Distributed Optimization for Machine Learning

Lecture 5 - Unconstrained Optimization: Gradient Descent

Tianyi Chen

School of Electrical and Computer Engineering
Cornell Tech, Cornell University

September 10, 2025

Differentiable unconstrained minimization

$$\begin{array}{ll} \min_{\mathbf{x}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathbb{R}^n \end{array}$$

- f (objective or cost function) is differentiable

Connecting abstract to concrete optimization

The notation $\min_{\mathbf{x}} f(\mathbf{x})$ can seem abstract. Let's explicitly map it to the machine learning training problem we've been discussing.

Model training problem

- **Parameters:** a huge set of weights and biases from all layers of our neural network.
  $\boldsymbol{\theta} = \{W_1, b_1, W_2, b_2, \ldots\}$
- **Loss function:** a measure of the average error over all data
  $L(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m}(h_{\boldsymbol{\theta}}(\text{data}_i) - \text{label}_i)^2$

Generic optimization problem

- **Variable:** a (very) long vector containing all the parameters flattened together.
  $\mathbf{x} \in \mathbb{R}^n$
- **Objective function:** a high-dimensional differentiable function to minimize $f(\mathbf{x})$

Training a model just means finding the variable $\mathbf{x}^*$ that minimizes $f(\mathbf{x})$. The number of parameters n can be in the millions or billions!

Connecting abstract to concrete optimization

We have m data points. For each data point $(\mathbf{x}^{(i)}, y^{(i)})$, the linear model predicts $\hat{y}^{(i)} = (\mathbf{x}^{(i)})^\top \boldsymbol{\theta}$. Our goal is to minimize the total squared error:

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{m} \left((\mathbf{x}^{(i)})^\top \boldsymbol{\theta} - y^{(i)} \right)^2$$

Data A (m × n features)

$$\mathbf{A} = \begin{pmatrix} — (\mathbf{x}^{(1)})^\top — \\ — (\mathbf{x}^{(2)})^\top — \\ \vdots \\ — (\mathbf{x}^{(m)})^\top — \end{pmatrix}$$

Params x

$$\mathbf{x} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{pmatrix}$$

Labels b (m samples)

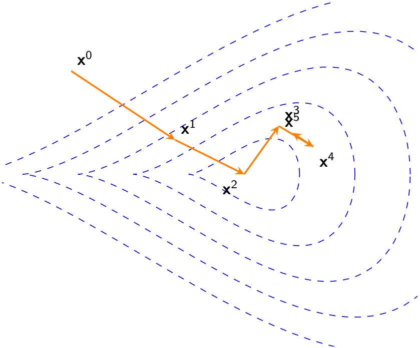
$$\mathbf{b} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix}$$

$$\mathbf{Ax} - \mathbf{b} = \begin{pmatrix} (\mathbf{x}^{(1)})^\top \boldsymbol{\theta} \\ \vdots \\ (\mathbf{x}^{(m)})^\top \boldsymbol{\theta} \end{pmatrix} - \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{pmatrix} = \begin{pmatrix} \text{error for sample 1} \\ \vdots \\ \text{error for sample m} \end{pmatrix}$$

Gradient descent (GD)

A building block of this course: **gradient descent**

$$\mathbf{x}^{t+1} = \mathbf{x}^{t} - \eta_t \nabla f(\mathbf{x}^{t})$$



- traced to Augustin Louis Cauchy '1847 ...

Table of Contents

Strongly convex and smooth problems

Now generalize quadratic minimization to a broader class of problems

$$\min_{\mathbf{x}} f(\mathbf{x})$$

Key assumption: $f(\cdot)$ is **strongly convex** and **smooth**.

- a twice-differentiable function f is said to be μ-strongly convex and L-smooth if the Hessian $\nabla^2 f(\mathbf{x})$ satisfies

$$\mathbf{0} \preceq \mu \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}, \quad \text{for all } \mathbf{x}$$

Strong convexity & smoothness in linear regression

To check the assumption, we first need to compute the Hessian matrix. The gradient is $\nabla f(\mathbf{x}) = \mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{b})$. Taking the derivative again gives:

$$\nabla^2 f(\mathbf{x}) = \mathbf{A}^\top \mathbf{A}$$

The condition $\mu \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$ means the eigenvalues of the Hessian are bounded between μ and L. For linear regression:

- **Strong convexity**: f is μ-strongly convex, where $\mu = \lambda_{\min}(\mathbf{A}^\top \mathbf{A})$, the smallest eigenvalue of $\mathbf{A}^\top \mathbf{A}$. We get strong convexity $(\mu > 0)$ if the data matrix $\mathbf{A}$ has linearly independent columns.

