Optimization for Machine Learning 50.579

Instructor: Ioannis Panageas Scribed by: Li Wei, Menglin Li, Panpan Li

Lecture 6. Accelerated Methods.

1 Introduction

In the previous lessons, we studied classic gradient descent, projected gradient descent and stochastic gradient descent. We then investigate online learning and non-convex optimization. Moreover, we'd like to know if we could do better in convex case from the perspective of convergence rate. This leads to accelerated gradient descent, which is first proposed by Nesterov in 1983.

2 Gradient Descent (Recap)

2.1 Gradient Descent for L-smooth

Theorem 2.1 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, convex (want to minimize) and L-smooth. Let $R = \left\|x_0 - x^*\right\|_2$. It holds for $T = \frac{2R^2L}{\epsilon}$*

$$f\left(x_{T+1}\right) - f\left(x^*\right) \leq \epsilon$$

with appropriately choosing $\alpha = \frac{1}{L}$.

Remarks:

- Speed of convergence is independent of dimension d.
- This result gives a rate of $\mathrm{O}\left(\frac{L}{\epsilon}\right)$.

2.2 Gradient Descent for μ-strongly Convex and L-smooth

Theorem 2.2 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, μ-strongly convex (want to minimize) and L-smooth. Let $R = \left\|x_0 - x^*\right\|_2$ It holds for $T = \frac{2L}{\mu} \ln\left(\frac{R}{\epsilon}\right)$*

$$\left\|x_T - x^*\right\|_2 \leq \epsilon$$

with appropriately choosing $\alpha = \frac{1}{L}$

Remarks:

- Speed of convergence is independent of dimension d.
- This result gives a rate of $\mathrm{O}\left(\frac{L}{\mu} \log \frac{1}{\epsilon}\right) \cdot \kappa := \frac{L}{\mu}$ is called condition number.

3 Definition

Definition 3.1 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. The Accelerated Gradient Descent is defined as follows:*

```
1. Initialization $x_1, y_1 = x_1,$ stepsize $\eta$
2. For $t = 1 \ldots\ T$ do
3. $y_{t+1} = x_t - \eta \nabla f(x_t)$
4. $x_{t+1} = (1 + \gamma_t) y_{t+1} - \gamma_t y_t = y_{t+1} + \gamma_t (y_{t+1} - y_t)$
5. End For
```

Remarks:

- This method was introduced by Nesterov in1983. $y_{t+1} - y_t$ is called momentum.
- γ_t is a sequence independent of x_t and $\gamma_t \geq 0$ for all t.

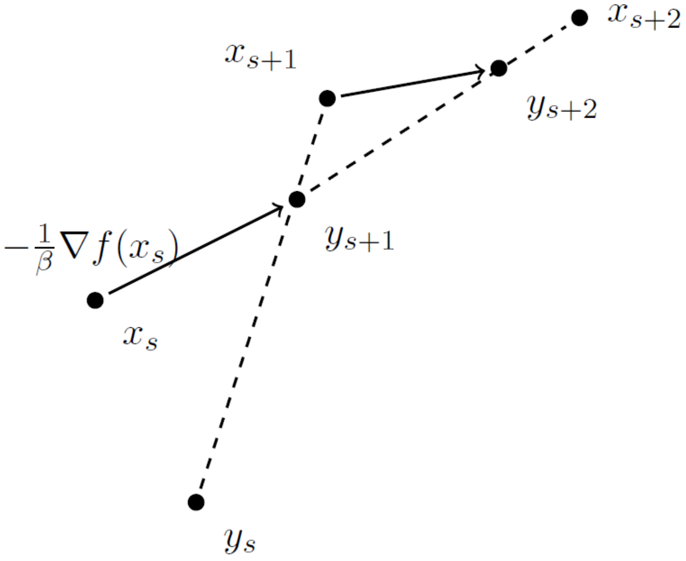

Figure 1: Illustration of Nesterovs accelerated gradient descent from [2].

Figure 1 intuitively presents why the accelerated method proposed by Nesterov could speed up the convergence of gradient descent.

