

Optimization In AI

Class 03: Gradient Descent

Instructor: Jianhao Ma

Department of Industrial Engineering, Tsinghua University

Announcements

1. Class Rescheduling

I will be traveling for INFORMS.

The **November 4** class will be rescheduled.

2. Homework 1

Released **today**.

Two weeks to complete.



[Rescheduling Survey](#)

Today's Topics

- 01 Descent And Stationarity
- 02 Smooth Convex Functions
- 03 Strong Convexity
- 04 Nonsmooth Convex Functions
- 05 Frontier: Large Steps

Convergence Rates And Error Metrics

Assumptions / Method	Step Size	Certified Quantity And Rate
Smooth, lower bounded	$1/L$	$\min_{0 \leq t < T} \ \nabla f(x_t)\ ^2 = O(1/T)$
Smooth convex, minimizer	$1/L$	Last gap: $f(x_T) - f^* = O(1/T)$
Smooth strongly convex	$1/L$	Last gap: $\Delta_T \leq (1 - \mu/L)^T \Delta_0$
Convex subgradient, fixed	$R/(G\sqrt{T})$	Averaged gap: $f(\bar{x}_T) - f^* \leq RG/\sqrt{T}$

$$\Delta_t = f(x_t) - f^*, \quad \bar{x}_T = T^{-1} \sum_{t=0}^{T-1} x_t.$$

PART 01

Descent And Stationarity

Minimize The Quadratic Upper Model

Recall: $f : \mathbb{R}^d \rightarrow \mathbb{R}$ has an L -Lipschitz gradient, $L > 0$.

Lemma (Descent Lemma)

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2.$$

Minimize The Model In y

$$0 = \nabla f(x) + L(y - x) \implies \boxed{x^+ = x - \frac{1}{L} \nabla f(x).}$$

Gradient Descent: One-Step Decrease

$$x^+ = x - \eta \nabla f(x)$$

$$f(x^+) \leq f(x) - \underbrace{\eta \left(1 - \frac{L\eta}{2}\right)}_{c(\eta)} \|\nabla f(x)\|^2.$$

Maximize $c(\eta)$ $\eta \geq 0$

$$c'(\eta) = 1 - L\eta = 0 \quad \implies \quad \eta = 1/L.$$

Lemma (One-Step Decrease Of Gradient Descent)

$$f(x^+) \leq f(x) - \frac{1}{2L} \|\nabla f(x)\|^2.$$

Nonconvex GD: Stationarity Rate

Theorem (Gradient Descent: Stationarity Bound)

Suppose f is L -smooth and $f^* := \inf_{x \in \mathbb{R}^d} f(x) > -\infty$. For $x_{t+1} = x_t - \nabla f(x_t)/L$ and every integer $T \geq 1$,

$$\min_{0 \leq t < T} \|\nabla f(x_t)\|^2 \leq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \leq \frac{2L(f(x_0) - f^*)}{T}.$$

$$\min_{t < T} \|\nabla f(x_t)\| = O(T^{-1/2}) \implies T = O(\epsilon^{-2}) \text{ for } \epsilon\text{-stationarity.}$$

Proof: Telescope The Decrease

One-Step Progress $\eta = 1/L$

$$\frac{1}{2L} \|\nabla f(x_t)\|^2 \leq f(x_t) - f(x_{t+1}).$$

Telescope

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

Minimum $\leq$ Average

$$\min_{0 \leq t < T} \|\nabla f(x_t)\|^2 \leq \frac{2L(f(x_0) - f^*)}{T}.$$

PART 02

Smooth Convex Functions

Smooth Convex GD: Convergence Rate

Theorem (Gradient Descent: Sublinear Convergence)

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex and L -smooth, $L > 0$, with minimizer x^* . For $x_{t+1} = x_t - \nabla f(x_t)/L$, let $R = \|x_0 - x^*\|$. For every integer $T \geq 1$,

$$f(x_T) - f^* \leq \frac{2LR^2}{T}.$$

Proof 1: Gap Recurrence

Proof 2: Potential

$$\frac{2LR^2}{T} \longrightarrow \frac{LR^2}{2T}.$$

Proof: From Descent To A Gap Recurrence

$$R > 0$$

Smoothness

$$\Delta_{t+1} \leq \Delta_t - \frac{1}{2L} \|\nabla f(x_t)\|^2.$$

Convexity

$$\|x_t - x^*\| \leq R$$

$$\Delta_t \leq \langle \nabla f(x_t), x_t - x^* \rangle \leq \|\nabla f(x_t)\| \|x_t - x^*\| \leq R \|\nabla f(x_t)\|.$$

Combine

$$\Delta_{t+1} \leq \Delta_t - \frac{\Delta_t^2}{2LR^2}.$$

Proof: Distance To A Minimizer Does Not Increase

Lemma (Nonincreasing Distance To A Minimizer)

For every $t \geq 0$, $\|x_{t+1} - x^*\| \leq \|x_t - x^*\| \leq R$.

Expand The Update

$$\begin{aligned}\|x_{t+1} - x^*\|^2 &= \|x_t - x^*\|^2 - \frac{2}{L} \langle \nabla f(x_t), x_t - x^* \rangle + \frac{\|\nabla f(x_t)\|^2}{L^2} \\ &\leq \|x_t - x^*\|^2 - \frac{2\Delta_t}{L} + \frac{2\Delta_t}{L} \\ &= \|x_t - x^*\|^2.\end{aligned}$$

Inequality: Convexity + Descent

$$\langle \nabla f(x_t), x_t - x^* \rangle \geq \Delta_t, \quad \frac{\|\nabla f(x_t)\|^2}{2L} \leq \Delta_t - \Delta_{t+1} \leq \Delta_t.$$

Induction: $\|x_t - x^*\| \leq R$.

Proof: Reciprocal Gap Growth

If $R = 0$ or $\Delta_T = 0$, the claim is immediate. Assume $\Delta_T > 0$; $C = 2LR^2 > 0$.