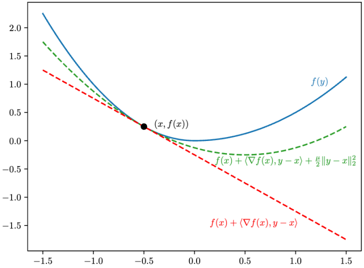
- **Smoothness**: f is L-smooth, where $L = \lambda_{\max}(\mathbf{A}^\top \mathbf{A})$, the largest eigenvalue of $\mathbf{A}^\top \mathbf{A}$. This is satisfied as long as our data is finite.

More on strong convexity

$f(\cdot)$ is said to be μ-**strongly convex** if

(i) $f(\mathbf{y}) \geq \underbrace{f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})}_{\text{first-order Taylor expansion}} + \frac{\mu}{2}||\mathbf{x} - \mathbf{y}||_2^2, \quad$ for all $\mathbf{x}, \mathbf{y}$

(ii) equivalently, $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu ||\mathbf{x} - \mathbf{y}||_2^2, \quad$ for all $\mathbf{x}, \mathbf{y}$

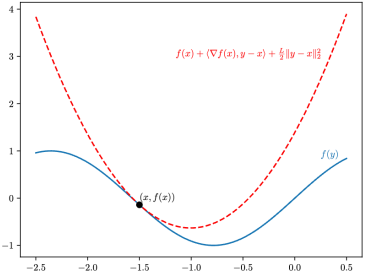


More on smoothness

A convex function $f(\cdot)$ is said to be ***L*-smooth** if

(i) $f(\mathbf{y}) \le \underbrace{f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y}-\mathbf{x})}_{\text{first-order Taylor expansion}} + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|_2^2, \quad$ for all $\mathbf{x}, \mathbf{y}$

(ii) $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \le L\|\mathbf{x}-\mathbf{y}\|_2,\;$ for all $\mathbf{x}, \mathbf{y}$ (**L-Lipschitz gradient**)



Convergence rate for strongly convex and smooth problems

Theorem 1 (GD for strongly convex and smooth functions)

Let f be μ-strongly convex and L-smooth. If $\eta_t \equiv \eta = \frac{2}{\mu+L}$, then

$$\|\mathbf{x}^t - \mathbf{x}^*\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^t \|\mathbf{x}^0 - \mathbf{x}^*\|_2,$$

where $\kappa := L/\mu$ is condition number; $\mathbf{x}^*$ is the minimizer.

- generalization of quadratic minimization problems
  - stepsize: $\eta = \frac{2}{\mu+L}$ (vs. $\eta = \frac{2}{\lambda_1(\mathbf{Q})+\lambda_n(\mathbf{Q})}$)
  - contraction rate: $\frac{\kappa-1}{\kappa+1}$ (vs. $\frac{\lambda_1(\mathbf{Q})-\lambda_n(\mathbf{Q})}{\lambda_1(\mathbf{Q})+\lambda_n(\mathbf{Q})}$)

Convergence rate for strongly convex and smooth problems

Theorem 1 (GD for strongly convex and smooth functions)

Let f be μ-strongly convex and L-smooth. If $\eta_t \equiv \eta = \frac{2}{\mu+L}$, then

$$\|\mathbf{x}^t - \mathbf{x}^*\|_2 \le \left(\frac{\kappa - 1}{\kappa + 1}\right)^t \|\mathbf{x}^0 - \mathbf{x}^*\|_2,$$

where $\kappa := L/\mu$ is condition number; $\mathbf{x}^*$ is the minimizer.

