Lec26

Friday, October 27, 2023 10:45 AM

Midterm will be released on Blackboard

Randomness

Wednesday, March 25, 2009 9:39 AM

- Until our discussion so far, the parameters of a convex optimization problem are assumed to be fixed & known
- We only need to optimize the (unknown) control variables
- Further, in our iterative algorithms, the value of the control variables in the previous iteration is also assumed to be known precisely.
- In reality, however, randomness in the system model & in observation may prevent us from knowing the precise value.

Randomness can exist due to the practical constraints in the protocol.

- e.g. In dual congestion controller each user chooses the rate by solving this problem

$$x_s(t) = \arg\max \; U_s(x_s) - x_s \sum_l H_s^l q_l(t)$$

- We have assumed that source s will know $q_l(t)$

- In practice, in order to avoid additional control messages, the source may need to learn the value of $q_l(t)$ through packet drops (REM).

  - the link drops/marks packet with probability $1-e^{-q_l}$

- $Y_0, Y_1, \cdots Y_n$

$$Y_i = \begin{cases} 1 & \text{if packet } i \text{ is dropped/marked by any link} \\ 0 & \text{otherwise} \end{cases}$$

then $P\{Y_i = 1\} = 1 - e^{-\sum_l H_{ls} q_l}$

$$\Rightarrow \quad \frac{\sum_{i=1}^{n} Y_i}{n} \to 1 - e^{-\sum_l H_{ls} q_l} \quad \text{as } n \to +\infty$$

- However, the source cannot wait for $n \to +\infty$. The control will be too slow.

- For any finite n, the source can only get an estimate of q_l (with random noise)

Randomness could also occur due to the inherent feature of the model.

e.g. The Water-filling problem.

$$\max \sum_{k=1}^{m} q_k \log\left(1 + \frac{g_k P_k}{N}\right)$$

$p_1, \dots, p_m$

$$\text{sub to} \quad \sum_{k=1}^{m} q_k p_k \le P_0$$

where q_k = Probability that the channel gain is g_k.

- In this problem formulation, we have assumed that we know the entire channel distribution.

- In reality, the channel distribution gained needs to be estimated through taking random samples

$\Rightarrow$ error/noise in the system model.

- In principle, we may resolve the randomness in the model by a two-step process:

$$\boxed{\text{Random Observation}} \xrightarrow[\text{as } n \to +\infty]{} \boxed{\text{Estimate quantities of interest}} \xrightarrow[\text{as } t \to +\infty]{} \boxed{\text{Solve the optimization problem}}$$

packet drops
channel condition

price
distribution

- Similarly, we can estimate q_k from measurements

- The problem with this approach is that we need $n \to +\infty$ for the estimate to be accurate (or noise-free); and then we need $t \to +\infty$ for the optimization algorithm to converge to the optimal solution.

- However, in practice it may be unreasonable to assume long estimation time.

  - There may be non-stationary changes in the system which does not allow us to use long estimation time

    e.g. the channel distribution may change.

  - The system may react very slowly.

    e.g. in dual congestion controller, the current price is used only once in each iteration. What is the point of making it very accurate?

Q If the estimation phase must be short, or perhaps need to be completely eliminated, can we still design an efficient algorithm?

$$\boxed{\text{randomness}} \xrightarrow[t \to +\infty]{?} \boxed{\text{optimal solution.}}$$

Estimating the mean

Wednesday, March 25, 2009 10:15 AM

- Let us motivate the proposed algorithm through the simplest estimation problem of estimating the mean of a sequence of i.i.d random variables.

$$\mu = E(X)$$

- Note that this is equivalent to solving the following optimization problem

$$\min_{\mu} E[(X-\mu)^2]$$

- Let

$$f(\mu) = E\left((X-\mu)^2\right) = EX^2 - 2\mu EX + \mu^2$$

- Let us consider an iterative algorithm for solving $f(\mu)$ (Even though it may seem redundant for such a simple problem...)

$$f'(\mu) = -2EX + 2\mu$$

$$\mu(t+1) = \mu(t) - \gamma f'(\mu(t))$$

$$= \mu(t) - 2\gamma(\mu(t) - E(X))$$

We know that $\mu(t) \to \mu = E(X)$ as $t \to +\infty$

- Of course, in reality we do not use the above iterative algorithm to estimate μ.

- Instead, we use

$$\mu(n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

when X_i's are i.i.d, $\mu(n) \to \mu$ as $n \to +\infty$

- Now let us look at this procedure as an iterative algorithm

$$\mu(n+1) = \frac{1}{n+1}\sum_{i=1}^{n+1} X_i$$

$$= \frac{1}{n+1}\cdot\left[n \cdot \mu(n) + X_{n+1}\right]$$

$$= \mu(n) - \frac{1}{n+1}\left[\mu(n) - X_{n+1}\right]$$

- Let us compare it with

$$\mu(t+1) = \mu(t) - 2\gamma\left[\mu(t) - E(X)\right]$$

① We replace the unknown $E(X)$ by the unbiased current observation.

