EECS 225A Statistical Signal Processing	Lecture 22 - 3/19/2020

Lecture 17: Stochastic Gradient Methods

Lecturer: Jiantao Jiao *Scribe: Yu Gai*

In this lecture, we present modern analysis of stochastic gradient methods (SGM), originally studied by Robbins and Monroe [1] in the literature of stochastic approximation. It is usually called "stochastic gradient descent" in optimization and machine learning literature, but it would be important to keep in mind that it is in general *not* a descent method.

1 Problem Setting

Given a function $F(x, \xi)$, where ξ is a random variable, we would like to solve

$$\min_x f(x) = \min_x \mathbb{E}_\xi F(x, \xi). \tag{1}$$

For each point x we query the oracle, the oracle would first draw a sample of ξ independent of all random variables observed and then output the gradient of $F(x, \xi)$ with respect to x, denoted as $g(x, \xi)$. Clearly we have

$$\begin{aligned} \mathbb{E}[g(x, \xi)] &= \mathbb{E}[\nabla F(x, \xi)] && (2)\\ &= \nabla \mathbb{E}[F(x, \xi)] && (3)\\ &= \nabla f(x). && (4) \end{aligned}$$

It is a quite general framework. In particular, it subsumes the following settings:

- Noisy gradient: suppose we want to minimize a smooth function $f(x)$. At each iteration, we have access to a *noisy* gradient at some point we can query $g_k = \nabla f(x_k) + \omega_k$, where ω_k is some zero-mean random variable that is independent of x_k.
- Incremental gradients: suppose we want to minimize

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x) \tag{5}$$

  then one can define

$$F(x, \xi) = f_\xi(x) \tag{6}$$

  where ξ is uniformly distributed over $[m] = \{1, \ldots, m\}$. It subsumes the case studied by the LMS algorithm in the i.i.d. data sample case, where we assume that the observations $\{(x_i, y_i)\}_{i=1}^m$ are i.i.d..
- Deterministic optimization: ξ is constant with probability one.

The stochastic gradient descent (SGD) algorithm starts with an initial point $x_0 \in \mathbb{R}^n$ and for each $k \geq 0$ computes the iterates

$$x_{k+1} = x_k - h_k g(x_k, \xi_k) \tag{7}$$

Denote by x_* an arbitrary optimal point of our problem and let $f_* = f(x_*)$. We also denote

$$G_k \triangleq g(x_k, \xi_k). \tag{8}$$

The SGD could require significantly less computational resources than GD. Indeed, if we consider the computational cost of evaluating the gradient of a specific f_i in the incremental gradient setting to be a unit, then for each iteration, SGD requires $O(1)$ computation, but GD requires $O(m)$ computation. We discussed the convergence theory of GD in the last lecture, and it would be interesting to investigate the convergence of SGD for the same problem. As we will see, SGD still performs quite well when the function is convex, but it loses the linear convergence property when the function f is assumed to be strongly convex. A good reference for this lecture is [2].

2 Convex function f

We assume that the function f satisfies the following properties:

1. f is convex:
2. F has bounded gradient on average:

$$\mathbb{E}_\xi \|g(x,\xi)\|^2 \leq M^2 \tag{9}$$

for all x.

Theorem 1. *Let $\bar{x} = \frac{1}{T+1}\sum_{k=0}^{T} x_k$ and $h_k \equiv h$. Then for any $h > 0$, we have*

$$\mathbb{E}[f(\bar{x})] - f(x_*) \leq \frac{\|x_0 - x_*\|^2}{2h(T+1)} + \frac{hM^2}{2}. \tag{10}$$

In particular, if $h = \frac{\|x_0 - x_*\|}{M\sqrt{T+1}}$, we have

$$\mathbb{E}[f(\bar{x})] - f(x_*) \leq \frac{\|x_0 - x_*\| M}{\sqrt{T+1}}. \tag{11}$$

A few remarks are in order.

