Distributed Optimization for Machine Learning

Lecture 7 - Gradient Methods for Constrained Problems

Tianyi Chen

School of Electrical and Computer Engineering
Cornell Tech, Cornell University

September 17, 2025

Constrained convex problems

$$\begin{array}{ll} \min_{\mathbf{x}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathcal{C} \end{array}$$

- $f(\cdot)$: convex function
- $\mathcal{C} \subseteq \mathbb{R}^n$: closed convex set

Example: Constrained logistic regression

Why constrained problems in ML?

- Standard logistic regression minimizes

$$\min_{\mathbf{w}} \ \frac{1}{N}\sum_{i=1}^{N} \log\big(1+\exp(-y_i\mathbf{w}^\top x_i)\big).$$

- To avoid overfitting, we may constrain the weights:

$$\|\mathbf{w}\|_2 \leq R.$$

- Interpretation:
  - Keeps parameters small $\Rightarrow$ better generalization
  - Equivalent to weight regularization but fits in constrained form

Example: Distributed learning with consensus

Why constrained problems in distributed systems?

- In distributed learning, the goal is to minimize a global loss function, which is the sum of local losses f_k from K different agents:

$$\min_{\{\mathbf{w}_k\},\mathbf{z}} \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i=1}^{N_k} \log\big(1 + \exp(-y_i \mathbf{w}_k^\top x_i)\big).$$

- Each agent k has its own local parameter $\mathbf{w}_k$. To solve the global problem, we must enforce a **consensus constraint**:

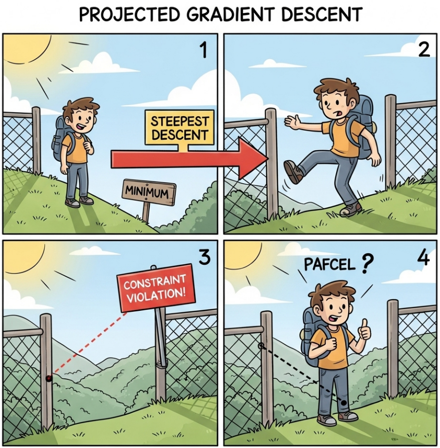
$$\mathbf{w}_k = \mathbf{z}, \quad \text{for all} \quad k \in \{1, \dots, K\}.$$

- Interpretation:
  - All agents agree on a single optimal parameter $\mathbf{z}$.
  - Solvable with only local communication (e.g., with neighbors).

Table of Contents

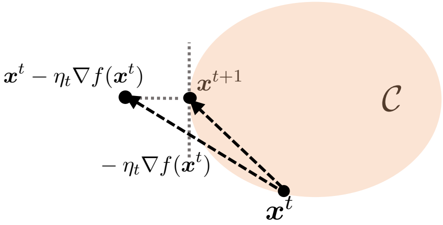
Everyday analogy

- Projected Gradient Descent =
  "*walk downhill, then return inside fence if you cross the boundary*".

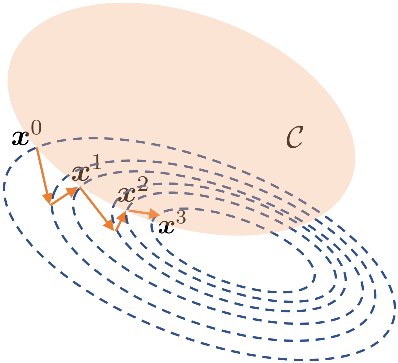


Why projection?

- A gradient step may take us **outside** the feasible set $\mathcal{C}$
- Projection brings us back to the **closest feasible point**
- If $\mathcal{C}$ is simple (ball, box, simplex), projection is cheap



Projected gradient descent



- works well if projection onto $\mathcal{C}$ can be computed efficiently

for $t = 0, 1, \cdots$:

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

where $\mathcal{P}_{\mathcal{C}}(\mathbf{x}) := \arg\min_{\mathbf{z}\in\mathcal{C}} \|\mathbf{x} - \mathbf{z}\|_2^2$ is $\underbrace{\text{Euclidean projection}}_{\text{quadratic minimization}}$ onto $\mathcal{C}$.

