CS2910 - Optimization Summer 2023

Lecture 9 — Momentum

Lecturer: Yunwei Ren *Scribed by Zhidan Li*

Contents

1 Overview

In this lecture, we turn our sight back to the gradient descent. We will show how to accelerate the gradient descent by properly choosing the step size.

2 Introduction

Recall the gradient descent:

$$x_{t+1} = x_t - \eta \nabla f(x_t).$$

The gradient descent lemma tells us, if we pick $\eta = 1/L$, or at least $\eta < 2/L$. Then there is a question: can we choose a larger η to make the gradient descent converge faster?

The answer to such a question varies. Here are two examples.

Example 1: Let $f(x) = x^2$. It is trivial that f is 2-smooth. If we choose $\eta = 2/L = 1$, we compute $x_{t+1} = x_t - 2x_t = -x_t$. When $x_0 = 1$, this means the sequence will be $1, -1, +1, \ldots$ Then the gradient descent will not converge.

Example 2: Consider the function f which is a smooth version of $x \mapsto |x|$, i.e., $f \approx |x|$ with L-smoothing around the origin for some large L. Around the origin, we see f is quite smooth so η

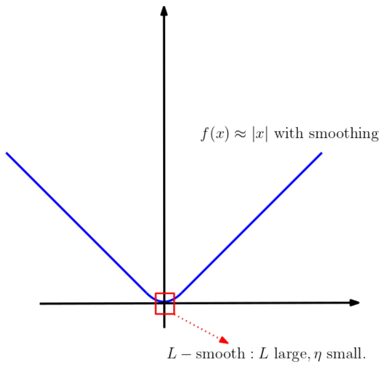
$f(x) \approx |x|$ with smoothing
$L-$smooth : L large, η small.

Figure 1: $f(x) \approx |x|$ with smoothing around $(0,0)$.

will be very small. However, away from the origin, we show f is ∞-smooth (linear function), which allows us to choose large η.

The observation gives rise to a natural question: how to adjust the step size automatically during the gradient descent? The answer is *momentum*. Roughly speaking, we will use 'weighted average of the past gradients', e.g., $\eta_t = \frac{1}{w}\sum_{s=t-q}^{t} \nabla f(x_s)$.

Definition 1 (Nesterov's acceleration gradient descent)**.** *Given $f : \mathbb{R}^d \to \mathbb{R}$, each iteration of **Nesterov's acceleration gradient descent** is as follow:*

$$Z_{t+1} = x_t - \eta \nabla f(x_t),\ x_{t+1} = (1-\gamma_t)Z_{t+1} + \gamma_t Z_t$$

where

$$\gamma_t = \frac{1-\lambda_t}{\lambda_{t+1}},\ \lambda_0 = 0,\ \lambda_{t+1} = \frac{1+\sqrt{1+4\lambda_t^2}}{2}.$$

Definition 2 (Polyak's heavy ball method)**.** *Given $f : \mathbb{R}^d \to \mathbb{R}$, each iteration of **Polyak's heavy ball method** is*

$$x_{t+1} = x_t - \eta g_t$$

where

$$g_t = (1-\gamma)g_{t-1} + \gamma\nabla f(x_t) = \gamma\sum_{s=0}^{t}(1-\gamma)^{t-s}\nabla f(x_s).$$

Remark 1. To compare Algorithm 1 and Algorithm 2, one can see: it is natural to understand Algorithm 2 intuitively (in fact, it is the default mode of PyTorch's momentum), and Algorithm 1 seems to make no sense (and nobody use it nowadays). However, Algorithm 1 provides the theoretically optimal convergence rate $O(1/t^2)$ for smooth convex functions while Algorithm 2 offers no theoretical guarantee.

3 Allen-Zhu and Orecchia's Accelerated Gradient Descent

Consider the question why it is hard to understand Nesterov's acceleration gradient descent. The intuition is clear but it is not easy to choose the descent. In this lecture, we will introduce a variant of the accelerated gradient descent introduced in [AZO14], which is easier to theoretically analyze.

Recall that how momentum works:

1. Choose a large $\eta \gg 1/L$.
2. If η is too large, it would be approximately equivalent to the phenomenon that x_t will start to bounce. The phenomenon implies the moving average g_t will automatically shrink. The shrink of g means the learning rate is decreasing.

Observe that, when x_t starts to bounce, the gradient is large. Then we will immediately decrease the learning rate.

Remark 2. We will not simply normalize the gradient or use the sign descent, since we need the argument that x_t close to a stationary point implies small $\nabla f(x_t)$ for a uniform step size scheme to converge.

3.1 Intuition of Allen-Zhu and Orecchia's acceleration

To simulate the momentum, we use the gradient norm as an indicator.

- If the norm of gradient is small, we pick large η.
- If the norm of gradient is large, we pick small η.

Then it gives rise to some questions:

1. How to determine the threshold?
2. How to reason about GD with $\eta \gg 1/L$.

3.2 A thought experiment

Now we do a thought experiment. Though the experiment and statements in it might be informal, it inspires us to understand A-Z and O's method. Without loss of generality assume that $f(x_*) = 0$. Let $K > 0$ be a parameter to be chosen. Consider the following two extreme cases.

Case 1: The gradient is always large. That is to say, for all t, $\|\nabla f(x_t)\|^2 \geq K$. Then it is safe for us to choose $\eta = 1/L$. By the gradient descent lemma,

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|^2 \\ &\leq f(x_t) - \frac{\|\nabla f(x_t)\|^2}{2L} \\ &\leq f(x_t) - \frac{K}{2L} \\ &\leq f(x_0) - \frac{tK}{2L}. \end{aligned}$$

Then it takes at most $\frac{Lf(x_0)}{K}$ to half the function value.