② We replace the constant stepsize by a sequence that diminishes to zero.

But then such an iterative algorithm will converge to the optimal solution!

- It turns out that the stepsize does not need to be $\frac{1}{n+1}$.

$$\mu(n+1) = \mu(n) - a_n\left(\mu(n) - \underline{X}_{n+1}\right)$$

will also work provided that $\{a_n\}$ satisfies certain conditions.

(10)

Stochastic approximation

Wednesday, March 25, 2009 10:26 AM

- This ideas form the basics of stochastic approximation algorithms

- Suppose that we want to minimize a function $f(x)$

  We may use an iterative algorithm

$$X_{n+1} = X_n - \gamma \nabla f(X_n)$$

  then x_n converges to a local minimum of f with appropriate step size.

- Consider now the case where $\nabla f(X_n)$ is corrupted by noise. We then use the following algorithm

$$X_{n+1} = X_n - a_n [\nabla f(X_n) + \underset{\substack{\uparrow \\ \text{noise}}}{W_n}]$$

- Under suitable conditions on a_n & W_n

  - $E[W_n^2] < +\infty$, $E[W_n] = 0$, W_n i.i.d (unbiased)

  - $\sum_{n=1}^{+\infty} a_n \to +\infty$, $\sum_{n=1}^{+\infty} a_n^2 < +\infty$

$n=1 \qquad\qquad n=1$

then $X_n \to$ a local minimum of f.

(Some of these conditions can be further relaxed.)

- Why it should work?
  - when a_n is small, the value of x will remain approximately the same over many time-slots
  - The stochastic approximation algorithm is then able to average out the "noise"
  - Convergence is more likely when the stepsize is small
    - it will however be slower!
  - More on the condition later.

(15)

Water-filling

Wednesday, March 25, 2009 10:31 AM

- Let us now return to the water-filling example and see how we can use the idea of stochastic approximation to develop a solution that combines estimation and optimization in a single step.

- Recall the problem

$$\max_{p_1, \cdots, p_M} \sum_{k=1}^{M} q_k \log\left(1 + \frac{g_k p_k}{N}\right)$$

$$\text{sub to} \quad \sum_{k=1}^{M} q_k p_k \leq P_0$$

- The algorithm (assuming perfect information)

① $$p_k(t) = \begin{cases} \frac{1}{\lambda(t)} - \frac{N}{g_k} & \text{if } \frac{1}{\lambda(t)} - \frac{N}{g_k} \geq 0 \\ 0 & \end{cases}$$

- Since at each time-slot, there is only one possible realization of g, we only need the value of $p_k(t)$ for the index k such that $g(t) = g_k$.

- Hence this equation can be simplified to

$$p(t) = \begin{cases} \frac{1}{\lambda(t)} - \frac{1}{g(t)}, & \text{if } \frac{1}{\lambda(t)} - \frac{1}{g(t)} = 0 \\ 0 & \text{o/w} \end{cases}$$

$$\begin{cases} \quad & 0 \quad , \quad o/w \end{cases}$$

② $$\lambda(t+1) = \left[\lambda(t) + \gamma\left(\sum_{k=1}^{M} p_k(t)\, g_k - P_0\right)\right]^+$$

- Note that it requires knowledge of the entire distribution.

- Instead, let us replace the gradient by an unbiased estimate

$$\Rightarrow \lambda(t+1) = \left[\lambda(t) + a_t\left(\sum_{k=1}^{M} p_k(t)\, \mathbb{1}_{\{g(t)=g_k\}} - P_0\right)\right]^+$$

$$= \left[\lambda(t) + a_t\left(p(t) - P_0\right)\right]^+$$

Ⓠ How does it work?

Ⓐ when $E[p(t)] = \sum_{k=1}^{M} p_k(t)\, g_k > P_0$
even though each iteration may go in the wrong direction, over bigger windows, the value of $\lambda \uparrow$

$$\Rightarrow p(t) \downarrow$$

Benefits

- No need to estimate the channel distributions before hand

- only need to measure current channel state

- If the channel distribution changes, the algorithm automatically adapts to the changes.

- online / adaptive solution.

- Use non-diminishing stepsize. 25

Rate control - skip

Wednesday, March 25, 2009 2:19 PM

- Recall in the dual controller, each user maximizes its net utility

$$x_s(t) = \arg\max \; U_s(x_s) - x_s \sum_l H_{ls} q_l \qquad (*)$$

- It can be implemented by a gradient-ascent iteration

$$\dot{x}_s = U_s'(x_s) - \sum_l H_{ls} q_l$$

- If we only have noise observations of q_l.