Examples of simple projections

- ℓ_2**-ball:** (just rescale if outside the ball)

$$\mathcal{C} = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq R\}, \quad \mathcal{P}_{\mathcal{C}}(\mathbf{y}) = \min\left(1, \frac{R}{\|\mathbf{y}\|_2}\right)\mathbf{y}$$

- **Box constraints:** (component-wise clipping)

$$\mathcal{C} = [l, u]^n, \quad \mathcal{P}_{\mathcal{C}}(\mathbf{y}) = \min\big(\max(\mathbf{y}, l), u\big)$$

- **Consensus constraint:** (average the disagreement)

$$\mathcal{C} = \{(\mathbf{x}_1, \ldots, \mathbf{x}_K) : \mathbf{x}_k = \mathbf{z} \text{ for all} k, \text{ for some } \mathbf{z}\}$$

The projection operator $\mathcal{P}_{\mathcal{C}}(\mathbf{x}_1, \ldots, \mathbf{x}_K)_k = \frac{1}{K}\sum_{j=1}^{K}\mathbf{x}_j$

Application: Distributed gradient descent (DGD)

- **Goal:** Minimize a global sum of functions $f(\mathbf{w}) = \sum_{k=1}^{K} f_k(\mathbf{w})$, where each f_k is known only to agent k.
- Each agent k maintains its own local estimate $\mathbf{w}_k$.
- **Feasible Set $\mathcal{C}$:** The consensus space.

$$\mathcal{C} = \{(\mathbf{w}_1, \ldots, \mathbf{w}_K) : \mathbf{w}_1 = \mathbf{w}_2 = \cdots = \mathbf{w}_K\}$$

The DGD Iteration (Conceptual)

At each time t, DGD performs two main steps for each agent k:

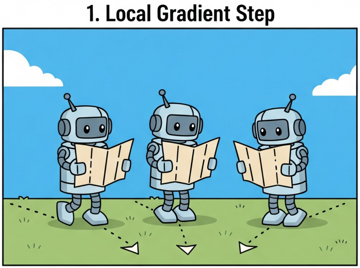
1. **Local gradient:** Take a step based on its own local objective f_k.
2. **Consensus:** Communicate with neighbors and average parameters.

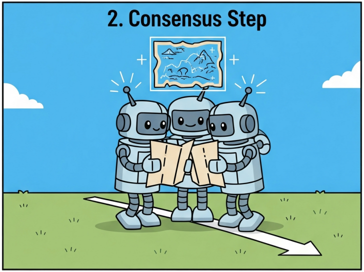
Everyday analogy

The DGD Iteration (Conceptual)

At each time t, DGD performs two main steps for each agent k:

1. **Local gradient:** Take a step based on its own local objective f_k.
2. **Consensus:** Communicate with neighbors and average parameters.





View DGD as Projected GD

Concatenate all agents' parameters: $\mathbf{W} = (\mathbf{w}_1^\top, \ldots, \mathbf{w}_K^\top)^\top$.

- **Unconstrained GD:** Each agent k runs a gradient step on f_k:

$$\mathbf{y}_k^{t+1} = \mathbf{w}_k^t - \eta \nabla f_k(\mathbf{w}_k^t)$$

  The combined unconstrained step is $\mathbf{Y}^{t+1} = (\mathbf{y}_1^{t+1\top}, \ldots, \mathbf{y}_K^{t+1\top})^\top$.

