IIIS 2018 Spring: ATCS - Convergence of GD/Stochastic GD/Variance Reduction
Lecture date: May 21, 2018
Instructor: Jian Li **Scribe: Khalid Usman**

1 Notions and Properties

Let us first recall some related notions and properties.

1.1 Convex Optimization

Consider a convex function $f : \mathbb{R}^n \to \mathbb{R}$, optimizer $x^* = arg\min_x f(x)$ satisfies that

$$\begin{cases} \nabla f(x^*) = 0 \\ \nabla^2 f(x^*) \leq 0 \end{cases} \tag{1}$$

1.2 Precedes and Succeeds

$A \preceq B$ means $A - B$ is positive semidefinite (PSD). $A \succeq B$ means $B - A$ is PSD. Notation $\prec$ and $\succ$ means the corresponding matrices are positive definite.

1.3 Taylor Expansion

A function can be Taylor expanded (to the second order)

$$f(x) = f(x_0) + \nabla f(x)^T(x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(y)(x - x_0) \quad \text{for some } y \in [x, x_0] \tag{2}$$

1.4 α-Strongly Convex

The following three definitions of α-strongly convex are equivalent.

1. $\nabla^2 f(x) \succeq \alpha I \quad \forall x$
2. $f(y) \geq f(x) + \nabla f(x)(y - x) + \frac{\alpha}{2}\|y - x\|_2^2$
3. $f(y) - \frac{\alpha}{2}\|y - x\|_2^2$ is convex for all x

1.5 L-Smooth

The following three definitions of L-smooth are equivalent.

1. $\nabla^2 f(x) \preceq LI \quad \forall x$
2. $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$
3. $f(y) \leq f(x) + \nabla f(x)(y - x) + \frac{L}{2}\|y - x\|_2^2$

2 Convergence of Gradient Descent

Theorem 1 (Convergence of fixed step size GD) *If function $f : \mathbb{R}^n \to \mathbb{R}$ is convex, α-strongly convex and L-smooth, when running GD $x^+ \leftarrow x - t\nabla f(x)$ and choose $t = \frac{1}{L}$,*

$$f(x^{(k)}) - f(x^*) \leq (1 - \frac{\alpha}{L})^k (f(x^{(0)}) - f(x^*)) \tag{3}$$

where x^* is minimizer of f.

Proof: The second difinite of α-strongly convex $f(y) \geq f(x) + \nabla f(x)(y-x) + \frac{\alpha}{2}\|y-x\|_2^2$ holds for any x and y, let $y = \tilde{y} = \frac{1}{\alpha}\nabla f(x)$ minimize the RHS. We have

$$f(y) \geq f(x) - \frac{1}{2\alpha}\|\nabla f(x)\|_2^2$$

choose $y = x^*$, we have

$$\|\nabla f(x)\|_2^2 \geq 2\alpha(f(x) - f(x^*)) \tag{4}$$

Consider $x^+ \leftarrow x - t\nabla f(x)$ and the third definition of L-smooth $f(y) \leq f(x) + \nabla f(x)(y-x) + \frac{L}{2}\|y-x\|_2^2$,

$$f(x^+) \leq f(x) - t\|\nabla f(x)\|_2^2 + \frac{L}{2}t^2\|\nabla f(x)\|_2^2$$

choose $t = \frac{1}{L}$, we have

$$f(x^+) - f(x^*) \leq f(x) - f(x^*) - \frac{1}{2L}t^2\|\nabla f(x)\|_2^2$$

Combine (4),

$$f(x^+) - f(x^*) \leq (1 - \frac{\alpha}{L})(f(x) - f(x^*))$$

Recursivly,

$$f(x^{(k)}) - f(x^*) \leq (1 - \frac{\alpha}{L})^k (f(x^{(0)}) - f(x^*)) \tag{5}$$

when $1 - \frac{\alpha}{L} < 1$, this is linear convergence.

□

If we want find $x^{(k)}$ such that $f(x^{(k)}) - f(x^*) < \epsilon$, we only need to run k iterations and $k = O(\log\frac{1}{\epsilon} / \log\frac{1}{1-\frac{\alpha}{L}}) = O(\frac{L}{\alpha}\log\frac{1}{\epsilon})$.