Case 2: The gradient is always small. That is to say, for all t, $\|\nabla f(x_t)\|^2 \leq K$. By the mirror descent lemma,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) &\leq f(x_*) + \frac{1}{2\eta T}\|x_* - x_0\|^2 + \frac{\eta}{2T}\sum_{t=0}^{T-1}\|\nabla f(x_t)^2\| \\ &\leq f(x_*) + \frac{1}{2\eta T}\|x_* - x_0\|^2 + \frac{\eta}{2}K.\end{aligned}$$

Then we choose η such that $\frac{1}{2\eta T}\|x_* - x_0\|^2$ and $\frac{\eta}{2}K$ are both not greater than $f(x_0)/4$, precisely $\eta = f(x_0)/2K$ and $T = \frac{4K\|x_* - x_0\|^2}{f(x_0)^2}$.

Now we want to determine K. Combining two cases, we have

$$T = \begin{cases} \dfrac{Lf(x_0)}{K} & \text{the gradient is always large;} \\ \dfrac{4K\|x_* - x_0\|^2}{f(x_0)^2} & \text{the gradient is always small.} \end{cases}$$

Make the two terms equal, and we set

$$K = \frac{L^{1/2} f(x_0)^{3/2}}{2\|x_* - x_0\|},\ T = \frac{2L^{1/2}\|x_* - x_0\|}{f(x_0)^{1/2}}.$$

Now we informally establish its convergence rate. Assume that $f(x_0) = C = \Theta(1)$ and $\|x_* - x_0\| = O(1)$. Then after $T = O(\sqrt{L})$ iterations, the value will be halved. Then to achieve $f(x_T) \leq \varepsilon$, we need $O(\sqrt{L/\varepsilon})$ iterations, which is better than the bound $O(L/\varepsilon)$ in the gradient descent.

3.3 Formal form of Allen-Zhu and Orecchia's method

Definition 3 (Allen-Zhu and Orecchia's accelerated GD)**.** *Suppose that we start at point x_0. Set $s_0 = \ell_0 = x_0$. We update x_t, s_t, ℓ_t as follows*

$$\begin{aligned} x_{t+1} &= (1-\tau)s_{t+1} + \tau\ell_{t+1}, \\ s_{t+1} &= x_t - \frac{1}{L}\nabla f(x_t), \\ \ell_{t+1} &= \ell_t - \eta\nabla f(x_t). \end{aligned}$$

where $\tau \in (0,1)$ is a hyperparameter and $\eta \gg 1/L$ is the step size for ℓ_t.

Remark 3. We give some rough interpretations to Definition 3. The update of s_t can be viewed as the gradient descent with size $\eta = \frac{1}{L}$, and ℓ_t can be viewed as the momentum update.

If f is flat, ℓ_t dominates ($\eta \gg 1/L$). This decreases the function value. Otherwise, x_t starts to bounce. Then ℓ_t becomes small, which means s_t dominates. This becomes the usual gradient descent update with a tight step size.

Now we formally establish the convergence rate of it.

Lemma 4 (modified mirror descent lemma)**.** *Under the above settings,*

$$f(x_t) - f(x_*) \leq \frac{1}{2\eta}\left(\|x_* - \ell_t\|^2 - \|x_* - \ell_{t+1}\|^2 + \|\ell_t - \ell_{t+1}\|^2\right) + \frac{1-\tau}{\tau}(f(s_t) - f(x_t)).$$

Based on Lemma 4, we will show the convergence rate of it.

Proposition 5. *Choose* $\eta = \frac{\|x_* - x_0\|}{\sqrt{2Lf(x_0)}}$ *and* $\tau = \frac{1}{L\eta+1}$*. Then we have*

$$\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) \leq f(x_*) + \frac{1}{T}\sqrt{\frac{Lf(x_0)\|x_* - x_0\|}{2}}.$$

Proof. By Lemma 4, it holds that

$$\begin{aligned} f(x_t) - f(x_*) &\leq \frac{1}{2\eta}\left(\|x_* - \ell_t\|^2 - \|x_* - \ell_{t+1}\|^2 + \|\ell_t - \ell_{t+1}\|^2\right) + \frac{1-\tau}{\tau}(f(s_t) - f(x_t)) \\ &= \frac{1}{2\eta}\left(\|x_* - \ell_t\|^2 - \|x_* - \ell_{t+1}\|^2 + \eta^2\|\nabla f(x_t)\|^2\right) + \frac{1-\tau}{\tau}(f(s_t) - f(x_t)) \\ &\overset{(i)}{\leq} \frac{1}{2\eta}\left(\|x_* - \ell_t\|^2 - \|x_* - \ell_{t+1}\|^2 + 2\eta^2 L(f(x_t) - f(s_{t+1}))\right) + \frac{1-\tau}{\tau}(f(s_t) - f(x_t)) \end{aligned}$$

where (i) comes from the update of gradient descent. Choose $\tau = \frac{1}{L\eta+1}$ such that $\eta L = \frac{1-\tau}{\tau}$, and we obtain

$$f(x_t) - f(x_*) \leq \frac{1}{2\eta}\left(\|x_* - \ell_t\|^2 - \|x_* - \ell_{t+1}\|^2 + 2\eta^2 L(f(s_t) - f(s_{t+1}))\right).$$

Summing over both sides, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) \leq f(x_*) + \frac{1}{2\eta T}\left(\|x_* - \ell_0\| + 2\eta^2 Lf(s_0)\right) = \frac{1}{2\eta}\|x_* - \ell_0\| + \frac{\eta L}{T}f(s_0).$$

Choose $\eta = \frac{\|x_* - x_0\|}{\sqrt{2Lf(x_0)}}$ and we prove the proposition. □

References

[AZO14] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *Information Technology Convergence and Services*, 2014.