- **Projection:** To enforce the consensus, DGD explicitly projects vectors $(\mathbf{y}_1, \ldots, \mathbf{y}_K)$ onto the consensus set $\mathcal{C}$ is given by **average**:

$$\mathcal{P}_{\mathcal{C}}(\mathbf{Y}^{t+1})_k = \frac{1}{K} \sum_{j=1}^{K} \mathbf{y}_j^{t+1} = \mathbf{y}^{t+1}$$

  This is the exact "averaging step" in many DGD algorithms.

The DGD update is precisely $(\nabla \mathbf{F}(\mathbf{W}^t) = (\nabla f_1(\mathbf{w}_1^t)^\top, \ldots, \nabla f_K(\mathbf{w}_K^t)^\top)^\top)$

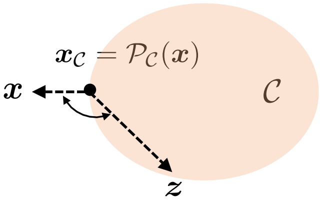
$$\mathbf{W}^{t+1} = \mathcal{P}_{\mathcal{C}}(\mathbf{W}^t - \eta \nabla \mathbf{F}(\mathbf{W}^t))$$

Key insights: Projection theorem

Fact (Projection theorem)

Let $\mathcal{C}$ be closed & convex. Then $\mathbf{x}_{\mathcal{C}}$ is the projection of $\mathbf{x}$ onto $\mathcal{C}$ iff

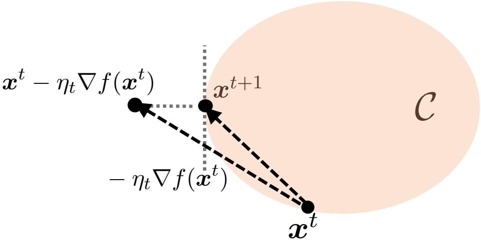
$$(\mathbf{x} - \mathbf{x}_{\mathcal{C}})^\top (\mathbf{z} - \mathbf{x}_{\mathcal{C}}) \leq 0, \quad \text{for all } \mathbf{z} \in \mathcal{C}$$



Intuition: The vector from $\mathbf{x}$ to its projection $\mathbf{x}_{\mathcal{C}}$ is always **orthogonal or pointing inward** relative to $\mathcal{C}$. This guarantees that projection moves us back into the feasible set without "losing descent information."

Aligned with descent direction

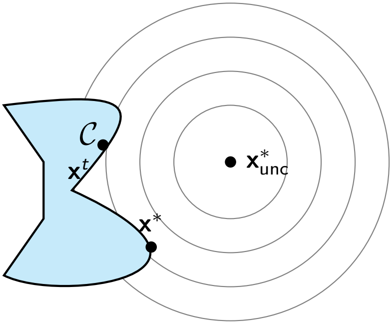
$$(\mathbf{x} - \mathbf{x}_{\mathcal{C}})^\top (\mathbf{z} - \mathbf{x}_{\mathcal{C}}) \leq 0, \quad \text{for } \mathbf{z} := \mathbf{x}^t, \ \mathbf{x} := \mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t), \ \mathbf{x}_{\mathcal{C}} := \mathbf{x}^{t+1}$$



$$-\nabla f(\mathbf{x}^t)^\top (\mathbf{x}^{t+1} - \mathbf{x}^t) \geq 0$$

$\Longrightarrow$ $\mathbf{x}^{t+1} - \mathbf{x}^t$ is positively correlated with the steepest descent direction

Why convexity is crucial?



1. The steepest descent direction $-\nabla f(\mathbf{x}^t)$ points across the "gap."
2. A normal gradient step takes us to an infeasible point $\mathbf{y}^{t+1}$.
3. The **closest** point in $\mathcal{C}$ is $\mathbf{x}^{t+1}$, which is on the other side of the gap.

In this case, $-\nabla f(\mathbf{x}^t)^\top(\mathbf{x}^{t+1} - \mathbf{x}^t) < 0$. Lead to an **ascent** direction.

