H(X) + 3$. To remove the constant 3, instead of applying the one-shot variable-length scheme on every symbol X_i, we apply the one-shot scheme on every group of k symbols, i.e., $X^k, X_{k+1}^{2k}, \ldots$. Now we can achieve a rate

$$R = \frac{1}{k}\Big(H(X^k) + 3\Big) = H(X) + \frac{3}{k}.$$

The proof can be completed by taking $k \to \infty$.

For the second proof, fix $R > H(X)$ and generate a random codebook $\mathfrak{X} := (x^n(w))_{w \in [\lfloor 2^{nR} \rfloor]}$, where $x_i(w) \sim P_X$ i.i.d. across $i \in [n]$, $w \in [\lfloor 2^{nR} \rfloor]$. Let $W \sim \mathrm{Unif}([\lfloor 2^{nR} \rfloor])$ and $\tilde{X}^n := x^n(W)$. Since $R > H(X)$, applying the soft covering lemma on the noiseless channel $P_{X|X}$, we have $\delta_{\mathrm{TV}}(\tilde{X}^n, P_X^n \,|\, \mathfrak{X}) \to 0$ in probability as $n \to \infty$. Hence, there exists a fixed choice $\mathfrak{x}$ of $\mathfrak{X}$ (that depends on n) such that $\delta_{\mathrm{TV}}(\tilde{X}^n, P_X^n \,|\, \mathfrak{X} = \mathfrak{x}) \to 0$. The result follows from taking $f_n(w) = x^n(w)$, where $x^n(w)$ is given by $\mathfrak{x}$.

For the converse result, consider a scheme with $\delta_{\mathrm{TV}}(\tilde{X}^n, P_X^n) \le \epsilon$. Since $\tilde{X}^n$ is a function of W_n, we have $H(\tilde{X}^n) \le H(W_n) \le nR$. Invoking the coupling lemma (Proposition 34) on $P_{\tilde{X}^n}$ and P_X^n, we can have a random sequence X^n with $X^n \sim P_X^n$ and $\mathbb{P}(X^n \neq \tilde{X}^n) \le \epsilon$. By Fano's inequality (Fano, 1961), $H(X^n|\tilde{X}^n) \le \epsilon n \log_2 |\mathcal{X}| + 1$. Therefore, $nH(X) = H(X^n) \le H(\tilde{X}^n) + H(X^n|\tilde{X}^n) \le nR + \epsilon n \log_2 |\mathcal{X}| + 1$. Taking $\epsilon \to 0$, we have $R \ge H(X)$. □

9.2 Distributed Source Simulation

9.2.1 One-shot Distributed Source Simulation

We then consider a multi-terminal generalization of the one-shot source simulation setting in Definition 54, called one-shot distributed source simulation (Wyner, 1975a; Kumar *et al.*, 2014; Li and El Gamal, 2017), where two terminals want to simulate a pair of correlated random variables X and Y respectively, such that $(X, Y) \sim P_{X,Y}$. They are allowed to access a common sequence of coin flips. The setting is depicted in Figure 9.3. We now state the setting in (Li and El Gamal, 2017).

Definition 58 (One-shot variable-length distributed source simulation)**.** Consider a joint distribution $P_{X,Y}$ over $\mathcal{X} \times \mathcal{Y}$. A one-shot variable-length distributed source simulation scheme is characterized by a pair $(\mathcal{C}, P_{X|W^N}, P_{Y|W^N})$ described below:

[3] Technically, since Definition 56 has $W_n \sim \mathrm{Unif}([\lfloor 2^{nR} \rfloor])$ instead of $W_n \sim \mathrm{Unif}(\{0,1\}^{\lfloor nR \rfloor})$, the common randomness cannot be treated as a fixed-length sequence of bits. Nevertheless, we can extract a fixed-length sequence of bits by taking $\tilde{W}_n = ((W_n - 1) \bmod 2^{\lfloor n(R-\epsilon) \rfloor}) + 1 \in [2^{\lfloor n(R-\epsilon) \rfloor}]$ (where $\epsilon > 0$), which can be treated as a sequence of $2^{\lfloor n(R-\epsilon) \rfloor}$ bits. It is straightforward to show that $\delta_{\mathrm{TV}}(\tilde{W}_n, \mathrm{Unif}([2^{\lfloor n(R-\epsilon) \rfloor}])) \to 0$.

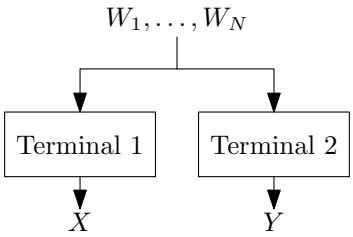


Figure 9.3: One-shot distributed source simulation.

- **Codebook.** $\mathcal{C} \subseteq \{0,1\}^*$ is a full prefix-free codebook (see Definition 54).

- **Terminals.** Given the common randomness as a sequence of coin flips $W_1, W_2, \ldots \overset{iid}{\sim}$ Bern(1/2), the two terminals read the coin flips one by one, and stop at $N := \min\{n \in \mathbb{N}_0 : W^n \in \mathcal{C}\}$.Terminal 1 then uses a stochastic decoder to output $X|W^N \sim P_{X|W^N}$, where $P_{X|W^N}$ is a Markov kernel from $\mathcal{C}$ to $\mathcal{X}$. Terminal 2 uses a stochastic decoder to output $Y|W^N \sim P_{Y|W^N}$, where $P_{Y|W^N}$ is a Markov kernel from $\mathcal{C}$ to $\mathcal{Y}$.

- **Requirement.** We require $(X, Y) \sim P_{X,Y}$ exactly.

- **Performance metric.** We are interested in the smallest expected number of common coin flips needed $\mathbb{E}[N]$. Let $L^* := \inf \mathbb{E}[N]$ be the optimal expected number of coin flips, where the infimum is over all schemes satisfying the requirement.

Note that here we allow unlimited local randomness at the terminals, and only limit the amount of common randomness. When $X = Y$, this setting reduces to the single-terminal source simulation setting in Definition 54, since the terminals cannot use any local randomness (i.e., $P_{X|W^N}$ and $P_{Y|W^N}$ are deterministic mappings) as an X that depends on the local randomness at Terminal 1 cannot agree with Y with probability 1, and hence the terminals can only use the common randomness to simulate $X = Y$.

Operationally, the two terminals would have a pre-shared stream of coin flips, either by actually sharing a long sequence of random bits, reading from the same publicly available randomness source (see Section 2.3), or by initializing their pseudorandom number generators (PRNGs) using the same seed. Then they would read a number of coin flips from the stream and output the correlated random variables X and Y. To allow the stream of coin flips to be reused for other tasks, the two terminals must be at the same position of the stream (or the PRNGs must be at the same state) after the scheme finishes, and hence the number

of coin flips used must be the same for the two terminals, and they must use the same prefix-free codebook $\mathcal{C}$.

Since $(W^N, X, Y) \sim P_{W^N} P_{X|W^N} P_{Y|W^N}$, we have $X \leftrightarrow W^N \leftrightarrow Y$. We also have $H(W^N) = \mathbb{E}[-\log_2 P_{W^N}(W^N)] = \mathbb{E}[N]$, and hence the problem becomes finding a random variable U that can be expressed in the form W^N for some codebook $\mathcal{C}$ and satisfies $X \leftrightarrow U \leftrightarrow Y$, such that $H(U)$ is minimized. If we remove the constraint that U that can be expressed in the form W^N and allow U to follow any distribution, we can still simulate U using approximately $H(U)$ number of coin flips by invoking Theorem 55. Therefore, similar to one-shot channel simulation without common randomness (Proposition 27), the answer is also given approximately by the common entropy (4.1)

$$G(X;Y) := \min_{P_{U|X,Y}: X \leftrightarrow U \leftrightarrow Y} H(U).$$

The following result has been given in (Kumar *et al.*, 2014; Li and El Gamal, 2017).

Proposition 59 (Kumar *et al.* 2014; Li and El Gamal 2017)**.** *For the one-shot variable-length distributed source simulation setting, the optimal expected number of coin flips is bounded by*

$$G(X;Y) \le L^* \le G(X;Y) + 2.$$

Proof. For the upper bound, given any U satisfying $X \leftrightarrow U \leftrightarrow Y$, we can generate U using an expected $\le H(U) + 2$ number of coin flips by Theorem 55, and have Terminal 1 generate X conditional on U, and Terminal 2 generate Y conditional on U. For the lower bound, since $(W^N, X, Y) \sim P_{W^N} P_{X|W^N} P_{Y|W^N}$, we have $X \leftrightarrow W^N \leftrightarrow Y$, and $\mathbb{E}[N] = \mathbb{E}[-\log_2 P_{W^N}(W^N)] = H(W^N) \ge G(X;Y)$. □

Alternatively, we may assume that the terminals share a common random variable $W \sim P_W$ instead of the sequence of coin flips. This is the original setting in (Kumar *et al.*, 2014). In this setting, the common randomness is a random variable W with a distribution P_W specially designed for this scheme (so we must design the scheme before the generation and sharing of W), instead of a general-purpose sequence of coin flips (where the scheme can be applied to any existing shared sequence of coin flips). If we study the smallest entropy $H(W)$, the answer is given by $G(X;Y)$ exactly. If we study the shortest expected length of a prefix-free encoding of W, the answer is between $G(X;Y)$ and $G(X;Y) + 1$ (Kumar *et al.*, 2014). If we instead study the smallest cardinality $|\mathcal{W}|$, the answer is given by the nonnegative rank $\text{rank}_+(\mathbf{P}_{X,Y})$ of the joint probability matrix $\mathbf{P}_{X,Y}$ (Zhang, 2012; Cubitt *et al.*, 2011), similar to Proposition 26.

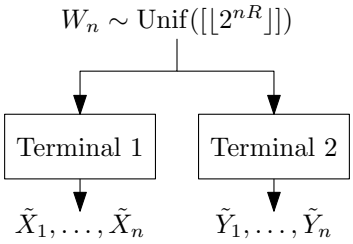


Figure 9.4: Asymptotic distributed source simulation.

9.2.2 Asymptotic Distributed Source Simulation

Distributed source simulation was originally considered in the asymptotic fixed-length approximate setting in (Wyner, 1975a), depicted in Figure 9.4. Unlike Definition 58, the terminals want to simulate two sequences $\tilde{X}^n, \tilde{Y}^n$ that approximately follow the i.i.d. distribution $P_{X,Y}^n$, using a fixed-length common randomness $W_n \sim \mathrm{Unif}([\lfloor 2^{nR} \rfloor])$. The setting is defined below.

Definition 60 (Asymptotic fixed-length distributed source simulation)**.** Consider a joint distribution $P_{X,Y}$ over $\mathcal{X} \times \mathcal{Y}$. An asymptotic fixed-length distributed source simulation scheme is characterized by a sequence of pairs $(P_{\tilde{X}^n|W_n}, P_{\tilde{X}^n|W_n})_{n \in \mathbb{N}^+}$ described below:

- **Terminals.** Given the common randomness $W_n \sim \mathrm{Unif}([\lfloor 2^{nR} \rfloor])$ (where $R \geq 0$ is the common randomness rate), Terminal 1 uses a stochastic decoder to output $\tilde{X}^n | W_n \sim P_{\tilde{X}^n|W_n}$, where $P_{\tilde{X}^n|W_n}$ is a Markov kernel from $[\lfloor 2^{nR} \rfloor]$ to $\mathcal{X}^n$. Terminal 2 uses a stochastic decoder to output $\tilde{Y}^n | W_n \sim P_{\tilde{Y}^n|W_n}$, where $P_{\tilde{Y}^n|W_n}$ is a Markov kernel from $[\lfloor 2^{nR} \rfloor]$ to $\mathcal{Y}^n$.

- **Requirement.** We require

$$\delta_{\mathrm{TV}}\big((\tilde{X}^n, \tilde{Y}^n),\, P_{X,Y}^n\big) \to 0 \;\text{ as } n \to \infty.$$

- **Performance metric.** We are interested in the smallest common randomness rate. Let R^* be the infimum of R over all schemes satisfying the requirement.

Similar to asymptotic channel simulation without common randomness (Theorem 40), the answer is given by Wyner's common information (Wyner, 1975a).

Theorem 61 (Wyner 1975a)**.** *For the asymptotic fixed-length distributed source simulation setting, where X, Y are finite discrete, the optimal common randomness rate is given by* Wyner's common information

$$J(X;Y) := \min_{P_{U|X,Y}: X \leftrightarrow U \leftrightarrow Y} I(X,Y;U).$$

Proof. We omit the subscript in W_n and simply write W. First prove the achievability part. Fix $P_{U|X,Y}$ satisfying $X \leftrightarrow U \leftrightarrow Y$. Fix $R > I(X,Y;U)$. Generate a random codebook $\mathfrak{U} := (u^n(m))_{m \in [\lfloor 2^{nR} \rfloor]}$, where $u_i(m) \sim P_U$ i.i.d. across $i \in [n]$, $m \in [\lfloor 2^{nR} \rfloor]$. Let $W \sim \mathrm{Unif}([\lfloor 2^{nR} \rfloor])$ and $\tilde{U}^n := u^n(W)$. Define random variables $(\tilde{X}^n, \tilde{Y}^n)$ with conditional distribution $(\tilde{X}^n, \tilde{Y}^n)|(\tilde{U}^n, W, \mathfrak{U}) \sim P^n_{X,Y|U}$. Since $X \leftrightarrow U \leftrightarrow Y$ forms a Markov chain, $P^n_{X,Y|U} = P^n_{X|U} P^n_{Y|U}$, and hence $\tilde{X}^n \leftrightarrow \tilde{U}^n \leftrightarrow \tilde{Y}^n$ forms a Markov chain. Since $\tilde{U}^n = u^n(W)$ is a function of $(W, \mathfrak{U})$, $\tilde{X}^n \leftrightarrow W \leftrightarrow \tilde{Y}^n$ forms a Markov chain conditional on $\mathfrak{U}$. Terminal 1 uses the Markov kernel $P_{\tilde{X}^n|W,\mathfrak{U}}(\cdot|\cdot, \mathfrak{U})$, and Terminal 2 uses the Markov kernel $P_{\tilde{Y}^n|W,\mathfrak{U}}(\cdot|\cdot, \mathfrak{U})$ (these kernels are "random" and depend on $\mathfrak{U}$, though we will later argue that there exists a fixed value of $\mathfrak{U}$ that gives good kernels). Applying the soft covering lemma on $P_{X,Y|U}$, we know that

$$\delta_{\mathrm{TV}}((\bar{X}^n, \bar{Y}^n),\, P^n_{X,Y} \,|\, \mathfrak{U}) \to 0$$

in probability as $n \to \infty$. There is a fixed choice $\mathfrak{u}$ of $\mathfrak{U}$ (which depends on n) that gives $\delta_{\mathrm{TV}}((X^n, \tilde{Y}^n),\, P^n_{X,Y} \,|\, \mathfrak{U} = \mathfrak{u}) \to 0$, which is the desired result.

We then prove the converse. Fix any scheme. Assume $\delta_{\mathrm{TV}}((\tilde{X}^n, \tilde{Y}^n),\, P^n_{X,Y}) \le \epsilon$. Since $(W, \tilde{X}^n, \tilde{Y}^n) \sim P_W P_{\tilde{X}^n|W} P_{\tilde{Y}^n|W}$, we have $\tilde{X}^n \leftrightarrow W \leftrightarrow \tilde{Y}^n$. Invoking the coupling lemma (Proposition 34) on $P_{\tilde{X}^n,\tilde{Y}^n}$ and $P^n_{X,Y}$, we can have random sequences X^n, Y^n with $(X^n, Y^n) \sim P^n_{X,Y}$ and $\mathbb{P}((X^n, Y^n) \neq (\tilde{X}^n, \tilde{Y}^n)) \le \epsilon$. We have

$$\begin{aligned}
nR &\ge H(W) \\
&\ge I(X^n, Y^n; W) \\
&= \sum_{i=1}^n I(X_i, Y_i; W | X^{i-1}, Y^{i-1}) \\
&\overset{(a)}{=} \sum_{i=1}^n I(X_i, Y_i; W, X^{i-1}, Y^{i-1}) \\
&\ge \sum_{i=1}^n I(X_i, Y_i; W) \\
&= nI(X_Q, Y_Q; W | Q) \\
&\overset{(b)}{=} nI(X_Q, Y_Q; W, Q),
\end{aligned}$$

where (a) is because $I(X_i, Y_i; X^{i-1}, Y^{i-1}) = 0$, and (b) is because $I(Q; X_Q, Y_Q) = 0$. The remaining obstacle is that $X_Q \leftrightarrow (W, Q) \leftrightarrow Y_Q$ does not hold, and we only have $\tilde{X}_Q \leftrightarrow (W, Q) \leftrightarrow \tilde{Y}_Q$ since Terminal 1 outputs $\tilde{X}_Q$ using W, Q, and Terminal 2 outputs $\tilde{Y}_Q$ using W, Q. Invoking Lemma 76 (proved in Appendix B), there exists V such that $X_Q \leftrightarrow (W, Q, V) \leftrightarrow Y_Q$ holds and $H(V) \leq \delta_{|\mathcal{X}|,|\mathcal{Y}|}(\epsilon)$, where $\delta_{|\mathcal{X}|,|\mathcal{Y}|}(\epsilon)$ is a function that tends to 0 as $\epsilon \to 0$ for any fixed $|\mathcal{X}|, |\mathcal{Y}|$. We have $R \geq I(X_Q, Y_Q; W, Q) \geq I(X_Q, Y_Q; W, Q, V) - \delta_{|\mathcal{X}|,|\mathcal{Y}|}(\epsilon)$. The result follows from letting $\epsilon \to 0$. □

Readers are referred to (Cuff *et al.*, 2010; Liu *et al.*, 2010; Kurri *et al.*, 2021) for generalizations of this setting to more than two terminals.

9.3 Local Channel Simulation

9.3.1 One-shot Local Channel Simulation

In the one-shot channel simulation setting without common randomness (Definition 25), the amount of communication is limited, though the amount of local randomness at the encoder and decoder is unlimited. We now modify Definition 25 so that the amount of communication is unlimited, but the total amount of local randomness at the encoder and decoder is limited. Since the encoder and decoder are linked by an infinite capacity channel, we may as well consider them as a single entity. This is a "local" channel simulation setting since it concerns a single terminal, unlike the "distributed" channel simulation settings with two terminals (encoder and decoder) that has been previously studied in this monograph. This setting (in the asymptotic case) is simply referred to as *channel simulation* in (Steinberg and Verdú, 1994), and referred to as *local synthesis* in (Cuff, 2013). Since we are using the term "channel simulation" instead of "channel synthesis" in this monograph, we call this setting "local channel simulation".