An intuitive view why we need smoothness and strongly convexity requirement is shown as figure 1. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is α-strongly convex and L-smooth, if the condition number $\kappa = \frac{L}{\alpha}$ is too large, the updating trajectory will be zigzag like with poor performance. You can consider a ill-conditioned function defined on $\mathbb{R}^2$ plane $f(x_1, x_2) = x_1^2 + 10000x_2^2$, where $L = 10000$, $\alpha = 1$ and $\kappa = 10000$.

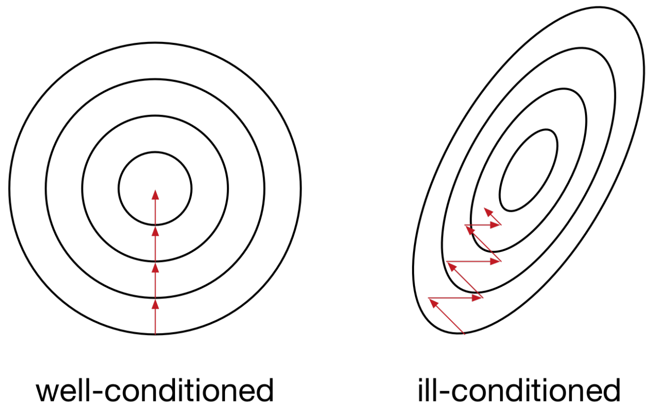

Figure 1: Ill-conditioned function causes poor gradient descent performance

Theorem 2 (Convergence of adaptive step size GD) *If function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-smooth, when running GD $x^{(k+1)} \leftarrow x^{(k)} - t^{(k)}\nabla f(x)$ and $\epsilon_0 \leq t^{(k)} \leq (2-\epsilon_0)\frac{1}{L}$,*

$$\frac{1}{\phi_k} \geq \frac{1}{LR} + k\frac{\epsilon_0^2}{2R^2} \tag{6}$$

where $\phi_k = f(x^{(k)}) - f(x^*)$, R is diameter of $f(x)$, i.e. $R = \max_x \|x - x^*\|$ $\epsilon_0 > 0$ is a constant and x^* is minimizer of f.

Proof: Given L-smoothness,

$$\begin{aligned}
f(x^{(k+1)}) - f(x^{(k)}) &\leq -t^{(k)}\|\nabla f(x)\|_2^2 + \frac{L}{2}\|{t^{(k)}}^2\nabla f(x)\|_2^2 \\
&= (\frac{L}{2}{t^{(k)}}^2 - t^{(k)})\|\nabla f(x)\|_2^2 \\
&= t^{(k)}(\frac{L}{2}t^{(k)} - 1)\|\nabla f(x)\|_2^2 \\
&\leq t^{(k)}(\frac{L}{2}(2-\epsilon_0)\frac{1}{L} - 1)\|\nabla f(x)\|_2^2 \\
&= t^{(k)}(-\frac{\epsilon_0}{2})\|\nabla f(x)\|_2^2 \\
&\leq -\frac{\epsilon_0^2}{2}\|\nabla f(x)\|_2^2
\end{aligned}$$

i.e.

$$\phi_k - \phi_{k+1} \geq \frac{\epsilon_0^2}{2}\|\nabla f(x)\|_2^2 \tag{7}$$

Consider

$$\begin{aligned}\phi_k &= f(x^{(k)}) - f(x^*) \\ &\le \langle \nabla f(x^{(k)}), x^{(k)} - x^* \rangle \\ &\le \|\nabla f(x^{(k)})\| \|x^{(k)} - x^*\| \\ &\le R\|\nabla f(x^{(k)})\|\end{aligned} \tag{8}$$

Put (8) into (7),

$$\phi_k - \phi_{k+1} \ge \frac{\epsilon_0^2}{2}\frac{\phi_k^2}{R^2} \tag{9}$$

$$\frac{1}{\phi_{k+1}} - \frac{1}{\phi_k} = \frac{\phi_k - \phi_{k+1}}{\phi_k \phi_{k+1}} \ge \frac{\phi_k - \phi_{k+1}}{\phi_k^2} \ge \frac{\epsilon_0^2}{2R^2}$$

Recursively,

$$\frac{1}{\phi_k} \ge \frac{1}{\phi_0} + k\frac{\epsilon_0^2}{2R^2} \ge \frac{1}{LR} + k\frac{\epsilon_0^2}{2R^2} \tag{10}$$

where $\phi_0 = f(x^{(0)}) - f(x^*) \le R\|\nabla f(x^{(0)})\| \le LR$ is used in the last inequality.