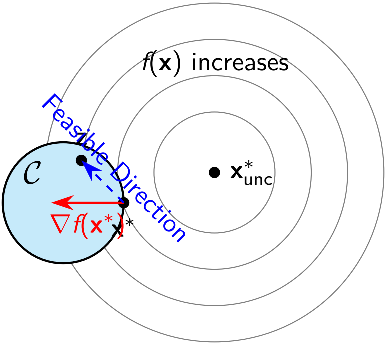
Table of Contents

Strongly convex and smooth problems

$$\begin{array}{ll} \min_{\mathbf{x}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathcal{C} \end{array}$$

- $f(\cdot)$: μ-strongly convex and L-smooth
- $\mathcal{C} \subseteq \mathbb{R}^n$: closed and convex

Optimality in constrained optimization

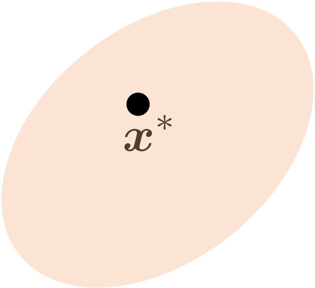


- For **unconstrained** problems, the optimality condition is that $\nabla f(\mathbf{x}^*) = \mathbf{0}$.
- In **constrained** problems, the true minimum $\mathbf{x}^*_{\text{unc}}$ might be outside the feasible set $\mathcal{C}$.
- The optimal feasible solution $\mathbf{x}^*$ is often on the boundary, at the point closest to the true minimum, but $\nabla f(\mathbf{x}^*) \neq \mathbf{0}$

At the optimal solution $\mathbf{x}^*$, any feasible step (from $\mathbf{x}^*$ to another point $\mathbf{z} \in \mathcal{C}$) cannot be a descent direction. Mathematically, this means:

$$\nabla f(\mathbf{x}^*)^\top (\mathbf{z} - \mathbf{x}^*) \geq 0, \quad \text{for all} \mathbf{z} \in \mathcal{C}$$

Convergence for strongly convex and smooth problems



Let's start with the case when $\mathbf{x}^*$ lies in the interior of $\mathcal{C}$ (so $\nabla f(\mathbf{x}^*) = 0$)

Convergence for strongly convex and smooth problems

Theorem 5

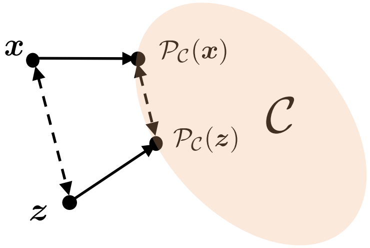
Suppose $\mathbf{x}^* \in \text{int}(\mathcal{C})$ such that $\nabla f(\mathbf{x}^*) = \mathbf{0}$, and let f be μ-strongly convex and L-smooth. If $\eta_t = \frac{2}{\mu+L}$, then

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

where $\kappa = L/\mu$ is condition number.

- the same convergence rate as for the unconstrained case

Aside: nonexpansiveness of projection operator



Fact 6 (Nonexpansiveness of projection)

For any $\mathbf{x}$ and $\mathbf{z}$, one has

$$\|\mathcal{P}_{\mathcal{C}}(\mathbf{x}) - \mathcal{P}_{\mathcal{C}}(\mathbf{z})\|_2 \leq \|\mathbf{x} - \mathbf{z}\|_2$$

Proof of Theorem 5

We have shown for the unconstrained case that

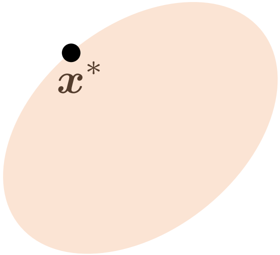
$$||\mathbf{x}^{t+1} - \mathbf{x}^*||_2 = ||\mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t) - \mathbf{x}^*||_2 \leq \frac{\kappa - 1}{\kappa + 1} ||\mathbf{x}^t - \mathbf{x}^*||_2$$

