Optimization Theory and Algorithm II **October 10, 2022**

Lecture 9

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

1 SGD

Non-convex and β-smooth objective functions:

SGD is a commonly accepted method for training deep neural networks, which are usually non-convex and smooth optimization problems. For GD, we have known that

$$\min_{0\leq t\leq T-1} \|\nabla f(\mathbf{x}^t)\| = O(\frac{1}{\sqrt{T}}).$$

What about SGD?

Theorem 1 *(Fixed Learning Rate)*

Suppose that A1 and A2 hold. Let $s_t = s \in (0, 1/\beta]$, then

$$\mathbb{E}[1/T \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}^t)\|^2] \leq s\beta\sigma^2 + \frac{2(f(\mathbf{x}^0) - f^*)}{Ts}.$$

Proof 1 *Based on the Lemma in Lecture 8,*

$$\begin{aligned}\mathbb{E}_{i_t}[f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t)] &\leq \frac{\beta s_t^2}{2}\sigma^2 - s_t(1 - \frac{\beta s_t}{2})\|\nabla f(\mathbf{x}^t)\|^2, \\ &\leq \frac{\beta s^2}{2}\sigma^2 - \frac{s}{2}\|\nabla f(\mathbf{x}^t)\|^2.\end{aligned}$$

Take the expectation over all indices, then

$$\mathbb{E}[f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t)] \leq \frac{\beta s^2}{2}\sigma^2 - \frac{s}{2}\mathbb{E}[\|\nabla f(\mathbf{x}^t)\|^2].$$

Thus,

$$f^* - f(\mathbf{x}^0) \leq \mathbb{E}[f(\mathbf{x}^T) - f(\mathbf{x}^0)] \leq -\frac{s}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\mathbf{x}^t)\|^2] + \frac{Ts^2\beta}{2}\sigma^2.$$

Then,

$$\mathbb{E}[1/T \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}^t)\|^2] \leq s\beta\sigma^2 + \frac{2(f(\mathbf{x}^0) - f^*)}{Ts}.$$

In addition, it has

$$\mathbb{E}[\min_{0\leq t\leq T-1} \|\nabla f(\mathbf{x}^t)\|^2] \leq s\beta\sigma^2 + \frac{2(f(\mathbf{x}^0) - f^*)}{sT}.$$

Remark 1 *Consider for SGD,*

$$\mathbb{E}[\min_{0\leq t\leq T-1} \|\nabla f(\mathbf{x}^t)\|] = O(\sigma + \sqrt{\frac{1}{T}}). \tag{1}$$

For GD, we has

$$\min_{0\le t\le T-1} \|\nabla f(\mathbf{x}^t)\| = O(\sqrt{\frac{1}{T}}). \tag{2}$$

Theorem 2 *(Non-fixed Learning Rate)*

Suppose that A1 and A2 hold. Let $s_t \in (0, 1/\beta]$ for all t, and $\sum_t s_t = \infty, \sum_t s_t^2 < \infty$. Then,

$$\mathbb{E}[\frac{1}{\sum_{t=0}^{T-1} s_t} \sum_{t=0}^{T-1} s_t \|\nabla f(\mathbf{x}^t)\|^2] \to 0,$$

as $T \to \infty$.

Proof 2 *Similar to the previous theorem,*

$$\mathbb{E}_{i_t}[f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t)] \le \frac{\beta s_t^2}{2}\sigma^2 - \frac{s_t}{2}\|\nabla f(\mathbf{x}^t)\|^2.$$

Then, take the expectation over all indices, then

$$\mathbb{E}[f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t)] \le \frac{\beta s_t^2}{2}\sigma^2 - \frac{s_t}{2}\|\mathbb{E}[\nabla f(\mathbf{x}^t)\|^2].$$