4 Analysis for Smooth, Strongly-convex Functions

Theorem 4.1 (Strongly convex case) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function, L-smooth and μ -strongly convex function. Assume that x^* is the minimizer and set $\gamma_t := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ and*

$\eta = \frac{1}{L}$. *Then it holds that*

$$f\left(y_{t+1}\right) - f\left(x^*\right) \leq \frac{L+\mu}{2} \left\|x_1 - x^*\right\|_2^2 e^{-\frac{t}{\sqrt{k}}},$$

hence we reach ϵ*-close in* ℓ_2 *after* $T := \sqrt{\frac{L}{\mu}} \log\left(\frac{R^2(L+\mu)}{\epsilon}\right)$ *iterations.*

Remarks:

- This result gives a rate of O $\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$.

We would introduce and prove two claims before the proof of the above theorem.

Claim 4.2 (Approximation of $f(x)$ from Below)

$$\Phi_{s+1} \leq f(x) + \left(1 - \frac{1}{\sqrt{\kappa}}\right)^s \left(\Phi_1(x) - f(x)\right)$$

Proof:

We first define the following sequence of functions:
$\Phi_1(x) = f\left(x_1\right) + \frac{\mu}{2}\left\|x - x_1\right\|_2^2$
$\Phi_{s+1}(x) = \left(1 - \frac{1}{\sqrt{\kappa}}\right)\Phi_s(x) + \frac{1}{\sqrt{k}}\left(f\left(x_s\right) + \nabla f\left(x_s\right)^\top\left(x - x_s\right) + \frac{\mu}{2}\left\|x - x_s\right\|_2^2\right)$
Then

$$\begin{aligned}\Phi_{s+1}(x) &= \left(1 - \frac{1}{\sqrt{\kappa}}\right)\Phi_s(x) + \frac{1}{\sqrt{\kappa}}\left(f\left(x_s\right) + \nabla f\left(x_s\right)^\top\left(x - x_s\right) + \frac{\mu}{2}\left\|x - x_s\right\|_2^2\right) \\ &\leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)\Phi_s(x) + \frac{1}{\sqrt{\kappa}} f(x) \text{ (from strong convexity)} \\ &= f(x) + \left(1 - \frac{1}{\sqrt{\kappa}}\right)\left(\Phi_s(x) - f(x)\right)\end{aligned}$$

Therefore

$$\begin{aligned}&\Phi_{s+1}(x) - f(x) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)\left(\Phi_s(x) - f(x)\right) \\ &\Rightarrow \Phi_{s+1}(x) - f(x) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^s \left(\Phi_1(x) - f(x)\right) \text{(telescopic product)} \\ &\Rightarrow \Phi_{s+1} \leq f(x) + \left(1 - \frac{1}{\sqrt{\kappa}}\right)^s \left(\Phi_1(x) - f(x)\right)\end{aligned}$$

■

Claim 4.3 (Approximation of $f(x)$ from Above)

$$f\left(y_s\right) \leq \min_x \Phi_s(x)$$

Proof: For $s = 1$ we have $f(y_1) \leq \min_x \Phi_1(x)$.

$$\begin{aligned}
&\Phi_1(x) = f(x_1) + \frac{\mu}{2}\|x - x_1\|_2^2, x_1 = y_1 \\
&\Rightarrow f(y_1) \leq \Phi_1(x) \\
&\Rightarrow f(y_1) \leq \min_x \Phi_1(x)
\end{aligned}$$

Set $\min_x \Phi_s(x) = \Phi_s^*$.

$$\begin{aligned}
f(y_{s+1}) &\leq f(x_s) - \frac{1}{2L}\|\nabla f(x_s)\|_2^2 \ \textit{(L-smoothness claim 2)} \\
&= \left(1 - \frac{1}{\sqrt{\kappa}}\right) f(y_s) + \left(1 - \frac{1}{\sqrt{\kappa}}\right)(f(x_s) - f(y_s)) + \frac{1}{\sqrt{\kappa}} f(x_s) - \frac{1}{2L}\|\nabla f(x_s)\|_2^2 \\
&\leq \left(1 - \frac{1}{\sqrt{\kappa}}\right) \Phi_s^* + \left(1 - \frac{1}{\sqrt{\kappa}}\right)(f(x_s) - f(y_s)) + \frac{1}{\sqrt{\kappa}} f(x_s) - \frac{1}{2L}\|\nabla f(x_s)\|_2^2 \ \textit{(FOC)} \\
&\leq \left(1 - \frac{1}{\sqrt{\kappa}}\right) \Phi_s^* + \left(1 - \frac{1}{\sqrt{\kappa}}\right)\nabla f(x_s)^\top (x_s - y_s) + \frac{1}{\sqrt{\kappa}} f(x_s) - \frac{1}{2L}\|\nabla f(x_s)\|_2^2
\end{aligned}$$