□

In order to make $\phi_k \le \epsilon$, it is easy to find $k = O(\frac{1}{\epsilon})$ but it's not optimal. When using Nesterov acceleration, $k = O(\frac{1}{\sqrt{\epsilon}})$ can be achieved. In addition if it's not smooth, $k = O(\frac{1}{\epsilon^2})$ can be achieved.

3 Convergence of Stochastice Gradient Descent

We proved the convergence of gradient descent. However, calculating gradient of a function usually involves visit all the data points which is expensive to calculate. The basic idea of *stochastic gradient descent* (SGD) is to use an estimator as a proxy of the gradient, which results in a significantly speed-up of per-iteration cost and does not hurt the number of iterations too much.

Theorem 3 (Convergence of SGD with fixed step size) *A fixed step size SGD* $x_{k+1} \leftarrow x_k - tg_k(x_k)$ *and* $\mathbb{E}[g_k(x)] = \nabla f(x) \quad \forall x$*. When function* $f : \mathbb{R}^n \to \mathbb{R}$ *is* α*-strongly convex and satisfies* $[\|g(x)\|^2] \le M^2$*, number of iterations* $k \sim O(\frac{1}{\epsilon^2})$ *is needed such that* $\mathbb{E}[f(\bar{x}) - f(x^*)] = \mathbb{E}[f(\frac{1}{k}\sum_{i=1}^k x_i) - f(x^*)] < \epsilon$*.*

Proof:

$$\begin{aligned}a_{k+1} \equiv \mathbb{E}[\|x_{k+1} - x^*\|^2] &= \mathbb{E}[\|x_k - tg_k(x_k) - x^*\|^2] \\ &= \mathbb{E}[\|x_k - x^*\|^2] - 2t\mathbb{E}[\langle g_k(x_k), x_k - x^*\rangle] + t^2\mathbb{E}[\|g_k(x_k)\|^2] \\ &\le a_k - 2t\langle \nabla f(x_k), x_k - x^*\rangle + t^2 M^2\end{aligned} \tag{11}$$

$$
\begin{aligned}
\mathbb{E}[f(\frac{1}{k}\sum_{i=1}^{k}x_i) - f(x^*)] &\le \mathbb{E}[\frac{1}{k}\sum_{i=1}^{k}(f(x_i) - f(x^*))] \quad \text{(convexity)} \\
&\le \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[\langle \nabla f(x_k), x_k - x^* \rangle] \quad \text{(convexity again)} \\
&\le \frac{1}{k}\sum_{i=1}^{k}\left(\frac{a_k - a_{k+1}}{2t} + \frac{t}{2}M^2\right) \quad \text{(inequality (11))} \\
&= \frac{a_0 - a_k}{2kt} + \frac{1}{2}tM^2 \\
&\le \frac{\mathbb{E}[\|x_0 - x^*\|^2]}{2kt} + \frac{1}{2}tM^2 \quad \text{(throw } a_k \text{ away)} \\
&\le \frac{f(x_0) - f(x^*)}{\alpha kt} + \frac{1}{2}tM^2 \quad (\alpha\text{-strongly convex})
\end{aligned} \tag{12}
$$

If we want $\mathbb{E}[f(\bar{x}) - f(x^*)] = \mathbb{E}[f(\frac{1}{k}\sum_{i=1}^{k}x_i) - f(x^*)] < \epsilon$, we simply set $t \sim O(\frac{1}{\epsilon})$ and $k \sim O(\frac{1}{\epsilon^2})$.

