ECE 273 Convex Optimization and Applications Instructor: Jun-Kun Wang
Scribe: Xiangyi Deng, Can Chen May 30, 2024
Editors/TAs: Marialena Sfyraki

Lecture 17: Acceleration via Chebyshev Polynomial

1 Solving strongly convex quadratic problems

Recall that we have proven in HW1 that the ridge linear regression problem

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^{n} \frac{1}{2}(y_i - x^\top z_i)^2 + \frac{\gamma}{2}\|x\|_2^2, \text{ where } \gamma > 0.$$

is equivalent to

$$\min_{x\in R^d} \frac{1}{2}x^\top Ax - b^\top x$$

where $A := \gamma I_d + \sum_{i=1}^n z_i z_i^\top$, $A \succ 0$.

1.1 Constant step size gradient descent

When we use gradient descent with constant step size η, let $x_* = \arg\min f(x)$, then by optimality condition we have $\nabla f(x_*) = 0 \implies Ax_* - b = 0$, thus the update can be written as

$$\begin{aligned}
x_{k+1} &= x_k - \eta\nabla f(x_k) \\
&= x_k - \eta(Ax_k - b) \\
&= x_k - \eta(Ax_k - Ax_* + Ax_* - b) \\
&= x_k - \eta A(x_k - x_*)
\end{aligned}$$

and we can form a k-degree polynomial of A

$$\begin{aligned}
x_{k+1} - x_* &= (I_d - \eta A)(x_k - x_*) \\
&= \underbrace{(I_d - \eta A)^k}_{k\text{-degree polynomial of } A} (x_1 - x_*), \text{ recursively apply the update step}
\end{aligned}$$

Before we can analyze the update step, we need some definitions of spectral norm.

Definition 1 (Spectral Norm $\|\cdot\|_2$)**.** *Given a matrix $B \in \mathbb{R}^{m\times n}$,*

- $\|B\|_2 := \sigma_{\max}(B)$ *is the largest singular value of B.*
- $\|B\|_2 = \max_{x:\|x\|_2=1} \|Bx\|_2$.

Using the fact that $\|B\|_2 = \sqrt{\lambda_{max}(B^\top B)}$, we have when matrix $B \in R^{d\times d}$ is symmetric and diagonalizable, i.e. $B = U\Lambda U^\top$, where U is orthogonal $U^\top = U^{-1}$ and Λ is diagonal, by the fact

$$\begin{aligned}\|B\|_2 = \sqrt{\lambda_{max}(B^\top B)} &= \sqrt{\lambda_{max}(U\Lambda U^\top)^\top(U\Lambda U^\top)}\\ &= \sqrt{\lambda_{max}(U\Lambda U^\top)^\top(U\Lambda U^\top)}\\ &= \sqrt{\lambda_{max}(U\Lambda^2 U^\top)}\end{aligned}$$

$$\|B\|_2 = \max(|\lambda_{max}(B)|, |\lambda_{min}(B)|)$$

By the definition of Spectral norm and the update equation, we have

$$\begin{aligned}\|x_{k+1} - x_*\|_2 &= \|(I - \eta A)(x_k - x_*)\|_2\\ &\le \|I - \eta A\|_2 \|x_k - x_*\|_2\end{aligned}$$

As A is definite and symmetric, i.e. $A \succ 0$, $A^\top = A$, the eigen-decomposition $A = U\Lambda U^{-1} = U\Lambda U^\top$. It follows that

$$\begin{aligned}\|I - \eta A\|_2 &= \|I - \eta U\Lambda U^\top\|_2\\ &= \|UU^\top - \eta U\Lambda U^\top\|_2\\ &= \|U(I - \eta\Lambda)U^\top\|_2\\ &= max_{i\in[d]}|I - \eta\lambda_i|\end{aligned}$$

Let $\mu = \lambda_{min}(A)$, $L = \lambda_{max}(A)$, we want to tune the learning rate η such that

$$\min_\eta \max_{i\in[d]} |I - \eta\lambda_i| \le \min_\eta \max_{\lambda\in[\mu,L]} |I - \eta\lambda|$$

The above inequality is true due to the right hand side having more degree of freedom when optimizing.