Let $\left(1 - \frac{1}{\sqrt{\kappa}}\right) \Phi_s^* + \left(1 - \frac{1}{\sqrt{\kappa}}\right)\nabla f(x_s)^\top (x_s - y_s) + \frac{1}{\sqrt{\kappa}} f(x_s) - \frac{1}{2L}\|\nabla f(x_s)\|_2^2 = A$. Now we get $f(y_{s+1}) \leq A$, then we would prove $A \leq \Phi_{s+1}^*$.
According to the definition of $\Phi_1(x)$ and $\Phi_s(x)$, we can get that

$$\begin{aligned}
&\nabla^2 \Phi_1(x) = \mu I_d \\
&\nabla^2 \Phi_s(x) = \left(1 - \frac{1}{\sqrt{\kappa}}\right) \nabla^2 \Phi_{s-1}(x) + \frac{1}{\sqrt{\kappa}} \mu I_d \\
&\Rightarrow \nabla^2 \Phi_s(x) = \mu I_d.
\end{aligned}$$

Therefore, for some v_s,

$$\Phi_s(x) = \Phi_s^* + \frac{\mu}{2}\|x - v_s\|_2^2. \tag{1}$$

Then

$$\begin{aligned}
\nabla \Phi_{s+1}(x) &= \left(1 - \frac{1}{\sqrt{\kappa}}\right) \nabla \Phi_s(x) + \frac{1}{\sqrt{\kappa}} \nabla f(x_s) + \frac{\mu}{\sqrt{\kappa}}(x - x_s) \\
&= \left(1 - \frac{1}{\sqrt{\kappa}}\right) \nabla(\Phi_s^* + \frac{2}{\mu}\|x - v_s\|_2^2) + \frac{1}{\sqrt{\kappa}} \nabla f(x_s) + \frac{\mu}{\sqrt{\kappa}}(x - x_s) \\
&= \left(1 - \frac{1}{\sqrt{\kappa}}\right) \mu(x - v_s) + \frac{1}{\sqrt{\kappa}} \nabla f(x_s) + \frac{\mu}{\sqrt{\kappa}}(x - x_s)
\end{aligned}$$

And v_{s+1} is a minimizer of Φ_{s+1} (that is, $\nabla \Phi_{s+1}(v_{s+1}) = 0$) We can find a relation for v_{s+1} and v_s by expanding $\nabla \Phi_{s+1}$ at v_{s+1}.

$$\nabla \Phi_{s+1}(v_{s+1}) = \left(1 - \frac{1}{\sqrt{\kappa}}\right) \mu(v_{s+1} - v_s) + \frac{1}{\sqrt{\kappa}} \nabla f(x_s) + \frac{\mu}{\sqrt{\kappa}}(v_{s+1} - x_s) = 0 \tag{2}$$

$$\Rightarrow \sqrt{\kappa}\mu v_{s+1} = (\sqrt{\kappa} - 1)\mu v_s + \mu x_s - \nabla f(x_s) \tag{3}$$

$$\Rightarrow v_{s+1} = \left(1 - \frac{1}{\sqrt{\kappa}}\right) v_s + \frac{1}{\sqrt{\kappa}} x_s - \frac{1}{\mu\sqrt{\kappa}} \nabla f(x_s) \tag{4}$$

Evaluating Φ_{s+1} at x_s we have

$$\Phi_{s+1}(x_s) = \left(1 - \frac{1}{\sqrt{\kappa}}\right)\Phi_s(x_s) + \frac{1}{\sqrt{\kappa}} f(x_s) \tag{5}$$

$$\Rightarrow \Phi^*_{s+1} + \frac{\mu}{2}\left\|x_s - v_{s+1}\right\|_2^2 = \left(1 - \frac{1}{\sqrt{\kappa}}\right)\Phi^*_s + \frac{\mu}{2}\left(1 - \frac{1}{\sqrt{\kappa}}\right)\left\|x_s - v_s\right\|_2^2 + \frac{1}{\sqrt{\kappa}} f(x_s) \textit{(using equation (1))} \tag{6}$$