Smoothness + Descent

$$0 < \Delta_t \leq \Delta_0 \leq \frac{L}{2}R^2 = \frac{C}{4}.$$

Reciprocals $1 - \Delta_t/C > 0$

$$\frac{1}{\Delta_{t+1}} \geq \frac{1}{\Delta_t(1 - \Delta_t/C)} \geq \frac{1}{\Delta_t} + \frac{1}{C}, \quad \frac{1}{1-z} \geq 1+z \quad (0 \leq z < 1).$$

Telescope

$$\frac{1}{\Delta_T} \geq \frac{1}{\Delta_0} + \frac{T}{2LR^2} \implies \boxed{\Delta_T \leq \frac{1}{\Delta_0^{-1} + T/(2LR^2)} \leq \frac{2LR^2}{T}}.$$

Proof 2: From Potential To Rate

Definition (Smooth Convex Potential)

$$\Phi_t = t\Delta_t + \frac{L}{2}D_t^2, \quad D_t = \|x_t - x^*\|.$$

If $\Phi_{t+1} \leq \Phi_t$ For $0 \leq t < T$

$$\Phi_T \leq \Phi_{T-1} \leq \dots \leq \Phi_0 = \frac{L}{2}R^2.$$

$$T\Delta_T \leq T\Delta_T + \frac{L}{2}D_T^2 = \Phi_T \leq \frac{L}{2}R^2.$$

Corollary (Gradient Descent: Last-Iterate Bound)

$$\Delta_T \leq \frac{LR^2}{2T} \quad (T \geq 1).$$

It Remains To Prove: $\Phi_{t+1} \leq \Phi_t$.

Proof 2: The Potential Does Not Increase

Expand The Squared Distance $\eta = 1/L$

$$\begin{aligned}\frac{L}{2}(D_t^2 - D_{t+1}^2) &= \langle \nabla f(x_t), x_t - x^* \rangle - \frac{\|\nabla f(x_t)\|^2}{2L} \\ &\geq \Delta_t - \frac{\|\nabla f(x_t)\|^2}{2L} && \text{convexity} \\ &\geq \Delta_{t+1} && \text{descent.}\end{aligned}$$

Expand The Potential Change And Substitute

$$\begin{aligned}\Phi_{t+1} - \Phi_t &= t(\Delta_{t+1} - \Delta_t) + \Delta_{t+1} + \frac{L}{2}(D_{t+1}^2 - D_t^2) \\ &\leq t(\Delta_{t+1} - \Delta_t) \leq 0 && \text{by descent.}\end{aligned}$$

Example: Linear Regression

$$A \in \mathbb{R}^{n \times d}, \quad b \in \mathbb{R}^n$$

$$f(x) = \frac{1}{2} \|Ax - b\|^2.$$

Gradient + Curvature

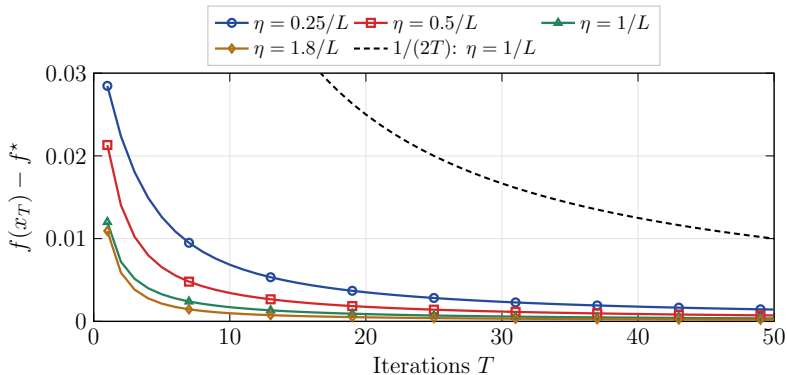
$$\nabla f(x) = A^\top (Ax - b), \quad \nabla^2 f(x) = A^\top A \succeq 0.$$

Smooth Convex Structure

$$L = \|A\|_2^2 \ (A \neq 0), \quad \arg \min f \neq \emptyset$$

Simulation: Linear Regression

$$d = 200, \quad A = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_d}), \quad b = 0$$
$$\lambda_j = 10^{-6(j-1)/(d-1)} \quad (j = 1, \dots, d), \quad x_0 = \mathbf{1}/\sqrt{d}, \quad L = 1$$



PART 03

Smooth Strongly Convex Functions

Strongly Convex GD: Geometric Convergence

Theorem (Gradient Descent: Linear Convergence)

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be L -smooth and μ -strongly convex, $0 < \mu \leq L$, with minimizer x^* . GD with $\eta = 1/L$ satisfies, for every integer $T \geq 1$,

$$f(x_T) - f^* \leq \left(1 - \frac{\mu}{L}\right)^T (f(x_0) - f^*).$$

Proof: The Same Descent, A Linear Recurrence

One-Step Decrease Of Gradient Descent $\eta = 1/L$

$$\begin{aligned}\Delta_{t+1} &\leq \Delta_t - \frac{1}{2L} \|\nabla f(x_t)\|^2 \\ &\leq \Delta_t - \frac{\mu}{L} \Delta_t = \left(1 - \frac{\mu}{L}\right) \Delta_t, \quad \|\nabla f(x_t)\|^2 \geq 2\mu \Delta_t.\end{aligned}$$

Iterate $\kappa = L/\mu$

$$\boxed{\Delta_T \leq (1 - \mu/L)^T \Delta_0} \leq e^{-T/\kappa} \Delta_0.$$

Geometry	Guaranteed Decrease	Gap Rate
Convex, $R > 0$	$\Delta_t^2/(2LR^2)$	$O(1/T)$
Strongly convex	$(\mu/L)\Delta_t$	Geometric

Proof: A Stronger Gap–Gradient Relation

Recall Convexity $\Delta_t \leq \|\nabla f(x_t)\| D_t$

Strong Convexity: Substitute $y = x^*$

$$\begin{aligned}\Delta_t &\leq \langle \nabla f(x_t), x_t - x^* \rangle - \frac{\mu}{2} D_t^2 && \text{strong convexity} \\ &\leq \|\nabla f(x_t)\| D_t - \frac{\mu}{2} D_t^2 && \text{Cauchy–Schwarz.}\end{aligned}$$