In the one-shot variable-length local channel simulation setting (which can be considered as a one-shot version of (Steinberg and Verdú, 1994)), the simulator observes X and the coin flips $W^N = (W_1, \ldots, W_N)$ obtained from a random number generator, and has to output Y such that $Y|X \sim P_{Y|X}$. One application is software simulation of a physical communication channel (e.g., a wireless fading channel), which is useful for measuring the performance of communication protocols (e.g., see (Mezzavilla *et al.*, 2015; Sun *et al.*, 2017)). In case if true randomness is expensive, the software will have to simulate the channel using as few random bits as possible (Steinberg and Verdú, 1994). Another interpretation of the setting is to regard the simulator as a "multi-purpose" random number generator, which is capable of generating a sample from any distribution in the family $(P_{Y|X}(\cdot|x))_{x \in \mathcal{X}}$. Upon receiving the input x, the simulator outputs a sample Y from the distribution $P_{Y|X}(\cdot|x)$. The goal is to use the least amount of randomness to sample from any distribution in this family.

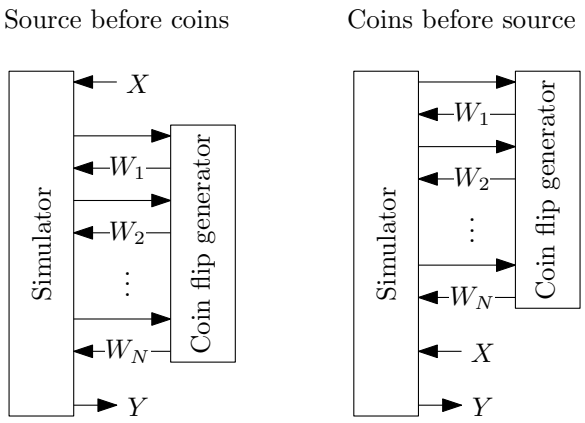


Figure 9.5: Left: One-shot local channel simulation, source before coins, where the simulator first observes X, and then interactively invokes the random number generator to obtain coin flips until the stopping time N (which can depend on X), and then outputs Y.
Right: One-shot local channel simulation, coins before source, where the simulator first interactively invokes the random number generator to obtain coin flips until the stopping time N (which cannot depend on X), and then observes X, and then outputs Y.

There are three options for whether the source distribution is known, and which of X and W^N is observed first:

- **Known source before coins.** $X \sim P_X$ follows a known distribution P_X and is observed before W^N, so the simulator can decide on a stopping time N based on X.
- **Arbitrary source before coins.** $X \in \mathcal{X}$ is arbitrary and is observed before W^N.
- **Coins before source.** W^N is observed before $X \in \mathcal{X}$, so the simulator must decide on the stopping time without knowing X. It does not matter whether the source distribution is known here.

The known source before coins setting has been studied, for example, in (Uyematsu and Kanaya, 1999) where a variable-length scheme based on the interval algorithm (Han and Hoshi, 1997) was proposed. The coins before source setting was discussed in (Cicalese *et al.*, 2019; Li, 2021). The settings are depicted in Figure 9.5, and are defined below.

Definition 62 (One-shot variable-length local channel simulation)**.** Consider a channel $P_{Y|X}$ (where Y is discrete), and a source distribution P_X (for the known source before coins case). A one-shot variable-length local channel simulation scheme is characterized by a pair $((\mathcal{C}_x)_{x\in\mathcal{X}}, f)$ described below:

- **Codebook.** $\mathcal{C}_x \subseteq \{0,1\}^*$ is a full prefix-free codebook for each $x \in \mathcal{X}$. For the coins before source case, we further require that $\mathcal{C}_x = \mathcal{C}$ are all the same and do not depend on x.

- **Simulator.** Given the source $X \in \mathcal{X}$ ($X \sim P_X$ for the known source before coins case) and a sequence of coin flips $W_1, W_2, \ldots \overset{iid}{\sim} \mathrm{Bern}(1/2)$, the simulator reads the coin flips one by one, and stops at

  $$N := \min\{n \in \mathbb{N}_0 : W^n \in \mathcal{C}_X\}.$$

  The simulator then outputs $Y := f(X, W^N)$, where $f : \mathcal{X} \times \{0,1\}^* \to \mathcal{Y}$ is the sampling function.

- **Requirement.** We require $Y|X \sim P_{Y|X}$ exactly.

- **Performance metric.** We are interested in the smallest expected number of coin flips needed $\mathbb{E}[N]$. For the known source before coin case, let $L^* := \inf \mathbb{E}[N]$ be the optimal expected number of coin flips, where the infimum is over all schemes satisfying the requirement. For the other two cases, we consider the worst case $L^* := \inf \sup_{x \in \mathcal{X}} \inf \mathbb{E}[N \mid X = x]$, where the infimum is over all schemes satisfying the requirement.

The two "source before coins" cases are straightforward. After knowing x, the simulator simply uses a source simulation scheme for $P_{Y|X}(\cdot|x)$. This is basically how a software library that supports generating random numbers from several different distributions works: when the user calls a function for generating a random number from a certain distribution, say the geometric distributrion $\mathrm{Geom}(1/2)$, the software library runs the best sampling scheme for $\mathrm{Geom}(1/2)$ and returns the result. The following results for the known source before coins setting and the arbitrary source before coins setting follow directly from Theorem 55.

Corollary 63 (Known/arbitrary source before coins (Knuth and Yao, 1976))**.** *For the one-shot variable-length local channel simulation setting, with known or arbitrary source before coins, the optimal expected number of coin flips is bounded by*

- *(Known source before coins)*

  $$H(Y|X) \le L^* \le H(Y|X) + 2.$$

- *(Arbitrary source before coins)*

  $$\sup_{x \in \mathcal{X}} H(Y|X = x) \le L^* \le \sup_{x \in \mathcal{X}} H(Y|X = x) + 2.$$

Proof. For the upper bounds of both cases, for each x, invoke Theorem 55 on $P_{Y|X}(\cdot|x)$ to construct $\mathcal{C}_x$ and $f(x,\cdot)$ so that if $N = \min\{n \in \mathbb{N}_0 : W^n \in \mathcal{C}_X\}$, then $f(x, W^N)|\{X = x\} \sim P_{Y|X}(\cdot|x)$, and $\mathbb{E}[N|X = x] \le H(Y|X = x) + 2$. For the known source before coins case, we have $\mathbb{E}[N] = \mathbb{E}[\mathbb{E}[N|X]] \le H(Y|X) + 2$.

For the lower bounds of both cases, for each x, since $f(x, W^N)|\{X = x\} \sim P_{Y|X}(\cdot|x)$, this is a scheme for source simulation for the distribution $P_{Y|X}(\cdot|x)$, and hence Theorem 55 gives $\mathbb{E}[N|X = x] \ge H(Y|X = x)$. □

We then consider the coins before source setting that has been discussed in (Cicalese *et al.*, 2019; Li, 2021). The simulator must read the coin flips $W_1, W_2, \ldots$ and decide on the stopping time $N = \min\{n \in \mathbb{N}_0 : W^n \in \mathcal{C}\}$ before it observes X. This is relevant in situations where the generator of the coin flips is slow (which may be the case for some hardware random number generators), and the simulator would like to invoke the coin flip generator beforehand and cache the result W^N, so that when it later receives X, it can respond quickly and output Y right away. If the generator is unreliable (the time it takes is unpredictable) or remote (the generator is located at another terminal connected by a slow communication link), such caching would be desirable for the reduction of the delay between the input of X into the simulator and the output of Y by the simulator.

Since W^N is independent of X, the problem becomes finding W^N with the smallest N such that $Y = f(X, W^N)$ has the desired conditional distribution $P_{Y|X}$. Readers may notice that this is the functional representation (El Gamal and Kim, 2011) discussed in Section 3.1, which is about finding a distribution P_W and a function $\phi : \mathcal{W} \times \mathcal{X} \to \mathcal{Y}$ such that if X is independent of $W \sim P_W$, and $Y = \phi(W, X)$, then $Y|X \sim P_{Y|X}$, i.e., $\phi(W, x) \sim P_{Y|X}(\cdot|x)$ for every $x \in \mathcal{X}$. However, unlike channel simulation with unlimited common randomness where we are interested in the smallest possible $H(Y|W)$ in (3.1), here we are interested in the smallest possible $H(W)$, which is approximately the number of coin flips needed to simulate W due to Theorem 55. Letting

$$H_W^* := \inf_{(P_W,\phi):\, \forall x:\, \phi(W,x) \sim P_{Y|X}(\cdot|x)} H(W), \tag{9.1}$$

the optimal expected number of coin flips for the coins before source setting is bounded by

$$H_W^* \le L^* \le H_W^* + 2. \tag{9.2}$$

This has been discussed in (Cicalese *et al.*, 2019; Li, 2021).

It was observed in (Kocaoglu *et al.*, 2017a) that the minimum entropy of functional representation in (9.1) is equivalent to the *minimum entropy coupling problem* (Vidyasagar, 2012; Painsky *et al.*, 2013; Kovačević *et al.*, 2015): given a collection of discrete distributions $(p_x)_{x\in\mathcal{X}}$, find the coupling $(Y_x)_{x\in\mathcal{X}}$ (i.e., $(Y_x)_x$ are jointly-distributed random variables with

marginals $Y_x \sim p_x$) with the smallest joint entropy $H((Y_x)_{x\in\mathcal{X}})$. Let

$$H^*((p_x)_{x\in\mathcal{X}}) := \inf_{P_{(Y_x)_x}:\, Y_x\sim p_x} H((Y_x)_x) \tag{9.3}$$

be the smallest joint entropy among couplings of $(p_x)_{x\in\mathcal{X}}$. Given a functional representation (P_W, ϕ), we can produce a coupling of $(P_{Y|X}(\cdot|x))_{x\in\mathcal{X}}$ by taking $Y_x = \phi(W, x)$, with joint entropy $H((Y_x)_{x\in\mathcal{X}}) \le H(W)$. For the other direction, given a coupling $(Y_x)_{x\in\mathcal{X}}$ of $(P_{Y|X}(\cdot|x))_{x\in\mathcal{X}}$, we can define a functional representation by $W = (Y_x)_{x\in\mathcal{X}}$ and $\phi(W, x) = Y_x$. Therefore, the minimum entropy coupling problem (9.3) is equivalent to the minimum entropy of functional representation (9.1) (as observed in (Kocaoglu *et al.*, 2017a)), which in turn is approximately equivalent to the local channel simulation problem due to (9.2) (as observed in (Cicalese *et al.*, 2019)). Hence, we can approximately characterize the optimal expected number of coin flips.

Theorem 64 (Coins before source (Kocaoglu *et al.*, 2017a; Cicalese *et al.*, 2019))**.** *We have*

$$H^*_W = H^*\big((P_{Y|X}(\cdot|x))_{x\in\mathcal{X}}\big).$$

Hence, the optimal expected number of coin flips for the local channel simulation problem (coins before source) is bounded by

$$H^*\big((P_{Y|X}(\cdot|x))_x\big) \le L^* \le H^*\big((P_{Y|X}(\cdot|x))_x\big) + 2.$$

Another equivalent way to state the minimum entropy coupling problem is through the concept of *aggregation* (Vidyasagar, 2012; Cicalese *et al.*, 2016). For two discrete distributions p (over $\mathcal{X}$) and q (over $\mathcal{Y}$), we say that p is an aggregation of q, written as $q \sqsubseteq p$ (adopting the notation in (Li, 2021)), if it is possible to form p by merging some of the masses of q, that is, if there exists a function $g : \mathcal{Y} \to \mathcal{X}$ such that if $Y \sim q$, then $g(Y) \sim p$. Then the minimum entropy coupling (9.3) can be equivalently stated as

$$H^*((p_x)_{x\in\mathcal{X}}) = \inf_{q:\, \forall x:\, q\sqsubseteq p_x} H(q). \tag{9.4}$$

Also note that $\sqsubseteq$ is a partial order, and entropy is nonincreasing with respect to $\sqsubseteq$ (i.e., $q \sqsubseteq p$ implies $H(q) \ge H(p)$). Therefore, if we could find q that is the greatest lower bound of $(p_x)_{x\in\mathcal{X}}$ (i.e., $q \sqsubseteq p_x$ for all x, and every q' satisfying $q' \sqsubseteq p_x$ for all x must also satisfy $q' \sqsubseteq q$), then this q would attain the infimum in (9.4).

Unfortunately, $\sqsubseteq$ is not a meet-semilattice, meaning that the greatest lower bound does not always exist. The minimum entropy coupling is difficult to compute. Even when $|\mathcal{X}| = 2$, the problem of finding $H^*(p_1, p_2)$ for two distributions p_1, p_2 is NP-hard (Vidyasagar, 2012; Kovačević *et al.*, 2015). The characterization in Theorem 64 is not quite helpful for the computation of the number of coin flips needed L^*.

To obtain an efficiently computable bound, (Cicalese *et al.*, 2019) studied another ordering—the *majorization order* (Marshall *et al.*, 2011). We say that q is majorized by p, written as $q \preceq p$, if $\sum_{y=1}^{k} q^{\downarrow}(y) \leq \sum_{y=1}^{k} p^{\downarrow}(y)$ for all $k \in \mathbb{N}^+$, where $q^{\downarrow}(1) \geq q^{\downarrow}(2) \geq \cdots$ are the entries of q sorted in nonascending order (append 0's at the end so $q^{\downarrow}(y)$ is defined for all $y \in \mathbb{N}^+$). Similar to the aggregation order $\sqsubseteq$, majorization $\preceq$ is also a partial order, and entropy is nonincreasing with respect to $\preceq$ (this is referred to as the *Schur-concavity* of entropy). Also, $q \sqsubseteq p$ implies $q \preceq p$ (Cicalese *et al.*, 2016).[4]

A nice property of majorization is that it forms a lattice (Marshall *et al.*, 2011; Cicalese and Vaccaro, 2002), and the greatest lower bound of a finite collection of discrete distributions $(p_x)_{x\in\mathcal{X}}$, written as $q = \bigwedge_{x\in\mathcal{X}} p_x$, can be given as a probability mass function $q : \mathbb{N}^+ \to \mathbb{R}$ where

$$q(k) := \inf_x \sum_{y=1}^{k} p_x^{\downarrow}(y) - \inf_x \sum_{y=1}^{k-1} p_x^{\downarrow}(y). \tag{9.5}$$

Note that the greatest lower bound (9.5) may or may not exist when the collection $(p_x)_{x\in\mathcal{X}}$ is infinite. Nevertheless, if $\lim_{k\to\infty} \inf_x \sum_{y=1}^{k} p_x^{\downarrow}(y) = 1$, then the greatest lower bound (9.5) is a valid distribution (Li, 2021).

As a result, we can obtain an efficiently computable lower bound of the minimum entropy coupling problem by

$$\begin{aligned} H^*((p_x)_{x\in\mathcal{X}}) &= \inf_{q:\, \forall x:\, q\sqsubseteq p_x} H(q) \\ &\geq \inf_{q:\, \forall x:\, q\preceq p_x} H(q) \\ &= H\Big(\bigwedge_x p_x\Big) \end{aligned}$$

if the greatest lower bound $\bigwedge_x p_x$ exists.[5] Furthermore, a bound in the other direction is also possible, showing that the solution to the minimum entropy coupling problem (and hence the local channel simulation problem) can be approximated by the greatest lower bound, giving an approximate solution that can be efficiently computed.

Theorem 65 (Coins before source (Cicalese *et al.*, 2019; Li, 2021; Compton, 2022))**.** *We have*

$$H\Big(\bigwedge_x P_{Y|X}(\cdot|x)\Big) \leq H^*((P_{Y|X}(\cdot|x))_x) \leq H\Big(\bigwedge_x P_{Y|X}(\cdot|x)\Big) + \log_2 e$$

[4] If $q \sqsubseteq p$, we can let $g : \mathcal{Y} \to \mathcal{X}$ such that if $Y \sim q$, then $g(Y) \sim p$. Let $y_1, y_2, \ldots$ be the entries of $\mathcal{Y}$ such that $q(y_1) \geq q(y_2) \geq \cdots$ are in nonincreasing order. Then $\sum_{y=1}^{k} q^{\downarrow}(y) = \sum_{i=1}^{k} q(y_i) \leq \sum_{x\in g(\{y_i:\, i\in[k]\})} p(x) \leq \sum_{x=1}^{k} p^{\downarrow}(x)$.

[5] If the greatest lower bound does not exist, the entropy of the minimum entropy coupling is infinite. To show this, note that if $\sum_{y=1}^{k} p^{\downarrow}(y) = a < 1$, then $p^{\downarrow}(y) \leq a/k$ for $y > k$, and hence $H(p) \geq (1-a)\log_2(k/a)$. Therefore, if $\lim_{k\to\infty} \inf_x \sum_{y=1}^{k} p_x^{\downarrow}(y) < 1$, then $\sup_x H(p_x) = \infty$, implying $H^*((p_x)_{x\in\mathcal{X}}) = \infty$.

if the greatest lower bound with respect to majorization $\bigwedge_x P_{Y|X}(\cdot|x)$ exists (otherwise H^* above is infinite). Hence, the optimal expected number of coin flips for the local channel simulation problem (coins before source) is bounded by

$$H\Big(\bigwedge_x P_{Y|X}(\cdot|x)\Big) \le L^* \le H\Big(\bigwedge_x P_{Y|X}(\cdot|x)\Big) + \log_2(4e) \text{ bits.}$$

The lower bound $H^* \ge H(\bigwedge P_{Y|X}(\cdot|x))$ was shown by (Cicalese *et al.*, 2019), which also proved an upper bound $H^* \le H(\bigwedge P_{Y|X}(\cdot|x)) + \lceil \log_2 |\mathcal{X}| \rceil$. The upper bound was improved to $H^* \le H(\bigwedge P_{Y|X}(\cdot|x)) + 2$ in (Li, 2021), and to $H^* \le H(\bigwedge P_{Y|X}(\cdot|x)) + \log_2 e$ in (Compton, 2022). Readers are referred to (Compton, 2022) for the proof. In particular, (Li, 2021) proved $H^* \le H(\bigwedge P_{Y|X}(\cdot|x)) + 2$ by showing that if $q \preceq p$, then

$$q \times \text{Geom}(1/2) \sqsubseteq p,$$

where $\text{Geom}(1/2)$ is the geometric distribution with parameter $1/2$, and $q \times \text{Geom}(1/2)$ denotes the product distribution, which implies that $(\bigwedge_{x'} P_{Y|X}(\cdot|x')) \times \text{Geom}(1/2) \sqsubseteq P_{Y|X}(\cdot|x)$ for all x. Also refer to (Shkel and Yadav, 2023; Shkel, 2024) for more bounds and discussions.

An interesting parallel between the arbitrary source before coins setting in Corollary 63 and the coins before source setting Theorem 65 is that for the arbitrary source before coins setting, the answer is approximately given by the least upper bound (with respect to the ordering $\le$ over $\mathbb{R}$) of the entropies of $P_{Y|X}(\cdot|x)$ for all $x \in \mathcal{X}$, whereas for the coins before source setting, the answer is approximately given by the entropy of the greatest lower bound (with respect to majorization) of $P_{Y|X}(\cdot|x)$ for all $x \in \mathcal{X}$. Swapping the order between the source and the coins corresponds to swapping the order between taking the entropy and taking the least upper bound / greatest lower bound in the answer.