From the nonexpansiveness of $\mathcal{P}_{\mathcal{C}}$, we know

$$\begin{aligned} ||\mathbf{x}^{t+1} - \mathbf{x}^*||_2 &= ||\mathcal{P}_{\mathcal{C}}(\mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t)) - \mathcal{P}_{\mathcal{C}}(\mathbf{x}^*)||_2 \\ &\leq ||\mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t) - \mathbf{x}^*||_2 \\ &= ||\mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t) - \mathbf{x}^* + \eta_t \nabla f(\mathbf{x}^*)||_2 \\ &\leq \frac{\kappa - 1}{\kappa + 1} ||\mathbf{x}^t - \mathbf{x}^*||_2 \end{aligned}$$

Apply it recursively to conclude the proof.

Convergence for strongly convex and smooth problems



What happens if we don't know whether $\mathbf{x}^* \in \text{int}(\mathcal{C})$?

- main issue: $\nabla f(\mathbf{x}^*)$ may not be $\mathbf{0}$ (so prior analysis might fail)

The fixed-point condition of optimality

An optimal point $\mathbf{x}^*$ is a **fixed point** of the projected gradient descent.

If you are at the optimum, it means:

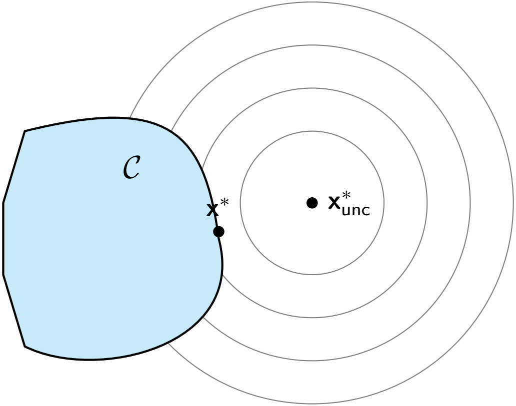
1. A gradient step $-\eta \nabla f(\mathbf{x}^*)$ points away from the feasible set $\mathcal{C}$.
2. Projecting this back onto $\mathcal{C}$ lands you **exactly where you started**.

A point $\mathbf{x}^*$ is optimal if and only if it satisfies:

$$\textbf{Fixed-point equation} \quad \mathbf{x}^* = \mathcal{P}_{\mathcal{C}}(\mathbf{x}^* - \eta \nabla f(\mathbf{x}^*)), \text{ for all } \eta \geq 0$$

This provides a clean way to analyze convergence.

The fixed-point condition of optimality



Fixed-point equation $\quad \mathbf{x}^* = \mathcal{P}_{\mathcal{C}}(\mathbf{x}^* - \eta \nabla f(\mathbf{x}^*)), \;$ for all $\eta \geq 0$

Convergence for strongly convex and smooth problems

Theorem 7 (projected GD for strongly convex and smooth)

Let f be μ-strongly convex and L-smooth. If $\eta_t \equiv \eta = \frac{1}{L}$, then

$$||\mathbf{x}^t - \mathbf{x}^*||_2^2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^t ||\mathbf{x}^0 - \mathbf{x}^*||_2^2$$

- same convergence guarantees as Theorem 5

Proof of Theorem 7

From the nonexpansiveness of $\mathcal{P}_{\mathcal{C}}$ and the fixed-point condition, we know

$$
\begin{aligned}
||\mathbf{x}^{t+1} - \mathbf{x}^*||_2 &= ||\mathcal{P}_{\mathcal{C}}(\mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t)) - \mathcal{P}_{\mathcal{C}}(\mathbf{x}^* - \eta_t \nabla f(\mathbf{x}^*))||_2 \\
&\leq ||\mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t) - (\mathbf{x}^* - \eta_t \nabla f(\mathbf{x}^*))||_2 \\
&= ||\mathbf{x}^t - \mathbf{x}^* - \eta_t(\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*))||_2 \\
&\leq \frac{\kappa - 1}{\kappa + 1}||\mathbf{x}^t - \mathbf{x}^*||_2.
\end{aligned}
$$

Apply it recursively to conclude the proof.