Complete The Square In D_t

$$\|\nabla f(x_t)\| D_t - \frac{\mu}{2} D_t^2 = \frac{\|\nabla f(x_t)\|^2}{2\mu} - \frac{\mu}{2} \left(D_t - \frac{\|\nabla f(x_t)\|}{\mu} \right)^2 \leq \frac{\|\nabla f(x_t)\|^2}{2\mu}.$$

$$\boxed{\|\nabla f(x_t)\|^2 \geq 2\mu \Delta_t.}$$

Potential View: An Exponential Weight

$$0 < \mu < L, \quad q = 1 - \mu/L \in (0, 1)$$

Definition (Strongly Convex Potential)

$$\Phi_t = q^{-t} \Delta_t.$$

If $\Phi_{t+1} \leq \Phi_t$ For $0 \leq t < T$

$$\Phi_T \leq \dots \leq \Phi_0 = \Delta_0 \implies \boxed{\Delta_T = q^T \Phi_T \leq q^T \Delta_0.}$$

Now Prove Monotonicity: Use $\Delta_{t+1} \leq q\Delta_t$

$$\begin{aligned} \Phi_{t+1} - \Phi_t &= q^{-(t+1)} \Delta_{t+1} - q^{-t} \Delta_t \\ &= q^{-(t+1)} (\Delta_{t+1} - q\Delta_t) \leq 0. \end{aligned}$$

$$\mu = L : \quad 0 \leq \Delta_{t+1} \leq 0 \implies \Delta_T = 0 \quad (T \geq 1).$$

Strongly Convex GD: Distance Convergence

Theorem (Gradient Descent: Distance Contraction)

Let $f \in C^2(\mathbb{R}^d)$ be μ -strongly convex and L -smooth, $0 < \mu \leq L$, with minimizer x^* . For GD with constant $0 < \eta < 2/L$,

$$r_\eta := \max\{|1 - \eta\mu|, |1 - \eta L|\} < 1,$$

$$\|x_T - x^*\| \leq r_\eta^T \|x_0 - x^*\|, \quad T \geq 1.$$

Step Size	Distance Factor r_η
$1/L$	$1 - \mu/L$
$2/(L + \mu)$	$(L - \mu)/(L + \mu)$

Proof: Average The Hessian Along A Segment

$$e_t = x_t - x^*, \quad z_t(s) = x^* + se_t, \quad 0 \leq s \leq 1$$

Fundamental Theorem Of Calculus $\nabla f(x^*) = 0$

$$\begin{aligned} \nabla f(x_t) &= \nabla f(x_t) - \nabla f(x^*) \\ &= \int_0^1 \nabla^2 f(z_t(s)) e_t ds = \bar{H}_t e_t, \quad \bar{H}_t := \int_0^1 \nabla^2 f(z_t(s)) ds. \end{aligned}$$

Strong Convexity + Smoothness $\forall x$

$$\mu I \preceq \nabla^2 f(x) \preceq LI.$$

Integrate The Bounds

$$\boxed{\mu I \preceq \bar{H}_t \preceq LI}, \quad \bar{H}_t = \bar{H}_t^\top.$$

Proof: Contract The Distance By A Spectral Bound

Exact Error Update

$$e_{t+1} = e_t - \eta \nabla f(x_t) = (I - \eta \bar{H}_t) e_t.$$

Symmetry + Eigenvalues In $[\mu, L]$

$$\begin{aligned} \|e_{t+1}\| &\leq \|I - \eta \bar{H}_t\|_2 \|e_t\| \\ &= \max_i |1 - \eta \lambda_i(\bar{H}_t)| \|e_t\| \\ &\leq \underbrace{\max\{|1 - \eta \mu|, |1 - \eta L|\}}_{r_\eta} \|e_t\|. \end{aligned}$$

Iterate The Same Uniform Bound

$$\|x_T - x^*\| \leq r_\eta^T \|x_0 - x^*\|.$$

$$\eta = \frac{2}{L+\mu} : r_\eta = \frac{L-\mu}{L+\mu}. \quad \mu = L : \eta = 1/L \Rightarrow x_1 = x^*.$$

Example: Ridge Regression

Quadratic Regularization

$$f_{\lambda}(x) = \frac{1}{2} \|Ax - b\|^2 + \frac{\lambda}{2} \|x\|^2, \quad \lambda > 0.$$

Update

$$x_{t+1} = x_t - \eta \left[A^{\top} (Ax_t - b) + \lambda x_t \right].$$

Curvature

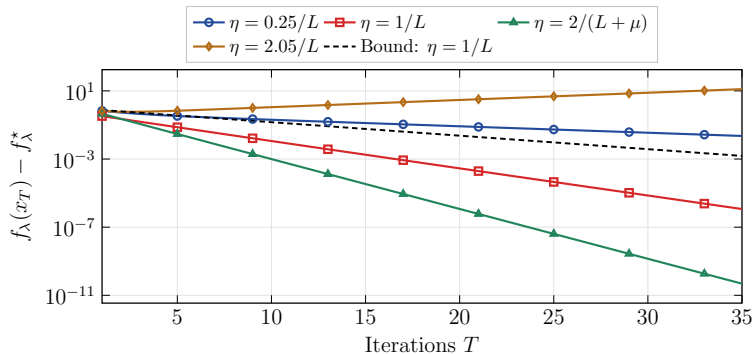
$$L = \lambda_{\max}(A^{\top}A) + \lambda, \quad \mu = \lambda_{\min}(A^{\top}A) + \lambda > 0.$$