1. In practice $\|x_0 - x_*\|, M$ may not be known, but there are many ways to choose h_k such that $\mathbb{E}[f(\bar{x})] \to f(x_*)$. For example, $h_k = \frac{h_0}{\sqrt{k+1}}$ would work, or a sequence that is square summable but not summable: $\sum_k h_k^2 < \infty, \sum_k h_k = \infty$.
2. In deterministic optimization, we usually would output the point among those we have searched with the smallest function value $\arg\min_k f(x_k)$, but in stochastic optimization we are not able to do that since we cannot evaluate $f(x_k)$ easily for any x_k. That is why in some sense the nice performance of SGD guaranteed by Theorem 1 is surprising, since it only requires $O(1/\epsilon^2)$ queries to achieve ϵ accuracy in expectation, but evaluating $f(x)$ for any single x up to accuracy ϵ would already take $O(1/\epsilon^2)$ queries.
3. As mentioned above, Theorem 1 works various types of step-size rules without knowing the parameters $\|x_0 - x_*\|^2$ or M. However, it is *only* true for the point $\bar{x}$ (and potentially other rules for producing the final estimate) but *not* the final search point x_T. To further illustrate the point, we would like to discuss the classical analysis of the LMS algorithm and show the if the learning rate is too big, then we even cannot guarantee that $\|\mathbb{E}[x_T] - x_*\| \to 0$ as $T \to \infty$.

   Consider $X \sim \mathcal{N}(0, R_X), Y = X^T w_* + v$, where $v \sim \mathcal{N}(0, \sigma_v^2)$, and X is independent of v. At each point $k \geq 0$, we obtain an independent sample (x_k, y_k) from this joint distribution. In population we would like to minimize $\mathbb{E}_{(X,Y)}[(Y - X^Y w)^2]$ as a function of w. Define $\Delta w_k = w_* - w_k$. The LMS algorithm's iterations are

$$w_{k+1} = w_k + 2hx_k(y_k - x_k^T w_k) \tag{12}$$

Then,

$$\begin{aligned}\Delta w_{k+1} &= w^* - (w_k + 2hx_k(y_k - x_k^T w_k)) && (13)\\ &= w^* - (w_k + 2hx_k(x_k^T w_* + v_k - x_k^T w_k)) && (14)\\ &= (I - 2hx_k x_k^T)\Delta w_k - 2hx_k v_k. && (15)\end{aligned}$$

Taking expectations on both sides, we have

$$\mathbb{E}[\Delta w_{k+1}] = (I - 2hR_X)\mathbb{E}[\Delta w_k]. \tag{16}$$

Hence, if $R_X > 0$, then to guarantee that for any initial $\mathbb{E}[\Delta w_0]$, we have $\|\mathbb{E}[\Delta w_k]\| \to 0$, we need to ensure that

$$0 < h < \frac{1}{\lambda_{\max}(R_X)}. \tag{17}$$

We say the LMS algorithm is *stable in the mean* if $\|\mathbb{E}[\Delta w_k]\| \to 0$. Note that in the literature of LMS usually $2h$ is called the learning rate and is denoted as μ.

Proof We expand

$$\begin{aligned}\mathbb{E}\|x_{k+1} - x_*\|^2 &= \mathbb{E}\|x_k - h_k G_k - x_*\|^2 && (18)\\ &= \mathbb{E}\|x_k - x_*\|^2 - 2h_k\mathbb{E}\langle G_k, x_k - x_*\rangle + h_k^2\mathbb{E}\|G_k\|^2 && (19)\\ &\le \mathbb{E}\|x_k - x_*\|^2 - 2h_k\mathbb{E}\langle G_k, x_k - x_*\rangle + h_k^2 M^2 && (20)\\ &\le \mathbb{E}\|x_k - x_*\|^2 - 2h_k\mathbb{E}\langle \nabla f(x_k), x_k - x_*\rangle + h_k^2 M^2. && (21)\end{aligned}$$

Here the last inequality is justified by the following. Because x_k only depends on $x_0, \xi_0, \xi_1, \ldots, \xi_{k-1}$ and $\xi_j \perp \xi_k$ for all $j < k$, $x_k \perp \xi_k$. By the law of iterated expectations,

$$\begin{aligned}\mathbb{E}\langle G_k, x_k - x_*\rangle &= \mathbb{E}_{x_0,\xi_0,\ldots,\xi_{k-1}}[\mathbb{E}_{\xi_k}[\langle G_k, x_k - x_*\rangle | x_0, \xi_0, \ldots, \xi_{k-1}]]\\ &= \mathbb{E}_{x_0,\xi_0,\ldots,\xi_{k-1}}[\langle \nabla f(x_k), x_k - x_*\rangle]\\ &= \mathbb{E}\langle \nabla f(x_k), x_k - x_*\rangle,\end{aligned}$$