Convex and smooth problems

$$\begin{array}{ll} \min_{\mathbf{x}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathcal{C} \end{array}$$

- $f(\cdot)$: convex and L-smooth
- $\mathcal{C} \subseteq \mathbb{R}^n$: closed and convex

Convergence for convex and smooth problems

Theorem 8 (projected GD for convex and smooth problems)

Let f be convex and L-smooth. If $\eta_t \equiv \eta = \frac{1}{L}$ then

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \leq \frac{3L\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + f(\mathbf{x}^0) - f(\mathbf{x}^*)}{t+1}$$

- similar convergence rate as for the unconstrained case
- a formal proof is provided for **ECE 7290 students**


DANGER
ZONE
HIGH VOLTAGE

Proof of Theorem 8*

We first recall our main steps when handling the unconstrained case:

1. **Step 1:** show cost improvement

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

2. **Step 2:** connect $||\nabla f(\mathbf{x}^t)||_2$ with $f(\mathbf{x}^t)$

$$||\nabla f(x^t)||_2 \geq \frac{f(x^t) - f(x^*)}{||x^t - x^*||_2} \geq \frac{f(x^t) - f(x^*)}{||x^0 - x^*||_2}$$

3. **Step 3:** let $\Delta_t := f(\mathbf{x}^t) - f(\mathbf{x}^*)$ to get

$$\Delta_{t+1} - \Delta_t \leq -\frac{\Delta_t^2}{2L||x^0 - x^*||_2^2}$$

and complete the proof by induction.

Proof of Theorem 8* (cont.)

We then modify these steps for the constrained case. As before, set $g_{\mathcal{C}}(\mathbf{x}^t) = L(\mathbf{x}^t - \mathbf{x}^{t+1})$, which generalizes $\nabla f(\mathbf{x}^t)$ in constrained case.

1. **Step 1:** show cost improvement

$$f(x^{t+1}) \leq f(x^t) - \frac{1}{2L}||g_{\mathcal{C}}(x^t)||_2^2$$

2. **Step 2:** connect $||g_{\mathcal{C}}(\mathbf{x}^t)||_2$ with $f(\mathbf{x}^t)$

$$||g_{\mathcal{C}}(x^t)||_2 \geq \frac{f(x^{t+1}) - f(x^*)}{||x^t - x^*||_2} \geq \frac{f(x^{t+1}) - f(x^*)}{||x^0 - x^*||_2}$$

3. **Step 3:** let $\Delta_t := f(\mathbf{x}^t) - f(\mathbf{x}^*)$ to get

$$\Delta_{t+1} - \Delta_t \leq -\frac{\Delta_{t+1}^2}{2L||x^0 - x^*||_2^2}$$

and complete the proof by induction.

Proof of Theorem 8* (cont.)

Generalize smoothness condition (under convexity) as follows

Lemma 9

Suppose f is convex and L-smooth. For any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, let

$$\mathbf{x}^+ = \mathcal{P}_{\mathcal{C}}(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x}))$$

and $g_{\mathcal{C}}(\mathbf{x}) = L(\mathbf{x} - \mathbf{x}^+)$. Then

$$f(\mathbf{y}) \geq f(\mathbf{x}^+) + g_{\mathcal{C}}(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{1}{2L}||g_{\mathcal{C}}(\mathbf{x})||_2^2$$

Proof of Theorem 8* (cont.)

Step 1: set $\mathbf{x} = \mathbf{y} = \mathbf{x}^t$ in Lemma 9 to reach

$$f(\mathbf{x}^t) \geq f(\mathbf{x}^{t+1}) + \frac{1}{2L} ||g_{\mathcal{C}}(\mathbf{x}^t)||_2^2$$

as desired.