$$\eta = 1/L \quad \Rightarrow \quad q = 1 - \mu/L$$

Simulation: Ridge Regression

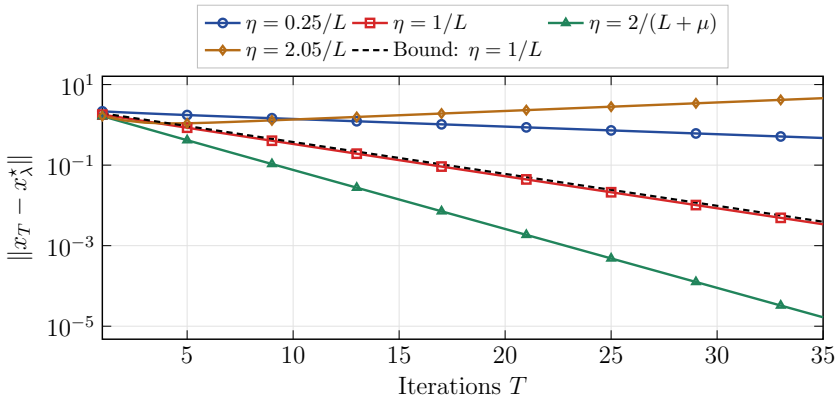
$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/5 & 0 \end{pmatrix}, \quad b = (1, 1)^\top, \quad x_0 = (0, 0, 2)^\top$$

$$\lambda = 1/5, \quad L = 6/5, \quad \mu = 1/5, \quad x_\lambda^* = (5/6, 5/6, 0)^\top$$



Simulation: Ridge Distance To The Minimizer

$$\lambda = 1/5, \quad L = 6/5, \quad \mu = 1/5, \quad x_0 = (0, 0, 2)^\top$$



PART 04

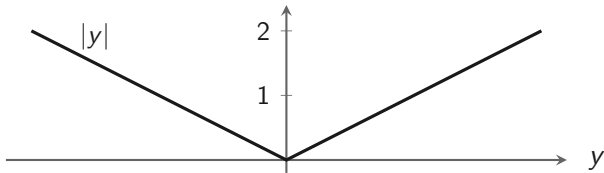
Nonsmooth Convex Functions

Subgradients Are Supporting Lower Bounds

Definition (Convex Subgradient)

For convex $f : \mathbb{R}^d \rightarrow \mathbb{R}$,

$$g \in \partial f(x) \iff f(y) \geq f(x) + \langle g, y - x \rangle \quad \forall y \in \mathbb{R}^d.$$



Subgradient Method

Setting

Minimize a convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$; start from $x_0 \in \mathbb{R}^d$.

Algorithm

For a constant stepsize $\eta > 0$,

$$x_{t+1} = x_t - \eta g_t, \quad g_t \in \partial f(x_t).$$

Output After T Steps

Uniform average $\bar{x}_T$ of $x_0, \dots, x_{T-1}$.

Subgradient: Convergence Rate

Assumptions

$f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex and has a minimizer x^* .

$$\|x_0 - x^*\| \leq R, \quad g_t \in \partial f(x_t), \quad \|g_t\| \leq G \quad (0 \leq t < T).$$

Theorem (Subgradient Method: Averaged-Iterate Bound)

For integer $T \geq 1$ and constant $\eta > 0$, use $x_{t+1} = x_t - \eta g_t$ and return the uniform average $\bar{x}_T$. Then

$$f(\bar{x}_T) - f^* \leq \frac{R^2}{2\eta T} + \frac{\eta G^2}{2}.$$

$$R, G > 0 : \quad \eta = \frac{R}{G\sqrt{T}} \quad \implies \quad f(\bar{x}_T) - f^* \leq \frac{RG}{\sqrt{T}}.$$

Proof: A One-Step Distance Bound

1. Expand The Squared Distance

$$D_{t+1}^2 = D_t^2 - 2\eta \langle g_t, x_t - x^* \rangle + \eta^2 \|g_t\|^2.$$

2. Apply The Subgradient Inequality

$$f^* \geq f(x_t) + \langle g_t, x^* - x_t \rangle \implies \langle g_t, x_t - x^* \rangle \geq \Delta_t.$$

Using $\|g_t\| \leq G$,

$$D_{t+1}^2 \leq D_t^2 - 2\eta \Delta_t + \eta^2 G^2.$$

Rearrange

$$\boxed{2\eta \Delta_t \leq D_t^2 - D_{t+1}^2 + \eta^2 G^2.}$$

Proof: Telescope And Average

3. Telescope The Distance Terms

$$2\eta \sum_{t=0}^{T-1} \Delta_t \leq D_0^2 - D_T^2 + T\eta^2 G^2 \leq R^2 + T\eta^2 G^2.$$

Jensen's Inequality

$$f(\bar{x}_T) - f^* \leq \frac{1}{T} \sum_{t=0}^{T-1} \Delta_t \leq \frac{R^2}{2\eta T} + \frac{\eta G^2}{2}.$$

Minimize The Bound: Balance Its Two Terms $R, G > 0$

$$\eta = \frac{R}{G\sqrt{T}} \implies \frac{R^2}{2\eta T} = \frac{\eta G^2}{2} = \frac{RG}{2\sqrt{T}}.$$

Diminishing Stepsizes: Weighted Averaging

$f : \mathbb{R}^d \rightarrow \mathbb{R}$ convex; $x^* \in \arg \min f$; $\|x_0 - x^*\| \leq R$; $R, G > 0$.

$x_{t+1} = x_t - \eta_t g_t$, $\eta_t > 0$; along this run, $g_t \in \partial f(x_t)$, $\|g_t\| \leq G$ for all $t \geq 0$.

Proof: Sum The Distance Bounds $T \geq 1$

$$2\eta_t \Delta_t \leq D_t^2 - D_{t+1}^2 + \eta_t^2 G^2,$$

$$2 \sum_{t=0}^{T-1} \eta_t \Delta_t \leq R^2 + G^2 \sum_{t=0}^{T-1} \eta_t^2.$$

Jensen's Inequality

$$\hat{x}_T = \frac{\sum_{t=0}^{T-1} \eta_t x_t}{\sum_{t=0}^{T-1} \eta_t}.$$

$$f(\hat{x}_T) - f^* \leq \frac{R^2 + G^2 \sum_{t=0}^{T-1} \eta_t^2}{2 \sum_{t=0}^{T-1} \eta_t}.$$

The $1/\sqrt{t+1}$ Schedule

Set $\eta_t = R/(G\sqrt{t+1})$ $t \geq 0$; every integer $T \geq 1$.