From the perspective of η, for fixed $\eta > 0$, $|1 - \eta\lambda| = \begin{cases} \eta\lambda - 1 & \lambda > \frac{1}{\eta}, \\ 1 - \eta\lambda & \lambda < \frac{1}{\eta}. \end{cases}$

$$\max_{\lambda\in[\mu,L]} |1 - \eta\lambda| = \max(\eta L - 1, 1 - \eta\mu)$$

$$\begin{aligned}\min_\eta \max_{i\in[d]} |I - \eta\lambda_i| &\le \min_\eta \max_{\lambda\in[\mu,L]} |I - \eta\lambda|\\ &= \min_\eta \max(\eta L - 1, 1 - \eta\mu)\end{aligned}$$

From the perspective of η, the optimal η is achieved when

$$\eta L - 1 = 1 - \eta\mu$$

$$\eta = \frac{2}{L+\mu}$$

$$\begin{aligned}
\|x_{k+1} - x_*\|_2 &\le \max_{i\in[d]} |1 - \eta\lambda_i| \|x_k - x_*\|_2 \\
&\le \max_{\lambda\in[\mu,L]} |1 - \eta\lambda| \|x_k - x_*\|_2 \\
&\le \max_{\lambda\in[\mu,L]} \left|1 - \frac{2\lambda}{L+\mu}\right| \|x_k - x_*\|_2 \\
&= \left(1 - \frac{2\mu}{L+\mu}\right) \|x_k - x_*\|_2, \text{optimal occurs at either } \lambda = \mu|L \\
&= \left(1 - \frac{2\mu}{L+\mu}\right)^k \|x_1 - x_*\|_2 \\
&= \left(1 - \frac{2}{1+\kappa}\right)^k \|x_1 - x_*\|_2
\end{aligned}$$

where $\kappa := \frac{L}{\mu}$ is the condition number.

1.2 General Gradient-based Methods

Now we consider a more general first order algorithm in the form

$$x_{k+1} = x_1 + span\{\nabla f(x_1), \nabla f(x_2), ..., \nabla f(x_k)\} \tag{1}$$

we have the following Lemma.

Lemma 1. *Consider solving*

$$\min_x \frac{1}{2} x^\top A x - b^\top x$$

Algorithms in the form of (1) have the following dynamics:

$$x_{k+1} - x_* = P_k(A)(x_1 - x_*),$$

where $P_k(A)$ is a k-degree polynomial of A and $P_0(A) = 1$.

Proof. We use induction to prove the lemma. Consider base case:

$$\begin{aligned}
x_1 - x_* &= 1(x_1 - x_*) \\
&= P_0(A)(x_1 - x_*)
\end{aligned}$$

Suppose at k

$$x_k - x_* = P_{k-1}(A)(x_1 - x_*)$$

Consider k+1

$$\begin{aligned}
x_{k+1} - x_* &= x_1 - x_* + span\{\nabla f(x_1), \nabla f(x_2), ..., \nabla f(x_k)\} \\
&= x_1 - x_* + \sum_{j=1}^{k} \alpha_j \nabla f(x_j) \\
&= x_1 - x_* + \sum_{j=1}^{k} \alpha_j (Ax_j - Ax_*) \\
&= x_1 - x_* + \sum_{j=1}^{k} \alpha_j A P_{j-1}(A)(x_1 - x_*) \\
&= (I_d + \sum_{j=1}^{k} \alpha_j A P_{j-1}(A))(x_1 - x_*) \\
&= P_k(A)(x_1 - x_*)
\end{aligned}$$

where $\{\alpha_j\}$ are some coefficients. □

Following **Lemma 1**, we have

$$\|x_{K+1} - x_*\|_2 \leq \|P_K(A)\|_2 \|x_1 - x_*\|_2$$

Our goal is to find the best k-degree polynomial that minimizes $P_K(A)$ for the worst case A,

$$P_K^* = \underset{P \in P_k; P_0(\cdot)=1}{\arg\min} \max_{A \in M} \|P_K(A)\|_2$$

where the set $M := \{A \succ 0 : \lambda_{min}(A) = \mu, \lambda_{max}(A) = L\}$. The solution is a "scaled-and-shifted" Chebyshev Polynomial.

