IIIS 2018 Spring: ATCS - Selected Topics in Optimization
Lecture date: May 28, 2018
Instructor: Jian Li **Scribe: Fanjin Zhang**

1 Introduction

In this lecture, we cover some optimization techniques with their theoretical analysis. We first introduce the ODE interpretation and convergence of Heavy ball method. Then we use construction method to show the lower bound of first-order method. Finally, we present ODE interpretation and analysis of Nesterov's acceleration.

2 Notation

We consider the objective function f is l-strongly convex and L-smooth. Thus $lI \preceq \nabla^2 f \preceq LI$. We use $\kappa = \frac{L}{l}$ to denote condition number of $\nabla^2 f$.

3 Heavy ball method

Heavy ball method is also called **chebyshev iterative method**. Its update rule is:

$$x^{k+1} = x^k - \gamma \nabla f(x) + \beta(x^k - x^{k-1}) \tag{1}$$

3.1 ODE interpretation

The corresponding ODE of heavy ball method is:

$$\mu \frac{\mathrm{d}^2 x}{\mathrm{d}t^2} = -\nabla f(x) - b\frac{\mathrm{d}x}{\mathrm{d}t} \tag{2}$$

We can regard $\frac{\mathrm{d}^2 x}{\mathrm{d}t^2}$ as acceleration, $-\nabla f(x)$ as force and $-b\frac{\mathrm{d}x}{\mathrm{d}t}$ as friction. By discretizing (2), we obtain:

$$\mu \frac{x(t+\Delta t) - 2x(t) + x(t-\Delta t)}{\Delta t^2} = -\nabla f(x(t)) - b\frac{x(t) - x(t-\Delta t)}{\Delta t} \tag{3}$$

(Note in the left we can use Tayler expansion to get the differential)
Thus, from (3) we have:

$$x(t+\Delta t) = x(t) - \frac{\Delta t^2}{\mu}\nabla f(x(t)) + (1 - \frac{b}{\mu}\Delta t)(x(t) - x(t-\Delta t)) \tag{4}$$

In the equation above, let $\gamma = \frac{\Delta t^2}{\mu}$ and $\beta = 1 - \frac{b}{\mu}\Delta t$ which is equivalent to (1).

3.2 Convergence analysis

The convergence rate of heavy ball method is $\mathcal{O}(\sqrt{\kappa}\log\frac{1}{\epsilon})$, where κ is condition number of $\nabla^2 f$ and ϵ is error tolerance. In the following, we first show that the convergence rate of gradient descent, which is $\mathcal{O}(\kappa\log\frac{1}{\epsilon})$. Then we show the improved convergence rate of heavy ball method.

Theorem 1. *The convergence rate of gradient descent is $\mathcal{O}(\kappa\log\frac{1}{\epsilon})$.*

Proof [1]. Denote the update rule of gradient descent $G_\alpha(x) = x - \alpha\nabla f(x)$. Assume $||G_\alpha(x) - G_\alpha(y)|| < L_G||x-y||$, where L_G is a constant less than 1. There exists the lemma below.

Lemma 2. $||x^{k+1}-x^*||_2 \le L_G^k||x^1-x^*||_2$*, where x^* is the optimal solution of f.*

Proof.

$$\begin{aligned}||x^{k+1}-x^*||_2 &= ||x^k - \alpha_k\nabla f(x_k) - (x^* - \alpha_k\nabla f(x^*))||_2 \\ &= ||G_\alpha(x^k) - G_\alpha(x^*)||_2 \\ &\le L_G||x^k - x^*||_2\end{aligned} \tag{5}$$

By using Eq (5) k times, the lemma is proved. □

Lemma 3. *Assume f is l-strongly convex and L-smooth. Therefore, $L_G \le \max\{|1-\alpha l|, |1-\alpha L|\}$.*

Proof.

$$\begin{aligned}||G_\alpha(x) - G_\alpha(y)||_2 &= ||x - \alpha\nabla f(x) - (y - \alpha\nabla f(y))||_2 \\ &= ||(x-y)(I - \alpha\nabla^2 f(z))||_2 \\ &\le ||x-y||_2||I-\alpha\nabla^2 f||_2 \\ &\le ||x-y||_2\max\{|1-\alpha l|, |1-\alpha L|\}.\end{aligned} \tag{6}$$

where $z \in [x,y]$ and the last line comes from the definition of matrix spectral norm. Thus, $L_G \le \max\{|1-\alpha l|, |1-\alpha L|\}$. □

In lemma 3, let $\alpha = \frac{2}{l+L}$, so $L_G = 1 - \mathcal{O}(\frac{1}{\kappa})$. Then let $(1-\mathcal{O}(\frac{1}{\kappa}))^t \le \epsilon$, we can obtain $t = \mathcal{O}(\kappa\log\frac{1}{\epsilon})$.