According to equation (4), we can get that

$$x_s - v_{s+1} = \left(1 - \frac{1}{\sqrt{\kappa}}\right)(x_s - v_s) + \frac{1}{\mu\sqrt{\kappa}}\nabla f(x_s) \tag{7}$$

$$\Rightarrow \left\|x_s - v_{s+1}\right\|_2^2 = \left(1 - \frac{1}{\sqrt{\kappa}}\right)\left\|x_s - v_s\right\|_2^2 + \frac{1}{\mu^2\kappa}\left\|\nabla f(x_s)\right\|_2^2 - \frac{2}{\mu\sqrt{\kappa}}\left(1 - \frac{1}{\sqrt{\kappa}}\right)\nabla f(x_s)^T(v_s - x_s) \tag{8}$$

Assume $v_s - x_s = \sqrt{\kappa}(x_s - y_s)$, then by induction we can get that

$$\begin{aligned}
v_{s+1} - x_{s+1} &= \left(1 - \frac{1}{\sqrt{\kappa}}\right)v_s + \frac{1}{\sqrt{\kappa}}x_s - \frac{1}{\mu\sqrt{\kappa}}\nabla f(x_s) - x_{s+1} \\
&= \left(1 - \frac{1}{\sqrt{\kappa}}\right)(v_s - x_s) + x_s - \frac{1}{\mu\sqrt{\kappa}}\nabla f(x_s) - x_{s+1} \\
&= \left(1 - \frac{1}{\sqrt{\kappa}}\right)\sqrt{\kappa}(x_s - y_s) + x_s - \frac{1}{\mu\sqrt{\kappa}}\nabla f(x_s) - x_{s+1} \\
&= \sqrt{\kappa}x_s - (\sqrt{\kappa} - 1)y_s - \frac{\sqrt{\kappa}}{L}\nabla f(x_s) - x_{s+1} \\
&= \sqrt{\kappa}y_{s+1} - (\sqrt{\kappa} - 1)y_s - x_{s+1} \textit{(def. of AGD)} \\
&= \sqrt{\kappa}y_{s+1} + (\sqrt{\kappa} + 1)x_{s+1} - 2\sqrt{\kappa}y_{s+1} - x_{s+1} \textit{(def. of AGD)} \\
&= \sqrt{\kappa}(x_{s+1} - y_{s+1}).
\end{aligned}$$

Therefore, it's true that

$$v_s - x_s = \sqrt{\kappa}(x_s - y_s). \tag{9}$$

Plug equation (8) and (9) into equation (6), we can get that

$$\Phi^*_{s+1} = A + \frac{1}{2\sqrt{\kappa}}\left(1 - \frac{1}{\sqrt{\kappa}}\right)\left\|x_s - y_s\right\|_2^2 \geq A$$

Combine with $f(y_{s+1}) \leq A$, one obtains that $f(y_{s+1}) \leq \Phi^*_{s+1}$. The proof of claim 4.3 is done. ■

Proof of Theorem 4.1:

According to claim 4.2 and 4.3, we have

$$\begin{aligned}
f(y_s) - f(x^*) &\leq \Phi_t(x^*) - f(x^*) \\
&\leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^t \left(\Phi_1(x^*) - f(x^*)\right) \\
&\leq \left(1 - \frac{1}{\sqrt{x}}\right)^t \left(f(x_1) - f(x^*) + \frac{\mu}{2}\left\|x_1 - x^*\right\|_2^2\right) \textit{(strong convexity)}
\end{aligned}$$

Since $f(x_1) - f(x^*) \leq \underbrace{\nabla f(x^*)^\top (x_1 - x^*)}_{=0} + \frac{L}{2}\|x_1 - x^*\|_2^2,$ we get

$$f(y_s) - f(x^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^t \frac{L+\mu}{2}\|x_1 - x^*\|_2^2 \leq \frac{L+\mu}{2}\|x_1 - x^*\|_2^2 e^{-\frac{t}{\sqrt{k}}}$$

■

5 Analysis for Smooth Convex Functions

Theorem 5.1 (Smooth case) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function, $L-$ smooth. Assume that x^* is the minimizer and set $\eta = \frac{1}{L}, \gamma_t := \frac{\lambda_t - 1}{\lambda_{t+1}}$ where $\lambda_0 = 0$ and $\lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$. Then it holds that*

$$f(y_t) - f(x^*) \leq \frac{2L\|x_1 - x^*\|_2^2}{t^2}$$

hence we reach ϵ -close in value after $T := (\sqrt{\frac{2LR^2}{\epsilon}})$ iterations.