There are polynomial time algorithms that can produce a coupling with an entropy within a constant gap away from the minimum (and hence an almost optimal scheme for the local channel simulation problem) (Kocaoglu *et al.*, 2017b; Rossi, 2019; Cicalese *et al.*, 2019; Li, 2021; Compton, 2022; Compton *et al.*, 2023). In particular, the greedy algorithm in (Kocaoglu *et al.*, 2017b) was shown to be within 1.22 bits from the optimum (Compton *et al.*, 2023). The asymptotic setting of coupling two i.i.d. distributions p^n, q^n has been studied in (Yu and Tan, 2018), where it was shown that $H^*(p^n, q^n) - n\max\{H(p), H(q)\} \to 0$ at least exponentially fast as $n \to \infty$ if $H(p) \ne H(q)$.

9.3.2 Asymptotic Local Channel Simulation

In this section, we consider the asymptotic local channel simulation setting in (Steinberg and Verdú, 1994), where the goal is to simulate the memoryless channel $P_{Y|X}^n$ approximately with the smallest amount of fixed-length local randomness $W_n \sim \text{Unif}([\lfloor 2^{nR} \rfloor])$. There are

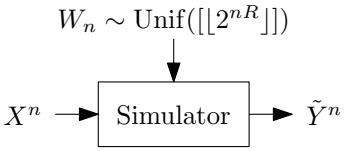


Figure 9.6: Asymptotic local channel simulation.

two options: known source distribution where $X^n \sim P_X^n$ is i.i.d. with a known distribution P_X, and arbitrary source where $X^n \in \mathcal{X}^n$ is arbitrary. Unlike Definition 62, there is no difference between whether the source is observed before the local randomness, since the local randomness is fixed to be a uniform integer, and does not depend on the sampler's choice. The setting is depicted in Figure 9.6, and is defined below.

Definition 66 (Asymptotic fixed-length local channel simulation)**.** Consider a channel $P_{Y|X}$ (where Y is discrete), and a source distribution P_X (for the known source distribution case). An asymptotic fixed-length local channel simulation scheme is characterized by a sequence $(f_n)_{n\in\mathbb{N}^+}$ described below:

- **Simulator.** Given the source $X^n \in \mathcal{X}^n$ ($X^n \sim P_X^n$ for the known source distribution case) and the local randomness $W_n \sim \mathrm{Unif}([\lfloor 2^{nR} \rfloor])$, the simulator outputs $\tilde{Y}^n := f(X^n, W_n)$, where $f : \mathcal{X}^n \times [\lfloor 2^{nR} \rfloor] \to \mathcal{Y}^n$ is the sampling function.

- **Requirement.**

  – For the known source distribution case, we require that $\tilde{Y}^n$ follows the conditional distribution $P_{Y|X}^n$ approximately, in the sense that

  $$\delta_{\mathrm{TV}}\big((X^n, \tilde{Y}^n),\, P_X^n P_{Y|X}^n\big) \to 0$$

  as $n \to \infty$. Or equivalently, $\delta_{\mathrm{TV}}(\tilde{Y}^n,\, P_{Y|X}^n \,|\, X^n) \to 0$ in probability as $n \to \infty$ by Lemma 36.

  – For the arbitrary source case, we require

  $$\sup_{x^n \in \mathcal{X}^n} \delta_{\mathrm{TV}}\big(\tilde{Y}^n,\, P_{Y|X}^n \,|\, X^n = x^n\big) \to 0$$

  as $n \to \infty$.

- **Performance metric.** We are interested in the smallest randomness rate. Let R^* be the infimum of R over all schemes satisfying the requirement.

The optimal randomness rate has been characterized in (Steinberg and Verdú, 1994). Note the similarity between this result and the one-shot local channel simulation (source before coin) result in Corollary 63.

Theorem 67 (Steinberg and Verdú 1994)**.** *For the asymptotic fixed-length local channel simulation of $P_{Y|X}$ where Y is finite discrete, the optimal randomness rate is:*

- *(Known source distribution)*
$$R^* = H(Y|X).$$
- *(Arbitrary source) If X is finite discrete,*
$$R^* = \max_{x \in \mathcal{X}} H(Y|X = x).$$

Proof. First prove the achievability of the known source distribution case. Invoke the one-shot variable-length scheme in Corollary 63 n times, we have a variable-length scheme that uses $\sum_{i=1}^n N_i$ coin flips, where N_i is the number of coin flips needed to simulate the channel $X_i \to Y_i$, satisfying $\mathbb{E}[N_i] \le H(Y|X) + 2$. By law of large numbers, $\mathbb{P}(\sum_{i=1}^n N_i > n(H(Y|X) + 3)) \to 0$ as $n \to \infty$. We then construct a fixed-length scheme by capping the number of coin flips to $n(H(Y|X)+3)$, where the simulator experiences a failure and outputs any $\tilde{Y}^n$ if the number of coin flips is insufficient.[6] The probability of failure approaches 0. Hence, we can achieve a rate $R = H(Y|X) + 3$. To remove the constant 3, instead of applying the one-shot variable-length scheme on every pair (X_i, Y_i), we apply the one-shot scheme on every group of k time slots, i.e., $(X^k, Y^k), (X_{k+1}^{2k}, Y_{k+1}^{2k}), \ldots$. Now we can achieve a rate $R = H(Y|X) + 3/k$. The proof can be completed by taking $k \to \infty$.

We then prove the arbitrary source case. Again invoke the one-shot variable-length scheme in Corollary 63 n times, and construct a fixed-length scheme by capping the number of coin flips to $n(1+\epsilon|\mathcal{X}|)(\max_x H(Y|X = x)+3)$ for some $\epsilon > 0$. Consider any fixed source sequence x^n. Fix $x \in \mathcal{X}$ and consider the time indices i where $x_i = x$. Let $n_x := |\{i : x_i = x\}|$ be the number of such time indices. By law of large numbers,
$$\mathbb{P}\Big(\sum_{i:\, x_i = x} N_i > n_x(H(Y|X = x) + 3)\Big) \le \delta_x(n_x),$$
where $\delta_x(\ell)$ is a non-increasing function satisfying $\lim_{\ell \to \infty} \delta_x(\ell) = 0$ for every x. We have
$$\mathbb{P}\Big(\sum_{i=1}^n N_i > n(1 + \epsilon|\mathcal{X}|)(H(Y|X = x) + 3)\Big)$$

[6] Technically, since Definition 66 has $W_n \sim \mathrm{Unif}([\lfloor 2^{nR} \rfloor])$ instead of $W_n \sim \mathrm{Unif}(\{0,1\}^{\lfloor nR \rfloor})$, the common randomness cannot be treated as a fixed-length sequence of bits. Refer to the proof of Theorem 57 for the conversion from W_n to a sequence of bits.

$$
\begin{aligned}
&\le \sum_x \mathbb{P}\Big(\sum_{i:\, x_i = x} N_i > (n_x + \lfloor \epsilon n \rfloor)(H(Y|X=x)+3) \Big) \\
&\le \sum_x \delta_x(n_x + \lfloor \epsilon n \rfloor) \\
&\le \sum_x \delta_x(\lfloor \epsilon n \rfloor) \\
&\to 0
\end{aligned}
$$

as $n \to \infty$. Hence, we can achieve a rate $(1+\epsilon|\mathcal{X}|)(\max_x H(Y|X=x)+3)$. The proof can be completed by considering groups of k time slots, and taking $k \to \infty$, $\epsilon \to 0$.

For the converse result for the known source distribution case, consider a scheme with $\delta_{\mathrm{TV}}((X^n, \tilde{Y}^n),\, P_X^n P_{Y|X}^n) \le \epsilon$. Since $\tilde{Y}^n$ is a function of (X^n, W_n), we have $H(\tilde{Y}^n|X^n) \le H(W_n) \le nR$. Invoking the coupling lemma (Proposition 34) on $P_{\tilde{Y}^n|X^n}(\cdot|X^n)$ and $P_{Y|X}^n(\cdot|X^n)$, we can have a random sequence Y^n with $Y^n|X^n \sim P_{Y|X}^n$ and $\mathbb{P}(Y^n \ne \tilde{Y}^n) \le \epsilon$. By Fano's inequality (Fano, 1961),

$$
H(Y^n|\tilde{Y}^n) \le \epsilon n \log_2 |\mathcal{Y}| + 1.
$$

Therefore,

$$
\begin{aligned}
nH(Y|X) &= H(Y^n|X^n) \\
&\le H(\tilde{Y}^n|X^n) + H(Y^n|\tilde{Y}^n) \\
&\le nR + \epsilon n \log_2 |\mathcal{Y}| + 1.
\end{aligned}
$$

Taking $\epsilon \to 0$, we have $R \ge H(Y|X)$. The converse result for the arbitrary source case follows from the known source distribution case by considering the degenerate source distribution $X = x$. □

The situation where the input process X^n is general (not necessarily i.i.d.) and the channel $P_{Y^n|X^n}$ is general (not necessarily memoryless) has also been studied in (Steinberg and Verdú, 1994), where the optimal rate is given as the conditional sup-entropy rate of Y given X. Algorithms based on the interval algorithm (Han and Hoshi, 1997) for computing the mapping f_n have been proposed in (Uyematsu and Kanaya, 1999). The situation where the source of randomness W_n is a general source (instead of being uniformly distributed) has been studied in (Altuğ and Wagner, 2012).

A related problem, called *local channel synthesis*, was studied in (Cuff, 2013), where the terminal observes $X^n \sim P_X^n$ and has to output Z^n as a function of X^n and local randomness, such that when (X^n, Z^n) is passed through the memoryless channel $P_{Y|X,Z}^n$, the output $\tilde{Y}^n$ approximately follows a prescribed conditional distribution $P_{Y|X}^n$, in the sense that $\delta_{\mathrm{TV}}((X^n, \tilde{Y}^n),\, P_X^n P_{Y|X}^n) \to 0$. It was shown in (Cuff, 2013) that the optimal local

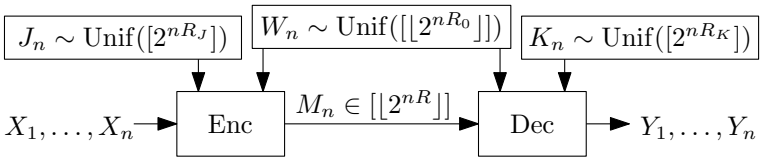


Figure 9.7: Asymptotic approximate fixed-length channel simulation with limited common and local randomness.

randomness rate is $I(Z;Y|X)$. Note that this result implies the known source distribution case of Theorem 67 by taking $Y = Z$.

9.4 Channel Simulation with Limited Common and Local Randomness

We now return to the "distributed" channel simulation setting with two terminals in Definition 37. The asymptotic fixed-length channel simulation setting where the local randomness at the encoder and the decoder is limited has been studied in (Cuff, 2008; Cuff, 2013). While Definition 37 allows the encoder and the decoder to be stochastic with unlimited local randomness, we modify the setting as follows:

- The encoder is given the local randomness $J_n \sim \mathrm{Unif}([2^{nR_J}])$ (where $R_J \geq 0$ is the encoder local randomness rate), and its output M_n must be a function of (W_n, X^n, J_n).
- The decoder is given the local randomness $K_n \sim \mathrm{Unif}([2^{nR_K}])$ (where $R_K \geq 0$ is the decoder local randomness rate), and its output $\tilde{Y}^n$ must be a function of (M_n, W_n, K_n).

Refer to Figure 9.7. We are interested in finding the optimal rate region, defined as the closure of achievable tuples (R, R_0, R_J, R_K) in a similar manner as Definition 37. The result in (Cuff, 2008; Cuff, 2013) is stated below.

Theorem 68 (D/∞/A/FL/KS/LCR/LLR (Cuff, 2008; Cuff, 2013))**.** *For the asymptotic approximate fixed-length channel simulation setting with known source distribution, limited common randomness (rate R_0), limited local randomness at the encoder (rate R_J) and the decoder (rate R_K), and finite discrete X, Y, the optimal rate region is given by*

$$\bigcup_{P_{U|X,Y}\colon X \leftrightarrow U \leftrightarrow Y} \left\{ \begin{array}{l} (R, R_0, R_J, R_K) \in \mathbb{R}^4 : \\ R \geq I(X;U), \\ R_0 + R \geq I(X,Y;U), \\ R_J \geq 0, \\ R_K \geq H(Y|U) \end{array} \right\}.$$

Note that the only constraint on R_J is $R_J \geq 0$, meaning that the encoder local randomness rate can be arbitrarily small. We may even move the encoder local randomness to the common randomness if we do not want any encoder local randomness.

Proof sketch. We will only show the achievability here. We adopt the strategy in (Cuff, 2013). Suppose we want to construct a scheme with rates R, R_0, R_J, R_K satisfying $R > I(X;U)$, $R_0 + R > I(X,Y;U)$, $R_J > 0$, $R_K > H(Y|U)$. It suffices to consider $R = I(X;U) + \epsilon$ for $0 < \epsilon < 1$, since if we have a communication rate greater than $I(X;U) + \epsilon$, we can use the excess communication rate to create common randomness instead, so there is no gain of generality by considering an R greater than $I(X;U) + \epsilon$. We use the same construction as in the proof of Theorem 45. Recall that $\mathfrak{U} = (u^n(m,w))_{m\in[\lfloor 2^{nR}\rfloor], w\in[\lfloor 2^{nR_0}\rfloor]}$ is a random codebook with i.i.d. entries following P_U, $\bar{M} \sim \mathrm{Unif}([\lfloor 2^{nR}\rfloor])$ is independent of $W \sim \mathrm{Unif}([\lfloor 2^{nR_0}\rfloor])$, $\bar{U}^n := u^n(\bar{M}, W)$, and $(\bar{X}^n, \bar{Y}^n)|\bar{U}^n \sim P_{X|U}^n P_{Y|U}^n$. First, we show that the encoding Markov kernel $P_{\bar{M}|\bar{X}^n,W,\mathfrak{U}}(\cdot|\cdot,\cdot,\mathfrak{U})$ can be made into a deterministic function with an additional input that is the local randomness $J_n \sim \mathrm{Unif}([2^{nR_J}])$ for arbitrarily small R_J. We have

$$\begin{aligned}
&H(\bar{M}|\bar{X}^n, W, \mathfrak{U}) \\
&= H(\bar{M}|W,\mathfrak{U}) - I(\bar{M};\bar{X}^n|W,\mathfrak{U}) \\
&\overset{(a)}{=} H(\bar{M}|W,\mathfrak{U}) - I(\bar{U}^n;\bar{X}^n|W,\mathfrak{U}) \\
&\overset{(b)}{\leq} H(\bar{M}|W,\mathfrak{U}) - I(\bar{U}^n,\mathfrak{U};\bar{X}^n|W) + \epsilon n \\
&\leq nR - I(\bar{U}^n;\bar{X}^n|W) + \epsilon n \\
&\overset{(c)}{=} n(R - I(U;X)) + \epsilon n, \\
&= 2\epsilon n,
\end{aligned}$$

where (a) is because $(\bar{M}, W, \mathfrak{U}) \leftrightarrow \bar{U}^n \leftrightarrow \bar{X}^n$, (c) is because the unconditional distribution of $\bar{U}^n$ is P_U^n (since each $u^n(m,w) \sim P_U^n$), independent of W. For (b), since $\mathbb{E}[\delta_{\mathrm{TV}}(\bar{X}^n, P_X^n \,|\, W, \mathfrak{U})] \to 0$ (7.3), by the coupling lemma (Proposition 34), we can have $X^n \sim P_X^n$ independent of $W, \mathfrak{U}$, with $\mathbb{P}(X^n \neq \bar{X}^n) \to 0$, and

$$I(\mathfrak{U};\bar{X}^n|W) \leq I(\mathfrak{U};X^n|W) + H(\bar{X}^n|X^n) \leq \epsilon n,$$

for large n due to Fano's inequality (Fano, 1961). Therefore, the encoder can use the one-shot variable-length local channel simulation scheme (known source before coins) in Corollary 63 to simulate the Markov kernel $P_{\bar{M}|\bar{X}^n,W,\mathfrak{U}}(\cdot|\cdot,\cdot,\mathfrak{U})$ using an expected $\leq n(R - I(U;X)) + o(n)$ number of coin flips.[7] By Markov's inequality, the probability that the encoder uses more

[7] Technically the input to the local channel simulation scheme should be X^n instead of $\bar{X}^n$, though this does not matter since $\mathbb{P}(X^n \neq \bar{X}^n) \to 0$.

than $2\sqrt{\epsilon}n$ coin flips is at most $\sqrt{\epsilon}$. Therefore, to turn the variable-length local randomness into fixed-length, we can cap the number of coin flips at $2\sqrt{\epsilon}n$, with an encoder local randomness rate $R_J = 2\sqrt{\epsilon}$, incurring an additive penalty $\sqrt{\epsilon}$ to the total variation distance. We obtain the desired result by taking $\epsilon \to 0$.

It remains to show that the decoding Markov kernel $P_{\bar{Y}^n|\bar{M},W,\mathfrak{U}}(\cdot|\cdot,\cdot,\mathfrak{U})$ (obtained from passing $u^n(m,w)$ through the memoryless channel $P_{Y|U}^n$) can be made into a deterministic function with an additional input that is the local randomness $K_n \sim \mathrm{Unif}([2^{nR_K}])$ as long as $R_K > H(Y|U)$. This follows directly from the asymptotic local channel simulation result in Theorem 67.

□

Also refer to (Hamdi *et al.*, 2024) for a related setting about rate-distortion-perception tradeoff, where encoder local randomness is not helpful if the compression rate is less than the entropy of the source.

On the other hand, works on one-shot channel simulation with limited common and local randomness appear to be limited. Theorem 4 and the cardinality bound in Theorem 13 implies that one-shot channel simulation is possible with $\le I(X;Y) + \log_2(I(X;Y)+2) + 3$ bits of communication and $\le \log_2(|\mathcal{X}|(|\mathcal{Y}|-1)+2)+2$ bits of common randomness, without any local randomness. Theorem 65 implies that one-shot channel simulation is possible with $\le H(X)+1$ bits of communication and $H(\bigwedge_x P_{Y|X}(\cdot|x)) + \log_2(4e)$ bits of local randomness at the decoder, without any common randomness (encoder simply transmits X and let the decoder perform local channel simulation). It also implies that one-shot channel simulation is possible with $\le H(Y)+1$ bits of communication and $H(\bigwedge_x P_{Y|X}(\cdot|x)) + \log_2(4e)$ bits of local randomness at the encoder, without any common randomness (encoder performs local channel simulation and transmits Y). Characterizing the tradeoff between communication, common and local randomness would be an interesting future direction.

10 Other Settings

10.1 Simulating a Channel with Feedback

For a channel $X \to Y$ from an encoder to a decoder, (perfect) feedback refers to the communication of the channel output Y from the decoder back to the encoder. To simulate a channel with feedback, in addition to allowing the decoder to output Y, we must also allow the encoder to know Y (Bennett *et al.*, 2014). This should be done without an actual feedback channel from the decoder to the encoder. For example, to add the feedback requirement to the one-shot unlimited-common-randomness setting (Definition 2), the code would consist of an additional function $g' : \mathcal{W} \times \mathcal{X} \times \{0,1\}^* \to \mathcal{Y}$ representing the encoder's output for Y, such that $g'(W, X, M) = g(W, M) = Y$ with probability 1. Letting the encoder know the output Y is useful, for example, if we are designing a component in a large stateful lossy compression algorithm (e.g. lossy compression with diffusion generative models in (Theis *et al.*, 2022)), where the decoder's reconstruction of the next piece of information depends on its reconstruction of the previous pieces, and hence the encoder should also know the decoder's reconstruction of the previous pieces, in order to be "synchronized" with the decoder and properly compress the next piece.

