Scribes: Brateesh Roy, Hargen Zheng May 30, 2024
Editor/TA: Marialena Sfyraki

Lecture 17: Acceleration via Chebyshev Polynomial

1 Gradient Descent in Strongly Convex Quadratic Problems

Let's recall the general quadratic form from HW1

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}x^\top Ax - b^\top x, \text{ where } A \succ 0,$$

which can be demonstrated to be equivalent to the problem

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

Let $f(x) = \frac{1}{2}x^\top Ax - b^\top x$, then $\nabla f(x) = Ax - b$. Consider

$$x^* = \arg\min_{x\in\mathbb{R}^d} \frac{1}{2}x^\top Ax - b^\top x.$$

Then, x^* satisfies

$$Ax^* - b = 0 \Leftrightarrow x^* = A^{-1}b.$$

Question: Now that we have obtained a closed-form solution to this problem, why do we need to concern ourselves with Gradient Descent?
Answer: Computing A^{-1} for $A \in \mathbb{R}^{d\times d}$ is $O(d^3)$ in time complexity.

The Gradient Descent step in this problem is given as:

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

The computation of $Ax_k - b$ is of complexity $O(d^2)$ (can be better if A is sparse). There are $O(log\frac{1}{\epsilon})$ number of iterations. That makes the time complexity of Gradient Descent $O(d^2log(\frac{1}{\epsilon}))$ which is better than the closed-form solution computation for large d.

Coming back to the problem,

$$\begin{aligned}
x_{k+1} &= x_k - \eta \nabla f(x_k) \\
&= x_k - \eta (A x_k - b) \\
&= x_k - \eta (A x - A x_*) \\
\Leftrightarrow x_{k+1} - x_* &= (I_d - \eta A)(x_k - x_*) \\
&= (I_d - \eta A)^k (x_1 - x_*)
\end{aligned}$$

Note that $(I_d - \eta A)^k$ is a k-th degree polynomial of matrix A. Before proceeding further, let's introduce the concept of the spectral norm of a matrix.

Definition 1. ***(Spectral Norm of a Matrix)*** *: For a matrix $B \in \mathbb{R}^{m\times n}$ its spectral norm $\|B\|_2$ is defined as the largest singular value of B, that is*

$$\|B\|_2 := \sigma_{max}(B) = \max_{x:\|x\|_2=1} \|Bx\|_2.$$

Fact: $\|B\|_2 = \sqrt{\lambda_{max}(B^\top B)}$
For a square matrix $B \in \mathbb{R}^{n\times n}$, if B is diagonalizable, i.e.,

$$\begin{gathered}
\exists\ U,\ \Lambda \in \mathbb{R}^{n\times n},\ U^\top U = I_n, \Lambda \ \textit{diagonal s.t.} \\
B = U \Lambda U^{-1},
\end{gathered}$$

then

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

Observe that

$$\begin{aligned}
B^\top B &= \left(U\Lambda U^{-1}\right)^\top \left(U \Lambda U^{-1}\right) \\
&= U^{-\top} \Lambda \underbrace{U^\top U}_{I_d} \Lambda U^{-1} \\
&= U^{-\top} \Lambda^2 U^{-1} \\
&= U \Lambda^2 U^{-1}.
\end{aligned}$$

Example: Let

$$\Lambda = \begin{bmatrix} 1 & 0 \\ 0 & -7 \end{bmatrix} \Rightarrow \Lambda^2 = \begin{bmatrix} 1 & 0 \\ 0 & 49 \end{bmatrix}.$$

Therefore,

$$\|B\|_2 = \sqrt{49}.$$

Now, we had

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

Taking the L_2 norm of both sides, we obtain:

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

Now, let's analyze the matrix $I_d - \eta A$. Since $A \succ 0$, A is diagonalizable as $A = U\Lambda U^\top$ where U is an orthonormal matrix and Λ is a diagonal matrix whose entries are the eigenvalues of A.

$$\begin{aligned} I_d - \eta A &= UU^\top - U\Lambda U^\top \\ &= U(I_d - \eta\Lambda)U^\top \end{aligned}$$