Thus,

$$\mathbb{E}[f(\mathbf{x}^T) - f(\mathbf{x}^0)] \le \frac{\beta\sigma^2}{2}\sum_{t=0}^{T-1} s_t^2 - \frac{1}{2}\sum_{t=0}^{T-1} s_t \mathbb{E}[\|\nabla f(\mathbf{x}^t)\|^2].$$

$$\begin{aligned}\frac{1}{2}\sum_{t=0}^{T-1} s_t \mathbb{E}[\|\nabla f(\mathbf{x}^t)\|^2] &\le \mathbb{E}[f(\mathbf{x}^0) - f(\mathbf{x}^T)] + \frac{\beta\sigma^2}{2}\sum_{t=0}^{T-1} s_t^2 \\ &\le f(\mathbf{x}^0) - f(\mathbf{x}^*) + \frac{\beta\sigma^2}{2}\sum_{t=0}^{T-1} s_t^2.\end{aligned}$$

Therefor,

$$\lim_{T\to\infty} \sum_{t=0}^{T-1} s_t \mathbb{E}[\|\nabla f(\mathbf{x}^t)\|^2] < \infty,$$

and

$$\mathbb{E}[\frac{1}{\sum_{t=0}^{T-1} s_t} \sum_{t=0}^{T-1} s_t \|\nabla f(\mathbf{x}^t)\|^2] \to 0.$$

Recall that, we have shown that GD for strong convex and smooth objective function has

$$\|\mathbf{x}^T - \mathbf{x}^*\|^2 = O(\exp(-T)), \text{ and } f(\mathbf{x}^T) - f(\mathbf{x}^*) = O(\exp(-T)).$$

What about SGD??

Theorem 3 *(Fixed Learning Rate)*

Assume that A1, A2 and A3 holds and $s_t = s \in (0, 1/\beta]$ for all t, then

$$\mathbb{E}[f(\mathbf{x}^T) - f^*] \le \frac{s\beta\sigma^2}{2\alpha} + \exp(-\alpha s T)(f(\mathbf{x}^0) - f(\mathbf{x}^*)).$$

Proof 3 *Based on Lemmas in lecture 8*

$$\begin{aligned}\mathbb{E}_{i_t}[f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t)] &\le \frac{\beta s_t^2}{2}\sigma^2 - s_t(1 - \frac{\beta s_t}{2})\|\nabla f(\mathbf{x}^t)\|^2, \\ &\le \frac{\beta s^2}{2}\sigma^2 - \frac{s}{2}\|\nabla f(\mathbf{x}^t)\|^2 \\ &\le \frac{\beta s^2}{2}\sigma^2 - \alpha s(f(\mathbf{x}^t) - f^*).\end{aligned}$$

Then,

$$\mathbb{E}_{i_t}[f(\mathbf{x}^{t+1}) - f^*] + f^* - f(\mathbf{x}^t) \le \frac{\beta s^2}{2}\sigma^2 - \alpha s(f(\mathbf{x}^t) - f^*),$$

thus,

$$\mathbb{E}_{i_t}[f(\mathbf{x}^{t+1}) - f^*] \le \frac{\beta s^2}{2}\sigma^2 + (1 - \alpha s)(f(\mathbf{x}^t) - f^*).$$

Moreover,

$$\begin{aligned}\mathbb{E}_{i_t}[f(\mathbf{x}^{t+1}) - f^*] - \frac{s\beta}{2\alpha}\sigma^2 &\le \frac{\beta s^2}{2}\sigma^2 - \frac{s\beta}{2\alpha}\sigma^2 + (1 - \alpha s)(f(\mathbf{x}^t) - f^*) \\ &= (1 - \alpha s)(f(\mathbf{x}^t) - f^* - \frac{s\beta}{2\alpha}\sigma^2).\end{aligned}$$

Take all expectation for the indices, then

$$\mathbb{E}[f(\mathbf{x}^{t+1}) - f^*] - \frac{s\beta}{2\alpha}\sigma^2 \le (1 - \alpha s)(\mathbb{E}[f(\mathbf{x}^t) - f^*] - \frac{s\beta}{2\alpha}\sigma^2).$$