Use $\sum_{k=1}^T 1/k \leq 1 + \log T$ and $\sum_{k=1}^T 1/\sqrt{k} \geq \sqrt{T}$.

Corollary

$$f(\hat{x}_T) - f^* \leq \frac{RG(1 + \sum_{k=1}^T 1/k)}{2 \sum_{k=1}^T 1/\sqrt{k}} \leq \frac{RG(2 + \log T)}{2\sqrt{T}}.$$

	Fixed Step	Diminishing Step
Learning rate	$R/(G\sqrt{T})$	$R/(G\sqrt{t+1})$
Needs horizon T ?	Yes	No
Output	Uniform average $\bar{x}_T$	Stepsize-weighted $\hat{x}_T$
Gap guarantee	$RG/\sqrt{T}$	$O(RG \log T/\sqrt{T})$

Potential View: Subgradient Convergence

Return To Constant $\eta > 0$ fixed $T \geq 1, 0 \leq t \leq T$.

Use the subgradient theorem's assumptions along this run.

Definition (Subgradient Potential)

$$\Phi_t = D_t^2 + 2\eta \sum_{s=0}^{t-1} \Delta_s + (T - t)\eta^2 G^2.$$

If $\Phi_{t+1} \leq \Phi_t$ For $0 \leq t < T$ $\Phi_T \leq \dots \leq \Phi_0$.

$$\Phi_0 = D_0^2 + T\eta^2 G^2, \quad \Phi_T = D_T^2 + 2\eta \sum_{t=0}^{T-1} \Delta_t.$$

$$2\eta \sum_{t=0}^{T-1} \Delta_t \leq \Phi_T \leq \Phi_0 \leq R^2 + T\eta^2 G^2.$$

Potential View: Rate And Monotonicity

First Recover The Rate: Jensen And Division By $2\eta T > 0$

$$f(\bar{x}_T) - f^* \leq \frac{1}{T} \sum_{t=0}^{T-1} \Delta_t \leq \frac{R^2}{2\eta T} + \frac{\eta G^2}{2}.$$

Substitute $\eta = R/(G\sqrt{T})$ $R, G > 0$

$$\boxed{f(\bar{x}_T) - f^* \leq \frac{RG}{2\sqrt{T}} + \frac{RG}{2\sqrt{T}} = \frac{RG}{\sqrt{T}}.}$$

Now Prove $\Phi_{t+1} \leq \Phi_t$: Use The Distance Bound

$$\begin{aligned}\Phi_{t+1} - \Phi_t &= D_{t+1}^2 - D_t^2 + 2\eta\Delta_t - \eta^2 G^2 \\ &\leq (-2\eta\Delta_t + \eta^2\|g_t\|^2) + 2\eta\Delta_t - \eta^2 G^2 \\ &= \eta^2(\|g_t\|^2 - G^2) \leq 0 \quad (0 \leq t < T).\end{aligned}$$

Example: Lasso

$$F_\lambda(x) = \frac{1}{2} \|Ax - b\|^2 + \lambda \|x\|_1, \quad \lambda > 0.$$

Ordinary Subgradient Update

$$g_t = A^\top (Ax_t - b) + \lambda s_t, \quad s_t \in \partial \|x_t\|_1, \quad x_{t+1} = x_t - \eta g_t.$$

A Sparse Optimizer: $A = I$

$$b = (2, 0.3)^\top, \quad \lambda = 0.5$$

$$x_j^* = \text{sign}(b_j) \max\{|b_j| - \lambda, 0\} \implies x^* = (1.5, 0)^\top.$$

Lasso: Bounded Iterates

Assumptions $A = I$, $\lambda > 0$, $0 < \eta \leq 1$, $s_t \in \partial \|x_t\|_1$.

Each coordinate of s_t lies in $[-1, 1]$, so $\|s_t\| \leq \sqrt{d}$.

Bound The Next Iterate Directly

$$x_{t+1} = (1 - \eta)x_t + \eta(b - \lambda s_t).$$

With $M = \|b\| + \lambda\sqrt{d}$,

$$\|x_{t+1}\| \leq (1 - \eta)\|x_t\| + \eta M.$$

Induction: Set $B = \max\{\|x_0\|, M\}$

Base: $\|x_0\| \leq B$. If $\|x_t\| \leq B$, then

$$\|x_{t+1}\| \leq (1 - \eta)B + \eta M \leq B.$$

$$\|x_t\| \leq B \quad \text{for all } t \geq 0.$$

Lasso: Bounded Subgradients

Same Setting $A = I$, $0 < \eta \leq 1$.

The Trajectory Bound Gives A Subgradient Bound

$$\|g_t\| = \|x_t - b + \lambda s_t\| \leq B + \|b\| + \lambda\sqrt{d} =: G \quad \forall t \geq 0.$$

From $x_0 = 0$, The Tuned Step Is Always Admissible

Soft-thresholding gives $|x_j^*| \leq |b_j|$, hence

$$R = \|x_0 - x^*\| = \|x^*\| \leq \|b\| \leq G.$$

For $R > 0$ and every integer $T \geq 1$, $0 < \eta = R/(G\sqrt{T}) \leq 1$, so

$$F_\lambda(\bar{x}_T) - F_\lambda^* \leq \frac{RG}{\sqrt{T}}.$$

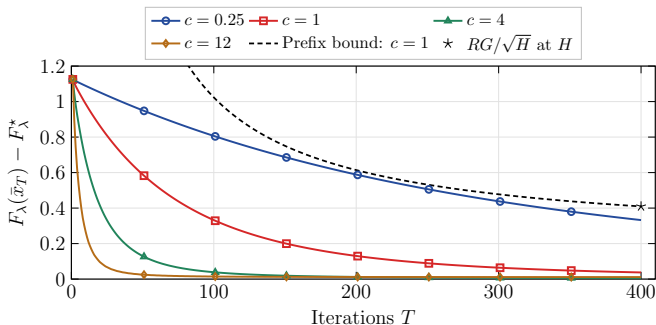
If $R = 0$, return the already optimal x_0 .