□

Theorem 4. *For heavy ball method, assume f is l-strongly convex and L-smooth and let $\gamma = \frac{4}{(\sqrt{l}+\sqrt{L})^2}$ and $\beta = \frac{\sqrt{L}-\sqrt{l}}{\sqrt{L}+\sqrt{l}}$. We have $||x^{k+1}-x^*||_2 \le (\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^k||x^1-x^*||_2$.*

Proof. In the following proof, instead of looking at $||x^{k+1}-x^*||_2$, we examine $||x^{k+1}-x^*||_2 + ||x^k - x^*||_2$:

$$
\begin{aligned}
\left\| \begin{bmatrix} x^{k+1} - x^* \\ x^k - x^* \end{bmatrix} \right\| &= \left\| \begin{bmatrix} x^k + \beta(x^k - x^{k-1}) - x^* \\ x^k - x^* \end{bmatrix} - \gamma \begin{bmatrix} \nabla f(x^k) \\ 0 \end{bmatrix} \right\| \\
&= \left\| \begin{bmatrix} (1+\beta)I & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} x^k - x^* \\ x^{k-1} - x^* \end{bmatrix} - \gamma \begin{bmatrix} \nabla^2 f(z^k)(x^k - x^*) \\ 0 \end{bmatrix} \right\| \\
&= \left\| \begin{bmatrix} (1+\beta)I - \gamma \nabla^2 f(z^k) & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} x^k - x^* \\ x^{k-1} - x^* \end{bmatrix} \right\| \\
&\leq \left\| \begin{bmatrix} (1+\beta)I - \gamma \nabla^2 f(z^k) & -\beta I \\ I & 0 \end{bmatrix} \right\| \left\| \begin{bmatrix} x^k - x^* \\ x^{k-1} - x^* \end{bmatrix} \right\|
\end{aligned} \tag{7}
$$

where $z^k \in [x^k, x^*]$(w.l.o.g. let $x^k < x^*$), and let

$$
T = \begin{bmatrix} (1+\beta)I - \gamma \nabla^2 f(z^k) & -\beta I \\ I & 0 \end{bmatrix} \tag{8}
$$

We introduce the following Proposition:

Lemma 5. *For $\beta \geq \max\{|1 - \sqrt{\gamma l}|^2, |1 - \sqrt{\gamma L}|^2\}$, $\rho(T) = \max_i |\lambda_i(T)| \leq \sqrt{\beta}$.*

Proof. Let $U \Lambda U^T$ be the eigendecomposition of $\nabla^2 f(z^k)$. Let Π be the $2n \times 2n$ matrix with entries

$$
\Pi_{i,j} = \begin{cases} 1 & i \text{ odd}, j = (i+1)/2 \\ 1 & i \text{ even}, j = n + i/2 \\ 0 & \text{otherwise} \end{cases} \tag{9}
$$

We have

$$
\begin{aligned}
&\Pi \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix} \begin{bmatrix} (1+\beta)I - \gamma \nabla^2 f(z_k) & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}^T \Pi^T \\
=&\Pi \begin{bmatrix} (1+\beta)I - \gamma \Lambda & -\beta I \\ I & 0 \end{bmatrix} \Pi^T \\
=&\begin{bmatrix} T_1 & 0 & \dots & 0 \\ 0 & T_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & T_n \end{bmatrix}
\end{aligned} \tag{10}
$$

where

$$
T_i = \begin{bmatrix} 1 + \beta - \gamma \lambda_i & -\beta \\ 1 & 0 \end{bmatrix} \tag{11}
$$