Conventional deterministic lossy compression schemes always satisfy the feedback requirement, since there is no randomness in the encoder and the decoder, and the information available at the encoder (source and description) is a superset of the information at the decoder (description only), implying that any information that can be deduced by the decoder can also be deduced by the encoder. For a similar reason, channel simulation schemes with unlimited common randomness (Definition 2) always satisfy the feedback requirement (or can easily be modified to satisfy the requirement). Due to the unlimited common randomness, we can assume that the encoding and decoding functions are deterministic, and the information at the encoder (W, X, M) is a superset of the information at the decoder (W, M).

However, the feedback requirement is not automatically satisfied when the common randomness is limited, where it may be reasonable to have a stochastic decoding function (e.g., Section 4), which has an output that is only known to the decoder. If the encoder must know the output of the decoder, then the decoder should not use a stochastic decoding function since the decoder's local randomness is unknown to the encoder. Therefore, the amount of common randomness may have to be increased so that the decoder can use it in place of local randomness.

The asymptotic optimal rate region with the feedback requirement has been characterized in (Bennett *et al.*, 2014).

Theorem 69 (D/∞/A/FL/KS/LCR/Feedback (Bennett *et al.*, 2014))**.** *For the asymptotic*

approximate fixed-length channel simulation setting (Definition 37) with known source distribution, limited common randomness and discrete X, Y, with the feedback requirement (i.e., the encoder needs to know the output $\tilde{Y}^n$ of the decoder),[1] *the optimal rate region is given by*

$$\left\{(R, R_0) \in \mathbb{R}^2 : \begin{array}{l} R \geq I(X;Y), \\ R_0 + R \geq H(Y) \end{array}\right\}. \tag{10.1}$$

Proof. For the achievability, we invoke the same arguments as in the proof of Theorem 45 applied on $U = Y$, giving $I(X;U) = I(X;Y)$ and $I(X,Y;U) = H(Y)$. The decoding Markov kernel $P_{\bar{Y}^n|\bar{M},W,\mathfrak{U}}(\cdot|\cdot,\cdot,\mathfrak{U})$ is a deterministic function since $\bar{Y}^n = \bar{U}^n = u^n(\bar{M}, W)$. Therefore, the encoder can use this decoding function on (M, W) to know the output of the decoder.

For the converse, let $\epsilon \leq 1/4$, and apply the same arguments as in the proof of Theorem 45 to define $Q \sim \mathrm{Unif}([n])$, Y^n with $Y^n|X^n \sim P_{Y|X}^n$ and $\mathbb{P}(\tilde{Y}^n \neq Y^n) \leq \epsilon$, and $U := (M, W, Q)$. We have

$$\begin{aligned} R &\geq I(X_Q; U), \\ R_0 + R &\geq I(X_Q, Y_Q; U). \end{aligned} \tag{10.2}$$

Let $\hat{Y}^n$ be the output of the encoder, with average symbol error probability $\mathbb{P}(\hat{Y}_Q \neq \tilde{Y}_Q) = n^{-1}\sum_{i=1}^n \mathbb{P}(\hat{Y}_i \neq \tilde{Y}_i) \leq \epsilon$. By Fano's inequality (Fano, 1961),

$$H(\hat{Y}_Q|\tilde{Y}_Q) \leq H_b(\epsilon) + \epsilon \log_2 |\mathcal{Y}|.$$

Also, since $\mathbb{P}(\hat{Y}_Q \neq Y_Q) \leq \mathbb{P}(\hat{Y}_Q \neq \tilde{Y}_Q) + \mathbb{P}(\tilde{Y}_Q \neq Y_Q) \leq 2\epsilon$,

$$H(Y_Q|\hat{Y}_Q) \leq H_b(2\epsilon) + 2\epsilon \log_2 |\mathcal{Y}|.$$

Hence,

$$\begin{aligned} H(Y_Q|U) &\leq H(Y_Q|\hat{Y}_Q) + H(\hat{Y}_Q|U) \\ &\overset{(a)}{\leq} H(Y_Q|\hat{Y}_Q) + H(\hat{Y}_Q|\tilde{Y}_Q) \\ &\leq 2H_b(2\epsilon) + 3\epsilon \log_2 |\mathcal{Y}|, \end{aligned}$$

where (a) is because $\hat{Y}_Q \leftrightarrow U \leftrightarrow \tilde{Y}_Q$ since the decoder outputs $\tilde{Y}_Q$ using only $U = (M, W, Q)$. Combining this with (10.2),

$$R \geq I(X_Q; Y_Q) - (2H_b(2\epsilon) + 3\epsilon \log_2 |\mathcal{Y}|),$$

[1] We can either require that the encoder know $\tilde{Y}^n$ exactly, or that the encoder outputs $\hat{Y}^n$ with vanishing block error probability $\mathbb{P}(\hat{Y}^n \neq \tilde{Y}^n) \to 0$ as $n \to \infty$, or that the encoder outputs $\hat{Y}^n$ with vanishing average symbol error probability $n^{-1}\sum_{i=1}^n \mathbb{P}(\hat{Y}_i \neq \tilde{Y}_i) \to 0$ as $n \to \infty$. All three options give the same optimal rate region.

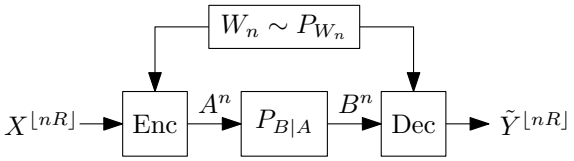


Figure 10.1: Asymptotic simulation of a channel using another channel.

$$R_0 + R \geq I(X_Q, Y_Q; Y_Q) - (2H_b(2\epsilon) + 3\epsilon \log_2 |\mathcal{Y}|).$$

The result follows from letting $\epsilon \to 0$. □

10.2 Simulating a Channel using Another Channel

In all of the previous discussions, the encoder and the decoder are connected by a noiseless link, where the description M can be sent without error. What if the encoder can only communicate to the decoder via another noisy link $P_{B|A}$? In this section, we will study the problem of simulating a noisy channel $P_{Y|X}$ using another noisy channel $P_{B|A}$.

Definition 70 (Asymptotic simulation of a channel using another channel)**.** Consider two general channels $P_{B|A}$, $P_{Y|X}$ and a source distribution P_X. An channel simulation scheme from $P_{B|A}$ to $P_{Y|X}$ at a rate $R > 0$ with common randomness rate $R_0 \in [0, \infty]$ is characterized by a sequence of tuples $(P_{W_n}, P_{A^n|W_n, X^{\lfloor nR \rfloor}}, P_{\tilde{Y}^{\lfloor nR \rfloor}|W_n, B^n})_{n \in \mathbb{N}^+}$ described below:

- **Common randomness.** There is a common random source $W_n \in \mathcal{W}_n$, $W_n \sim P_{W_n}$ available to the encoder and the decoder. If $R_0 = \infty$ (unlimited common randomness) we can choose an arbitrary distribution P_{W_n} as a part of the coding scheme. If $R_0 \neq \infty$, P_{W_n} is fixed to be $\text{Unif}([\lfloor 2^{nR_0} \rfloor])$.

- **Encoder.** Let $m := \lfloor nR \rfloor$. The encoder observes W_n and a source sequence $X^m \sim P_X^m$ (for the arbitrary source case, we can have any $X^m \in \mathcal{X}^m$), and sends $A^n|(W_n, X^m) \sim P_{A^n|W_n, X^m}$ produced by passing W_n, X^m through an encoding Markov kernel $P_{A^n|W_n, X^m}$ from $\mathcal{W}_n \times \mathcal{X}^m$ to $\mathcal{A}^n$.

- **Decoder.** The decoder observes the channel output sequence $B^n|A^n \sim P_{B|A}^n$, and then outputs $\tilde{Y}^m|(W_n, B^n) \sim P_{\tilde{Y}^m|W_n, B^n}$ produced by passing W_n, B^n through a decoding Markov kernel $P_{\tilde{Y}^m|W_n, B^n}$ from $\mathcal{W}_n \times \mathcal{B}^n$ to $\mathcal{Y}^m$.

- **Requirement.**
  - For the known source distribution case, we require that $\tilde{Y}^m$ follows the conditional distribution $P_{Y|X}^m$ approximately, in the sense that
$$\delta_{\mathrm{TV}}\big((X^m, \tilde{Y}^m),\, P_X^m P_{Y|X}^m\big) \to 0 \tag{10.3}$$
as $n \to \infty$, where $m = \lfloor nR \rfloor$.
  - For the arbitrary source case, we require
$$\sup_{x^m \in \mathcal{X}^m} \delta_{\mathrm{TV}}\big(\tilde{Y}^m,\, P_{Y|X}^m \,\big|\, X^m = x^m\big) \to 0 \tag{10.4}$$
as $n \to \infty$.
- **Performance metrics.** We say that the rate pair (R, R_0) is achievable if there exists a scheme at rate R with common randomness rate R_0 satisfying the above requirement. For $R_0 \in [0, \infty]$, write $R^*(R_0)$ for the supremum of R such that (R, R_0) is achievable.

While channel simulation can be regarded as "lossy source coding with a requirement on the conditional distribution of Y^n given X^n", Definition 70 can be regarded as "joint source channel coding with a requirement on the conditional distribution of Y^n given X^n".

The reverse Shannon theorem (Bennett *et al.*, 2002; Bennett *et al.*, 2014) shows that, under the presence of unlimited common randomness, one can asymptotically simulate a channel using another channel at a rate given by the ratio between the two capacities.

Corollary 71 (Bennett *et al.* 2002; Bennett *et al.* 2014)**.** *Consider two channels $P_{B|A}$, $P_{Y|X}$ where A, B, X, Y are discrete and finite. Assume unlimited common randomness $R_0 = \infty$. We have:*

- *(Known source distribution)*
$$R^*(\infty) = \frac{\max_{P_A} I(A;B)}{I(X;Y)}.$$
- *(Arbitrary source)*
$$R^*(\infty) = \frac{\max_{P_A} I(A;B)}{\max_{P_X} I(X;Y)}.$$

Proof. The intuition is that, in the presence of unlimited common randomness, n copies of the channel $P_{B|A}$ is approximately equivalent to $n \max_{P_A} I(A;B)$ noiseless bits due to the channel coding theorem (for converting $P_{B|A}$ to noiseless bits) and the channel simulation result in Theorem 38 (for converting noiseless bits to $P_{B|A}$). If we replace the n copies of $P_{B|A}$ by $n \max_{P_A} I(A;B)$ bits, then Definition 70 reduces to the original asymptotic channel simulation setting in Definition 37, and we can invoke Theorem 38 to obtain the optimal rate. We now present the proof.

For the achievability part and the known source distribution case, fix any $R < (\max_{P_A} I(A;B))/I(X;Y)$. Assume $\epsilon > 0$ satisfies $R + \epsilon < (\max_{P_A} I(A;B))/I(X;Y)$. Applying the channel coding theorem, we can transmit a message $M \in [\lfloor 2^{n(R+\epsilon)I(X;Y)} \rfloor]$ through n uses of the channel $P_{B|A}$ asymptotically. The result follows from applying Theorem 38 to simulate $\lfloor nR \rfloor$ copies of $P_{Y|X}$ using the description M. The arbitrary source case is similar.

For the converse part and the known source distribution case, assume the contrary that it is possible to simulate $P_{Y|X}$ from $P_{B|A}$ at a rate $R > (\max_{P_A} I(A;B))/I(X;Y)$ with common randomness W_n. We call this Scheme 1. Assume $\epsilon > 0$ satisfies $R > (\max_{P_A} I(A;B) + \epsilon)/I(X;Y)$. Fix a channel simulation scheme for the channel $P_{B|A}$ with arbitrary source and unlimited common randomness W_n' at a description rate $\max_{P_A} I(A;B) + \epsilon$ (Definition 37), which we call Scheme 2. We now construct a channel simulation scheme for $P_{Y|X}$ with common randomness (W_n, W_n'), which we call Scheme 3. In Scheme 3, the encoder observes X^m, W_n, W_n', applies Scheme 1 using W_n to produce A^n, and then applies Scheme 2 using W_n' to produce a description $M \in [\lfloor 2^{n(\max_{P_A} I(A;B)+\epsilon)} \rfloor]$. The decoder observes M, W_n, W_n', applies Scheme 2 to produce B^n, and then applies Scheme 1 to produce $\tilde{Y}^m$. Scheme 3 uses a description rate $(\max_{P_A} I(A;B) + \epsilon)/R < I(X;Y)$ bits per channel simulated, which is impossible due to Theorem 38. The arbitrary source case follows from applying the known source distribution case on the P_X achieving $\max_{P_X} I(X;Y)$. □

The problem becomes harder when R_0 is finite. We state an achievability result in (Haddadpour *et al.*, 2016). Refer to (Haddadpour *et al.*, 2016) for the proof.

Theorem 72 (Haddadpour *et al.* 2016)**.** *Consider two channels $P_{B|A}$, $P_{Y|X}$ where A, B, X, Y are discrete and finite. Then $P_{Y|X}$ can be asymptotically simulated from $P_{B|A}$ at rate $R = 1$ with common randomness rate R_0 (i.e., the rate pair $(1, R_0)$ is achievable) and known source distribution P_X if there exists random variables X, Y, A, B, U satisfying:*

$$(X,Y) \sim P_X P_{Y|X},$$

$$B|A \sim P_{B|A},$$

$$(X,U) \leftrightarrow A \leftrightarrow B,$$

$$(X,A) \leftrightarrow (B,U) \leftrightarrow Y,$$

$$R_0 + I(U;B) > I(U;X,Y),$$
$$I(U;B) > I(U;X).$$

We may also impose an exact condition on the distribution of Y^m, i.e., the TV distances in (10.3), (10.4) are zero. The case where the channels are binary-input binary-output channels (i.e., A, B, X, Y are binary), $R = 1$ with unlimited common randomness has been fully characterized in (Haddadpour *et al.*, 2016). Readers are referred to (Cubitt *et al.*, 2011; Haddadpour *et al.*, 2016) for further discussions. A setting where local randomness is also limited has been investigated in (Obead *et al.*, 2021), where schemes which utilize the randomness in the channel $P_{B|A}$ to reduce local randomness were proposed. A setting where the joint distribution of $A^n, B^n, X^m, \tilde{Y}^m$ is controlled (not only the joint distribution of $X^m, \tilde{Y}^m$) has been studied in (Cervia *et al.*, 2020).

10.3 Interactive Channel Simulation

In the interactive channel simulation setting (Gohari and Anantharam, 2011; Yassaee *et al.*, 2015), the two terminals have A^n and B^n respectively where $(A_i, B_i) \overset{iid}{\sim} P_{A,B}$, and share the common randomness $W \sim \mathrm{Unif}([\lfloor 2^{nR_0} \rfloor])$. The interactive communication consists of r rounds. In the j-th round where j is an odd number, Terminal 1 generates M_j by applying a stochastic mapping on W, A^n, M^{j-1}, and sends it to Terminal 2. In the j-th round where j is an even number, Terminal 2 generates M_j by applying a stochastic mapping on W, B^n, M^{j-1}, and sends it to Terminal 1. The communication rate constraints are

$$\frac{1}{n} \sum_{j \text{ odd}} H(M_j) \le R_{12}, \ \frac{1}{n} \sum_{j \text{ even}} H(M_j) \le R_{21},$$

where R_{12} and R_{21} are the communication rate from Terminal 1 to Terminal 2 and that from Terminal 2 to Terminal 1 respectively. After the r rounds interactive communication, Terminal 1 outputs $\tilde{X}^n$ by applying a stochastic mapping on W, A^n, M^r, and Terminal 2 outputs $\tilde{Y}^n$ by applying a stochastic mapping on W, B^n, M^r. We require that $\tilde{X}^n$ and $\tilde{Y}^n$ approximately follow a prescribed conditional distribution $P_{X,Y|A,B}$. More precisely, we require

$$\delta_{\mathrm{TV}}\Big((A^n, B^n, \tilde{X}^n, \tilde{Y}^n),\, P_{A,B}^n P_{X,Y|A,B}^n\Big) \to 0$$

as $n \to \infty$. The optimal rate region is the closure of the set of tuples (R_{12}, R_{21}, R_0) such that there exists a valid channel simulation scheme. Note that this reduces to the setting in Definition 37 when $r = 1$ and $B = X = \emptyset$. When X, Y are functions of A, B, this becomes the interactive function computation problem (Yao, 1979; Ma and Ishwar, 2011), which has been briefly discussed in Section 1.9. This setting is also related to the information of

formation (Renner and Wolf, 2003), which concerns the approximate distributed simulation of X^n, Y^n through interactive communication, under a constraint that the interactive communication cannot reveal more information than a sequence Z^n jointly distributed with X^n, Y^n.

The optimal rate region is characterized in (Yassaee *et al.*, 2015). Please refer to (Yassaee *et al.*, 2015) for the proof.

Theorem 73 (Yassaee *et al.* 2015)**.** *The optimal rate region of the interactive channel simulation setting is*

$$\bigcup_{U^r} \left\{ \begin{array}{l} (R_{12}, R_{21}, R_0) \in \mathbb{R}^2 : \\ R_{12} \geq I(A; U^r|B), \\ R_{21} \geq I(B; U^r|A), \\ R_0 + R_{12} \geq I(A; U^r|B) + I(U_1; X, Y|A, B), \\ R_0 + R_{12} + R_{21} \\ \quad \geq I(A; U^r|B) + I(B; U^r|A) + I(U^r; X, Y|A, B) \end{array} \right\},$$

where the union is over $P_{U^r|A,B,X,Y}$ satisfying that

$$\begin{gathered} (A, B, X, Y) \sim P_{A,B} P_{X,Y|A,B}, \\ B \leftrightarrow (A, U^{j-1}) \leftrightarrow U_j \ \ \textit{for odd } j, \\ A \leftrightarrow (B, U^{j-1}) \leftrightarrow U_j \ \ \textit{for even } j, \\ (B, Y) \leftrightarrow (A, U^r) \leftrightarrow X, \\ (A, X) \leftrightarrow (B, U^r) \leftrightarrow Y. \end{gathered}$$

10.4 Secure Channel Simulation

A channel simulation setting with a secrecy constraint was studied in (Cuff, 2013), where we impose an additional constraint in Definition 37 where the communication M_n occurs over a public channel, and we require that $(X^n, \tilde{Y}^n)$ is approximately independent of M_n, so an eavesdropper observing M_n cannot learn about $X^n, \tilde{Y}^n$. Note that the common randomness W_n is assumed to be secret. More specifically, we need

$$\delta_{\text{TV}}\big((X^n, \tilde{Y}^n, M_n),\ P_X^n P_{Y|X}^n P_{M_n}\big) \to 0 \tag{10.5}$$

as $n \to \infty$. The distribution on the left is the actual joint distribution of $(X^n, \tilde{Y}^n, M_n)$, and the distribution on the right is the ideal joint distribution where the channel simulation is exact, and M_n is independent of $(X^n, \tilde{Y}^n)$.