It can be seen that the eigenvalues of $I_d - \eta A$ are given by the entries of $I_d - \eta\Lambda$ which are equal to $\big(1 - \eta\lambda_i(A)\big)_{i=1}^d$. Thus,

$$\begin{aligned} \|x_{k+1} - x_*\|_2 &\leq \|I_d - \eta A\|_2 \|x_k - x_*\|_2 \\ &= \max_{i \in [d]} \big|1 - \eta\lambda_i(A)\big| \, \|x_k - x_*\|_2 \end{aligned}$$

Let $\mu = \lambda_{min}(A)$ and $L = \lambda_{max}(A)$. Now, the previous inequality holds for any η. We would like to choose such a value for η as to tighten down the upper bound on the R.H.S., i.e. :

$$\min_\eta \max_{i \in [d]} \big|1 - \eta\lambda_i(A)\big|$$

Thus, we have a min-max problem.

1.1 Finding the Optimal η

Now, we have that:

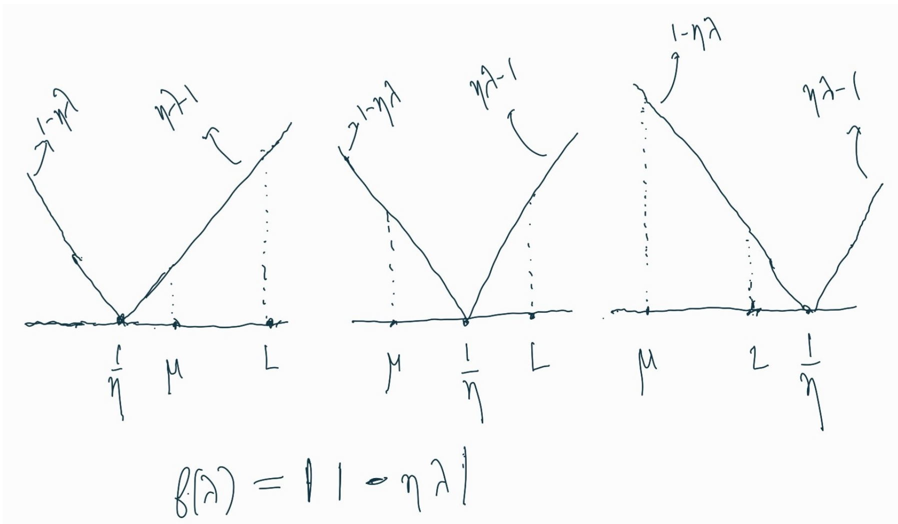
$$\min_\eta \max_{i \in [d]} \big|1 - \eta\lambda_i(A)\big| \leq \min_\eta \max_{\lambda \in [\mu, L]} |1 - \eta\lambda|$$

For a fixed value of η, let's analyze the function $|1 - \eta\lambda|$ to identify where the max lies and what it evaluates to.

$$|1 - \eta\lambda| = \begin{cases} 1 - \eta\lambda & , \text{ if } \lambda \leq \frac{1}{\eta} \\ \eta\lambda - 1 & , \text{ if } \lambda \geq \frac{1}{\eta} \end{cases}$$

This is a scaled and shifted version of the V-shaped modulus function, with the tip of the V at $\frac{1}{\eta}$. Now, depending on where $\frac{1}{\eta}$ lies w.r.t. μ and L, we can have three cases:

(i) $\frac{1}{\eta} \leq \mu$, (ii) $\mu \leq \frac{1}{\eta} \leq L$, (iii) $L \leq \frac{1}{\eta}$



Case 1: $\frac{1}{\eta} \leq \mu$. Since $\lambda \in [\mu, L]$, $\lambda \geq \frac{1}{\eta}$. Therefore,

$$|1 - \eta\lambda| = \eta\lambda - 1.$$

The max occurs at $\lambda = L$, that is

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

The max evaluates out to be $\eta L - 1$
However,

$$\frac{1}{\eta} \leq \mu \leq L \implies 1 - \eta\mu \leq 0 \leq \eta L - 1.$$

Therefore:

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

Case 2: $\mu \leq \frac{1}{\eta} \leq L$. Since $\lambda \in [\mu, L]$, $\lambda \geq \frac{1}{\eta}$. Therefore,

$$|1 - \eta\lambda| = \eta\lambda - 1$$

The max occurs at the boundaries, either $\lambda = L$ or $\lambda = \mu$.