- dimension-free: iteration complexity is $\mathcal{O}\left(\frac{\log \frac{1}{\epsilon}}{\log \frac{\kappa+1}{\kappa-1}}\right)$, which is independent of the problem size n if κ does not depend on n

Proof of Theorem 1

To mimic the analysis of quadratic case (cf. $\nabla f(\mathbf{x}^t) = \mathbf{Q}(\mathbf{x}^t - \mathbf{x}^*)$)

$$
\begin{aligned}
\mathbf{x}^{t+1} - \mathbf{x}^* &= \mathbf{x}^t - \mathbf{x}^* - \eta_t \nabla f(\mathbf{x}^t) = (\mathbf{I} - \eta_t \mathbf{Q})(\mathbf{x}^t - \mathbf{x}^*) \\
\Longrightarrow \; ||\mathbf{x}^{t+1} - \mathbf{x}^*||_2 &\leq ||\mathbf{I} - \eta_t \mathbf{Q}|| \cdot ||\mathbf{x}^t - \mathbf{x}^*||_2
\end{aligned}
$$

for strongly convex cases, we have

$$
||\mathbf{x}^{t+1} - \mathbf{x}^*||_2 = ||\mathbf{x}^t - \mathbf{x}^* - \eta \nabla f(\mathbf{x}^t)||_2.
$$

We can "generate" $(\mathbf{x}^t - \mathbf{x}^*)$ from the fundamental theorem of calculus

$$
\nabla f(\mathbf{x}^t) = \nabla f(\mathbf{x}^t) - \underbrace{\nabla f(\mathbf{x}^*)}_{=0} = \left(\int_0^1 \nabla^2 f(\mathbf{x}_\tau) d\tau \right) (\mathbf{x}^t - \mathbf{x}^*),
$$

where $\mathbf{x}_\tau := \mathbf{x}^t + \tau(\mathbf{x}^* - \mathbf{x}^t)$ lies on a **line segment** between $\mathbf{x}^t$ and $\mathbf{x}^*$.

Proof of Theorem 1 (cond't)

Building upon this connection, we have

$$\begin{aligned}
||\mathbf{x}^{t+1} - \mathbf{x}^*||_2 &= ||\mathbf{x}^t - \mathbf{x}^* - \eta \nabla f(\mathbf{x}^t)||_2 \\
&= \left\| \left(\mathbf{I} - \eta \int_0^1 \nabla^2 f(\mathbf{x}_\tau) d\tau \right) (\mathbf{x}^t - \mathbf{x}^*) \right\|_2 \\
&\leq \sup_{0 \leq \tau \leq 1} ||\mathbf{I} - \eta \nabla^2 f(\mathbf{x}_\tau)|| \cdot ||\mathbf{x}^t - \mathbf{x}^*||_2 \\
&\leq \frac{L - \mu}{L + \mu} ||\mathbf{x}^t - \mathbf{x}^*||_2
\end{aligned}$$

where we first choose the constant stepsize as $\eta = \frac{2}{\mu + L}$, and then use the fact that f be μ-strongly convex and L-smooth.

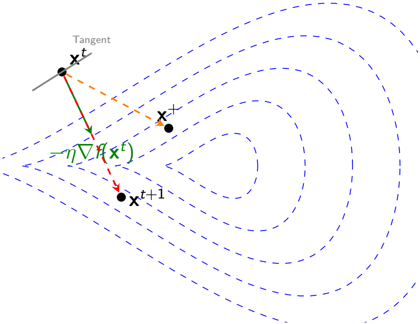
Repeat this argument for all iterations to conclude the proof.

Hint: The spectral norm of $\mathbf{I} - \eta \nabla^2 f(\mathbf{x}_\tau)$ is its largest eigenvalue.

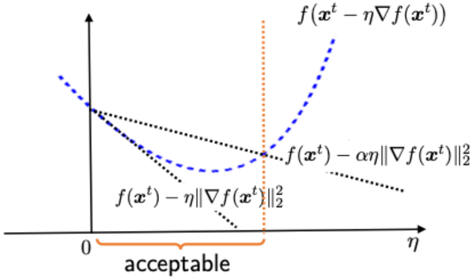
Backtracking line search

Practically, one often performs line searches rather than adopting constant stepsizes. Most line searches in practice are, however, *inexact*.