That is, T is similar to the block diagonal matrix with 2×2 diagonal blocks T_i. To compute the eigenvalues of T, it suffices to compute the eigenvalues of all of the T_i. For fixed i, the eigenvalues of the 2×2 matrix are roots of the equation

$$
u^2 - (1 + \beta - \gamma \lambda_i)u + \beta = 0 \tag{12}
$$

In the cases that $\beta \geq (1-\sqrt{\gamma\lambda_i})^2$, the roots of the characteristic equations are imaginary, and both have magnitude $\sqrt{\beta}$. Note that by assumption

$$(1-\sqrt{\gamma\lambda_i})^2 \leq \max\{|1-\sqrt{\gamma l}|^2, |1-\sqrt{\gamma L}|^2\} \tag{13}$$

and letting $\beta = \max\{|1-\sqrt{\gamma l}|^2, |1-\sqrt{\gamma L}|^2\}$ completes the proof. □

Hence, setting $\gamma = \frac{4}{(\sqrt{l}+\sqrt{L})^2}$ and $\beta = \max\{|1-\sqrt{\gamma l}|^2, |1-\sqrt{\gamma L}|^2\} = (\frac{\sqrt{L}-\sqrt{l}}{\sqrt{L}+\sqrt{l}})^2$. Thus $\rho(T) \leq \sqrt{\beta} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$.

$$\begin{aligned}
\left\|\begin{bmatrix} x^{k+1}-x^* \\ x^k - x^* \end{bmatrix}\right\| &\leq \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\left\|\begin{bmatrix} x^{k}-x^* \\ x^{k-1} - x^* \end{bmatrix}\right\| \\
&\leq \Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^2\left\|\begin{bmatrix} x^{k-1}-x^* \\ x^{k-2} - x^* \end{bmatrix}\right\| \\
&\leq \dots \\
&\leq \Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^k\left\|\begin{bmatrix} x^{1}-x^* \\ x^{0} - x^* \end{bmatrix}\right\|
\end{aligned} \tag{14}$$

Or, in other words,

$$\left\|x^{k+1}-x^*\right\| \leq \Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^k \|x^1 - x^*\| \tag{15}$$

□

4 Lower bound of first order method

Theorem 6. *There exists a L-smooth and l-strongly convex function $f\colon l_2 \to \mathbb{R}$ with condition number $\kappa = \frac{L}{l}$ such that for any $k \geq 1$ and any black box first order method, the following lower bound holds.*

$$f(x^k) - f(x^*) \geq \frac{l}{2}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^{2(k-1)}\left\|x^1 - x^*\right\|^2 \tag{16}$$

Proof [3]. As is typical of lower bound proofs, we prove this theorem by constructing an example. The example function we construct is an l_2 function. Informally speaking, l_2 functions are vectors with infinitely many coordinates that are also square summable. Formally,

$$l_2 = \{x = (x(n)), n \in \mathbb{N}, \sum_{i=1}^{\infty} x(i)^2 < +\infty\} \tag{17}$$

We define an operator that assumes the form of a tridiagonal matrix. Let

$$A = \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots \end{bmatrix} \tag{18}$$

Using the operator above, we can define the following quadratic function.

$$f(x) = \frac{l(\kappa - 1)}{8}(x^T Ax - 2e_1^T x) + \frac{l}{2}\|x\|^2 \tag{19}$$

Here, e_1 denotes the first vector of the canonical basis, i.e., $e_1 = [1, 0, ..., 0]^T$. We compute the gradient of f.

$$\nabla f(x) = \frac{l(\kappa - 1)}{4}(Ax - e_1) + lx \tag{20}$$

We assume that the starting point for our gradient descent routine will be $x^1 = 0$. Plugging that into the expression above, we get $\nabla f(x)_{x=x^1} = -\frac{l(\kappa-1)}{4}e_1$.

Since x^k is the linear combination of x^{k-1} and $\nabla f(x^{k-1})$, it is easy to show (by mathematical induction) that if x^k has non-zero entries upto element at index $k-1$, then x^{k+1} will have non-zero entries upto k. The way the Hessian A is designed, the non-zero values propogate linearly across the dimensions, one dimension per each step of the gradient descent routine.

That is, $x^k(i) = 0, \forall t \geq k$. Let's now consider the norm

$$\begin{aligned} \left\| x^k - x^* \right\| &= \sum_{i=1}^{\infty}(x^k(i) - x^*(i))^2 \\ &\geq \sum_{i=k}^{\infty}(x^k(i) - x^*(i))^2 \\ &= \sum_{i=k}^{\infty}(x^*(i))^2 \end{aligned} \tag{21}$$

Because f is l-strongly convex, it gives us

$$f(x^k) - f(x^*) \geq \frac{l}{2}\left\| x^k - x^* \right\|^2 \geq \frac{l}{2}\sum_{i=k}^{\infty}(x^*(i))^2 \tag{22}$$

