ECE 901: Large-scale Machine Learning and Optimization **Spring 2018**

Lecture 9 — 02/22

Lecturer: Dimitris Papailiopoulos *Scribe: Guangtong Bai, Yuan-Ting Hsieh*

Note: These lecture notes are still rough, and have only have been mildly proofread.

9.1 Introduction

So far, we have seen the convergence properties of a handful of gradient-based methods on **convex** problems. However, in practice, we need to deal with problems with **non-convexity**, which implies hardness. In this lecture, we will mainly look at the convergence of Stochastic Gradient Descent (SGD) method on non-convex problems.

Q: What can we say about general non-convex problems?
A: Not much. As $f(\mathbf{w})$ can have exponential number of local minimas and saddle points. However, we can show the convergence to critical points, or add structure and prove convergence bounds for some non-convex losses.

Q: Why can not we show a global convergence bound?
A: In general, $\min_{\mathbf{w}} f(\mathbf{w})$ can be a *NP-Hard* problem.

9.2 SGD on Smooth Non-convex Functions

As we pointed out in the previous question, we can add structures to non-convex problems to give some guarantees. We start with adding β-smoothness.

9.2.1 Reminder: β-smooth Functions

Theorem 9.1 (β-smoothness). *f is β-smooth if*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \beta \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \tag{9.1}$$

This implies that:

$$|f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle| \leq \frac{\beta}{2} \|\mathbf{x} - \mathbf{y}\|^2, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \tag{9.2}$$

9.2.2 Convergence of SGD on β-smooth Functions

Let $\mathbf{x} = \mathbf{w}_{k+1}$, $\mathbf{y} = \mathbf{w}_k$, and update rule of SGD be:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \gamma \nabla f_{s_k}(\mathbf{w}_k), s_k \sim uniform(1, 2, .., n) \tag{9.3}$$

Then the smoothness implies that:

$$f(\mathbf{w}_{k+1}) - f(\mathbf{w}_k) - \langle \nabla f(\mathbf{w}_k), \mathbf{w}_{k+1} - \mathbf{w}_k \rangle \;\; \leq \;\; \frac{\beta}{2} \left\| \mathbf{w}_k - \mathbf{w}_{k+1} \right\| \tag{9.4}$$

Let f_k denotes $f(\mathbf{w}_k)$ and substitute Eq. 9.3 into above we get:

$$f_{k+1} - f_k + \gamma \langle \nabla f(\mathbf{w}_k), \nabla f_{s_k}(\mathbf{w}_k) \rangle \;\; \leq \;\; \frac{\beta}{2} \gamma^2 \left\| \nabla f_{s_k}(\mathbf{w}_k) \right\|^2 \tag{9.5}$$

Take expected values on both sides and assume $\left\| \nabla f_{s_k}(\mathbf{w}_k) \right\|^2 \leq M^2$ then we have:

$$\mathbb{E}[f_{k+1} - f_k] + \gamma \mathbb{E} \left\| \nabla f(\mathbf{w}_k) \right\|^2 \;\; \leq \;\; \gamma^2 M^2 \frac{\beta}{2} \tag{9.6}$$

$$\mathbb{E} \left\| \nabla f(\mathbf{w}_k) \right\|^2 \;\; \leq \;\; \frac{\mathbb{E}[f_k - f_{k+1}]}{\gamma} + \gamma M^2 \frac{\beta}{2} \tag{9.7}$$

Now we apply Eq. 9.7 over and over again:

$$\mathbb{E} \left\| \nabla f(\mathbf{w}_1) \right\|^2 \;\; \leq \;\; \frac{\mathbf{E}(f_1 - f_2)}{\gamma} + \frac{\gamma M^2 \beta}{2} \tag{9.8}$$

$$\mathbb{E} \left\| \nabla f(\mathbf{w}_2) \right\|^2 \;\; \leq \;\; \frac{\mathbf{E}(f_2 - f_3)}{\gamma} + \frac{\gamma M^2 \beta}{2} \tag{9.9}$$

$$\vdots \tag{9.10}$$

Sum the above up, and ignore the denominator 2, we have:

$$\sum_{t=1}^{T} \mathbb{E} \left\| \nabla f(\mathbf{w}_t) \right\|^2 \;\; \leq \;\; \frac{f_1 - f^*}{\gamma} + T \gamma M^2 \beta \tag{9.11}$$

Divide both sides by T, we have:

$$\min_{t=1:T} \mathbb{E} \left\| \nabla f(\mathbf{w}_t) \right\|^2 \leq \frac{\sum_{t=1}^{T} \mathbb{E} \left\| \nabla f(\mathbf{w}_t) \right\|^2}{T} \;\; \leq \;\; \frac{f_1 - f^*}{\gamma T} + \gamma M^2 \beta \tag{9.12}$$

Let us assume that $f_1 - f^* \leq D$. Then if we set $\gamma = \sqrt{\frac{D}{\beta M^2 T}}$, we can get:

$$\min_{t} \mathbb{E} \left\| \nabla f(\mathbf{w}_t) \right\|^2 \leq 2 \sqrt{\frac{D \beta M^2}{T}} \tag{9.13}$$

The above convergence bound tells that: no matter what $f(\mathbf{w})$ looks like, if it is smooth, it can reach an "approximate" critical point in $\mathcal{O}(\frac{1}{\sqrt{T}})$ iterations.