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

However,

$$\mu \leq \frac{1}{\eta} \leq L \implies 0 \leq 1 - \eta\mu, \ 0 \leq \eta L - 1$$
$$\implies |1 - \eta\mu| = 1 - \eta\mu, \text{ and } |\eta L - 1| = \eta L - 1.$$

Therefore:

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

Case 3 (Similar to Case 1): $L \leq \frac{1}{\eta}$. Since $\lambda \in [\mu, L]$, $\lambda \leq \frac{1}{\eta}$. Therefore,

$$|1 - \eta\lambda| = 1 - \eta\lambda.$$

The max occurs at $\lambda = \mu$.

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

The max evaluates out to be $1 - \eta\mu$. However,

$$\mu \leq L \leq \frac{1}{\eta} \implies 1 - \eta\mu \geq 0 \geq \eta L - 1.$$

Therefore:

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

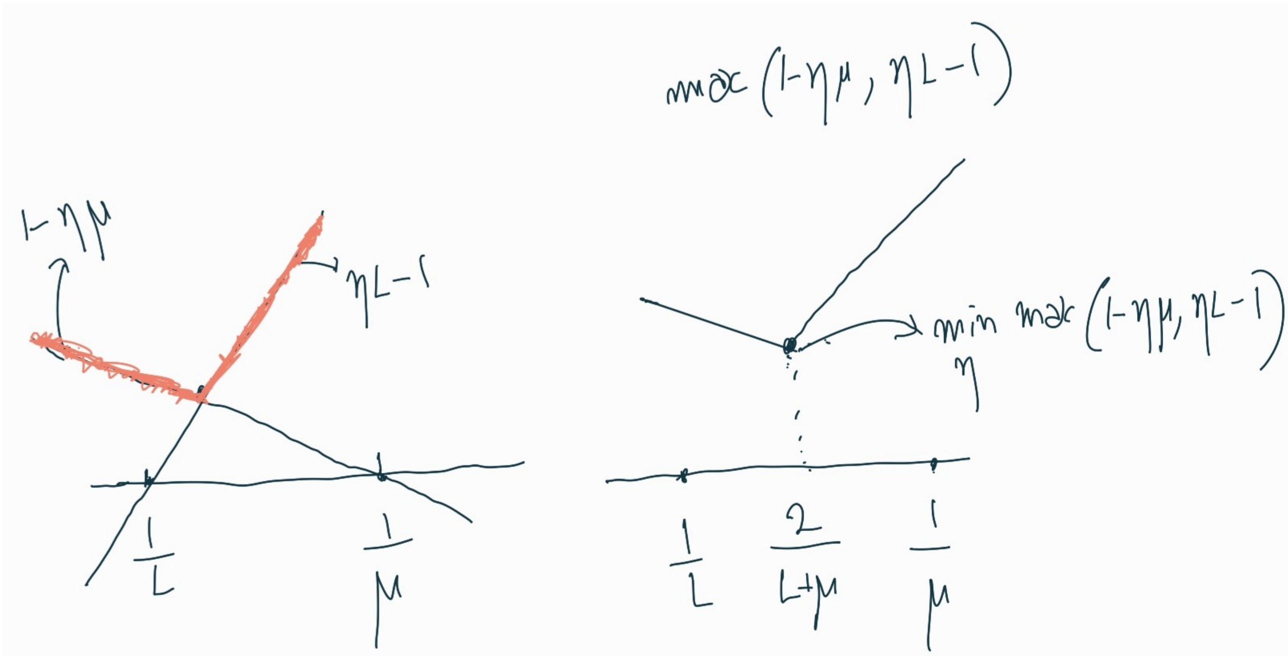
As it turns out, in all cases the max evaluates out to be:

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

Therefore, the min-max problem evaluates to:

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

Now, let's see from the η-player's perspective. The value of η that minimizes this max function happens when the two lines cross each other:



$$
\begin{aligned}
1-\eta\mu &= \eta L - 1 \\
\Leftrightarrow \eta &= \frac{2}{\mu+L}
\end{aligned}
$$