A simple and effective scheme: **backtracking line search**



Backtracking line search

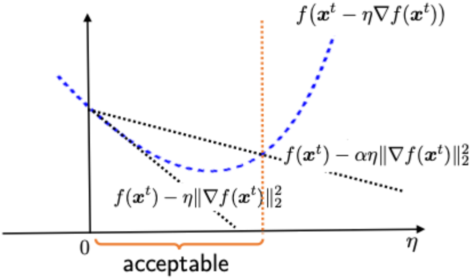


Armijo condition: for some $0 < \alpha < 1$

$$f(\mathbf{x}^t - \eta \nabla f(\mathbf{x}^t)) < f(\mathbf{x}^t) - \alpha\eta ||\nabla f(\mathbf{x}^t)||_2^2 \tag{5}$$

- $f(\mathbf{x}^t) - \alpha\eta||\nabla f(\mathbf{x}^t)||_2^2$ lies above $f(\mathbf{x}^t - \eta\nabla f(\mathbf{x}^t))$ for small η
- ensures **sufficient decrease** of objective values

Backtracking line search



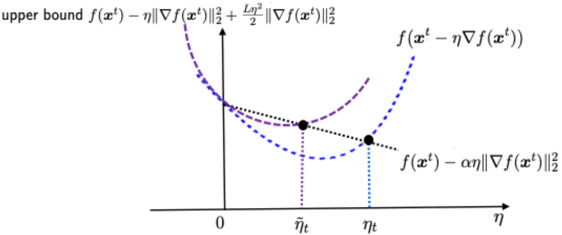
Algorithm 2 - Backtracking line search for GD

1: Initialize $\eta = 1$, $0 < \alpha \leq 1/2$, $0 < \beta < 1$

2: **while** $f(\mathbf{x}^t - \eta\nabla f(\mathbf{x}^t)) > f(\mathbf{x}^t) - \alpha\eta||\nabla f(\mathbf{x}^t)||_2^2$ **do**

3: $\quad \eta \leftarrow \beta\eta$

Backtracking line search



Practically, backtracking line search often (but not always) provides good estimates on the **local Lipschitz constants** of gradients.

Convergence for backtracking line search

Theorem 2 (Boyd, Vandenberghe '04)

Let f be μ-strongly convex and L-smooth. With backtracking line search, the objective function satisfies

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \leq \left(1 - \min\{2\mu\alpha, \frac{2\beta\alpha\mu}{L}\}\right)^t (f(\mathbf{x}^0) - f(\mathbf{x}^*))$$

where $\mathbf{x}^*$ is the minimizer.

Table of Contents

Dropping strong convexity

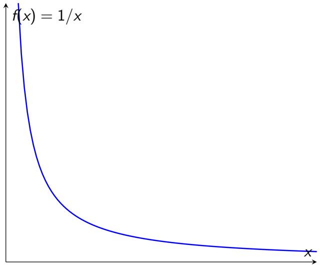
What happens if we completely drop (local) strong convexity?

$$\min_{\mathbf{x}} f(\mathbf{x})$$

- **Key assumption:** $f(\mathbf{x})$ is **convex** and **smooth**

Dropping strong convexity

Without strong convexity, it may often be better to focus on objective improvement (rather than improvement on estimation error).



Example: consider $f(x) = 1/x \; (x > 0)$. GD iterates $\{\mathbf{x}^t\}$ might never converge to $x^* = \infty$. In comparison, $f(\mathbf{x}^t)$ might approach $f(x^*) = 0$.

Objective improvement and stepsize

Question:

- can we ensure reduction of the objective value (i.e. $f(\mathbf{x}^{t+1}) < f(\mathbf{x}^t)$) without strong convexity?
- what stepsizes guarantee sufficient decrease?

Key idea: **majorization-minimization**

- find a *simple* majorizing function of $f(\mathbf{x})$ and optimize it instead

Objective improvement and stepsize

From the smoothness assumption,

$$f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t) \le \nabla f(\mathbf{x}^t)^\top (\mathbf{x}^{t+1} - \mathbf{x}^t) + \frac{L}{2}||\mathbf{x}^{t+1} - \mathbf{x}^t||_2^2$$

$$= \underbrace{-\eta_t ||\nabla f(\mathbf{x}^t)||_2^2 + \frac{\eta_t^2 L}{2}||\nabla f(\mathbf{x}^t)||_2^2}_{\text{majorizing function of objective reduction due to smoothness}}$$