The optimal rate region is given in (Cuff, 2013).

Theorem 74 (D/∞/A/FL/KS/LCR/Secure (Cuff, 2013))**.** *For the secure asymptotic approximate fixed-length channel simulation setting (Definition 37) with known source distribution, limited common randomness, the secrecy constraint in (10.5), and finite discrete X, Y, the optimal rate region is given by*

$$\bigcup_{P_{U|X,Y}\colon X\leftrightarrow U\leftrightarrow Y} \left\{ (R, R_0) \in \mathbb{R}^2 : \begin{array}{l} R \geq I(X;U), \\ R_0 \geq I(X,Y;U) \end{array} \right\}. \tag{10.6}$$

Proof sketch. We will only prove the achievability. The construction in (Cuff, 2013) is to invoke Theorem 74 to construct a scheme with $R = I(X;U) + \epsilon$ and $R_0 = I(Y;U|X) + \epsilon$ without the secrecy constraint (let its communication and common randomness be $M_n \in [\lfloor 2^{nR} \rfloor]$ and $W_n \in [\lfloor 2^{nR_0} \rfloor]$ respectively), and then use an additional common randomness rate of R to apply the one-time pad (Shannon, 1949) on M_n. More specifically, we use an additional common randomness $V_n \sim \mathrm{Unif}([\lfloor 2^{nR} \rfloor])$, and the encoder transmits $M_n' := (M_n + V_n - 1 \bmod \lfloor 2^{nR} \rfloor) + 1$ instead. It is straightforward to check that M_n' is independent of M_n, and hence is independent of $(X^n, \tilde{Y}^n)$. The total common randomness rate is $I(Y;U|X) + \epsilon + R = I(X,Y;U) + 2\epsilon$. □

An interactive channel simulation setting (see Section 10.3) with a secrecy constraint was investigated in (Gohari *et al.*, 2012), where the optimal rate region was given.

10.5 Channel Simulation over Networks

This monograph focuses on "point-to-point" channel simulation where there is one encoder and one decoder. Channel simulation over networks with three or more terminals has also been studied in the literature.

Cascade networks. The extension of the channel simulation setting to cascade networks has been investigated by Bloch and Kliewer (2013), Satpathy and Cuff (2016), and Vellambi *et al.* (2017). Consider three nodes sharing nR_0 bits of common randomness. Node 1 observes $X^n \sim P_X^n$ and transmits nR_1 bits to Node 2. Node 2 then outputs Y^n and transmits nR_2 bits to Node 3. Finally, Node 3 outputs Z^n. We require (X^n, Y^n, Z^n) to follow a prescribed i.i.d. distribution $P_{X,Y,Z}^n$ approximately, in the sense that the total variation distance between $P_{X,Y,Z}^n$ and the distribution of (X^n, Y^n, Z^n) approaches 0 as $n \to \infty$. In (Satpathy and Cuff, 2016), a security constraint is also imposed, where we require the two messages to be approximately independent of (X^n, Y^n, Z^n).

Multiple encoders. The extension of the channel simulation setting to multiple encoders (i.e., the simulation of a multiple access channel) has been studied by Kurri *et al.* (2022), Atif *et al.* (2022), and Atif *et al.* (2021). Encoder 1 and Encoder 2 observe X_1^n and X_2^n respectively, where $(X_{1,i}, X_{2,i}) \overset{iid}{\sim} P_{X_1,X_2}$ for $i \in [n]$. Encoder 1 can transmit nR_1 bits to the decoder, and Encoder 2 can transmit nR_2 bits to the decoder. The decoder then outputs Y^n. We require (X_1^n, X_2^n, Y^n) to follow a prescribed i.i.d. distribution $P_{X_1,X_2,Y}^n$ approximately, in the sense that the total variation distance between $P_{X_1,X_2,Y}^n$ and the distribution of (X_1^n, X_2^n, Y^n) approaches 0 as $n \to \infty$. A one-shot channel simulation result was given by Nema *et al.* (2024). An extension of this setting with a secrecy constraint was studied by Ramachandran *et al.* (2024).

Multiple decoders. The extension of the channel simulation setting to multiple decoders (i.e., the simulation of a broadcast channel) has been studied by Cuff (2013), Haddadpour *et al.* (2016), Cao *et al.* (2023), Cao *et al.* (2022a), and Managoli and Prabhakaran (2024), where the encoder observes X and sends a message to each of the two decoders. Decoder 1 and Decoder 2 outputs Y and Z respectively. The goal is to have (Y, Z) follow a prescribed conditional distribution given X approximately.

General networks. The most general setting would be to have a general network of nodes connected by noiseless and noisy links. Coordination problems over networks has been studied in (Cuff *et al.*, 2010). In (Lee and Chung, 2015; Lee and Chung, 2018), an achievability result for general network with an empirical coordination constraint is studied. It can be checked that (Lee and Chung, 2015; Lee and Chung, 2018) also gives a channel simulation result if the nodes share unlimited common randomness. A one-shot result for channel simulation over a general network with unlimited common randomness is given in (Liu and Li, 2024). An automated theorem proving framework, that can automatically compute inner and outer bounds of the optimal rate region of a general network for channel simulation with unlimited common randomness (as well as source and channel coding problems), was given in (Li, 2023). The framework is implemented in a computer program called Python Symbolic Information Theoretic Inequality Prover (Li, 2020).

10.6 Single-Input Multiple-Output Channel Simulation

Multiple-output channel simulation (Choi and Li, 2021) concerns the setting where the encoder observes a single input symbol X, and sends a prefix-free codeword to the decoder, which will output $Y_1, \ldots, Y_n$ which are conditionally i.i.d. following a prescribed conditional distribution $P_{Y|X}$ given X. An equivalent formulation of the setting is that, given a certain

parametric family of distributions $(Q_\theta)_{\theta\in\mathcal{A}}$, the encoder observes a parameter $\theta \in \mathcal{A}$, and sends a prefix-free codeword to the decoder, which will output $Y_1, \ldots, Y_n$ which are i.i.d. following Q_θ given θ. This is useful for conveying information about a distribution to the decoder, where the decoder does not require the exact analytical formula of the distribution, but only requires samples following that distribution, so that statistical inference for the distribution can be conducted on those samples. Interested readers are referred to (Choi and Li, 2021) for schemes based on dyadic decomposition (Li and El Gamal, 2017; Li and El Gamal, 2018a), and to (Kobus *et al.*, 2024b) for a sampling scheme based on iteratively updating the reference distribution given the previous samples.

		Unlimited common randomness	No common randomness	Limited common randomness
One-shot	Exact VL	$L^* \le I(X;Y)+$ $\log_2(I(X;Y)+2)+3$ Sec. 3, Thm. 4	$H^* = G(X;Y)$ $:= \min_{X \leftrightarrow W \leftrightarrow Y} H(W)$ Sec. 4, Prop. 27	See Sec. 9.4
	Exact FL	$\mathsf{N}^* = \min\left\{k : \mathbf{P}_{Y\|X} \in \operatorname{conv}(\{\mathbf{Q}_{Y\|X} : \|\mathbf{1}^T\mathbf{Q}_{Y\|X}\|_0 \le k\})\right\}$ Sec. 3.7, Thm. 24	$\mathsf{N}^* = \operatorname{rank}_+(\mathbf{P}_{Y\|X})$ Sec. 4, Prop. 26	
	Approx. FL	Sec. 8, Thm. 50	Sec. 8, Thm. 51	Sec. 8, Thm. 53
Asymptotic	Exact VL	$R^* = I(X;Y)$ Sec. 5, Thm. 32	$R^* = \overline{G}(X;Y)$ $:= \lim_{n\to\infty} \frac{1}{n} G(X^n;Y^n)$ Sec. 6, Prop. 44	$R^* \le \min_{X \leftrightarrow U \leftrightarrow Y} \max\{I(X;U), H(U) - R_0\}$ Sec. 7, Thm. 47
	Approx. FL	$R^* = I(X;Y)$ Sec. 5, Thm. 38	$R^* = J(X;Y)$ $:= \min_{X \leftrightarrow U \leftrightarrow Y} I(X,Y;U)$ Sec. 6, Thm. 40	$R^* = \min_{X \leftrightarrow U \leftrightarrow Y} \max\{I(X;U), I(X,Y;U) - R_0\}$ Sec. 7, Thm. 45

Table 11.1: Channel simulation results under different assumptions: whether we consider one-shot (1) or asymptotic (∞), whether we require the output Y to follow the desired conditional distribution exactly (E) or approximately (A), whether the description M is fixed-length (FL) or variable-length (VL), and whether we allow common randomness between the encoder and the decoder. We assume the source distribution P_X is known (KS) in this table.

11 Conclusions and Future Directions

We briefly summarize the results on channel simulation and related problems discussed in this monograph. An overview of the optimal description lengths or rates of various settings is given in Table 11.1.

One-shot exact variable-length channel simulation. Techniques applicable to the one-shot exact simulation of general channels are of practical interest since they do not require any assumption on the source and channel structure. They include greedy rejection sampling (Section 3.2.2), Poisson functional representation (Section 3.3) and other sampling-based methods (Section 3.5). They can achieve an expected description length close to the capacity of the channel simulated, within a logarithmic gap (Theorem 4 and Corollary 8). Dithering-based schemes (Section 3.6) are applicable to additive noise channels. Schemes based on dyadic decomposition (Section 4.2) have the advantage that they do not require common randomness.

One-shot approximate fixed-length channel simulation. Likelihood encoder and minimal random coding (Sections 3.4, 5.6 and 8.2) only requires a fixed-length description, though they can only simulate the channel approximately. They achieve a description length close to the KL divergence between the target distribution and the reference distribution (Theorem 15), similar to greedy rejection sampling and Poisson functional representation.

Asymptotic channel simulation. Letting the blocklength approach infinity and tolerating a vanishing total variation distance allows us to characterize the optimal description rate and common randomness rate precisely (Theorem 45), giving elegant results of theoretical interest. Most of the asymptotic approximate results in this monograph are proved using the soft covering lemma (Section 5.5). Asymptotic exact results can also be derived using various techniques (Sections 5.1, 5.2, 6.2 and 7.2).

Other results. Other notable results including the source and local channel simulation results (Section 9), the simulation of a channel with feedback (Section 10.1), the simulation of a channel using another channel (Section 10.2), interactive channel simulation (Section 10.3), secure channel simulation (Section 10.4), and various results on channel simulation over networks (Section 10.5).

We now discuss several future research directions.

Gaussian channel simulation. With applications to neural compression (Havasi *et al.*, 2019; Flamich *et al.*, 2022; He *et al.*, 2024a) and differential privacy (Hasırcıoğlu and Gündüz, 2024; Hegazy *et al.*, 2024; Yan *et al.*, 2023), the additive Gaussian noise channel appears to be one of the most popular channels to be simulated. Finding the right balance between short description length and algorithmic efficiency is therefore of practical interest. Apart from general-purpose channel simulation schemes such as minimal random coding, schemes based on vector quantization (Ling and Li, 2024; Kobus *et al.*, 2024a) are also promising.

Simulation of differential privacy mechanisms. The simulation of a differential privacy mechanism has an additional constraint that the scheme must be differentially private against the decoder. While approximate simulation schemes for general mechanisms have been studied in (Bassily and Smith, 2015; Bun *et al.*, 2019; Shah *et al.*, 2022), and an exact scheme has been studied in (Liu *et al.*, 2024), our understanding on the fundamental limits of privacy mechanism simulation is still rather limited. For example, it is unknown whether the two-fold increase of the privacy budget in (Shah *et al.*, 2022; Liu *et al.*, 2024) (Theorem 16 and Section 3.3.5) is fundamental to general simulation schemes with pure differential privacy that are exact or approximate within a small total variation distance.

One-shot variable-length channel simulation with limited common randomness. Readers may notice that this monograph discusses the "unlimited common randomness" and "no common randomness" cases of one-shot variable-length channel simulation, but not the "limited common randomness" case (except the brief mention at the end of Section 9.4). It would be interesting to investigate whether we can prove a one-shot variable-length version of the trade-off region between description length and common randomness in Theorem 45 and Theorem 47. For the common randomness, we can consider the case where the encoder and decoder generate the coin flips before observing the source and the description (similar to the coins before source setting in Section 9.3.1), and the case where they can generate the coin flips after observing the source and the description, but must keep the two coin flip sequences synchronized (similar to the source before coins setting in Section 9.3.1).

We would also like to mention a recent work (Sriramu *et al.*, 2024) on an efficient algorithm for channel simulation via polar codes (Arikan, 2009). The use of structured codes for channel simulation is a promising future direction.

Acknowledgements

The author was supported in part by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No.s: CUHK 24205621 (ECS), CUHK 14209823 (GRF)]. The author would like to thank Prof. Aaron Wagner, Prof. Alexander Barg, and the anonymous reviewers for their invaluable suggestions and advice, which greatly improved this monograph. The author would also like to thank Mike Casey and Mark de Jongh from Now Publishers for their help throughout the publication process, and Prof. Amin Aminzadeh Gohari, Prof. Ayfer Özgür, Dr. Lucas Theis, Chih Wei Ling, Yanxiao Liu and Wei-Ning Chen for the insightful discussions.

A Zipf Distribution

The *Zipf distribution* (Ross, 2019) (also known as *zeta distribution*) with parameter $s > 1$ is a distribution over $\mathbb{N}^+$ with probability mass function

$$\text{Zipf}(k; s) := \frac{k^{-s}}{\zeta(s)},$$

where $\zeta(s) = \sum_{k=1}^{\infty} k^{-s}$ is the Riemann zeta function. It is the maximum entropy distribution for $K \in \mathbb{N}^+$ when $\mathbb{E}[\log_2 K]$ is fixed. We can use the Zipf distribution to show the following bound (e.g., see (Li and El Gamal, 2018b)).

Proposition 75. *For random variable $K \in \mathbb{N}^+$ following the distribution P_K, its cross entropy with* $\text{Zipf}(s)$ *is bounded by*

$$H(P_K, \text{Zipf}(s)) \le s\mathbb{E}[\log_2 K] + \log_2 \frac{s}{s-1}. \tag{A.1}$$

Therefore, if $\mathbb{E}[\log_2 K] \le \ell$, letting $s = 1 + 1/\ell$, we have

$$H(K) \le H(P_K, \text{Zipf}(s)) \le \ell + \log_2(\ell + 1) + 1.$$

Proof. We have

$$\begin{aligned} H(P_K, \text{Zipf}(s)) &= \sum_{k=1}^{\infty} P_K(k) \log_2 \frac{\zeta(s)}{k^{-s}} \\ &= s\mathbb{E}[\log_2 K] + \log_2 \zeta(s), \end{aligned}$$

where

$$\zeta(s) \le 1 + \int_1^{\infty} \kappa^{-s} \mathrm{d}\kappa = \frac{s}{s-1}. \tag{A.2}$$

The result follows. □

Proposition 75 suggests that, if we know that $\mathbb{E}[\log_2 K] \le \ell$, then we can use the Shannon code (Shannon, 1948) designed for the distribution $\text{Zipf}(s)$ where $s = 1 + 1/\ell$ to encode K, to obtain a codeword with expected length upper-bounded by

$$H(P_K, \text{Zipf}(s)) + 1 \le \ell + \log_2(\ell + 1) + 2 \text{ bits}.$$

Refer to Section 1.12. The downside is that we need to know ℓ when we construct the code, and the Shannon code over an infinite alphabet can be hard to construct.

In contrast, if we do not know the bound $\mathbb{E}[\log_2 K] \le \ell$ when we design the code, we can still use the Elias delta code (Elias, 1975) to encode K, which will result in a codeword length upper-bounded by

$$\ell + 2\log_2(\ell + 1) + 1 \text{ bits}$$

if $\mathbb{E}[\log_2 K] \le \ell$. While it is possible to improve this bound to $\ell+(1+\epsilon)\log_2(\ell+1)+O(1)$, for example, by using the Elias omega code (Elias, 1975), it is impossible to design a prefix-free code over $\mathbb{N}^+$ that achieves an expected length upper-bounded by $\ell + \log_2(\ell+1) + O(1)$ for every ℓ and random variable K with $\mathbb{E}[\log_2 K] \le \ell$.[1] Therefore, although using a "universal" code such as the Elias delta code has the advantage that we do not need to know the bound $\mathbb{E}[\log_2 K] \le \ell$ beforehand, it comes with a small penalty on the expected length.

Practically, if we are given the bound $\mathbb{E}[\log_2 K] \le \ell$, then there are several options for the encoding of $K \in \mathbb{N}^+$:

- Shannon code (Shannon, 1948) for the distribution $\mathrm{Zipf}(1+1/\ell)$, or any prefix-free code $f : \mathbb{N}^+ \to \{0,1\}^*$ with $|f(k)| \le \lceil \mathrm{Zipf}(k; 1+1/\ell) \rceil$ for $k \in \mathbb{N}^+$. The expected length is upper-bounded by $\ell+\log_2(\ell+1)+2$. Nevertheless, it can be hard to construct.

- A code over positive integers with efficient encoding and decoding algorithms such as the Elias delta code (Elias, 1975), with a slight penalty on the expected length. The advantage is that the code does not depend on ℓ.

- Use a "hybrid" approach: first construct the Shannon code $f_{\mathrm{S}} : [k_0+1] \to \{0,1\}^*$ for the distribution of $\tilde{K} := \min\{K, k_0+1\}$ where $K \sim \mathrm{Zipf}(1+1/\ell)$ and k_0 is a large fixed integer (but not too large so it is viable to construct the Shannon code), and then encode $k \in \mathbb{N}^+$ into $f_{\mathrm{S}}(k)$ if $k \le k_0$, or $f_{\mathrm{S}}(k_0+1) \| f_\delta(k-k_0)$ if $k > k_0$, where $f_\delta : \mathbb{N}^+ \to \{0,1\}^*$ is the Elias delta code, and "$\|$" stands for concatenation.

- A suitable comma code such as the Fibonacci code (Fraenkel and Kleinb, 1996) (which is optimal for a Zipf distribution with a certain parameter).

[1] This is because $\sum_{k=1}^{\infty} 2^{-\log_2 k - \log_2(\log_2 k+1) - c} = \sum_{k=1}^{\infty} \frac{1}{2^c k(\log_2 k+1)} = \infty$, violating Kraft's inequality (Kraft, 1949).

B Turning Approximate Markov Chains into Exact Markov Chains

The following lemma shows that if the Markov chain "$X \leftrightarrow U \leftrightarrow Y$" almost holds, that is, there exists random variables $\tilde{X}, \tilde{Y}$ with $\tilde{X} \leftrightarrow U \leftrightarrow \tilde{Y}$ and $\mathbb{P}((X,Y) \neq (\tilde{X},\tilde{Y})) \approx 0$, then there exists a random variable V with small entropy such that $X \leftrightarrow (U,V) \leftrightarrow Y$ holds exactly.