For the optimal $\eta = \frac{2}{L+\mu}$,

$$
\begin{aligned}
\|x_{k+1}-x_*\|_2 &\le \max_{i\in[d]} |1-\eta\lambda_i| \, \|x_k - x_*\|_2 \\
&\le \max_{\lambda\in[\mu,L]} |1-\eta\lambda| \, \|x_k - x_*\|_2 \\
&\le \max_{\lambda\in[\mu,L]} \left|1-\frac{2\lambda}{L+\mu}\right| \|x_k - x_*\|_2 \\
&= \left(1-\frac{2\mu}{L+\mu}\right) \|x_k - x_*\|_2 \quad \text{(piecewise linear function)} \\
&= \left(1-\frac{2\mu}{L+\mu}\right)^k \|x_1 - x_*\|_2 \quad \text{(by recursive expansion)}
\end{aligned}
$$

Note that $\left|1-\frac{2\lambda}{L+\mu}\right|$ is a piece-wise linear function. The argmax of $\left|1-\frac{2\lambda}{L+\mu}\right|$ would be either μ or L and it turns out it would be μ in this case. That how we obtained $\max_{\lambda\in[\mu,L]}\left|1-\frac{2\lambda}{L+\mu}\right| = \left(1-\frac{2\mu}{L+\mu}\right)$.

We can get convergence rate as follows:

$$\begin{aligned}
\|x_{k+1} - x_*\|_2 &\le \left(1 - \frac{2\mu}{L+\mu}\right)^k \|x_1 - x_*\|_2 \\
&= \left(1 - \frac{2}{\kappa+1}\right)^k \|x_1 - x_*\|_2 \\
&= \left(1 - \Theta\left(\frac{1}{\kappa}\right)\right)^k \|x_1 - x_*\|_2
\end{aligned}$$

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

2 Chebyshev Polynomials

Consider any algorithm in the form:

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

Lemma 1. *Consider solving* $\min_x \frac{1}{2}x^\top Ax - b^\top x$*. Algorithms in the form of* (1) *has 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 will use induction.
Base case:

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

where $P_0(A) = 1$. Suppose at k, we have

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

Consider $k+1$,

$$x_{k+1} - x_* = x_1 - x_* + \underbrace{\sum_{j=1}^{k} \alpha_j \nabla f(x_j)}_{\text{span of gradients}},$$

where $\{\alpha_j\}$ are some co-efficients.

We can expand as follows:

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

□

Here, given

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

our goal is to find the best K-degree polynomial:

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

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

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

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

Here is an equivalent definition:

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

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

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

where $h(\cdot)$ is the mapping $h(\lambda) := \frac{L+\mu-2\lambda}{L-\mu}$.
Observe that the mapping $h(\cdot)$ maps all $\lambda \in [\mu, L]$ into the interval $[-1, 1]$:

- $h(\mu) = \frac{L+\mu-2\mu}{L-\mu} = 1$.
- $h(L) = \frac{L+\mu-2L}{L-\mu} = -1$.

As a result, by the definition of K-degree Chebyshev Polynomial of the first kind, we have

$$\Phi_K(h(\lambda)) \leq 1.$$

Also, we have

$$h(0) = \frac{L+\mu}{L-\mu} = 1 + \frac{2\mu}{L-\mu} > 1,$$

so by the properties of Chebyshev Polynomial, $\Phi_K(h(0))$ would grow exponentially.

Lemma 2. *(see e.g., Lemma 3 in [Wang (2023)] and Section 2.3 in [dAspremont et al. (2021)]) For any positive integer K, we have*

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

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:

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

$$= O\left(\left(1 - \Theta\left(\sqrt{\frac{\mu}{L}}\right)\right)^K\right).$$

□

We have derived the dynamic of gradient descent as

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

For Chebyshev method, we have

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

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

For example, suppose $\kappa = 100$. Then, $1 - \frac{2}{\kappa+1} \cong 0.98$ and $1 - \frac{2}{\sqrt{\kappa}+1} \cong 1 - \frac{2}{11} \approx 0.8$.

Having a dependency of square root of condition number κ is considered to be better than having a linear dependency of the condition number because $1 - \frac{2}{\sqrt{\kappa}+1} \le 1 - \frac{2}{k+1}$ as $\kappa \ge 1$.