Now we have

$$\begin{aligned}\mathbb{E}\|x_{k+1} - x_*\|^2 &\le \mathbb{E}\|x_k - x_*\|^2 - 2h_k\mathbb{E}\langle \nabla f(x_k), x_k - x_*\rangle + h_k^2 M^2 && (22)\\ &\le \mathbb{E}\|x_k - x_*\|^2 - 2h_k\mathbb{E}[f(x_k) - f(x_*)] + h_k^2 M^2. && (23)\end{aligned}$$

where the last inequality follows from the convexity of f.

Consequently,

$$2h_k\mathbb{E}(f(x_k) - f(x_*)) \le \mathbb{E}\|x_k - x_*\|^2 - \mathbb{E}\|x_{k+1} - x_*\|^2 + h_k^2 M^2. \tag{24}$$

Summing both sides over k from 0 to T, we have

$$\sum_{k=0}^{T} 2h_k\mathbb{E}(f(x_k) - f(x_*)) \le \|x_0 - x_*\|^2 + \sum_{k=0}^{T} h_k^2 M^2. \tag{25}$$

Dividing both sides by $\sum_{k=0}^{T} 2h_k$, we have

$$\frac{1}{\sum_{k=0}^{T} h_k}\left(\sum_{k=0}^{T} h_k\mathbb{E}[f(x_k)]\right) - f(x_*) \le \frac{\|x_0 - x_*\|^2 + M^2\sum_{k=0}^{T} h_k^2}{2\sum_{k=0}^{T} h_k}. \tag{26}$$

If we define

$$\bar{x} = \frac{\sum_{k=0}^{T} h_k x_k}{\sum_{k=0}^{T} h_k}, \tag{27}$$

then it follows from the convexity of f and Jensen's inequality that

$$\mathbb{E}[f(\bar{x})] - f(x_*) \leq \frac{\|x_0 - x_*\|^2 + M^2 \sum_{k=0}^{T} h_k^2}{2\sum_{k=0}^{T} h_k}. \tag{28}$$

Now using the assumption that $h_k \equiv h$ and denoting $D = \|x_0 - x_*\|$, we have

$$\mathbb{E}[f(\bar{x})] - f(x_*) \leq \frac{\|x_0 - x_*\|^2 + M^2(T+1)h^2}{2h(T+1)} \tag{29}$$

$$= \frac{D^2}{2h(T+1)} + \frac{M^2 h}{2}. \tag{30}$$

Because of AM-GM inequality $a + b \geq 2\sqrt{ab}$ for any $a, b \geq 0$,

$$\frac{D^2}{2h(T+1)} + \frac{hM^2}{2} \geq \frac{DM}{\sqrt{T+1}}$$

where equality is attained at $h = h^* = \frac{D}{M\sqrt{T+1}}$, one can use h^* to obtain

$$\mathbb{E}f(\bar{x}) - f(x_*) \leq \frac{DM}{\sqrt{T+1}} \tag{31}$$

□

3 Strongly convex function f

As we saw in the last lecture, GD for strongly convex functions could converge much faster compared to the convex case. Unfortunately, in SGD strong convexity does not help as much as we have hoped: it only improves the rate from $O(1/\sqrt{T})$ to $O(1/T)$.

We assume that the function f satisfies the following properties:

1. f is strongly convex:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2. \tag{32}$$

2. F has bounded gradient on average:

$$\mathbb{E}_\xi \left\|g(x, \xi)\right\|^2 \leq M^2 \tag{33}$$

for all x.