Thus,

$$\begin{aligned}\mathbb{E}[f(\mathbf{x}^T) - f^*] &\le \frac{s\beta}{2\alpha}\sigma^2 + (1 - \alpha s)^T(f(\mathbf{x}^0) - f^* - \frac{s\beta}{2\alpha}\sigma^2) \\ &\le \frac{s\beta\sigma^2}{2\alpha} + \exp(-\alpha sT)(f(\mathbf{x}^0) - f(\mathbf{x}^*)).\end{aligned}$$

Theorem 4 *(SGD with diminishing learning rate)*

Suppose that A1, A2 and A3 hold, and s_t satisfies $\sum_t s_t = \infty$ and $\sum_t s_t^2 < \infty$. For example, we set $s_t = \frac{\ell}{\gamma+t}, \ell > 1/\alpha, \gamma > 0$ and $s_0 = \frac{\ell}{\gamma} \le 1/\beta$. Then

$$\mathbb{E}[f(\mathbf{x}^T) - f^*] \le \frac{\nu}{\gamma + T}, \tag{3}$$

where $\nu = \max\{\gamma(f(\mathbf{x}^0) - f^*), \frac{\ell^2\beta\sigma^2}{2(\ell\alpha-1)}\}$.

Proof 4 *Based on lemmas in lecture 8 and fact $1 - \frac{\beta s_t^2}{2} \le 1 - \frac{\beta s_0^2}{2} = 1/2$, then*

$$\mathbb{E}_{i_t}[f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t)] \le \frac{\beta s_t^2}{2}\sigma^2 - \alpha s_t(f(\mathbf{x}^t) - f^*).$$

Then,

$$\mathbb{E}_{i_t}[f(\mathbf{x}^{t+1}) - f^*] \le \frac{\beta s_t^2}{2}\sigma^2 + (1 - \alpha s_t)(f(\mathbf{x}^t) - f^*).$$

Take all expectations, it has

$$\mathbb{E}[f(\mathbf{x}^{t+1}) - f^*] \le \frac{\beta s_t^2}{2}\sigma^2 + (1 - \alpha s_t)\mathbb{E}[(f(\mathbf{x}^t) - f^*)].$$

Let us prove the final results by induction, for $t = 0$

$$\mathbb{E}[f(\mathbf{x}^0) - f^*] = \frac{\gamma}{\gamma+0}(f(\mathbf{x}^0) - f^*) \leq \frac{\nu}{\gamma+0},$$

by the definition of ν.

Suppose that holds for $t > 0$, then

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}^{t+1}) - f^*] &\leq \frac{\beta s_t^2}{2}\sigma^2 + (1-\alpha s_t)\mathbb{E}[(f(\mathbf{x}^t) - f^*)] \\
&\leq \frac{\beta s_t^2}{2}\sigma^2 + (1-\alpha s_t)\frac{\nu}{\gamma+t} \\
&= \frac{\beta\sigma^2\ell^2}{2(\gamma+t)^2} + (1 - \frac{\alpha\ell}{\gamma+t})\frac{\nu}{\gamma+t} \\
&= \frac{(\gamma+t-1)\nu}{(\gamma+t)^2} - \frac{(\alpha\ell-1)\nu}{(\gamma+t)^2} + \frac{\beta\sigma^2\ell^2}{2(\gamma+t)^2}.
\end{aligned}$$

Due to the facts

$$\frac{\beta\sigma^2\ell^2}{2} - (\alpha\ell-1)\nu \leq \frac{\beta\sigma^2\ell^2}{2} - \frac{\beta\sigma^2\ell^2(\alpha\ell-1)}{2(\ell\alpha-1)} = 0,$$

and

$$(\gamma+t)^2 \geq (\gamma+t+1)(\gamma+t-1) = (\gamma+t)^2 - 1,$$

then

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}^{t+1}) - f^*] &\leq \frac{(\gamma+t-1)\nu}{(\gamma+t)^2} \\
&\leq \frac{\nu}{\gamma+t+1}.
\end{aligned}$$