Simulation: Subgradient Method For Lasso

$$A = I, \quad b = (2, 0.3)^\top, \quad \lambda = 0.5, \quad x_0 = 0, \quad H = 400$$

$$R = \|x^*\|, \quad G = 2(\|b\| + \lambda\sqrt{2}), \quad \eta = cR/(G\sqrt{H}) \leq 1 \quad \text{for all four runs}$$

$$s_t = \text{sign}(x_t), \quad \text{sign}(0) = 0$$

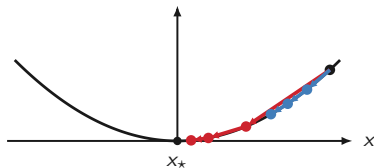


PART 05

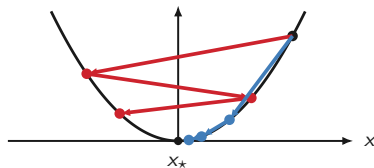
Stepsize-Based Acceleration

Main Idea: Stepsize Hedging

Flat direction



Steep direction



- Large stepsizes converge faster along low-curvature (flat) directions.
- Small stepsizes converge faster along high-curvature (steep) directions.
- Hedging both long and short stepsizes leads to acceleration.

Chebyshev Steps: Endpoint Convergence

Theorem (Euclidean Error In Exact Arithmetic)

Let $f(x) = f^* + \frac{1}{2}(x - x^*)^\top Q(x - x^*)$, where

$$Q = Q^\top, \quad 0 < \mu < L, \quad \mu I \preceq Q \preceq LI.$$

For every integer $T \geq 1$, there are explicit positive steps $\eta_0, \dots, \eta_{T-1}$ depending only on μ, L, T such that, for every x_0 , GD satisfies

$$\|x_T - x^*\| \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^T \|x_0 - x^*\|, \quad \kappa = \frac{L}{\mu}.$$

Best constant-step GD

$$O(\kappa \log(1/\epsilon))$$

Chebyshev steps

$$O(\sqrt{\kappa} \log(1/\epsilon))$$

Step 1: Gradient Descent As A Polynomial

Let $e_t = x_t - x^*$. Since $\nabla f(x_t) = Qe_t$,

$$e_{t+1} = (I - \eta_t Q)e_t, \quad e_T = p_T(Q)e_0, \quad p_T(\lambda) = \prod_{t=0}^{T-1} (1 - \eta_t \lambda).$$

The Spectral Theorem Controls Every Direction

$$\|e_T\| \leq \max_{\lambda \in [\mu, L]} |p_T(\lambda)| \|e_0\|.$$

Design Problem: Minimize Worst-Case Error Amplification

$$\rho_T^* = \min_{\substack{p \in \mathbb{R}[\lambda] \\ \deg(p) \leq T, \, p(0)=1}} \max_{\lambda \in [\mu, L]} |p(\lambda)|.$$

We optimize across all possible eigenvalues in $[\mu, L]$.

This first allows *any* polynomial with degree $\leq T$ and $p(0) = 1$.

Step 3 will verify positive real roots, hence positive GD stepsizes.

Step 2: A Polynomial Maximization Problem

Map The Interval And The Normalization Point

$$z = \frac{L + \mu - 2\lambda}{L - \mu}, \quad c = \frac{L + \mu}{L - \mu} > 1, \quad [\mu, L] \mapsto [-1, 1], \quad 0 \mapsto c.$$

Write $p(\lambda) = r(z)$ and $\|r\|_\infty = \max_{z \in [-1, 1]} |r(z)|$. Then

$$\rho_T^* = \min_{\substack{\deg(r) \leq T \\ r(c)=1}} \|r\|_\infty.$$

Rescale: Small On The Interval $\iff$ Large At c

$$M_T = \max_{\substack{\deg(s) \leq T \\ |s(z)| \leq 1 \ \forall z \in [-1, 1]}} s(c), \quad \rho_T^* = \frac{1}{M_T}.$$

Indeed, $s = r/\|r\|_\infty$ gives $s(c) = 1/\|r\|_\infty$; conversely,

$r = s/s(c)$ gives $r(c) = 1$ and $\|r\|_\infty \leq 1/s(c)$ when $s(c) > 0$.

Stay between -1 and 1 on the interval; grow largest at $c > 1$.

Step 2: Alternating Extrema

What Polynomial Grows Largest Outside $[-1, 1]$?

Try alternating between $+1$ and -1 at $T + 1$ points:
balance the worst-case deviations across the interval.

Definition: Chebyshev Polynomials

$$C_T(\cos \theta) = \cos(T\theta), \quad |C_T(z)| \leq 1 \quad (z \in [-1, 1]).$$

The cosine addition formula gives the recurrence

$$C_0(z) = 1, \quad C_1(z) = z, \quad C_{m+1}(z) = 2zC_m(z) - C_{m-1}(z).$$

Thus C_T is a degree- T polynomial; for example, $C_2(z) = 2z^2 - 1$.

Exactly $T + 1$ Alternating Extrema

$$z_j = \cos \frac{j\pi}{T}, \quad C_T(z_j) = (-1)^j, \quad j = 0, \dots, T.$$

Step 2: Why Chebyshev Is Optimal

Proof By Contradiction $\deg(s) \leq T, \quad \|s\|_\infty \leq 1$

Suppose $s(c) > C_T(c) > 0$ (positivity: the recurrence identity in Step 4).

$$\alpha = \frac{C_T(c)}{s(c)} \in (0, 1), \quad h(z) = C_T(z) - \alpha s(z).$$

Alternating Signs Force T Interior Zeros

$$h(z_j) = (-1)^j - \alpha s(z_j), \quad |\alpha s(z_j)| < 1.$$

Thus $h(z_j)$ has sign $(-1)^j$: one zero in each (z_{j+1}, z_j) ,
 $j = 0, \dots, T-1$, by continuity. The T intervals are disjoint.