(**pick** $\eta_t = 1/L$ to minimize the majorizing function)

$$= -\frac{1}{2L}||\nabla f(\mathbf{x}^t)||_2^2$$

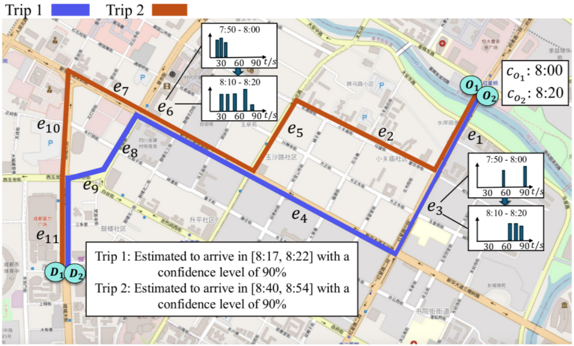
Objective improvement

Fact 7 Suppose f is L-smooth. Then GD with $\eta_t = 1/L$ obeys

$$f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t) - \frac{1}{2L}\|\nabla f(\mathbf{x}^t)\|_2^2$$

- for η_t sufficiently small, GD results in improvement in the objective
- does **NOT** rely on convexity!

Make connections to ETA



- How many miles I can drive per hour given the total distance?

A byproduct under additional curvature conditions

From the per-iteration objective improvement

$$
\begin{aligned}
f(\mathbf{x}^{t+1}) - f(\mathbf{x}^*) &\overset{\text{(i)}}{\leq} f(\mathbf{x}^t) - f(\mathbf{x}^*) - \frac{1}{2L}\|\nabla f(\mathbf{x}^t)\|_2^2 \\
&\overset{\text{(ii)}}{\leq} f(\mathbf{x}^t) - f(\mathbf{x}^*) - \frac{\mu}{L}(f(\mathbf{x}^t) - f(\mathbf{x}^*)) \\
&= \left(1 - \frac{\mu}{L}\right)(f(\mathbf{x}^t) - f(\mathbf{x}^*))
\end{aligned}
$$

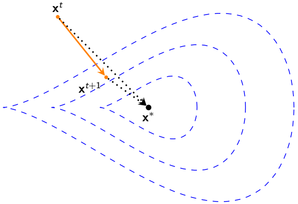
where (i) follows from Fact 7, and (ii) comes from the so-called Polyak-Lojasiewicz (PL) condition (implied by strong convexity)

$$
\|\nabla f(\mathbf{x})\|_2^2 \geq 2\mu(f(\mathbf{x}) - f(\underbrace{\mathbf{x}^*}_{\text{minimizer}})), \quad \text{for all } \mathbf{x}.
$$

Apply it recursively to obtain the linear convergence of $f(\mathbf{x}^t) - f(\mathbf{x}^*)$.

Improvement in estimation accuracy

GD is not only improving the objective value, but is also dragging the iterates towards minimizer(s), as long as η_t is not too large.



$||\mathbf{x}^t - \mathbf{x}^*||_2$ is **monotonically nonincreasing** in t

Treating f as 0-strongly convex, we can see from our previous analysis for strongly convex problems that

$$||\mathbf{x}^{t+1} - \mathbf{x}^*||_2 \leq ||\mathbf{x}^t - \mathbf{x}^*||_2$$

Improvement in estimation accuracy

One can further show that $||\mathbf{x}^t - \mathbf{x}^*||_2$ is strictly decreasing unless $\mathbf{x}^t$ is already the minimizer.

Fact 8 Let f be convex and L-smooth. If $\eta_t \equiv \eta = 1/L$, then

$$||\mathbf{x}^{t+1} - \mathbf{x}^*||_2^2 \leq ||\mathbf{x}^t - \mathbf{x}^*||_2^2 - \frac{1}{L^2}||\nabla f(\mathbf{x}^t)||_2^2$$

where $\mathbf{x}^*$ is any minimizer of $f(\cdot)$.

Proof of Fact 8*

It follows that

$$
\begin{aligned}
\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2^2 &= \|\mathbf{x}^t - \mathbf{x}^* - \eta(\nabla f(\mathbf{x}^t) - \underbrace{\nabla f(\mathbf{x}^*)}_{=0})\|_2^2 \\
&= \|\mathbf{x}^t - \mathbf{x}^*\|_2^2 - 2\eta\langle \mathbf{x}^t - \mathbf{x}^*, \nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\rangle \\
&\qquad + \eta^2\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\|_2^2 \\
&\le \|\mathbf{x}^t - \mathbf{x}^*\|_2^2 - \underbrace{\frac{2\eta}{L}\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\|_2^2}_{\ge \text{ (smooth+cvx)}} + \eta^2\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\|_2^2 \\
&= \|\mathbf{x}^t - \mathbf{x}^*\|_2^2 - \frac{1}{L^2}\|\nabla f(\mathbf{x}^t) - \underbrace{\nabla f(\mathbf{x}^*)}_{=0}\|_2^2 \quad (\text{since } \eta = 1/L)
\end{aligned}
$$

Monotonicity of gradient sizes

When $\eta_t = 1/L$, gradient sizes are also monotonically non-increasing.