Lemma 76. *For finite discrete random variables $X, Y, \tilde{X}, \tilde{Y}, U$ ($X, \tilde{X} \in \mathcal{X}$ and $Y, \tilde{Y} \in \mathcal{Y}$) with $\tilde{X} \leftrightarrow U \leftrightarrow \tilde{Y}$, there exists a random variable $V \in \mathcal{V}$ with $X \leftrightarrow (U,V) \leftrightarrow Y$, $|\mathcal{V}| \leq \min\{|\mathcal{X}|, |\mathcal{Y}|\} + 1$, and*

$$H(V) \leq H_b(\min\{\eta, 1/2\}) + \eta \log_2 \min\{|\mathcal{X}|, |\mathcal{Y}|\},$$

where H_b is the binary entropy function, and

$$\eta := 2(|\mathcal{X}||\mathcal{Y}|)^{1/4} \sqrt{\mathbb{P}((X,Y) \neq (\tilde{X},\tilde{Y}))}.$$

Proof. We first prove the following claim, which basically states that if two random variables has a small TV distance from being independent, then they are conditionally independent given a random variable that is close to being degenerate:

For finite discrete random variables $X, Y, \tilde{X}, \tilde{Y}$ with $\tilde{X}$ independent of $\tilde{Y}$, there exists a random variable $V \in [0..\min\{|\mathcal{X}|, |\mathcal{Y}|\}]$ with $X \leftrightarrow V \leftrightarrow Y$ and

$$P_V(0) \geq 1 - 2(|\mathcal{X}||\mathcal{Y}|)^{1/4} \sqrt{\delta_{\mathrm{TV}}((X,Y),(\tilde{X},\tilde{Y}))}.$$

We now show this claim. Assume $\mathcal{X} = [|\mathcal{X}|]$ and $|\mathcal{X}| \leq |\mathcal{Y}|$. Applying the coupling lemma (Proposition 34), we can assume $\delta_{\mathrm{TV}}((X,Y),(\tilde{X},\tilde{Y})) = \mathbb{P}(E)$, where E is the event $(X,Y) \neq (\tilde{X},\tilde{Y})$. Define $V \in [0..|\mathcal{X}|]$ with

$$\begin{aligned} &P_{V,X,Y}(0,x,y) \\ &:= \left[P_{\tilde{X}}(x) - \left(\frac{|\mathcal{Y}|}{|\mathcal{X}|}\right)^{\frac{1}{4}} \sqrt{\mathbb{P}(E,\ \tilde{X} = x)}\right]_+ \left[P_{\tilde{Y}}(y) - \left(\frac{|\mathcal{X}|}{|\mathcal{Y}|}\right)^{\frac{1}{4}} \sqrt{\mathbb{P}(E,\ \tilde{Y} = y)}\right]_+, \end{aligned}$$

where $[t]_+ := \max\{t, 0\}$, $P_{V,X,Y}(x,x,y) := P_{X,Y}(x,y) - P_{V,X,Y}(0,x,y)$, and $P_{V,X,Y}(v,x,y) := 0$ for $v \neq x$. To check that this is a valid distribution,

$$\begin{aligned} &P_{V,X,Y}(0,x,y) \\ &= \left[P_{\tilde{X}}(x) - \left(\frac{|\mathcal{Y}|}{|\mathcal{X}|}\right)^{\frac{1}{4}} \sqrt{\mathbb{P}(E,\ \tilde{X} = x)}\right]_+ \left[P_{\tilde{Y}}(y) - \left(\frac{|\mathcal{X}|}{|\mathcal{Y}|}\right)^{\frac{1}{4}} \sqrt{\mathbb{P}(E,\ \tilde{Y} = y)}\right]_+ \end{aligned}$$

$$
\begin{aligned}
&\le \left[P_{\tilde{X}}(x) - \left(\frac{|\mathcal{Y}|}{|\mathcal{X}|}\right)^{\frac{1}{4}} \sqrt{\mathbb{P}(E,\, (\tilde{X},\tilde{Y}) = (x,y))}\right]_+ \\
&\quad \cdot \left[P_{\tilde{Y}}(y) - \left(\frac{|\mathcal{X}|}{|\mathcal{Y}|}\right)^{\frac{1}{4}} \sqrt{\mathbb{P}(E,\, (\tilde{X},\tilde{Y}) = (x,y))}\right]_+ \\
&\overset{(a)}{\le} P_{\tilde{X}}(x) P_{\tilde{Y}}(y) - \mathbb{P}(E,\, (\tilde{X},\tilde{Y}) = (x,y)) \\
&= \mathbb{P}((\tilde{X},\tilde{Y}) = (x,y)) - \mathbb{P}((X,Y) \ne (x,y),\, (\tilde{X},\tilde{Y}) = (x,y)) \\
&\le P_{X,Y}(x,y),
\end{aligned}
$$

where (a) is due to the inequality $[a-s]_+[b-t]_+ \le [ab-st]_+$ for $a,b,s,t \ge 0$.[1] We have the Markov chain $X \leftrightarrow V \leftrightarrow Y$. We also have

$$
\begin{aligned}
\sum_x \left(\frac{|\mathcal{Y}|}{|\mathcal{X}|}\right)^{1/4} \sqrt{\mathbb{P}(E,\, \tilde{X} = x)} &\le \left(\frac{|\mathcal{Y}|}{|\mathcal{X}|}\right)^{1/4} |\mathcal{X}| \sqrt{\frac{1}{|\mathcal{X}|} \sum_x \mathbb{P}(E,\, \tilde{X} = x)} \\
&= (|\mathcal{X}||\mathcal{Y}|)^{1/4} \sqrt{\mathbb{P}(E)}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
P_V(0) &= \left(\sum_x \left[P_{\tilde{X}}(x) - \left(\frac{|\mathcal{Y}|}{|\mathcal{X}|}\right)^{1/4} \sqrt{\mathbb{P}(E,\, \tilde{X} = x)}\right]_+\right) \\
&\qquad \cdot \left(\sum_y \left[P_{\tilde{Y}}(y) - \left(\frac{|\mathcal{X}|}{|\mathcal{Y}|}\right)^{1/4} \sqrt{\mathbb{P}(E,\, \tilde{Y} = y)}\right]_+\right) \\
&\ge \left[1 - (|\mathcal{X}||\mathcal{Y}|)^{1/4} \sqrt{\mathbb{P}(E)}\right]_+^2 \\
&\ge 1 - 2(|\mathcal{X}||\mathcal{Y}|)^{1/4} \sqrt{\mathbb{P}(E)},
\end{aligned}
$$

which is the desired claim.

We now prove Lemma 76. Applying the claim on $P_{X,Y,\tilde{X},\tilde{Y}|U}(\cdot|u)$ for each u, there exists $V \in [0..\min\{|\mathcal{X}|,|\mathcal{Y}|\}]$ with $X \leftrightarrow V \leftrightarrow Y$ conditional on $U = u$ (and hence $X \leftrightarrow (U,V) \leftrightarrow Y$) and

$$
P_{V|U}(0|u) \ge 1 - 2(|\mathcal{X}||\mathcal{Y}|)^{1/4} \sqrt{\mathbb{P}((X,Y) \ne (\tilde{X},\tilde{Y}) \,|\, U = u)}.
$$

We have

$$
\begin{aligned}
P_V(0) &\ge 1 - \mathbb{E}_U \left[2(|\mathcal{X}||\mathcal{Y}|)^{1/4} \sqrt{\mathbb{P}((X,Y) \ne (\tilde{X},\tilde{Y}) \,|\, U)}\right] \\
&\ge 1 - 2(|\mathcal{X}||\mathcal{Y}|)^{1/4} \sqrt{\mathbb{P}((X,Y) \ne (\tilde{X},\tilde{Y}))}.
\end{aligned}
$$

[1] Assume $a > s$, $b > t$ (otherwise the inequality is trivial). We have $[a-s]_+[b-t]_+ = ab - at - bs + st < ab - st \le [ab-st]_+$.

Hence $P_V(0) \geq 1 - \eta$, and

$$\begin{aligned} H(V) &= H_b(P_V(0)) + (1 - P_V(0))H(V \mid V \neq 0) \\ &\leq H_b(\min\{\eta, 1/2\}) + \eta \log_2 \min\{|\mathcal{X}|, |\mathcal{Y}|\}. \end{aligned}$$

□

References

Abadi, M., A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. (2016). "Deep learning with differential privacy". In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.

Agustsson, E. and L. Theis. (2020). "Universally quantized neural compression". In: *Advances in Neural Information Processing Systems*. Vol. 34. 12367–12376.

Altuğ, Y. and A. B. Wagner. (2012). "Source and channel simulation using arbitrary randomness". *IEEE Transactions on Information Theory*. 58(3): 1345–1360.

Amiri, S., A. Belloum, S. Klous, and L. Gommans. (2021). "Compressive differentially private federated learning through universal vector quantization". In: *AAAI Workshop on Privacy-Preserving Artificial Intelligence*. 2–9.

Anantharam, V. and V. Borkar. (2007). "Common randomness and distributed control: A counterexample". *Systems & control letters*. 56(7-8): 568–572.

Andrés, M. E., N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. (2013). "Geo-indistinguishability: Differential privacy for location-based systems". In: *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. 901–914.

Angel, O. and Y. Spinka. (2019). "Pairwise optimal coupling of multiple random variables". *arXiv preprint arXiv:1903.00632*.

Arikan, E. (2009). "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels". *IEEE Transactions on Information Theory*. 55(7): 3051–3073.

Atif, T. A., A. Padakandla, and S. S. Pradhan. (2021). "Synthesizing correlated randomness using algebraic structured codes". In: *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2417–2422.

Atif, T. A., A. Padakandla, and S. S. Pradhan. (2022). "Source coding for synthesizing correlated randomness". *IEEE Transactions on Information Theory*. 69(1): 626–649.

Ballé, J., P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici. (2020). "Nonlinear transform coding". *IEEE Journal of Selected Topics in Signal Processing*. 15(2): 339–353.

Ballé, J., V. Laparra, and E. P. Simoncelli. (2017). "End-to-end optimized image compression". In: *5th International Conference on Learning Representations, ICLR 2017*.

Bao, J., P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, and J. A. Hendler. (2011). "Towards a theory of semantic communication". In: *2011 IEEE Network Science Workshop*. IEEE. 110–117.

Barak, B., M. Braverman, X. Chen, and A. Rao. (2010). "How to compress interactive communication". In: *Proceedings of the forty-second ACM symposium on Theory of computing*. 67–76.

Barnum, H., C. M. Caves, C. A. Fuchs, R. Jozsa, and B. Schumacher. (2001). "On quantum coding for ensembles of mixed states". *Journal of Physics A: Mathematical and General*. 34(35): 6767.

Bassily, R., S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff. (2018). "Learners that use little information". In: *Algorithmic Learning Theory*. PMLR. 25–55.

Bassily, R. and A. Smith. (2015). "Local, private, efficient protocols for succinct histograms". In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 127–135.

Bell, J. S. (1964). "On the Einstein Podolsky Rosen paradox". *Physics Physique Fizika*. 1(3): 195.

Bennett, C. H., G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters. (1993). "Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels". *Physical review letters*. 70(13): 1895.

Bennett, C. H., P. W. Shor, J. Smolin, and A. V. Thapliyal. (2002). "Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem". *IEEE Transactions on Information Theory*. 48(10): 2637–2655.

Bennett, C. H. and S. J. Wiesner. (1992). "Communication via one-and two-particle operators on Einstein-Podolsky-Rosen states". *Physical review letters*. 69(20): 2881.

Bennett, C. H., I. Devetak, A. W. Harrow, P. W. Shor, and A. Winter. (2014). "The quantum reverse Shannon theorem and resource tradeoffs for simulating quantum channels". *IEEE Transactions on Information Theory*. 60(5): 2926–2959. ISSN: 0018-9448. DOI: 10.1109/TIT.2014.2309968.

Berger, T. (1971). *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, NJ, USA.

Berger, T. (2003). "Rate-distortion theory". *Wiley Encyclopedia of Telecommunications*.

Berman, A. and R. J. Plemmons. (1994). *Nonnegative matrices in the mathematical sciences*. SIAM.

Berta, M., J. M. Renes, and M. M. Wilde. (2014). "Identifying the information gain of a quantum measurement". *IEEE Transactions on Information Theory*. 60(12): 7987–8006.

Blau, Y. and T. Michaeli. (2018). "The perception-distortion tradeoff". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6228–6237.

Blau, Y. and T. Michaeli. (2019). "Rethinking lossy compression: The rate-distortion-perception tradeoff". In: *International Conference on Machine Learning*. PMLR. 675–685.

Bloch, M. R. and J. Kliewer. (2013). "Strong coordination over a line network". In: *2013 IEEE International Symposium on Information Theory*. IEEE. 2319–2323.

Block, A. and Y. Polyanskiy. (2023). "The sample complexity of approximate rejection sampling with applications to smoothed online learning". In: *The Thirty Sixth Annual Conference on Learning Theory*. PMLR. 228–273.

Bowe, S., A. Gabizon, and I. Miers. (2017). "Scalable multi-party computation for zk-SNARK parameters in the random beacon model". *Cryptology ePrint Archive*.

Brassard, G., R. Cleve, and A. Tapp. (1999). "Cost of exactly simulating quantum entanglement with classical communication". *Physical Review Letters*. 83(9): 1874.

Braun, G., R. Jain, T. Lee, and S. Pokutta. (2017). "Information-theoretic approximations of the nonnegative rank". *computational complexity*. 26: 147–197.

Braverman, M. and A. Garg. (2014). "Public vs private coin in bounded-round information". In: *International Colloquium on Automata, Languages, and Programming*. Springer. 502–513.

Brody, J., H. Buhrman, M. Koucký, B. Loff, F. Speelman, and N. Vereshchagin. (2016). "Towards a reverse Newman's theorem in interactive information complexity". *Algorithmica*. 76: 749–781.

Bun, M., J. Nelson, and U. Stemmer. (2019). "Heavy hitters and the structure of local privacy". *ACM Transactions on Algorithms (TALG)*. 15(4): 1–40.

Cao, M. X., N. Ramakrishnan, M. Berta, and M. Tomamichel. (2022a). "Channel simulation: finite blocklengths and broadcast channels". *arXiv preprint arXiv:2212.11666*.

Cao, M. X., N. Ramakrishnan, M. Berta, and M. Tomamichel. (2022b). "One-shot point-to-point channel simulation". In: *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 796–801.

Cao, M. X., N. Ramakrishnan, M. Berta, and M. Tomamichel. (2023). "Broadcast channel simulation". In: *2023 IEEE International Symposium on Information Theory (ISIT)*. 1430–1435. DOI: 10.1109/ISIT54713.2023.10206649.

Cerf, N. J., N. Gisin, and S. Massar. (2000). "Classical teleportation of a quantum bit". *Physical Review Letters*. 84(11): 2521.

Cervia, G., L. Luzzi, M. Le Treust, and M. R. Bloch. (2020). "Strong coordination of signals and actions over noisy channels with two-sided state information". *IEEE Transactions on Information Theory*. 66(8): 4681–4708.

Chakrabarti, A., Y. Shi, A. Wirth, and A. Yao. (2001). "Informational complexity and the direct sum problem for simultaneous message complexity". In: *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*. IEEE. 270–278.

Chatterjee, S. and P. Diaconis. (2018). "The sample size required in importance sampling". *The Annals of Applied Probability*. 28(2): 1099–1135.

Chen, J., L. Yu, J. Wang, W. Shi, Y. Ge, and W. Tong. (2022). "On the rate-distortion-perception function". *IEEE Journal on Selected Areas in Information Theory*.

Choi, C. F. and C. T. Li. (2021). "Multiple-output channel simulation and lossy compression of probability distributions". In: *2021 IEEE Information Theory Workshop (ITW)*. IEEE.

Choi, Y., M. El-Khamy, and J. Lee. (2019). "Variable rate deep image compression with a conditional autoencoder". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3146–3154.

Choi, Y., M. El-Khamy, and J. Lee. (2020). "Universal deep neural network compression". *IEEE Journal of Selected Topics in Signal Processing*. 14(4): 715–726.

Cicalese, F., L. Gargano, and U. Vaccaro. (2016). "Approximating probability distributions with short vectors, via information theoretic distance measures". In: *2016 IEEE ISIT*. IEEE. 1138–1142.

Cicalese, F., L. Gargano, and U. Vaccaro. (2019). "Minimum-entropy couplings and their applications". *IEEE Transactions on Information Theory*. 65(6): 3436–3451.

Cicalese, F. and U. Vaccaro. (2002). "Supermodularity and subadditivity properties of the entropy on the majorization lattice". *IEEE Transactions on Information Theory*. 48(4): 933–938.

Clark, J. and U. Hengartner. (2010). "On the use of financial data as a random beacon". In: *2010 Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE 10)*.

Cohen, J. E. and U. G. Rothblum. (1993). "Nonnegative ranks, decompositions, and factorizations of nonnegative matrices". *Linear Algebra and its Applications*. 190: 149–168.

Compton, S. (2022). "A tighter approximation guarantee for greedy minimum entropy coupling". In: *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 168–173.

Compton, S., D. Katz, B. Qi, K. Greenewald, and M. Kocaoglu. (2023). "Minimum-entropy coupling approximation guarantees beyond the majorization barrier". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 10445–10469.

Conway, J. H. and N. J. A. Sloane. (2013). *Sphere Packings, Lattices and Groups*. Vol. 290. Springer Science & Business Media.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Cover, T. M. and H. H. Permuter. (2007). "Capacity of coordinated actions". In: *2007 IEEE International Symposium on Information Theory*. IEEE. 2701–2705.

Cover, T. M. and J. A. Thomas. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience. ISBN: 0471241954.

Csirik, J. A. (2002). "Cost of exactly simulating a bell pair using classical communication". *Physical Review A*. 66(1): 014302.

Csiszár, I. (1998). "The method of types". *IEEE Transactions on Information Theory*. 44(6): 2505–2523.

Csiszár, I. and J. Körner. (2011). *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press.

Cubitt, T. S., D. Leung, W. Matthews, and A. Winter. (2011). "Zero-error channel capacity and simulation assisted by non-local correlations". *IEEE Transactions on Information Theory*. 57(8): 5509–5523.

Cuff, P. (2008). "Communication requirements for generating correlated random variables". In: *2008 IEEE International Symposium on Information Theory*. IEEE. 1393–1397.

Cuff, P. (2013). "Distributed channel synthesis". *IEEE Transactions on Information Theory*. 59(11): 7071–7096.

Cuff, P. (2016). "Soft covering with high probability". In: *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2963–2967.

Cuff, P., H. Permuter, and T. M. Cover. (2010). "Coordination capacity". *IEEE Transactions on Information Theory*. 56(9): 4181–4206.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.

Duchi, J. C., M. I. Jordan, and M. J. Wainwright. (2013). "Local privacy and statistical minimax rates". In: *2013 IEEE 54th annual symposium on foundations of computer science*. IEEE. 429–438.

Dupont, E., H. Loya, M. Alizadeh, A. Golinski, Y. Teh, and A. Doucet. (2022). "COIN++: Neural compression across modalities". *Transactions on Machine Learning Research*. 2022(11).