Step 2: set $\mathbf{x} = \mathbf{x}^t$ and $\mathbf{y} = \mathbf{x}^*$ in Lemma 9 to get

$$\begin{aligned} 0 \geq f(\mathbf{x}^*) - f(\mathbf{x}^{t+1}) &\geq g_{\mathcal{C}}(\mathbf{x}^t)^\top (\mathbf{x}^* - \mathbf{x}^t) + \frac{1}{2L} ||g_{\mathcal{C}}(\mathbf{x}^t)||_2^2 \\ &\geq g_{\mathcal{C}}(\mathbf{x}^t)^\top (\mathbf{x}^* - \mathbf{x}^t) \end{aligned}$$

which together with Cauchy-Schwarz yields

$$||g_{\mathcal{C}}(\mathbf{x}^t)||_2 \geq \frac{f(\mathbf{x}^{t+1}) - f(\mathbf{x}^*)}{||\mathbf{x}^t - \mathbf{x}^*||_2} \tag{7}$$

Proof of Theorem 8* (cont.)

It also follows from our analysis for the strongly convex case that (by taking $\mu = 0$ in Theorem 7)

$$\|\mathbf{x}^t - \mathbf{x}^*\|_2 \leq \|\mathbf{x}^0 - \mathbf{x}^*\|_2$$

which combined with (7) reveals

$$\|g_{\mathcal{C}}(\mathbf{x}^t)\|_2 \geq \frac{f(\mathbf{x}^{t+1}) - f(\mathbf{x}^*)}{\|\mathbf{x}^0 - \mathbf{x}^*\|_2}$$

Step 3: letting $\Delta_t = f(\mathbf{x}^t) - f(\mathbf{x}^*)$, the previous bounds together give

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

Use induction to finish the proof (which we omit here).

Proof of Lemma 9*

$$
\begin{aligned}
f(\mathbf{y}) - f(\mathbf{x}^+) &= f(\mathbf{y}) - f(\mathbf{x}) - (f(\mathbf{x}^+) - f(\mathbf{x})) \\
&\geq \underbrace{\nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})}_{\text{convexity}} - \underbrace{\left(\nabla f(\mathbf{x})^\top (\mathbf{x}^+ - \mathbf{x}) + \frac{L}{2} ||\mathbf{x}^+ - \mathbf{x}||_2^2 \right)}_{\text{smoothness}} \\
&= \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}^+) - \frac{L}{2} ||\mathbf{x}^+ - \mathbf{x}||_2^2 \\
&\geq g_{\mathcal{C}}(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}^+) - \frac{L}{2} ||\mathbf{x}^+ - \mathbf{x}||_2^2 \quad \text{(by (6))} \\
&= g_{\mathcal{C}}(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + g_{\mathcal{C}}(\mathbf{x})^\top \underbrace{(\mathbf{x} - \mathbf{x}^+)}_{= \frac{1}{L} g_{\mathcal{C}}(\mathbf{x})} - \frac{L}{2} || \underbrace{\mathbf{x}^+ - \mathbf{x}}_{= -\frac{1}{L} g_{\mathcal{C}}(\mathbf{x})} ||_2^2 \\
&= g_{\mathcal{C}}(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2L} ||g_{\mathcal{C}}(\mathbf{x})||_2^2
\end{aligned}
$$

Summary

- **projected gradient descent**

	stepsize rule	**convergence rate**
convex & smooth	$\eta_t = \frac{1}{L}$	$\mathcal{O}(\frac{1}{t})$
strongly convex & smooth	$\eta_t = \frac{1}{L}$	$\mathcal{O}((1 - \frac{1}{\kappa})^t)$

Recap and fine-tuning

- What we have talked about today?
  - ⇒ What are important constraints in distributed ML?
  - ⇒ How and why projected gradient descent works?
  - ⇒ How fast it converges compared to gradient descent?

Welcome anonymous survey!