Lemma 9 Let f be convex and smooth. If $\eta_t \equiv \eta = 1/L$, then GD obeys

$$\|\nabla f(\mathbf{x}^{t+1})\|_2 \leq \|\nabla f(\mathbf{x}^t)\|_2$$

As a result, GD enjoys at least 3 types of monotonicity as t grows:

- objective value $f(\mathbf{x}^t) \searrow$
- estimation error $\|\mathbf{x}^t - \mathbf{x}^*\|_2 \searrow$
- gradient size $\|\nabla f(\mathbf{x}^t)\|_2 \searrow$

Proof of Lemma 9

Recall that the fundamental theorem of calculus gives

$$\begin{aligned}\nabla f(\mathbf{x}^{t+1}) &= \nabla f(\mathbf{x}^t) + \int_0^1 \nabla^2 f(\mathbf{x}_\tau)(\mathbf{x}^{t+1} - \mathbf{x}^t)d\tau \\ &= \underbrace{\left(\mathbf{I} - \eta \int_0^1 \nabla^2 f(\mathbf{x}_\tau)d\tau\right)}_{=:\mathbf{B}} \nabla f(\mathbf{x}^t),\end{aligned}$$

where $\mathbf{x}_\tau := \mathbf{x}^t + \tau(\mathbf{x}^{t+1} - \mathbf{x}^t)$. When $\eta \leq 1/L$, it is easily seen that

$$\mathbf{0} \preceq \mathbf{B} \preceq \mathbf{I} \implies \mathbf{0} \preceq \mathbf{B}^2 \preceq \mathbf{I}$$

We can thus derive

$$||\nabla f(\mathbf{x}^{t+1})||_2^2 - ||\nabla f(\mathbf{x}^t)||_2^2 = \nabla f(\mathbf{x}^t)^\top(\mathbf{B}^2 - \mathbf{I})\nabla f(\mathbf{x}^t) \leq 0$$

Convergence rate for convex and smooth problems

However, without strong convexity, convergence is typically much slower than linear (or geometric) convergence.

Theorem 10 (GD for convex and smooth problems)
Let f be convex and L-smooth. If $\eta_t \equiv \eta = 1/L$, then GD obeys

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \leq \frac{2L\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{t}$$

where $\mathbf{x}^*$ is any minimizer of $f(\cdot)$.

- attains ϵ-accuracy within $\mathcal{O}(1/\epsilon)$ iterations (vs. $\mathcal{O}(\log \frac{1}{\epsilon})$ iterations for linear convergence)

Proof of Theorem 10 (cont.)

From Fact 7,

$$f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t) \le -\frac{1}{2L}||\nabla f(\mathbf{x}^t)||_2^2$$

To infer $f(\mathbf{x}^t)$ recursively, it is often easier to replace $||\nabla f(\mathbf{x}^t)||_2$ with simpler functions of $f(\mathbf{x}^t)$. Use convexity and Cauchy-Schwarz to get

$$f(\mathbf{x}^*) - f(\mathbf{x}^t) \ge \nabla f(\mathbf{x}^t)^\top (\mathbf{x}^* - \mathbf{x}^t) \ge -||\nabla f(\mathbf{x}^t)||_2||\mathbf{x}^t - \mathbf{x}^*||_2$$

$$\Longrightarrow \; ||\nabla f(\mathbf{x}^t)||_2 \ge \frac{f(\mathbf{x}^t) - f(\mathbf{x}^*)}{||\mathbf{x}^t - \mathbf{x}^*||_2} \overset{\text{Fact 8}}{\ge} \frac{f(\mathbf{x}^t) - f(\mathbf{x}^*)}{||\mathbf{x}^0 - \mathbf{x}^*||_2}$$