The Outside Point Gives One Zero Too Many

$$h(c) = C_T(c) - \alpha s(c) = 0, \quad c > 1.$$

So $h \not\equiv 0$ has $T+1$ distinct zeros, but $\deg(h) \leq T$: impossible.

$M_T = C_T(c).$

Attained by $s = C_T$.

Step 2: The Optimal Spectral Polynomial

Normalize The Maximizer At c

$$r^*(z) = \frac{C_T(z)}{C_T(c)}, \quad r^*(c) = 1, \quad \|r^*\|_\infty = \frac{1}{C_T(c)}.$$

Return To The Spectral Interval

$$p_T^*(\lambda) = \frac{C_T\left(\frac{L + \mu - 2\lambda}{L - \mu}\right)}{C_T\left(\frac{L + \mu}{L - \mu}\right)}.$$

$$p_T^*(0) = 1, \quad \max_{\lambda \in [\mu, L]} |p_T^*(\lambda)| = \frac{1}{C_T(c)} = \rho_T^*.$$

Can this optimal polynomial be factored into positive GD steps?

Step 3: From Roots To Stepsizes

Can Ordinary GD Implement This Polynomial?

The cosine identity gives T distinct roots of C_T :

$$\zeta_j = \cos\left(\frac{(2j-1)\pi}{2T}\right), \quad j = 1, \dots, T.$$

Mapping them back gives the roots of p_T^* :

$$\gamma_j = \frac{L + \mu}{2} - \frac{L - \mu}{2} \cos\left(\frac{(2j-1)\pi}{2T}\right) \in (\mu, L).$$

The Roots And The Value At Zero Determine The Polynomial

$$p_T^*(\lambda) = \prod_{j=1}^T \left(1 - \frac{\lambda}{\gamma_j}\right), \quad \boxed{\eta_{j-1} = \frac{1}{\gamma_j} > 0.}$$

Each factor is one GD update; their product gives $e_T = p_T^*(Q)e_0$.

Thus positive GD steps attain the relaxed minimax optimum.

Step 4: Bounding The Chebyshev Polynomial

$$q = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \in (0, 1), \quad c = \frac{L + \mu}{L - \mu} = \frac{1}{2}(q + q^{-1}).$$

Compute The Denominator Using The Same Recurrence

Set $u_m = (q^{-m} + q^m)/2$. It satisfies

$$u_0 = 1, \quad u_1 = c, \quad u_{m+1} = 2cu_m - u_{m-1}.$$

The initial values and recurrence agree with $C_m(c)$, so

$$C_T(c) = \frac{1}{2}(q^{-T} + q^T) > 0.$$

Combine With $|C_T(z)| \leq 1$ On $[-1, 1]$

$$\max_{\lambda \in [\mu, L]} |p_T^*(\lambda)| = \frac{1}{C_T(c)} = \frac{2q^T}{1 + q^{2T}} \leq 2q^T.$$

Step 5: From Polynomial To Convergence

Symmetry gives $Q = V\Lambda V^\top$, $V^\top V = I$, $\lambda_i \in [\mu, L]$. Hence

$$\begin{aligned}\|e_T\| &= \|p_T^*(Q)e_0\| \leq \|p_T^*(Q)\|_2 \|e_0\| \\ &= \max_i |p_T^*(\lambda_i)| \|e_0\| \leq 2q^T \|e_0\|.\end{aligned}$$

From Contraction To Iteration Complexity $s = 1/\sqrt{\kappa} \in (0, 1)$

$$\log(1/q) = \log \frac{1+s}{1-s} = \int_0^s \frac{2}{1-u^2} du \geq 2s \quad \implies \quad q^T \leq e^{-2T/\sqrt{\kappa}}.$$

For relative error $0 < \epsilon \leq 1/2$, it suffices to take

$$T \geq \frac{\sqrt{\kappa}}{2} \log \frac{2}{\epsilon}, \quad T = O(\sqrt{\kappa} \log(1/\epsilon)).$$

Best constant step: $\eta = 2/(L + \mu)$, factor $\frac{\kappa - 1}{\kappa + 1}$.

A designed sequence improves the condition-number dependence: $\kappa \rightarrow \sqrt{\kappa}$.

From Quadratic To General Convex Objectives

Chebyshev

Fixed quadratic Hessian Q .

Known spectral bounds $0 < \mu < L$.

A single polynomial $p_T(Q)$ controls the endpoint.

When the Hessian changes, the fixed- Q polynomial argument no longer applies.

Silver Requires A Different Analysis Of The Whole Schedule.

Silver

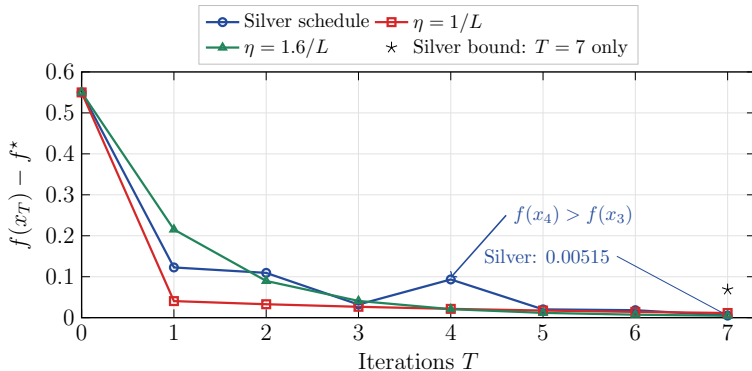
General convex, L -smooth f .

A minimizer exists.

An accelerated objective-gap guarantee at selected horizons.