2 Chebyshev Polynomial and the Chebyshev method

Definition 2 (K-degree Chebyshev Polynomial of the first kind)**.** *We denote $\Phi_K(\cdot)$ the degree-K hebyshev polynomial of the first kind, which is defined by:*

$$\Phi_K(x) = \begin{cases} \cos(K \arccos(x)) & \textit{if } x \in [-1, 1], \\ \cosh(K \, arccosh(x)) & \textit{if } x > 1, \\ (-1)^K \cosh(K \, arccosh(x)) & \textit{if } x < 1. \end{cases}$$

Equivalent definition is the following

$$\begin{aligned}
&\Phi_0(x) = 1, \\
&\Phi_1(x) = x, \\
&\Phi_k(x) = 2x\Phi_{k-1}(x) - \Phi_{k-2}(x), \ \textit{for } k \geq 2
\end{aligned}$$

Consider a scaled-and-shifted K-degree Chebyshev polynomial,

Definition 3 (Scaled-and-shifted Chebyshev Polynomial)**.**

$$\bar{\Phi}_K(\lambda) := \frac{\Phi_K(h(\lambda))}{\Phi_K(h(0))}$$

where $h(\cdot)$ *is the mapping* $h(\lambda) := \frac{L+\mu-2\lambda}{L-\mu}$.

Remark. Observe that the mapping $h(\cdot)$ maps all $\lambda \in [\mu, L]$ into the interval $[-1, 1]$.

Lemma 2. *For any positive integer* K*, we have*

$$\max_{\lambda \in [\mu, L]} \left|\bar{\Phi}_K(\lambda)\right| \leq 2\left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^K.$$

Remark. Condition number $\kappa := \frac{L}{\mu} \geq 1 \Rightarrow 1 - \frac{2}{\sqrt{\kappa}+1} \leq 1$

Proof. Observe that the numerator of $\bar{\Phi}_K(\lambda) = \frac{\Phi_K(h(\lambda))}{\Phi_K(h(0))}$ satisfies $|\Phi_K(h(\lambda))| \leq 1$, since $h(\lambda) \in [-1, 1]$ for $\lambda \in [\mu, L]$ and that the Chebyshev polynomial satisfies $|\Phi_K(\cdot)| \leq 1$ when its argument is in $[-1, 1]$ by the definition. It remains to bound the denominator, which is $\Phi_K(h(0)) = \cosh\left(K \ \text{arccosh}\left(\frac{L+\mu}{L-\mu}\right)\right)$. Since

$$\text{arccosh}\left(\frac{L+\mu}{L-\mu}\right) = \log\left(\frac{L+\mu}{L-\mu} + \sqrt{\left(\frac{L+\mu}{L-\mu}\right)^2 - 1}\right) = \log(\theta), \ \text{where } \theta := \frac{\sqrt{L}+\sqrt{\mu}}{\sqrt{L}-\sqrt{\mu}},$$

we have

$$\Phi_K(h(0)) = \cosh\left(K \ \text{arccosh}\left(\frac{L+\mu}{L-\mu}\right)\right) = \frac{\exp(K\log(\theta))+\exp(-K\log(\theta))}{2} = \frac{\theta^K+\theta^{-K}}{2} \geq \frac{\theta^K}{2}.$$

Combing the above inequalities, we obtain the desired result:

$$\begin{aligned}
\max_{\lambda \in [\mu, L]} \left|\bar{\Phi}_K(\lambda)\right| = \max_{\lambda \in [\mu, L]} \left|\frac{\Phi_K(h(\lambda))}{\Phi_K(h(0))}\right| \leq \frac{2}{\theta^K} &= 2\left(1 - 2\frac{\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^K \\
&= O\left(\left(1 - \Theta\left(\sqrt{\frac{\mu}{L}}\right)\right)^K\right)
\end{aligned}$$

$\square$

Recall the convergence analysis for the following algorithms.