Theorem 2. *Suppose f satisfies the assumptions above and $h_k = \frac{1}{\mu(k+1)}$, Then the SGD algorithm satisfies*

$$\frac{1}{T+1}\sum_{k=0}^{T} \mathbb{E}[f(x_k) - f(x_*)] \leq \frac{M^2}{2\mu(T+1)}(1 + \ln(T+1)) \tag{34}$$

$$\mathbb{E}[\|x_k - x_*\|^2] \leq \frac{Q}{k+1} \tag{35}$$

for any $k \geq 0$, where $Q = \max\{\|x_0 - x_*\|^2, \frac{M^2}{\mu^2}\}$.

Unlike Theorem 1, the choice of step-size in Theorem 2 is highly sensitive to misspecifications. We would like to first convince the readers that in general a constant step size (no matter how small) would fail to achieve vanishing $\mathbb{E}[\|x_k - x_*\|^2]$, thereby justifying the difference between the stochastic optimization problem and the deterministic optimization problem.

Again, the classical literature of the LMS algorithm would serve as a good playground. Recall the step (15) in the analysis of the LMS algorithm:

$$\Delta w_{k+1} = (I - 2hx_kx_k^T)\Delta w_k - 2hx_kv_k. \tag{36}$$

Then,

$$\|\Delta w_{k+1}\|^2 = \|(I - 2hx_kx_k^T)\Delta w_k\|^2 + 4h^2v_k^2\|x_k\|^2 - 4hv_kx_k^T(I - 2hx_kx_k^T)\Delta w_k. \tag{37}$$

Taking expectations on both sides, we have

$$\mathbb{E}[\|\Delta w_{k+1}\|^2] = \mathbb{E}[\|(I - 2hx_kx_k^T)\Delta w_k\|^2] + 4h^2\sigma_v^2 \operatorname{Tr}(R_X). \tag{38}$$

One can compute the limiting behavior of $\mathbb{E}[\|\Delta w_{k+1}\|^2]$ exactly [3, Page 462]. We define

$$A(h) = h\sum_{m=1}^{n} \frac{\lambda_m}{1 - 2h\lambda_m}, \tag{39}$$

where $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ are the n eigenvalues of the matrix R_X. It was shown in [3, Page 465][4] that as $k \to \infty$, we have

$$\mathbb{E}[(Y - X^Tw_\infty)^2] - \mathbb{E}[(Y - X^Tw_*)^2] = \mathbb{E}[(Y - X^Tw_\infty)^2] - \sigma_v^2 \tag{40}$$

$$= \sigma_v^2 \frac{A(h)}{1 - A(h)}. \tag{41}$$

Hence, no matter how small h is, as long as it is a constant, we cannot achieve vanishing error. In the LMS literature, people call the LMS filter *mean-square stable* if $\mathbb{E}[(Y - X^Tw_\infty)^2] < \infty$, and it corresponds to the condition

$$A(h) < 1 \tag{42}$$

in this analysis setting. Again, we remind the reader that usually the step-size in LMS literature is denoted by μ and is equal to our $2h$. If the stochastic process (x_k, y_k) is not i.i.d., then the analysis is quite complicated and the readers are referred to [5].

Proof We first prove the bound on function values and then the bound on $\mathbb{E}\|x_k - x_*\|^2$. It follows from the proof of the previous theorem, in particular (22) that

$$\mathbb{E}\,\|x_{k+1} - x_*\|^2 \le \mathbb{E}\,\|x_k - x_*\|^2 - 2h_k\mathbb{E}\,\langle \nabla f(x_k), x_k - x_*\rangle + h_k^2M^2. \tag{43}$$

It follows from the definition of strong convexity that

$$\langle \nabla f(x_k), x_k - x_*\rangle \ge f(x_k) - f(x_*) + \frac{\mu}{2}\|x_k - x_*\|^2. \tag{44}$$

Hence we have

$$\mathbb{E}[f(x_k) - f(x_*)] \le \mathbb{E}\,\langle \nabla f(x_k), x_k - x_*\rangle - \frac{\mu}{2}\mathbb{E}[\|x_k - x_*\|^2] \tag{45}$$

$$\le \frac{\mathbb{E}\|x_k - x_*\|^2 - \mathbb{E}\|x_{k+1} - x_*\|^2}{2h_k} - \frac{\mu}{2}\mathbb{E}[\|x_k - x_*\|^2] + \frac{h_kM^2}{2} \tag{46}$$

$$= \frac{\mu k}{2}\mathbb{E}\|x_k - x_*\|^2 - \frac{\mu(k+1)}{2}\mathbb{E}\|x_{k+1} - x_*\|^2 + \frac{h_kM^2}{2}. \tag{47}$$