For example, in REM, let $Y_n = 1$ if packet n is marked.

$$P\{Y_n = 1\} = 1 - e^{-\sum_l H_{ls} q_l} \approx \sum_l H_{ls} q_l$$

Hence, we can replace the iteration by

$$x_s(n+1) = x_s(n) + a_n \left[U_s'(x_s(n)) - Y_n \right]$$

Note: Such kinds of "hill-climbing" algorithm tend to be more robust to error that the one-time update like (*).

35

Proof of convergence

Wednesday, March 25, 2009 2:23 PM

Theorem: Let $f(x)$ be a convex and differentiable function and θ is the minimum point of f. Assume X_n, W_n, H_n, V_n with

$$X_{n+1} = X_n - a_n(\nabla f(X_n) + W_n)$$

and $W_n = \underbrace{H_n}_{\text{biased noise}} + \underbrace{V_n}_{\text{unbiased noise}}$

where $a_n \in (0,1)$, $a_n \downarrow 0$, $\sum_{n=1}^{+\infty} a_n = +\infty$, and $\sum_{n=1}^{+\infty} a_n^2 < +\infty$. Assume

1. ∇f is bounded
2. $\forall k: \inf\{\langle \nabla f(x), x-\theta\rangle;\ \frac{1}{k} \le \|x-\theta\| \le k\} > 0$

   (strong convexity)
3. $\sum_{n=1}^{+\infty} a_n E\|H_n\| < +\infty$, $\sum_{n=1}^{+\infty} a_n^2 E\|H_n\|^2 < +\infty$

   (biased noise eventually die out)
4. $E[V_n \mid X_1, H_1, V_1, \cdots, X_{n-1}, H_{n-1}, V_{n-1}] = 0$

   $\sum_{n=1}^{+\infty} a_n^2 E\|V_n\|^2 < +\infty$

   (unbiased noise has bounded second moments)

Then $\underline{X}_n \to \theta$ almost surely.

Note: The additional conditions (3) & (4) hold if $H_n = 0$ & V_n is i.i.d with bounded variance.

Sketch of proof:

- For simplicity, consider only the case where $H_n = 0$

- Goal is to separate out the error term due to noise and show that it is small compared to the gradient descent.

Since $X_{n+1} = X_n - a_n(\nabla f(\underline{X}_n) + V_n)$

$$\|X_{n+1} - x^*\|^2 = \|X_n - x^*\|^2 + a_n^2 \|\nabla f(\underline{X}_n)\|^2 + a_n^2 \|V_n\|^2$$

$$- 2 a_n \langle \nabla f(\underline{X}_n), \underline{X}_n - x^* \rangle$$

$$- 2 a_n \langle \underline{X}_n - x^*, V_n \rangle$$

$$+ 2 a_n^2 \langle \nabla f(\underline{X}_n), V_n \rangle$$

Taking expectation conditioned on $\underline{X}_n$

$$E\left(\|X_{n+1} - x^*\|^2 \,\middle|\, X_n\right)$$

$$\leq \|X_n - x^*\|^2 - 2a_n \langle \nabla f(X_n), X_n - x^* \rangle$$

descent

$$+ \quad a_n^2 \|\nabla f(X_n)\|^2 + a_n^2 \|V_n\|^2$$

Taking another expectation

$$E\left(\|X_{n+1} - x^*\|^2\right) \leq E\left(\|X_n - x^*\|^2\right) - 2a_n E\langle \nabla f(X_n), X_n - x^* \rangle$$
$$+ a_n^2 E\|\nabla f(X_n)\|^2 + a_n^2 E\|V_n\|^2$$

Recall that

$$\langle \nabla f(X_n), X_n - x^* \rangle \geq 0$$

Further, since $\sum a_n^2 < +\infty$, $\|\nabla f(X_n)\|$ is bounded

$$\sum a_n^2 E\left[\|\nabla f(X_n)\|^2\right] < +\infty$$

Finally $\qquad \sum a_n^2 E\left(\|V_n\|^2\right) < +\infty$

Hence, using a telescoping argument, we must have

$$E\left[\|X_n - x^*\|\right] \text{ converges to a limit}$$

Since $\langle \nabla f(X_n), X_n - x^* \rangle \geq 0$ & $\sum a_n = +\infty$, we must have

$$\langle \nabla f(X_n), X_n - X^* \rangle \to 0 \quad \text{as } n \to +\infty.$$

This will eventually leads to $X_n \to X^*$.

(See handout StochasticApprox.pdf.)

In summary.

- Need $\sum a_n^2 < +\infty$ so that the noise can eventually be averaged out
    - Step size must be small!
- Need $\sum a_n = +\infty$ so that the iteration will move close to the optimal
    - Step size cannot be too small!

50