Gradient Descent

$$\|x_{K+1} - x_*\|_2 \leq \left(1 - \frac{2}{\kappa + 1}\right)^K \|x_1 - x_*\|_2.$$

Chebyshev method

$$\begin{aligned}\|x_{K+1} - x_*\|_2 &\leq \min_{P \in P_K; P_0(\cdot)=1} \max_{A \in M} \|P_K(A)\|_2 \|x_1 - x_*\|_2 \\ &\leq 2\left(1 - \frac{2}{\sqrt{\kappa} + 1}\right)^K \|x_1 - x_*\|_2.\end{aligned}$$

where the set $M := \{A \succ 0 : \lambda_{\min}(A) = \mu, \lambda_{\max}(A) = L\ \}$

From the above inequalities, we know that the convergence rate of Chebyshev method is faster than Gradient Descent. Exactly how much faster depends on the size of the condition number.

Q: What is the optimal algorithm implied by the scaled-and-shifted K-degree Chebyshev polynomial?

Consider a scaled-and-shifted K-degree Chebyshev Polynomial

$$\bar{\Phi}_K(\lambda) := \frac{\Phi_K(h(\lambda))}{\Phi_K(h(0))}, \tag{1}$$

where $h(\cdot)$ is the mapping $h(\lambda) := \frac{L+\mu-2\lambda}{L-\mu}$.

$$\bar{\Phi}_0(\lambda) = \frac{\Phi_0(h(\lambda))}{\Phi_0(h(0))} = 1.$$

Since $(x_1 - x_*) = \Phi_0(h(\lambda))(x_1 - x_*)$, we can pick any $x_1 \in \mathbb{R}^d$.

From above, we have

$$\Phi_0(x) = 1, \tag{2}$$
$$\Phi_1(x) = x, \tag{3}$$
$$\Phi_k(x) = 2x\Phi_{k-1}(x) - \Phi_{k-2}(x), \text{for } k \geq 2 \tag{4}$$

By (1) and (3), we get

$$\bar{\Phi}_1(\lambda) = \frac{\Phi_1(h(\lambda))}{\Phi_0(h(0))} = \frac{h(\lambda)}{h(0)} = \frac{L + \mu - 2\lambda}{L + \mu} = 1 - \frac{2\lambda}{L + \mu}.$$

From above, we get

$$x_2 = x_1 - \frac{2}{L + \mu}\nabla f(x_1),$$

and we know that

$$x_2 - x_* = \left(1 - \frac{2A}{L+\mu}\right)(x_1 - x_*)$$

For $k \geq 2$, we have

$$\bar{\Phi}_k(\lambda) = \frac{2\theta_k}{L-\mu}(L+\mu-2\lambda)\bar{\Phi}_{k-1}(\lambda) + \left(1 - \frac{2\theta_k(L+\mu)}{L-\mu}\right)\bar{\Phi}_{k-2}(\lambda),$$

where $\theta_k = \frac{1}{2\frac{L+\mu}{L-\mu} - \theta_{k-1}}$ and $\theta_1 = \frac{L-\mu}{L+\mu}$.

Now we are ready to derive the update of the Chebyshev method when $K > 2$:

$$\begin{aligned}
x_{K+1} - x_* &= \bar{\Phi}_K(A)(x_1 - x_*) \\
&= \frac{2\theta_K}{L-\mu}((L+\mu)I_d - 2A)\bar{\Phi}_{K-1}(A)(x_1 - x_*) + \left(1 - \frac{2\theta_K(L+\mu)}{L-\mu}\right)\bar{\Phi}_{K-2}(A)(x_1 - x_*) \\
&= \frac{2\theta_K}{L-\mu}((L+\mu)I_d - 2A)(x_K - x_*) + \left(1 - \frac{2\theta_K(L+\mu)}{L-\mu}\right)(x_{K-1} - x_*) \\
&= \beta_K(x_K - x_*) - \frac{4\theta_K}{L-\mu}\nabla f(x_K) + (1-\beta_K)(x_{K-1} - x_*)
\end{aligned}$$

From the above, we can conclude that when $k > 2$, the update is

$$x_{K+1} = x_K - \frac{4\theta_K}{L-\mu}\nabla f(x_K) + \beta_K(x_K - x_{K-1}),$$

where the momentum $\beta_K(x_K - x_{K-1})$ is the weighted sum of previous gradients.