□

Theorem 4 (Convergence of SGD with adaptive step size) *A SGD $x_{k+1} \leftarrow x_k - t_k g_k(x_k)$ and $\mathbb{E}[g_k(x)] = \nabla f(x) \quad \forall x$ with adaptive step size $t_k = \frac{1}{\alpha k}$. When function $f : \mathbb{R}^n \to \mathbb{R}$ is α-strongly convex and satisfies $[\|g(x)\|^2] \le M^2$, number of iterations $k \sim O(\frac{1}{\epsilon})$ is needed such that $\mathbb{E}[\|x_k - x^*\|^2] < \epsilon$.*

Proof:

Due to α-strongly convexity,

$$f(x^*) - f(x_k) \ge \langle \nabla f(x_k), x^* - x_k \rangle + \frac{\alpha}{2}\|x_k - x^*\|^2 \tag{13}$$

$$f(x_k) - f(x^*) \ge \langle \nabla f(x^*), x_k - x^* \rangle + \frac{\alpha}{2}\|x_k - x^*\|^2 \tag{14}$$

Add (14) to (13), we have

$$\langle \nabla f(x_k) - \nabla f(x^*), x^* - x_k \rangle + \alpha\|x_k - x^*\|^2 \le 0 \tag{15}$$

$$\langle \nabla f(x_k), x_k - x^* \rangle \ge \alpha\|x_k - x^*\|^2 \tag{16}$$

Like what we do in (11), we can get

$$
\begin{aligned}
\mathbb{E}[\|x_{k+1} - x^*\|^2] &\le \mathbb{E}[\|x_k - x^*\|^2] - 2t_k\langle \nabla f(x_k), x_k - x^* \rangle + t_k^2 M^2 \\
&\le \mathbb{E}[\|x_k - x^*\|^2] - 2\alpha t_k\|x_k - x^*\|^2 + t_k^2 M^2 \quad \text{(use (16))} \\
&= (1 - 2\alpha t_k)\mathbb{E}[\|x_k - x^*\|^2] + t_k^2 M^2
\end{aligned} \tag{17}
$$

Define $H = \max(\|x_0 - x^*\|, \frac{M^2}{\alpha^2})$. We want to proove $\mathbb{E}[\|x_k - x^*\|^2] \leq \frac{H}{k}$ by induction. Obviously, $\mathbb{E}[\|x_0 - x^*\|^2]$ satisfies. Suppose $\mathbb{E}[\|x_k - x^*\|^2] \leq \frac{H}{k}$. (17) can be written as

$$\begin{aligned}\mathbb{E}[\|x_{k+1} - x^*\|^2] &\leq (1 - 2\alpha t_k)\mathbb{E}[\|x_k - x^*\|^2] + t_k^2 M^2 \\ &\leq (1 - 2\frac{1}{k})\frac{H}{k} + \frac{H}{k} \quad (\text{use } t_k = \frac{1}{\alpha k} \text{ and definition of } H) \\ &= \frac{k-1}{k^2} H \leq \frac{H}{k+1}\end{aligned} \tag{18}$$

By induction, we can conclude that $\mathbb{E}[\|x_k - x^*\|^2] \leq \frac{H}{k}$. There exists $k \sim O(\frac{1}{\epsilon})$, such that $\mathbb{E}[\|x_k - x^*\|^2] \leq \frac{H}{k} < \epsilon$.

□

4 Stochastic Variance Reduced Gradient (SVRG)

Suppose a function to optimize is the sum of n functions $P(w) = \frac{1}{n}\sum_{i=1}^{n} \psi_i(w)$. SGD updating rule is $w_{t+1} \leftarrow w_t - \eta \nabla \psi_{i_t}(w_t)$, where $i_t \sim [n]$ is random chosen which induces large variance and subsequently slows down the convergence.

The idea of SVRG[1] is to use a occasionally updated estimate $\tilde{w}$ to compensate for the randomness of choosing ψ_{i_t}. Figure 2 sketches the intuition of variance reduction.

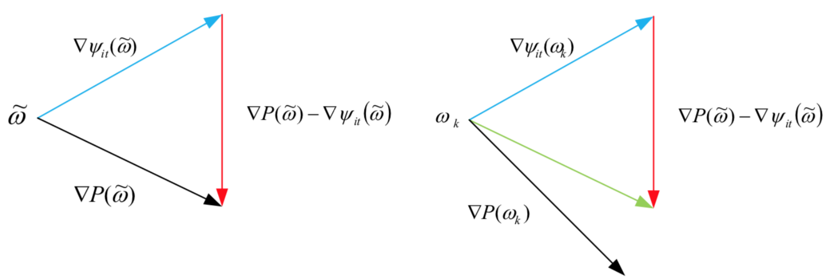

Figure 2: Intuition of SVRG (from [2])

The algorithm is shown as follows.