Silver Schedule: A Seven-Step Trajectory

$$f(x) = \frac{1}{2}(0.1x_1^2 + x_2^2), \quad x_0 = (1, 1)^\top, \quad L = 1, \quad f^* = 0$$



Recursive Construction Of The Silver Schedule

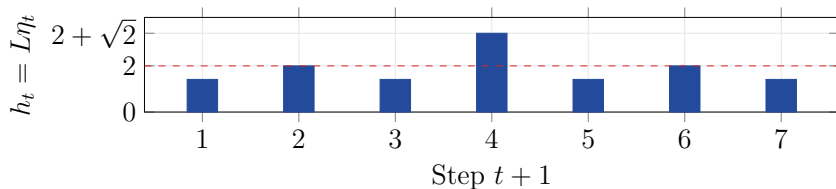
$$h_t = L\eta_t, \quad \rho = 1 + \sqrt{2}$$

$$H_1 = [\sqrt{2}], \quad H_{k+1} = [H_k, 1 + \rho^{k-1}, H_k], \quad k \geq 1.$$

$$H_1 = [\sqrt{2}],$$

$$H_2 = [\sqrt{2}, 2, \sqrt{2}],$$

$$H_3 = [\sqrt{2}, 2, \sqrt{2}, 2 + \sqrt{2}, \sqrt{2}, 2, \sqrt{2}].$$



$$|H_k| = 2^k - 1.$$

Convergence Rate Of The Silver Schedule

Setting And Algorithm

$f : \mathbb{R}^d \rightarrow \mathbb{R}$ convex, L -smooth, $L > 0$; $x^* \in \arg \min f$; $R = \|x_0 - x^*\|$.

$$x_{t+1} = x_t - (h_t/L)\nabla f(x_t), \quad (h_0, \dots, h_{T-1}) = H_k.$$

Theorem (Altschuler–Parrilo (2023), Theorem 1.1)

For every integer $k \geq 1$, at the block endpoint $T = 2^k - 1$,

$$f(x_T) - f^* \leq \frac{LR^2}{1 + \sqrt{4\rho^{2k} - 3}}.$$

$$\rho^k = (T + 1)^{p_{\text{sil}}}, \quad p_{\text{sil}} = \log_2(1 + \sqrt{2}) \approx 1.2716.$$

Altschuler–Parrilo, Theorem 1.1 and Definition 2.1.

Worst-Case Error Of Predetermined Schedules

Algorithm Class Fix $L > 0$ and an integer horizon T .

$$x_{t+1} = x_t - \eta_t \nabla f(x_t), \quad \eta_0, \dots, \eta_{T-1} \geq 0.$$

Steps may depend on L, T , but are fixed before the objective and observed gradients.

Definition (Worst-Case Normalized Last-Iterate Error)

$$\mathcal{R}_T(\eta) = \sup_{\substack{d \geq 1, f: \mathbb{R}^d \rightarrow \mathbb{R} \text{ convex, } L\text{-smooth} \\ x^* \in \arg \min f, x_0 \neq x^*}} \frac{f(x_T) - f^*}{(L/2) \|x_0 - x^*\|^2}.$$

Ma–Chen Lower Bound For Predetermined Steps

Same Class: Predetermined, Nonnegative GD Steps

Theorem (Ma–Chen (2026), Theorem 2.1)

Let $p_{\text{MC}} = \sqrt{2 + \sqrt{3}} \approx 1.9319$. For every $p \in (p_{\text{MC}}, 2)$ there is $c_p > 0$ such that, for every $T \geq 1$ and every such schedule,

$$\mathcal{R}_T(\eta) \geq \frac{c_p}{(T+1)^p}.$$

Consequence

No uniform $O(T^{-2})$ last-iterate guarantee in this algorithm class.

Ma–Chen, A Lower Bound For Stepsize-Based Acceleration Of Gradient Descent.

Ye-Liu Lower Bound And The Silver Exponent

Same Class And Normalization

Theorem (Ye-Liu (2026), Theorem 1.1)

There is an absolute $C_0 > 0$ such that, for every sufficiently large T and every predetermined nonnegative schedule,

$$\mathcal{R}_T(\eta) \geq T^{-(p_{\text{sil}} + \delta_T)}, \quad \delta_T = C_0 \sqrt{\frac{\log \log T}{\log T}} \rightarrow 0.$$

Consequence For Uniform Convergence Rates

For every fixed $\epsilon > 0$, no uniform $O(T^{-(p_{\text{sil}} + \epsilon)})$ guarantee.

Ye-Liu, Silver Rate Is (Almost) Optimal For Gradient Descent.

Convergence Rates: Scheduling And Momentum

Common Setting General convex, L -smooth f with a minimizer.

Metric: last-iterate objective gap; scale: LR^2 .

Method	Gap Upper Bound	Mechanism / Endpoint
GD, constant $1/L$	$O(T^{-1})$	One-step descent; every T
GD, silver schedule	$O(T^{-1.2716\dots})$	Long steps; $T = 2^k - 1$
Nesterov acceleration	$O(T^{-2})$	Momentum update; every T

Summary Of Convergence Guarantees

Setting	Step Choice	Output And Certificate
Smooth, lower bounded	$1/L$	$\min_{t < T} \ \nabla f(x_t)\ ^2 = O(1/T)$
Smooth convex	$1/L$	Last gap $\leq LR^2/(2T)$
Smooth strongly convex	$1/L$	Last gap $\leq (1 - \mu/L)^T \Delta_0$
Subgradient (fixed)	$R/(G\sqrt{T})$	Averaged gap $\leq RG/\sqrt{T}$
Smooth convex, silver	H_k/L	Last gap $O(LR^2/T^{1.2716\dots})$

References

Classical Theory And Examples

Yuxin Chen, ELE522: gradient methods, pp. 44–47, 52–61, 66–69.

Dmitriy Drusvyatskiy: *Optimization For Learning; Convex Analysis And Nonsmooth Optimization*.

Polynomial And Stepsize-Based Acceleration

d'Aspremont–Scieur–Taylor: Acceleration Methods, Chapter 2.

Agarwal–Goel–Zhang: Acceleration Via Fractal Learning Rate Schedules.

Altschuler–Parrilo: Silver Stepsize Schedule For Smooth Convex Optimization.

Ma–Chen: A Lower Bound For Stepsize-Based Acceleration Of Gradient Descent.

Ye–Liu: Silver Rate Is (Almost) Optimal For Gradient Descent.