Remarks:

- This result gives a rate of O$(\sqrt{\frac{L}{\epsilon}})$

Proof: Using the unconstrained version of Lemma 3.6 from [2] one obtains

$$\begin{aligned} & f(y_{s+1}) - f(y_s) && (10)\\ & \leq \nabla f(x_s)^T(x_s - y_s) - \frac{1}{2L}\|\nabla f(x_s)\|_2^2 && (11)\\ & = L(x_s - y_{s+1})^T(x_s - y_s) - \frac{L}{2}\|x_s - y_{s+1}\|_2^2 && (12) \end{aligned}$$

Similarly we also get

$$f(y_{s+1}) - f(x^*) \leq L(x_s - y_{s+1})^T(x_s - x^*) - \frac{L}{2}\|x_s - y_{s+1}\|_2^2 \tag{13}$$

Now multiplying (12) by $(\lambda_s - 1)$ and adding the results to (13), one obtains with $\delta_s = f(y_s) - f(x^*)$.

$$\lambda_s\delta_{s+1} - (\lambda_s - 1)\delta_s \leq L(x_s - y_{s+1})^T(\lambda_s x_s - (\lambda_s - 1)y_s - x^*) - \frac{L}{2}\lambda_s\|x_s - y_{s+1}\|_2^2$$

Multiplying this equality by λ_s and using that by definition $\lambda_{s-1}^2 = \lambda_s^2 - \lambda_s$, as well as the elementary identity $2a^Tb - \|a\|_2^2 = \|b\|_2^2 - \|b-a\|_2^2$, one obtains

$$\begin{aligned} & \lambda_s^2\delta_{s+1} - \lambda_{s-1}^2\delta_s && (14)\\ & \leq \frac{L}{2}(2\lambda_s(x_s - y_{s+1})^T(\lambda_s x_s - (\lambda_s - 1)y_s - x^*) - \|\lambda_s(y_{s+1} - x_s)\|_2^2) && (15)\\ & = \frac{L}{2}(\|\lambda_s x_s - (\lambda_s - 1)(y_s - x^*)\|_2^2 - \|\lambda_s y_{s+1} - (\lambda_s - 1)(y_s - x^*)\|_2^2) && (16) \end{aligned}$$

Now remark that, by definition, one has

$$x_{s+1} = y_{s+1} + \gamma(y_s - y_{s+1}) \tag{17}$$

$$\Leftrightarrow \lambda_{s+1} x_{s+1} = \lambda_{s+1} y_{s+1} + (1 - \lambda_s)(y_s - y_{s+1}) \tag{18}$$

$$\Leftrightarrow \lambda_{s+1} x_{s+1} - (\lambda_{s+1} - 1) y_{s+1} = \lambda_s y_{s+1} - (\lambda_s - 1) y_s \tag{19}$$

Putting together (16) and (19) one gets that $\mu_s = \lambda_s x_s - (\lambda_s - 1) y_s - x^*$,

$$\lambda_s^2 \delta_{s+1} - \lambda_{s-1}^2 \delta_s^2 \leq \frac{L}{2}(\|\mu_s\|_2^2 - \|\mu_{s+1}\|_2^2)$$

Summing these inequalities from $s = 1$ to $s = t - 1$ one obtains:

$$\delta_t \leq \frac{L}{2\lambda_{t-1}^2} \|\mu_1\|_2^2 .$$

By induction it is easy to see that $\lambda_{t-1} \geq \frac{t}{2}$ which concludes the proof. ■

References

[1] Sebastian Ruder. . *An overview of gradient descent optimization algorithm.* . arXiv preprint arXiv:1609.04747, 20.

[2] Bubeck S. . *Convex optimization: Algorithms and Complexity* . . Foundations and Trends in Machine Learning, 2015, 8(3-4): 231-357.