Firstly, we can see that the expectation of the new gradient term equals to the true gradient.

$$\mathbb{E}_{i_t}[(\nabla \psi_{i_t}(w_{t-1}) - \nabla \psi_{i_t}(\tilde{w}) + \tilde{\mu})] = \nabla P(w_{t-1}) - \nabla P(\tilde{w}) + \nabla P(\tilde{w}) = \nabla P(w_{t-1})$$

Then we can briefly bound the variance (for complete proof, see [2]). Let $v_t = \nabla \psi_{i_t}(w_{t-1}) - \nabla \psi_{i_t}(\tilde{w}) + \tilde{\mu}$.

Algorithm 1: SVRG

for $s = 1, 2, \dots$ *epochs* **do**
  $\tilde{w} = \tilde{w}_{s-1}$;
  $\tilde{\mu} = \frac{1}{n} \sum_{i=1}^{k} \psi_i(\tilde{w}) = \nabla P(\tilde{w})$;
  $w_0 = \tilde{w}$;
  for $t = 1, 2, \dots, m$ **do**
    $w_t = w_{t-1} - \eta(\nabla \psi_{i_t}(w_{t-1}) - \nabla \psi_{i_t}(\tilde{w}) + \tilde{\mu})$ where $i_t \sim [n]$;
  option 1: $\tilde{w}_s = w_t$ where $t \overset{unif.}{\sim} [m]$;
  option 2: $\tilde{w}_s = w_m$;

$$\begin{aligned}
\mathbb{E}[\|v_t\|^2] &= \mathbb{E}[\|\nabla \psi_{i_t}(w_{t-1}) - \nabla \psi_{i_t}(\tilde{w}) + \tilde{\mu}\|^2] \\
&= \mathbb{E}[\|\nabla \psi_{i_t}(w_{t-1}) - \nabla \psi_{i_t}(w^*) + \nabla \psi_{i_t}(w^*) - \nabla \psi_{i_t}(\tilde{w}) + \tilde{\mu}\|^2] \\
&\leq 2\mathbb{E}[\|\nabla \psi_{i_t}(w_{t-1}) - \nabla \psi_{i_t}(w^*)\|^2] + 2\mathbb{E}[\|\nabla \psi_{i_t}(\tilde{w}) - \nabla \psi_{i_t}(w^*) - \tilde{\mu}\|^2] \\
&\leq 2\mathbb{E}[\|\nabla \psi_{i_t}(w_{t-1}) - \nabla \psi_{i_t}(w^*)\|^2] + 2\mathbb{E}[\|\nabla \psi_{i_t}(\tilde{w}) - \nabla \psi_{i_t}(w^*) - \mathbb{E}[\nabla \psi_{i_t}(\tilde{w}) - \psi_{i_t}(w^*)]\|^2] \\
&\leq 2\mathbb{E}[\|\nabla \psi_{i_t}(w_{t-1}) - \nabla \psi_{i_t}(w^*)\|^2] + 2\mathbb{E}[\|\nabla \psi_{i_t}(\tilde{w}) - \nabla \psi_{i_t}(w^*)\|^2] \\
&\leq 4L[P(w_{t-1}) - P(w^*) + P(\tilde{w}) - P(w^*)]
\end{aligned} \tag{19}$$

The first inequality uses $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$; the second inequality uses $\mathbb{E}[\|X - \mathbb{E}X\|^2] \leq \mathbb{E}[\|X\|^2]$; the last is due to L-smoothness of function $\psi_i(w)$. When w gets closer to w^*, the variance gets closer to zero.

Acknowledgement

Revision by Zhang Chuheng based on Niu Hui's handwritten lecture notes.

References

[1] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[2] Tong Zhang Rie Johnson. Stochastic gradient descent with variance reduction. `http://ranger.uta.edu/~heng/CSE6389_15_slides/SGD2.pdf`, 2015.