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

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

where β_K is called the momentum parameter and $\beta_K(x_K - x_{K-1})$ is the momentum term (weighted average of previous gradients).
If we set a constant step size for gradient descent, we have

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

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

$$x_{k+1} = x_k - \eta_k \nabla f(x_k).$$

Answer: Here, we have $x_{k+1} = x_k - \eta_k(Ax_k - Ax_*) \Rightarrow x_{k+1} - x_* = (I_d - \eta_k A)(x_k - x_*)$. The dynamic becomes

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

Hence

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

Chebyshev roots are given as

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

and

$$\bar{\Phi}_k(r_k^{(K)}) = 0.$$

The equivalent form of $\bar{\Phi}_K(\lambda)$ is given as

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

The convergence rate thus becomes

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

where the inequality is by Lemma 2.

To go beyond quadratic, we have the following two results:

Negative result: Gradient descent with Chebyshev step size fails to converge [Agarwal et al. (2021)]

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

Positive result: Gradient descent with a scheme of non-constant step size converges at a rate [Altschuler et al. (2023)]

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

Bibliographic notes

More prelimiaries of calculus and linear algebra can be found in Chapter 1 of [Drusvyatskiy (2020)] and Chapter 2 of [Vishnoi (2021)].

References

[Drusvyatskiy (2020)] Dmitriy Drusvyatskiy. Convex Analysis and Nonsmooth Optimization. 2020.

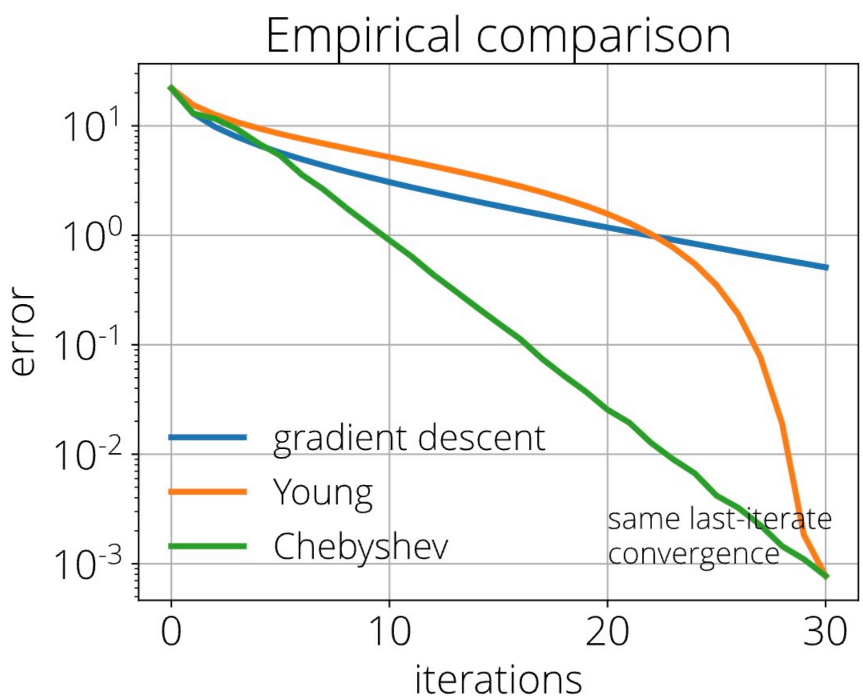


Figure 1: Comparison of GD with a constant step size, GD with Chebyshev step size (Young's method), and Chebyshev method. Picture taken from [Pedregosa (2021)].

[Vishnoi (2021)] Nisheeth K. Vishnoi. Algorithms for Convex Optimization. Cambridge University Press, 2021.

[Agarwal et al. (2021)] Naman Agarwal, Surbhi Goel, Cyril Zhang. Acceleration via Fractal Learning Rate Schedules. ICML 2021.

[Altschuler et al. (2023)] Jason M. Altschuler, Pablo A. Parrilo. Acceleration by Stepsize Hedging I: Multi-Step Descent and the Silver Stepsize Schedule. arXiv:2309.07879. 2023

[Wang (2023)] Jun-Kun Wang and Andre Wibisono. Accelerating Hamiltonian Monte Carlo via Chebyshev Integration Time. ICLR 2023.

[dAspremont et al. (2021)] Alexandre dAspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. Foundations and Trends in Optimization. 2021.

[Pedregosa (2021)] Fabian Pedregosa. Acceleration without Momentum. `http://fabianp.net/blog/2021/no-momentum/` 2021.