Summing both sides over k from 0 to T, we obtain

$$\sum_{k=0}^{T} \mathbb{E}[f(x_k) - f(x_*)] \leq \sum_{k=0}^{T} h_k \frac{M^2}{2} \tag{48}$$

$$= \frac{M^2}{2\mu} \sum_{k=0}^{T} \frac{1}{k+1} \tag{49}$$

$$\leq \frac{M^2}{2\mu}(1 + \ln(T+1)). \tag{50}$$

We now prove the bound on $\mathbb{E}\|x_k - x_*\|^2$.

$$\mathbb{E}\left\|x_{k+1} - x_*\right\|^2 = \mathbb{E}\left\|x_k - h_k G_k - x_*\right\|^2 \tag{51}$$

$$= \mathbb{E}\left\|x_k - x_*\right\|^2 - 2h_k \mathbb{E}\left\langle G_k, x_k - x_*\right\rangle + h_k^2 \mathbb{E}\left\|G_k\right\|^2 \tag{52}$$

$$\leq \mathbb{E}\left\|x_k - x_*\right\|^2 - 2h_k \mathbb{E}\left\langle G_k, x_k - x_*\right\rangle + h_k^2 M^2 \tag{53}$$

$$= \mathbb{E}\left\|x_k - x_*\right\|^2 - 2h_k \mathbb{E}\left\langle \nabla f(x_k), x_k - x_*\right\rangle + h_k^2 M^2 \tag{54}$$

Here the last equality follows from the law of iterated expectations, the independence between x_k and ξ_k for all k, and the fact that G_k is unbiased.

As we have shown in the last lecture regarding strongly convex functions, one has

$$\langle \nabla f(x_k), x_k - x_* \rangle = \langle \nabla f(x_k) - f(x_*), x_k - x_* \rangle \tag{55}$$

$$\geq \mu \|x_k - x_*\|^2. \tag{56}$$

Hence, we have

$$\mathbb{E}\left\|x_{k+1} - x_*\right\|^2 \leq (1 - 2h_k\mu)\|x_k - x_*\|^2 + h_k^2 M^2. \tag{57}$$

We now show that by appropriately choosing the step size h_k we can achieve $O(1/T)$ convergence rate. We claim that

$$\mathbb{E}\|x_k - x_*\|^2 \leq \frac{Q}{k+1}, \tag{58}$$

where $Q = \max\{\|x_0 - x_*\|^2, \frac{M^2}{\mu^2}\}$.

It is clearly true when $k = 0$. Suppose it is true for some $k \geq 0$, then we have

$$\mathbb{E}\|x_{k+1} - x_*\|^2 \leq (1 - 2h_k\mu)\frac{Q}{k+1} + \frac{M^2}{\mu^2(k+1)^2} \tag{59}$$

$$\leq \left(1 - \frac{2}{k+1}\right)\frac{Q}{k+1} + \frac{Q}{(k+1)^2} \tag{60}$$

$$\leq \frac{Q}{k+1}\left(1 - \frac{2}{k+1} + \frac{1}{k+1}\right) \tag{61}$$

$$= \frac{Q}{k+1}\frac{k}{k+1} \tag{62}$$

$$\leq \frac{Q}{k+2}, \tag{63}$$

where in the last step we used the inequality that

$$k(k+2) \leq (k+1)^2. \tag{64}$$

□

References

[1] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[2] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[3] A. H. Sayed, *Fundamentals of adaptive filtering*. John Wiley & Sons, 2003.

[4] J. Foley and F. M. Boland, "A note on the convergence analysis of LMS adaptive filters with Gaussian data," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 7, pp. 1087–1089, 1988.

[5] S. C. Douglas and W. Pan, "Exact expectation analysis of the lms adaptive filter," *IEEE Transactions on Signal Processing*, vol. 43, no. 12, pp. 2863–2871, 1995.