Dupont, E., A. Golinski, M. Alizadeh, Y. W. Teh, and A. Doucet. (2021). "COIN: COmpression with Implicit Neural representations". In: *Neural Compression: From Information Theory to Applications–Workshop@ ICLR 2021*.

Dwork, C., F. McSherry, K. Nissim, and A. Smith. (2006). "Calibrating noise to sensitivity in private data analysis". In: *Theory of cryptography conference*. Springer. 265–284.

Dwork, C. and A. Roth. (2014). "The algorithmic foundations of differential privacy". *Foundations and Trends® in Theoretical Computer Science*. 9(3–4): 211–407.

El Gamal, A. and Y.-H. Kim. (2011). *Network Information Theory*. Cambridge University Press.

Elias, P. (1972). "The efficient construction of an unbiased random sequence". *The Annals of Mathematical Statistics*. 43(3): 865–870.

Elias, P. (1975). "Universal codeword sets and representations of the integers". *IEEE Transactions on Information Theory*. 21(2): 194–203.

Erdemir, E., T.-Y. Tung, P. L. Dragotti, and D. Gündüz. (2023). "Generative joint source-channel coding for semantic image transmission". *IEEE Journal on Selected Areas in Communications*.

Evfimievski, A., J. Gehrke, and R. Srikant. (2003). "Limiting privacy breaches in privacy preserving data mining". In: *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 211–222.

Fano, R. M. (1949). *The transmission of information*. Vol. 65. Massachusetts Institute of Technology, Research Laboratory of Electronics.

Fano, R. M. (1961). *Transmission of Information*. The MIT Press.

Feldman, V. and K. Talwar. (2021). "Lossless compression of efficient private local randomizers". In: *International Conference on Machine Learning*. PMLR. 3208–3219.

Feldmann, M. (1995). "New loophole for the Einstein-Podolsky-Rosen paradox". *Foundations of Physics Letters*. 8: 41–53.

Flamich, G. (2023). "Greedy Poisson rejection sampling". *Advances in Neural Information Processing Systems*. 36.

Flamich, G., M. Havasi, and J. M. Hernández-Lobato. (2020). "Compressing images by encoding their latent representations with relative entropy coding". *Advances in Neural Information Processing Systems*. 33: 16131–16141.

Flamich, G., S. Markou, and J. M. Hernández-Lobato. (2022). "Fast relative entropy coding with A* coding". In: *International Conference on Machine Learning*. PMLR. 6548–6577.

Flamich, G., S. Markou, and J. M. Hernández-Lobato. (2024). "Faster relative entropy coding with greedy rejection coding". *Advances in Neural Information Processing Systems*. 36.

Flamich, G. and L. Theis. (2023). "Adaptive greedy rejection sampling". In: *2023 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 454–459.

Flamich, G. and L. Wells. (2024). "Some Notes on the Sample Complexity of Approximate Channel Simulation". *arXiv preprint arXiv:2405.04363*.

Fraenkel, A. S. and S. T. Kleinb. (1996). "Robust universal complete codes for transmission and compression". *Discrete Applied Mathematics*. 64(1): 31–55.

Frenkel, P. E. and M. Weiner. (2015). "Classical information storage in an n-level quantum system". *Communications in Mathematical Physics*. 340(2): 563–574.

Gish, H. and J. Pierce. (1968). "Asymptotically efficient quantizing". *IEEE Transactions on Information Theory*. 14(5): 676–683.

Goc, D. and G. Flamich. (2024). "On Channel Simulation with Causal Rejection Samplers". In: *2024 IEEE International Symposium on Information Theory (ISIT)*. 1682–1687. DOI: 10.1109/ISIT57864.2024.10619339.

Gohari, A., M. H. Yassaee, and M. R. Aref. (2012). "Secure channel simulation". In: *2012 IEEE Information Theory Workshop*. IEEE. 406–410.

Gohari, A. A. and V. Anantharam. (2011). "Generating dependent random variables over networks". In: *2011 IEEE Information Theory Workshop*. IEEE. 698–702.

Gossner, O., P. Hernandez, and A. Neyman. (2006). "Optimal use of communication resources". *Econometrica*. 74(6): 1603–1636.

Gray, R. M. and T. G. Stockham. (1993). "Dithered quantizers". *IEEE Transactions on Information Theory*. 39(3): 805–812.

Gray, R. M. and D. L. Neuhoff. (1998). "Quantization". *IEEE Transactions on Information Theory*. 44(6): 2325–2383.

Gündüz, D., Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae. (2022). "Beyond transmitting bits: Context, semantics, and task-oriented communications". *IEEE Journal on Selected Areas in Communications*. 41(1): 5–41.

Guo, Z., G. Flamich, J. He, Z. Chen, and J. M. Hernández-Lobato. (2023). "Compression with Bayesian implicit neural representations". *Advances in Neural Information Processing Systems*. 36: 1938–1956.

Haddadpour, F., M. H. Yassaee, S. Beigi, A. Gohari, and M. R. Aref. (2016). "Simulation of a channel with another channel". *IEEE Transactions on Information Theory*. 63(5): 2659–2677.

Hajek, B. and M. Pursley. (1979). "Evaluation of an achievable rate region for the broadcast channel". *IEEE Transactions on Information Theory*. 25(1): 36–46. ISSN: 0018-9448. DOI: 10.1109/TIT.1979.1055989.

Hamdi, Y., A. B. Wagner, and D. Gündüz. (2024). "The Rate-Distortion-Perception Trade-off: The Role of Private Randomness". *arXiv preprint arXiv:2404.01111*.

Han, T. S. and S. Verdú. (1993). "Approximation theory of output statistics". *IEEE Transactions on Information Theory*. 39(3): 752–772. ISSN: 0018-9448. DOI: 10.1109/18.256486.

Han, T. S. and M. Hoshi. (1997). "Interval algorithm for random number generation". *IEEE Transactions on Information Theory*. 43(2): 599–611. ISSN: 0018-9448. DOI: 10.1109/18.556116.

Haramoto, H., M. Matsumoto, and P. L'Ecuyer. (2008a). "A fast jump ahead algorithm for linear recurrences in a polynomial space". In: *International Conference on Sequences and Their Applications*. Springer. 290–298.

Haramoto, H., M. Matsumoto, T. Nishimura, F. Panneton, and P. L'Ecuyer. (2008b). "Efficient Jump Ahead for F2-Linear Random Number Generators". *INFORMS Journal on Computing*. 20(3): 385–390.

Harsha, P., R. Jain, D. McAllester, and J. Radhakrishnan. (2010). "The communication complexity of correlation". *IEEE Transactions on Information Theory*. 56(1): 438–449.

Hasırcıoğlu, B. and D. Gündüz. (2024). "Communication efficient private federated learning using dithering". In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 7575–7579.

Havasi, M., R. Peharz, and J. M. Hernández-Lobato. (2019). "Minimal random code learning: Getting bits back from compressed model parameters". In: *7th International Conference on Learning Representations, ICLR 2019*.

Hayashi, M. (2006). "General nonasymptotic and asymptotic formulas in channel resolvability and identification capacity and their application to the wiretap channel". *IEEE Transactions on Information Theory*. 52(4): 1562–1575.

Hayashi, M. (2008). "Second-order asymptotics in fixed-length source coding and intrinsic randomness". *IEEE Transactions on Information Theory*. 54(10): 4619–4637.

Hayashi, M. (2009). "Information spectrum approach to second-order coding rate in channel coding". *IEEE Transactions on Information Theory*. 55(11): 4947–4966.

He, J., G. Flamich, Z. Guo, and J. M. Hernández-Lobato. (2024a). "RECOMBINER: Robust and Enhanced Compression with Bayesian Implicit Neural Representations". In: *The Twelfth International Conference on Learning Representations*.

He, J., G. Flamich, and J. M. Hernández-Lobato. (2024b). "Accelerating Relative Entropy Coding with Space Partitioning". *arXiv preprint arXiv:2405.12203*.

Hegazy, M., R. Leluc, C. T. Li, and A. Dieuleveut. (2024). "Compression with exact error distribution for federated learning". In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. Vol. 238. *Proceedings of Machine Learning Research*. PMLR. 613–621.

Hegazy, M. and C. T. Li. (2022). "Randomized quantization with exact error distribution". In: *2022 IEEE Information Theory Workshop (ITW)*. IEEE. 350–355.

Hoeffding, W. and G. Simons. (1970). "Unbiased coin tossing with a biased coin". In: *The Collected Works of Wassily Hoeffding*. Springer. 501–512.

Holevo, A. S. (1973). "Bounds for the quantity of information transmitted by a quantum communication channel". *Problemy Peredachi Informatsii*. 9(3): 3–11.

Huffman, D. A. (1952). "A method for the construction of minimum-redundancy codes". *Proceedings of the IRE*. 40(9): 1098–1101.

Huijben, I. A., W. Kool, M. B. Paulus, and R. J. Van Sloun. (2022). "A review of the gumbel-max trick and its extensions for discrete stochasticity in machine learning". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 45(2): 1353–1371.

Isik, B., F. Pase, D. Gunduz, S. Koyejo, T. Weissman, and M. Zorzi. (2024). "Adaptive Compression in Federated Learning via Side Information". In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. Ed. by S. Dasgupta, S. Mandt, and Y. Li. Vol. 238. *Proceedings of Machine Learning Research*. PMLR. 487–495.

Itô, K. (1984). *An Introduction to Probability Theory*. Cambridge University Press.

Jain, R., J. Radhakrishnan, and P. Sen. (2003). "A direct sum theorem in communication complexity via message compression". In: *Automata, Languages and Programming: 30th International Colloquium, ICALP 2003 Eindhoven, The Netherlands, June 30–July 4, 2003 Proceedings 30*. Springer. 300–315.

Jain, R., Y. Shi, Z. Wei, and S. Zhang. (2013). "Efficient protocols for generating bipartite classical distributions and quantum states". *IEEE Transactions on Information Theory*. 59(8): 5171–5178.

Jayant, N. and L. Rabiner. (1972). "The application of dither to the quantization of speech signals". *Bell System Technical Journal*. 51(6): 1293–1304.

Kailath, T. (1967). "The divergence and Bhattacharyya distance measures in signal selection". *IEEE transactions on communication technology*. 15(1): 52–60.

Kairouz, P., H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, *et al.* (2021). "Advances and open problems in federated learning". *Foundations and trends® in machine learning*. 14(1–2): 1–210.

Kallenberg, O. (2002). *Foundations of Modern Probability*. Springer Science & Business Media.

Kasiviswanathan, S. P., H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. (2011). "What can we learn privately?" *SIAM Journal on Computing*. 40(3): 793–826.

Kemperman, J. (1974). "On the Shannon capacity of an arbitrary channel". In: *Indagationes Mathematicae (Proceedings)*. Vol. 77. No. 2. North-Holland. 101–115.

Khisti, A., A. Behboodi, G. Cesa, and P. Kumar. (2024). "Unequal message protection: one-shot analysis via Poisson matching lemma". In: *2024 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 629–634.

Kingma, D. P. and M. Welling. (2013). "Auto-encoding variational Bayes". *arXiv preprint arXiv:1312.6114*.

Kirac, A. and P. Vaidyanathan. (1996). "Results on lattice vector quantization with dithering". *IEEE Transactions On Circuits and Systems II: Analog and Digital Signal Processing*. 43(12): 811–826.

Kloek, T. and H. K. Van Dijk. (1978). "Bayesian estimates of equation system parameters: an application of integration by Monte Carlo". *Econometrica: Journal of the Econometric Society*: 1–19.

Kneusel, R. T. (2018). *Random numbers and computers*. Vol. 239. Springer.

Knuth, D. E. and A. C. Yao. (1976). "The complexity of nonuniform random number generation". *Algorithms and Complexity: New Directions and Recent Results*: 357–428. Ed. by J. F. Traub.

Kobus, S., L. Theis, and D. Gündüz. (2024a). "Gaussian Channel Simulation with Rotated Dithered Quantization". In: *2024 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 1907–1912.

Kobus, S., T.-Y. Tung, and D. Gündüz. (2024b). "Universal Sample Coding". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Kocaoglu, M., A. G. Dimakis, S. Vishwanath, and B. Hassibi. (2017a). "Entropic causal inference". In: *Thirty-First AAAI Conference on Artificial Intelligence*.

Kocaoglu, M., A. G. Dimakis, S. Vishwanath, and B. Hassibi. (2017b). "Entropic causality and greedy minimum entropy coupling". In: *2017 IEEE ISIT*. IEEE. 1465–1469.

Kolmogorov, A. N. (1956). "On certain asymptotic characteristics of completely bounded metric spaces". In: *Dokl. Akad. Nauk SSSR*. Vol. 108. No. 3. 385–388.

Kovačević, M., I. Stanojević, and V. Šenk. (2015). "On the entropy of couplings". *Information and Computation*. 242: 369–382.

Kraft, L. G. (1949). "A device for quantizing, grouping, and coding amplitude-modulated pulses". *MA thesis*. Massachusetts Institute of Technology.

Kramer, G. and S. A. Savari. (2007). "Communicating probability distributions". *IEEE Transactions on Information Theory*. 53(2): 518–525.

Kumar, G. R., C. T. Li, and A. El Gamal. (2014). "Exact common information". In: *Proc. IEEE Int. Symp. Inf. Theory*. 161–165.

Kurri, G. R., V. M. Prabhakaran, and A. D. Sarwate. (2021). "Coordination through shared randomness". *IEEE Transactions on Information Theory*. 67(8): 4948–4974.

Kurri, G. R., V. Ramachandran, S. R. B. Pillai, and V. M. Prabhakaran. (2022). "Multiple access channel simulation". *IEEE Transactions on Information Theory*. 68(11): 7575–7603.

Lang, N., E. Sofer, T. Shaked, and N. Shlezinger. (2023). "Joint privacy enhancement and quantization in federated learning". *IEEE Transactions on Signal Processing*. 71: 295–310.

Last, G. and M. Penrose. (2017). *Lectures on the Poisson Process*. Vol. 7. Cambridge University Press.

Le Treust, M. and T. Tomala. (2018). "Strategic coordination with state information at the decoder". In: *International Zurich Seminar on Information and Communication (IZS 2018). Proceedings*. ETH Zurich. 30–34.

Lee, D. D. and H. S. Seung. (1999). "Learning the parts of objects by non-negative matrix factorization". *Nature*. 401(6755): 788–791.

Lee, S.-H. and S.-Y. Chung. (2015). "A unified approach for network information theory". In: *2015 IEEE ISIT*. IEEE. 1277–1281.

Lee, S.-H. and S.-Y. Chung. (2018). "A unified random coding bound". *IEEE Transactions on Information Theory*. 64(10): 6779–6802.

Lei, E., H. Hassani, and S. S. Bidokhti. (2022). "Neural estimation of the rate-distortion function with applications to operational source coding". *IEEE Journal on Selected Areas in Information Theory*. 3(4): 674–686.

Li, C. T., X. Wu, A. Ozgur, and A. El Gamal. (2018). "Minimax learning for remote prediction". In: *2018 IEEE ISIT*. 541–545. DOI: 10.1109/ISIT.2018.8437318.

Li, C. T. (2020). "PSITIP - Python Symbolic Information Theoretic Inequality Prover". URL: https://github.com/cheuktingli/psitip.

Li, C. T. (2021). "Efficient approximate minimum entropy coupling of multiple probability distributions". *IEEE Transactions on Information Theory*. 67(8): 5259–5268. DOI: 10.1109/TIT.2021.3076986.

Li, C. T. (2023). "An automated theorem proving framework for information-theoretic results". *IEEE Transactions on Information Theory*. 69(11): 6857–6877. DOI: 10.1109/TIT.2023.3296597.

Li, C. T. (2024). "Pointwise redundancy in one-shot lossy compression via Poisson functional representation". In: *arXiv preprint; short version presented at 2024 International Zurich Seminar on Information and Communication*. 28–29. URL: https://arxiv.org/pdf/2401.14805.pdf.

Li, C. T. and V. Anantharam. (2019). "Pairwise multi-marginal optimal transport and embedding for earth mover's distance". *arXiv preprint arXiv:1908.01388*.

Li, C. T. and V. Anantharam. (2021). "A unified framework for one-shot achievability via the Poisson matching lemma". *IEEE Transactions on Information Theory*. 67(5): 2624–2651.

Li, C. T. and A. El Gamal. (2017). "Distributed simulation of continuous random variables". *IEEE Transactions on Information Theory*. 63(10): 6329–6343.

Li, C. T. and A. El Gamal. (2018a). "A universal coding scheme for remote generation of continuous random variables". *IEEE Transactions on Information Theory*. 64(4): 2583–2592. DOI: 10.1109/TIT.2018.2803752.

Li, C. T. and A. El Gamal. (2018b). "Strong functional representation lemma and applications to coding theorems". *IEEE Transactions on Information Theory*. 64(11): 6967–6978. DOI: 10.1109/TIT.2018.2865570.

Li, M., J. Klejsa, and W. B. Kleijn. (2010). "Distribution preserving quantization with dithering and transformation". *IEEE Signal Processing Letters*. 17(12): 1014–1017.

Li, M., J. Klejsa, and W. B. Kleijn. (2011). "On distribution preserving quantization". *arXiv preprint arXiv:1108.3728*.

Lindvall, T. (2002). *Lectures on the coupling method*. Courier Corporation.

Ling, C. W. and C. T. Li. (2023). "Vector quantization with error uniformly distributed over an arbitrary set". In: *2023 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 856–861.

Ling, C. W. and C. T. Li. (2024). "Rejection-Sampled Universal Quantization for Smaller Quantization Errors". In: *2024 IEEE International Symposium on Information Theory (ISIT)*. 1883–1888. DOI: 10.1109/ISIT57864.2024.10619181.

Liu, J. and S. Verdú. (2018). "Rejection sampling and noncausal sampling under moment constraints". In: *2018 IEEE ISIT*. 1565–1569. DOI: 10.1109/ISIT.2018.8437857.

Liu, J., M. H. Yassaee, and S. Verdú. (2019). "Sharp bounds for mutual covering". *IEEE Transactions on Information Theory*. 65(12): 8067–8083.

Liu, J. S. (2004). *Monte Carlo Strategies in Scientific Computing*. New York, NY, USA: Springer.

Liu, W., G. Xu, and B. Chen. (2010). "The common information of N dependent random variables". In: *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 836–843.

Liu, Y., W.-N. Chen, A. Ozgur, and C. T. Li. (2024). "Universal Exact Compression of Differentially Private Mechanisms". In: *Advances in Neural Information Processing Systems*. Vol. 37. Curran Associates, Inc. 91492–91531.

Liu, Y. and C. T. Li. (2024). "One-Shot Coding over General Noisy Networks". In: *2024 IEEE International Symposium on Information Theory (ISIT)*. 3124–3129.

Ma, N. and P. Ishwar. (2011). "Some results on distributed source coding for interactive function computation". *IEEE Transactions on Information Theory*. 57(9): 6180–6195.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

Maddison, C. J. (2016). "A Poisson process model for Monte Carlo". *Perturbation, Optimization, and Statistics*: 193–232.

Maddison, C. J., D. Tarlow, and T. Minka. (2014). "A* sampling". *Advances in neural information processing systems*. 27.