If we differentiate f and set ∇f to 0, we obtain an infinite linear system, of the following form.

$$\begin{aligned} &1 - 2\frac{\kappa + 1}{\kappa - 1}x^*(1) + x^*(2) = 0, \\ &x^*(k-1) - 2\frac{\kappa + 1}{\kappa - 1}x^*(k) + x^*(k+1), \forall k \geq 2 \end{aligned} \tag{23}$$

The solution of the above system is given by

$$x^*(i) = \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^i \tag{24}$$

Now, we plug this into the above expression, which gives us

$$
\begin{aligned}
f(x^k) - f(x^*) &\geq \frac{l}{2}\left\|x^k - x^*\right\|^2 \\
&\geq \frac{l}{2}\sum_{i=k}^{\infty}(x^*(i))^2 \\
&= \frac{l}{2}\sum_{i=k}^{\infty}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2i} \\
&= \frac{l}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2(k-1)}\left\|x^1 - x^*\right\|^2
\end{aligned}
\tag{25}
$$

This proves the theorem. □

5 Nesterov's acceleration

The update rule of Nesterov's acceleration is

$$
\begin{aligned}
x^{k+1} &= y^k - s\nabla f(y^k) \\
y^k &= x^k + \frac{k-1}{k+2}(x^k - x^{k-1})
\end{aligned}
\tag{26}
$$

where $y^0 = x^0$. The related second-order ODE takes the following form:

$$
\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} + \frac{3}{t}\frac{\mathrm{d}x}{\mathrm{d}t} + \nabla f(x) = 0 \tag{27}
$$

For the derivation of the ODE, please refer to [6].

As for the convergence rate, there exists the theorem below.

Theorem 7. *If f is convex, $f(x^k) - f(x^*) \leq \frac{2\left\|x^0 - x^*\right\|}{k^2}$.*

Proof. Consider the energy functional $\varepsilon(k) = k^2(f(x^k) - f(x^*)) + 2\left\|x^k + \frac{1}{2}k\frac{\mathrm{d}x}{\mathrm{d}k} - x^*\right\|$.

$$
\begin{aligned}
\frac{\mathrm{d}\varepsilon}{\mathrm{d}k} &= 2k(f(x^k) - f(x^*)) + k^2\langle\nabla f(x^k), \frac{\mathrm{d}x}{\mathrm{d}k}\rangle + 4\langle x + \frac{k}{2}\frac{\mathrm{d}x}{\mathrm{d}k} - x^*, \frac{3}{2}\frac{\mathrm{d}x}{\mathrm{d}k} + \frac{k}{2}\frac{\mathrm{d}^2 x}{\mathrm{d}k^2}\rangle \\
&= 2k(f(x^k) - f(x^*)) + 4\langle x^k - x^*, -k\frac{\nabla f(x^k)}{2}\rangle \text{ (by using (27))} \\
&\leq 0 \text{ (by convexity)}
\end{aligned}
\tag{28}
$$

Thus, $f(x^k) - f(x^*) \leq \frac{\varepsilon(k)}{k^2} \leq \frac{\varepsilon(0)}{k^2} = \frac{2\left\|x^0 - x^*\right\|}{k^2}$. □

6 Brief Summary

In this section, we briefly summarize some topics (or taxonomy) of optimization techniques. Derivative based optimization can be mainly divided into two categories: **full gradient methods** and **stochastic gradient methods**. Full gradient methods require full batch samples to update at each step, including Gradient Descent (GD), Heavy Ball method, Nesterov Accelerated Gradient

Descent (NAGD)[5], etc.. However, stochastic gradient methods sample a subset samples at every step, which include Stochastic Gradient Descent (SGD), Stochastic Variance Reduced Gradient (SVRG)[4] and so on. Besides, distributed optimization is another emerging topic which considers updating in parallel each time, related works including ADMM[2], AsySVRG[7] and so on.

References

[1] CS726 - Lyapunov analysis and the heavy ball method. `http://pages.cs.wisc.edu/~brecht/cs726docs/HeavyBallLinear.pdf`. Accessed: 2018-06-26.

[2] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[3] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

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

[5] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[6] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterovs accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

[7] Shen-Yi Zhao and Wu-Jun Li. Fast asynchronous parallel stochastic gradient descent: A lock-free approach with convergence guarantee. In *AAAI*, pages 2379–2385, 2016.