3 Gradient Descent with the Chebyshev step size

Gradient Descent with a constant step size has the following dynamic:

$$x_{k+1} - x_* = (I_d - \eta A)(I_d - \eta A)\ldots(I_d - \eta A)(x_1 - x_*).$$

Gradient Descent

$$\|x_{K+1} - x_*\|_2 \leq \left(1 - \frac{2}{\kappa+1}\right)^K \|x_1 - x_*\|_2.$$

Q: What if we specify a scheme of non-constant step size in GD?

$$\begin{aligned} x_{k+1} &= x_k - \eta_k \nabla f(x_k) \\ &= x_k - \eta_k (A x_k - A x_*) \\ \Rightarrow x_{k+1} - x_* &= (I_d - \eta_k A)(x_k - x_*) \\ \Rightarrow x_{k+1} - x_* &= (I_d - \eta_k A)(I_d - \eta_{k-1} A)(x_{k-1} - x_*) \\ &= \cdots \end{aligned}$$

The dynamic becomes

$$x_{k+1} - x_* = (I_d - \eta_k A)(I_d - \eta_{k-1} A) \dots (I_d - \eta_1 A)(x_1 - x_*).$$

Hence

$$\|x_{K+1} - x_*\|_2 \leq \max_{i \in [d]} \left| \Pi_{k=1}^{K} (1 - \eta_k \lambda_i) \right| \|x_1 - x_*\|_2$$

We are going to use the following result:

(Chebyshev roots)

$$r_k^{(K)} := \frac{L+\mu}{2} - \frac{L-\mu}{2} \cos\left(\frac{(k - \frac{1}{2})\pi}{K} \right).$$

Equivalent form of $\bar{\Phi}_K(\lambda)$

$$\bar{\Phi}_K(\lambda) = \Pi_{k=1}^{K} \left(1 - \frac{\lambda}{r_k^{(K)}} \right).$$

Notice that if we set $\eta_k = \frac{1}{r_k^{(K)}}$, then we have $\min \eta_k \approx \frac{1}{L}$ and $\max \eta_k \approx \frac{1}{\mu}$.

Accelerating GD by the Chebyshev step size

Recall

$$\|x_{K+1} - x_*\|_2 \leq \max_{i \in [d]} \left| \Pi_{k=1}^{K} (1 - \eta_k \lambda_i) \right| \|x_1 - x_*\|_2.$$

Denote $\sigma(k)$ the k_{th} element of the array $[1, 2, \ldots, K]$ *after* an arbitrary permutation σ. Set $\eta_k = \frac{1}{r^{(K)}_{\sigma(k)}}$. Then, we have

$$\begin{aligned}
\|x_{K+1} - x_*\|_2 &\leq \max_{i \in [d]} \left| \Pi_{k=1}^{K} (1 - \eta_k \lambda_i) \right| \|x_1 - x_*\|_2 \\
&\leq \min_{\lambda \in [\mu, L]} |\bar{\Phi}_K(\lambda)| \|x_1 - x_*\|_2 \\
&\leq 2 \left(1 - \frac{2}{\sqrt{\kappa} + 1} \right)^K \|x_1 - x_*\|_2.
\end{aligned}$$

4 Going beyond quadratic?

(Negative result) Gradient descent with Chebyshev step size fails to converge [1]

$$f(x) = \log\cosh(x) + 0.01x^2$$

(Positive result) Gradient descent with a scheme of non-constant step size converges at a rate [2]

$$\|x_{k+1} - x_*\|_2 \leq \left(1 - \Theta\left(\frac{1}{\kappa^{0.7864}} \right) \right)^k \|x_1 - x_*\|_2.$$

Please refer to the reference for more.

5 Bibliographic notes

More materials about acceleration methods can be found in [1], [2], [3], and [4]

References

[1] Naman Agarwal, Surbhi Goel, Cyril Zhang, *Acceleration via Fractal Learning Rate Schedules*, arXiv:2103.01338

[2] Jason M. Altschuler, Pablo A. Parrilo, *Acceleration by Stepsize Hedging I: Multi-Step Descent and the Silver Stepsize Schedule*, arXiv:2309.07879

[3] Jun-Kun Wang, Andre Wibisono, *Accelerating Hamiltonian Monte Carlo via Chebyshev Integration Time*, arXiv:2207.02189

[4] Fabian Pedregosa. On the Link Between Optimization and Polynomials, Part 4 `https://fa.bianp.net/blog/2021/no-momentum/#sec2`