However, the above bound seems to be too **pessimistic** to explain the practical performance of SGD. As in real-life problems, both β and M are usually proportional to D, making the bound proportional to D^2, which can be large.

Fortunately, in real life, non-convex functions are usually easier. In next section, we will show that by adding a bit more structure, we can get better results.

9.3 More Structure: Polyak-Łojasiewicz Functions

9.3.1 Polyak-Łojasiewicz Functions

Theorem 9.2 (Polyak-Lojasiewicz (PL) Condition). *A function is μ-PL if*

$$\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*) \tag{9.14}$$

Remarks:

- This implies that if $\nabla f(x) = 0$ then we are at the global minima.
- If a function is both PL and convex then it is strongly convex.
- For PL functions, all local minima are global minima.
- Usually, Neural Networks' loss functions, when near the global minima, are PL.

9.3.2 Convergence of SGD on PL functions

Starting from 9.6, substitute PL-condition 9.2 into the equation, we get:

$$\begin{aligned}
\mathbb{E}[f_{k+1} - f^*] &\leq \mathbb{E}[f_k - f^*] - \gamma\mu\mathbb{E}[f_k - f^*] + \gamma^2 M^2 \frac{\beta}{2} && (9.15)\\
&\leq (1 - \gamma\mu)\mathbb{E}[f_k - f^*] + \gamma^2 M^2 \frac{\beta}{2} && (9.16)\\
&\leq (1 - \gamma\mu)^2\mathbb{E}[f_{k-1} - f^*] + \sum_{t=0}^{1}(1 - \gamma\mu)^t\gamma^2 M^2 \frac{\beta}{2} && (9.17)\\
&\ldots && (9.18)\\
&\leq (1 - \gamma\mu)^{k+1}\mathbb{E}[f_0 - f^*] + \sum_{t=0}^{k}(1 - \gamma\mu)^t\gamma^2 M^2 \frac{\beta}{2} && (9.19)\\
&\leq (1 - \gamma\mu)^{k+1}\mathbb{E}[f_0 - f^*] + \frac{\gamma M^2 \beta}{2\mu} && (9.20)
\end{aligned}$$

To make this value less than or equal to ϵ, one way is to let both terms less than or equal to $\frac{\epsilon}{2}$. This gives us:

$$\gamma = \frac{\epsilon\mu}{\beta M^2} \tag{9.21}$$

$$T_\epsilon = O(\frac{\beta M^2}{\mu^2}\frac{1}{\epsilon} log \frac{\mathbb{E}[f_0 - f^*]}{\epsilon}) \tag{9.22}$$

This convergence rate is much faster than if we only have smoothness.

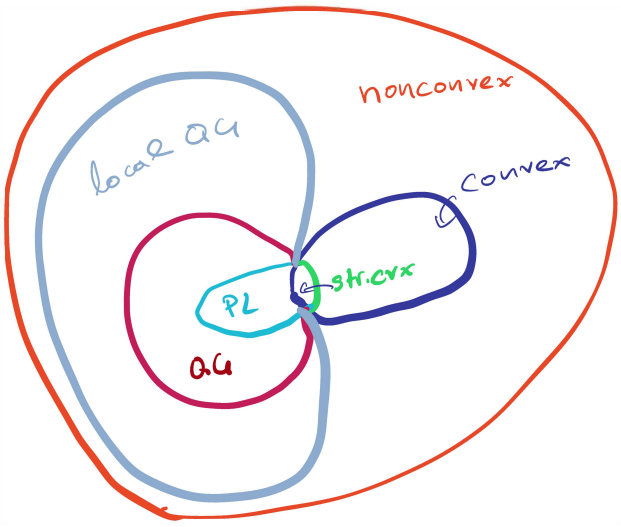
9.4 Quadratic Growth Functions and Beyond

9.4.1 Quadratic Growth Functions

Theorem 9.3 (Quadratic Growth). *A function is quadratic growth if*

$$f(w) - f^* \geq \mu \|w - \prod_{w^*}(w)\|^2 \tag{9.23}$$

Unfortunately, no convergence guarantees known for quadratic growth. Also, it is unclear how relevant it is in practice.
Figure below roughly shows the relationship of functions with different convexity.



Some open problems related to this lecture are:

1. What are the interesting non-convex functions where we can show their convergence bounds?
2. Do Neural Networks have as many local minima as global minima?
3. How large is the sub-optimality gap between local and global minima?

In next lecture, we will be talking about how to choose step size in practice.