Setting $\Delta_t := f(\mathbf{x}^t) - f(\mathbf{x}^*)$ and combining the above bounds yield

$$\Delta_{t+1} - \Delta_t \le -\frac{1}{2L||\mathbf{x}^0 - \mathbf{x}^*||_2^2}\Delta_t^2 =: -\frac{1}{w_0}\Delta_t^2$$

Proof of Theorem 10 (cont.)

$$\Delta_{t+1} \leq \Delta_t - \frac{1}{w_0}\Delta_t^2$$

Dividing both sides by $\Delta_t \Delta_{t+1}$ and rearranging terms give

$$\begin{aligned}
& \frac{1}{\Delta_{t+1}} \geq \frac{1}{\Delta_t} + \frac{1}{w_0}\frac{\Delta_t}{\Delta_{t+1}} \\
\Longrightarrow\ & \frac{1}{\Delta_{t+1}} \geq \frac{1}{\Delta_t} + \frac{1}{w_0} \quad \text{(since } \Delta_t \geq \Delta_{t+1} \text{ (Fact 7))} \\
\Longrightarrow\ & \frac{1}{\Delta_t} \geq \frac{1}{\Delta_0} + \frac{t}{w_0} \geq \frac{t}{w_0} \\
\Longrightarrow\ & \Delta_t \leq \frac{w_0}{t} = \frac{2L\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{t}
\end{aligned}$$

as claimed.

Table of Contents

In-Class Lab: The nonconvex case - a bumpy road

Goal: See how the starting point leads to different local minima.

The Setup

- **Our function:** $f(x) = \frac{1}{4}x^4 - 2x^2$. This function has two minima.
- **Its gradient:** $f'(x) = x^3 - 4x$.
- **The GD update rule:** $x_{t+1} = x_t - \eta \cdot f'(x_t)$.
- We will use a learning rate of $\eta = \mathbf{0.1}$.

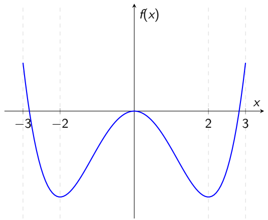


Figure: You will plot your GD steps on a graph like this.

Part 1: Starting at $\mathbf{x_0 = 3.0}$

Instructions: The gradient values are provided. Calculate the 'Update' and the 'Next Point' for each step.

t	$\mathbf{x_t}$	**Gradient** $f'(x_t)$ **(Given)**	**Update** $\eta \cdot f'(x_t)$	**Next Point** x_{t+1}
0	3.0	15.0	1.5	1.5
1	1.5	-2.625		
2				

Questions

1. Plot your points $(x_0, x_1, x_2, \dots)$ on the graph.
2. Which minimum does the path seem to be approaching ($x = 2$ or $x = -2$)?

Part 2: Starting at $\mathbf{x_0 = -3.0}$

Instructions: Now, start from the left side and repeat the process.

t	$\mathbf{x_t}$	Gradient $f'(x_t)$ (Given)	Update $\eta \cdot f'(x_t)$	Next Point x_{t+1}
0	-3.0	-15.0	-1.5	-1.5
1	-1.5	2.625		
2				

The Final Question

Based on your two experiments, what is the most important factor in determining which minimum GD finds in a nonconvex problem?

Solutions: The Importance of Initialization

Part 1: Starting at $x_0 = 3.0$

t	x_t	$f'(x_t)$	$\eta \cdot f'(x_t)$	x_{t+1}
0	3.0	15.0	1.5	1.5
1	1.5	-2.625	-0.263	1.763
2	1.763	-1.565	-0.157	1.920

$\rightarrow$ Converges to $\mathbf{x = 2}$

Part 2: Starting at $x_0 = -3.0$

t	x_t	$f'(x_t)$	$\eta \cdot f'(x_t)$	x_{t+1}
0	-3.0	-15.0	-1.5	-1.5
1	-1.5	2.625	0.263	-1.763
2	-1.763	1.565	0.157	-1.920

$\rightarrow$ Converges to $\mathbf{x = -2}$

Key Takeaway

For nonconvex problems, the algorithm is only guaranteed to find a local minimum, and the one it finds is determined by the starting point.

Recap and fine-tuning

- What we have talked about today?
  - $\Rightarrow$ How GD performs in strongly convex and smooth problems?
  - $\Rightarrow$ Without strong convexity, the rate slows to **sublinear**, $\mathcal{O}(1/t)$.

Welcome anonymous survey!