Managoli, M. A. and V. M. Prabhakaran. (2024). "Broadcast Channel Synthesis from Shared Randomness". In: *2024 IEEE International Symposium on Information Theory (ISIT)*. 1919–1924. DOI: 10.1109/ISIT57864.2024.10619094.

Marshall, A. W., I. Olkin, and B. C. Arnold. (2011). *Inequalities: theory of Majorization and its Applications*. New York, Dordrecht, Heidelberg, London: Springer.

Massar, S., D. Bacon, N. J. Cerf, and R. Cleve. (2001). "Classical simulation of quantum entanglement without local hidden variables". *Physical Review A*. 63(5): 052305.

Matsumoto, R. (2018). "Introducing the perception-distortion tradeoff into the rate-distortion theory of general information sources". *IEICE Communications Express*. 7(11): 427–431.

Matsumoto, R. (2019). "Rate-distortion-perception tradeoff of variable-length source coding for general information sources". *IEICE Communications Express*. 8(2): 38–42.

Maudlin, T. (1992). "Bell's inequality, information transmission, and prism models". In: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*. Vol. 1992. No. 1. Cambridge University Press. 404–417.

McMahan, B., E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. (2017). "Communication-efficient learning of deep networks from decentralized data". In: *Artificial intelligence and statistics*. PMLR. 1273–1282.

McMahan, H. B., D. Ramage, K. Talwar, and L. Zhang. (2018). "Learning Differentially Private Recurrent Language Models". In: *International Conference on Learning Representations*.

Mezzavilla, M., S. Dutta, M. Zhang, M. R. Akdeniz, and S. Rangan. (2015). "5G mmWave module for the ns-3 network simulator". In: *Proceedings of the 18th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. 283–290.

Mildenhall, B., P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. (2021). "NeRF: Representing scenes as neural radiance fields for view synthesis". *Communications of the ACM*. 65(1): 99–106.

Mitzenmacher, M. and E. Upfal. (2017). *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press.

Nema, A., S. Sreekumar, and M. Berta. (2024). "One-Shot Multiple Access Channel Simulation". In: *2024 IEEE International Symposium on Information Theory (ISIT)*. 2981–2986. DOI: 10.1109/ISIT57864.2024.10619283.

NumPy Developers. (2024). "Mersenne Twister (MT19937)". https://numpy.org/doc/stable/reference/random/bit_generators/mt19937.html. URL: https://numpy.org/doc/stable/reference/random/bit_generators/mt19937.html.

O'Neill, M. E. (2014). "PCG: A family of simple fast space-efficient statistically good algorithms for random number generation". *ACM Transactions on Mathematical Software*.

Obead, S. A., B. N. Vellambi, and J. Kliewer. (2021). "Strong coordination over noisy channels". *IEEE Transactions on Information Theory*. 67(5): 2716–2738.

Oded, G. (2004). *Foundations of Cryptography: Basic Tools*. Cambridge: Cambridge university press.

Oohama, Y. (2011). "Performance analysis of the interval algorithm for random number generation based on number systems". *IEEE Transactions on Information Theory*. 57(3): 1177–1185.

Ornstein, D. S. and P. C. Shields. (1990). "Universal almost sure data compression". *The Annals of Probability*: 441–452.

Painsky, A., S. Rosset, and M. Feder. (2013). "Memoryless representation of Markov processes". In: *2013 IEEE ISIT*. IEEE. 2294–298.

Park, J. J., P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. (2019). "DeepSDF: Learning continuous signed distance functions for shape representation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 165–174.

Pase, F., S. Kobus, D. Gündüz, and M. Zorzi. (2023). "Semantic communication of learnable concepts". In: *2023 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 731–736.

Peres, Y. (1992). "Iterating von Neumann's procedure for extracting random bits". *The Annals of Statistics*: 590–597.

Phan, B., A. Khisti, and C. Louizos. (2024). "Importance matching lemma for lossy compression with side information". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 1387–1395.

Pirandola, S., S. L. Braunstein, R. Laurenza, C. Ottaviani, T. P. Cope, G. Spedalieri, and L. Banchi. (2018). "Theory of channel simulation and bounds for private communication". *Quantum Science and Technology*. 3(3): 035009.

Polyanskiy, Y., H. V. Poor, and S. Verdú. (2010). "Channel coding rate in the finite blocklength regime". *IEEE Transactions on Information Theory*. 56(5): 2307–2359.

Polyanskiy, Y. and Y. Wu. (2024). *Information theory: From coding to learning*. Cambridge University Press.

Propp, J. and D. Wilson. (1998). "Coupling from the past: a user's guide". *Microsurveys in Discrete Probability*. 41: 181–192.

Propp, J. G. and D. B. Wilson. (1996). "Exact sampling with coupled Markov chains and applications to statistical mechanics". *Random Structures & Algorithms*. 9(1-2): 223–252.

Rabin, M. O. (1983). "Transaction protection by beacons". *Journal of Computer and System Sciences*. 27(2): 256–267.

Ramachandran, V., T. J. Oechtering, and M. Skoglund. (2024). "Multi-terminal strong coordination over noiseless networks with secrecy constraints". In: *International Zurich Seminar on Information and Communication (IZS 2024). Proceedings*. ETH Zürich. 159–163.

RAND Corporation. (2001). *A Million Random Digits with 100,000 Normal Deviates*. Santa Monica, CA: RAND Corporation. DOI: 10.7249/MR1418.

Rao, A. and A. Yehudayoff. (2020). *Communication Complexity and Applications*. Cambridge University Press.

Renner, R. and S. Wolf. (2003). "New bounds in secret-key agreement: The gap between formation and secrecy extraction". In: *Advances in Cryptology–EUROCRYPT 2003: International Conference on the Theory and Applications of Cryptographic Techniques, Warsaw, Poland*. Springer. 562–577.

Rényi, A. (1961). "On measures of entropy and information". In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Roberts, L. (1962). "Picture coding using pseudo-random noise". *IRE Transactions on Information Theory*. 8(2): 145–154.

Roche, J. R. (1991). "Efficient generation of random variables from biased coins". In: *Proceedings. 1991 IEEE International Symposium on Information Theory*. IEEE. 169–169.

Ross, S. (2019). *A First Course in Probability*. Pearson Higher Ed.

Rossi, M. (2019). "Greedy additive approximation algorithms for minimum-entropy coupling problem". In: *2019 IEEE ISIT*. IEEE. 1127–1131.

Saldi, N., T. Linder, and S. Yüksel. (2013). "Randomized quantization and optimal design with a marginal constraint". In: *2013 IEEE International Symposium on Information Theory*. IEEE. 2349–2353.

Saldi, N., T. Linder, and S. Yüksel. (2014). "Randomized quantization and source coding with constrained output distribution". *IEEE Transactions on Information Theory*. 61(1): 91–106.

Saldi, N., T. Linder, and S. Yüksel. (2015). "Output constrained lossy source coding with limited common randomness". *IEEE Transactions on Information Theory*. 61(9): 4984–4998.

Salmon, J. K., M. A. Moraes, R. O. Dror, and D. E. Shaw. (2011). "Parallel random numbers: as easy as 1, 2, 3". In: *Proceedings of 2011 international conference for high performance computing, networking, storage and analysis*. 1–12.

Satpathy, S. and P. Cuff. (2014). "Secure coordination with a two-sided helper". In: *2014 IEEE International Symposium on Information Theory*. IEEE. 406–410.

Satpathy, S. and P. Cuff. (2016). "Secure cascade channel synthesis". *IEEE Transactions on Information Theory*. 62(11): 6081–6094.

Sayood, K. (2018). *Introduction to data compression*. Morgan Kaufmann.

Schuchman, L. (1964). "Dither signals and their effect on quantization noise". *IEEE Transactions on Communication Technology*. 12(4): 162–165.

Sefidgaran, M., A. Zaidi, and P. Krasnowski. (2024). "Minimal Communication-Cost Statistical Learning". In: *2024 IEEE International Symposium on Information Theory (ISIT)*. 777–782. DOI: 10.1109/ISIT57864.2024.10619653.

Shah, A., W.-N. Chen, J. Balle, P. Kairouz, and L. Theis. (2022). "Optimal compression of locally differentially private mechanisms". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 7680–7723.

Shahmiri, A. M., C. W. Ling, and C. T. Li. (2024). "Communication-efficient Laplace mechanism for differential privacy via random quantization". In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 4550–4554.

Shannon, C. E. (1948). "A mathematical theory of communication". *Bell system technical journal*. 27(3): 379–423.

Shannon, C. E. (1949). "Communication theory of secrecy systems". *The Bell system technical journal*. 28(4): 656–715.

Shao, Y., Q. Cao, and D. Gunduz. (2022). "A theory of semantic communication". *arXiv preprint arXiv:2212.01485*.

Shkel, Y. (2024). "Functional Representation Lemma: Algorithms and Applications". In: *International Zurich Seminar on Information and Communication (IZS 2024)*. 26.

Shkel, Y. Y. and A. K. Yadav. (2023). "Information spectrum converse for minimum entropy couplings and functional representations". In: *2023 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 66–71.

Shlezinger, N., M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui. (2020). "UVeQFed: Universal vector quantization for federated learning". *IEEE Transactions on Signal Processing*. 69: 500–514.

Sitzmann, V., J. Martel, A. Bergman, D. Lindell, and G. Wetzstein. (2020). "Implicit neural representations with periodic activation functions". *Advances in Neural Information Processing Systems*. 33: 7462–7473.

Song, E. C., P. Cuff, and H. V. Poor. (2016). "The likelihood encoder for lossy compression". *IEEE Transactions on Information Theory*. 62(4): 1836–1849.

Sriramu, S. M., R. Barsz, E. Polito, and A. B. Wagner. (2024). "Fast Channel Simulation via Error-Correcting Codes". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Sriramu, S. M. and A. B. Wagner. (2024). "Optimal Redundancy in Exact Channel Synthesis". In: *2024 IEEE International Symposium on Information Theory (ISIT)*. 1913–1918. DOI: 10.1109/ISIT57864.2024.10619703.

Stanley, K. O. (2007). "Compositional pattern producing networks: A novel abstraction of development". *Genetic programming and evolvable machines*. 8: 131–162.

Steinberg, Y. and S. Verdú. (1994). "Channel simulation and coding with side information". *IEEE Transactions on Information Theory*. 40(3): 634–646.

Steinberg, Y. and S. Verdú. (1996). "Simulation of random processes and rate-distortion theory". *IEEE Transactions on Information Theory*. 42(1): 63–86.

Steiner, M. (2000). "Towards quantifying non-local information transfer: finite-bit non-locality". *Physics Letters A*. 270(5): 239–244.

Sun, S., G. R. MacCartney, and T. S. Rappaport. (2017). "A novel millimeter-wave channel simulator and applications for 5G wireless communications". In: *2017 IEEE international conference on communications (ICC)*. IEEE. 1–7.

Tancik, M., P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng. (2020). "Fourier features let networks learn high frequency functions in low dimensional domains". *Advances in Neural Information Processing Systems*. 33: 7537–7547.

Theis, L., T. Salimans, M. D. Hoffman, and F. Mentzer. (2022). "Lossy compression with Gaussian diffusion". *arXiv preprint arXiv:2206.08889*.

Theis, L. and A. B. Wagner. (2021). "A coding theorem for the rate-distortion-perception function". In: *Neural Compression: From Information Theory to Applications–Workshop@ ICLR 2021*.

Theis, L. and N. Yosri. (2022). "Algorithms for the communication of samples". In: *International Conference on Machine Learning*. PMLR. 21308–21328.

Triastcyn, A., M. Reisser, and C. Louizos. (2021). "DP-REC: Private & communication-efficient federated learning". *arXiv preprint arXiv:2111.05454*.

Tschannen, M., E. Agustsson, and M. Lucic. (2018). "Deep generative models for distribution-preserving lossy compression". *Advances in neural information processing systems*. 31.

Uyematsu, T. and F. Kanaya. (1999). "Channel simulation by interval algorithm: A performance analysis of interval algorithm". *IEEE Transactions on Information Theory*. 45(6): 2121–2129.

Vandaele, A., N. Gillis, F. Glineur, and D. Tuyttens. (2016). "Heuristics for exact nonnegative matrix factorization". *Journal of Global Optimization*. 65: 369–400.

Vavasis, S. A. (2010). "On the complexity of nonnegative matrix factorization". *SIAM Journal on Optimization*. 20(3): 1364–1377.

Vellambi, B. N., J. Kliewer, and M. R. Bloch. (2017). "Strong coordination over multi-hop line networks using channel resolvability codebooks". *IEEE Transactions on Information Theory*. 64(2): 1132–1162.

Vellambi, B. N. and J. Kliewer. (2018). "New results on the equality of exact and Wyner common information rates". In: *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 151–155.

Vidyasagar, M. (2012). "A metric between probability distributions on finite sets of different cardinalities and applications to order reduction". *IEEE Transactions on Automatic Control*. 57(10): 2464–2477.

Von Neumann, J. (1963). "Various techniques used in connection with random digits". *John von Neumann, Collected Works*. 5: 768–770.

Wagner, A. B. (2022). "The rate-distortion-perception tradeoff: The role of common randomness". *arXiv preprint arXiv:2202.04147*.

Walker, S. (1999). "The uniform power distribution". *Journal of Applied Statistics*. 26(4): 509–517.

Wannamaker, R. A., S. P. Lipshitz, J. Vanderkooy, and J. N. Wright. (2000). "A theory of nonsubtractive dither". *IEEE Transactions on Signal Processing*. 48(2): 499–516.

Watanabe, S., S. Kuzuoka, and V. Y. F. Tan. (2015). "Nonasymptotic and second-order achievability bounds for coding with side-information". *IEEE Transactions on Information Theory*. 61(4): 1574–1605. ISSN: 0018-9448. DOI: 10.1109/TIT.2015.2400994.

Watanabe, S. and T. S. Han. (2020). "Interval algorithm for random number generation: information spectrum approach". *IEEE Transactions on Information Theory*. 66(3): 1691–1701. DOI: 10.1109/TIT.2019.2946235.

Watanabe, S. and M. Hayashi. (2014). "Strong converse and second-order asymptotics of channel resolvability". In: *2014 IEEE International Symposium on Information Theory*. IEEE. 1882–1886.

Weaver, W. (1953). "Recent contributions to the mathematical theory of communication". *ETC: a Review of General Semantics*: 261–281.

Weissman, T. and E. Ordentlich. (2005). "The empirical distribution of rate-constrained source codes". *IEEE Transactions on Information Theory*. 51(11): 3718–3733.

Willems, F. and E. van der Meulen. (1985). "The discrete memoryless multiple-access channel with cribbing encoders". *IEEE Transactions on Information Theory*. 31(3): 313–327. ISSN: 0018-9448. DOI: 10.1109/TIT.1985.1057042.

Wilson, D. B. (2000). "Layered multishift coupling for use in perfect sampling algorithms (with a primer on CFTP)". *Monte Carlo Methods*. 26: 141–176.

Winter, A. (2002). "Compression of sources of probability distributions and density operators". *arXiv preprint quant-ph/0208131*.

Wyner, A. and J. Ziv. (1994). "The sliding-window Lempel-Ziv algorithm is asymptotically optimal". *Proceedings of the IEEE*. 82(6): 872–877. DOI: 10.1109/5.286191.

Wyner, A. D. (1975a). "The common information of two dependent random variables". *IEEE Transactions on Information Theory*. 21(2): 163–179.

Wyner, A. D. (1975b). "The wire-tap channel". *Bell system technical journal*. 54(8): 1355–1387.

Yagli, S. and P. Cuff. (2019). "Exact exponent for soft covering". *IEEE Transactions on Information Theory*. 65(10): 6234–6262.

Yan, G., T. Li, T. Lan, K. Wu, and L. Song. (2023). "Layered randomized quantization for communication-efficient and privacy-preserving distributed learning". *arXiv preprint arXiv:2312.07060*.

Yang, Y., R. Bamler, and S. Mandt. (2020). "Improving inference for neural image compression". *Advances in Neural Information Processing Systems*. 33: 573–584.

Yang, Y., S. Mandt, and L. Theis. (2023). "An introduction to neural data compression". *Foundations and Trends® in Computer Graphics and Vision*. 15(2): 113–200.

Yao, A. C.-C. (1979). "Some complexity questions related to distributive computing". In: *Proceedings of the eleventh annual ACM symposium on Theory of computing*. 209–213.

Yassaee, M. H., M. R. Aref, and A. Gohari. (2013). "Non-asymptotic output statistics of random binning and its applications". In: *2013 IEEE ISIT*. 1849–1853. DOI: 10.1109/ISIT.2013.6620547.

Yassaee, M. H. (2019). "Almost exact analysis of soft covering lemma via large deviation". In: *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 1387–1391.

Yassaee, M. H. (2015). "One-shot achievability via fidelity". In: *Proc. IEEE Int. Symp. Inf. Theory*. IEEE. 301–305.

Yassaee, M. H., M. R. Aref, and A. Gohari. (2014). "Achievability proof via output statistics of random binning". *IEEE Transactions on Information Theory*. 60(11): 6760–6786.

Yassaee, M. H., A. Gohari, and M. R. Aref. (2015). "Channel simulation via interactive communications". *IEEE Transactions on Information Theory*. 61(6): 2964–2982.

Yeung, R. W. (2008). *Information theory and network coding*. New York: Springer Science & Business Media.

Yu, L. and V. Y. Tan. (2018). "Asymptotic coupling and its applications in information theory". *IEEE Transactions on Information Theory*. 65(3): 1321–1344.

Yu, L. and V. Y. Tan. (2019). "Exact channel synthesis". *IEEE Transactions on Information Theory*. 66(5): 2799–2818.

Yu, L. and V. Y. Tan. (2020). "On exact and ∞-Rényi common informations". *IEEE Transactions on Information Theory*. 66(6): 3366–3406.

Yu, L. and V. Y. Tan. (2022). "Common information, noise stability, and their extensions". *Foundations and Trends® in Communications and Information Theory*. 19(2): 107–389.

Zamir, R. (2014). *Lattice Coding for Signals and Networks: A Structured Coding Approach to Quantization, Modulation and Multiuser Information Theory*. Cambridge University Press. DOI: 10.1017/CBO9781139045520.

Zamir, R. and M. Feder. (1992). "On universal quantization by randomized uniform/lattice quantizers". *IEEE Transactions on Information Theory*. 38(2): 428–436.

Zhang, G., J. Qian, J. Chen, and A. Khisti. (2021). "Universal rate-distortion-perception representations for lossy compression". *Advances in Neural Information Processing Systems*. 34: 11517–11529.

Zhang, S. (2012). "Quantum strategic game theory". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 39–59.

Zhang, Z., E.-H. Yang, and V. K. Wei. (1997). "The redundancy of source coding with a fidelity criterion. 1. Known statistics". *IEEE Transactions on Information Theory*. 43(1): 71–91.

Ziv, J. (1985). "On universal quantization". *IEEE Transactions on Information Theory*. 31(3): 344–347.

Ziv, J. and A. Lempel. (1977). "A universal algorithm for sequential data compression". *IEEE Transactions on Information Theory*. 23(3): 337–343.