Remark 2

- *From the result, we see that choosing a decreasing learning rate results in a sublinear convergence rate, which is worse that is worse than the SGD with constant learning rate. However, note that such a choice enables to reach any neighborhood of the optimal values.*
- *The similar result*

  $$\mathbb{E}[f(\mathbf{x}^T) - f^*] \leq O(\|\mathbf{x}^0 - \mathbf{x}^*\| \exp(-\frac{\alpha T}{\beta}) + \frac{\sigma^2}{\alpha^2 T})$$

  can be found in [1].
- *For only the convex function, SGD has the property*

  $$\mathbb{E}[f(\mathbf{x}^T) - f^*] = O(1/\sqrt{T}).$$

  See Theorem 8.18 on Page 475 of Textbook.

1.0.1 Extensions

- Momentum Method:

  $$\begin{aligned}
  \mathbf{x}^{t+1} &= \mathbf{x}^t + \mathbf{v}^{t+1}, \\
  \mathbf{v}^{t+1} &= \mu_t\mathbf{v}^t - s_t\nabla f_{i_t}(\mathbf{x}^t).
  \end{aligned}$$

  This means

  $$\mathbf{x}^{t+1} = \mathbf{x}^t - s_t\nabla f_{i_t}(\mathbf{x}^t) + \mu_t\underbrace{(\mathbf{x}^t - \mathbf{x}^{t-1})}_{\text{momentum}}.$$

- Nesterov Accelerate Method:

$$\begin{aligned}\mathbf{y}^{t+1} &= \mathbf{x}^t + \mu_t(\mathbf{x}^t - \mathbf{x}^{t-1}),\\ \mathbf{x}^{t+1} &= \mathbf{y}^{t+1} - s_t \nabla f_{i_t}(\mathbf{y}^{t+1}).\end{aligned}$$

  This means

$$\mathbf{x}^{t+1} = \mathbf{x}^t - s_t \nabla f_{i_t}(\mathbf{y}^{t+1}) + \mu_t \underbrace{(\mathbf{x}^t - \mathbf{x}^{t-1})}_{\text{momentum}}$$

  and $\mu_t = \frac{t-1}{t+2}$.

- AdaGrad:

$$\begin{aligned}\mathbf{x}^{t+1} &= \mathbf{x}^t - \frac{s_t}{\sqrt{G^t + \epsilon \mathbb{1}_n}} \otimes \mathbf{g}^t,\\ G^{t+1} &= G^t + \mathbf{g}^t \otimes \mathbf{g}^t,\end{aligned}$$

  where $\mathbf{g}^t = \nabla f_{i_t}(\mathbf{x}^t)$.

- RMSProp:

$$\begin{aligned}\mathbf{x}^{t+1} &= \mathbf{x}^t - \frac{s_t}{R^t} \otimes \mathbf{g}^t,\\ M^{t+1} &= \rho M^t + (1-\rho)\mathbf{g}^t \otimes \mathbf{g}^t,\\ R^{t+1} &= \sqrt{M^{t+1} + \epsilon \mathbb{1}_n}.\end{aligned}$$

- Adam:

$$\begin{aligned}S^{t+1} &= \rho_1 S^t + (1-\rho_1)\mathbf{g}^t,\\ M^{t+1} &= \rho_2 M^t + (1-\rho_2)\mathbf{g}^t \otimes \mathbf{g}^t,\\ \mathbf{x}^{t+1} &= \mathbf{x}^t - \frac{s_t}{\sqrt{\widetilde{M}^t + \epsilon \mathbb{1}_n}} \otimes \widetilde{S}^t,\end{aligned}$$

  where $\widetilde{S}^t = \frac{S^t}{1-\rho_1}$ and $\widetilde{M}^t = \frac{M^t}{1-\rho_2}$.

References

[1] Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.