Distributed Optimization for Machine Learning

Lecture 5 - Unconstrained Optimization: Gradient Descent

Tianyi Chen

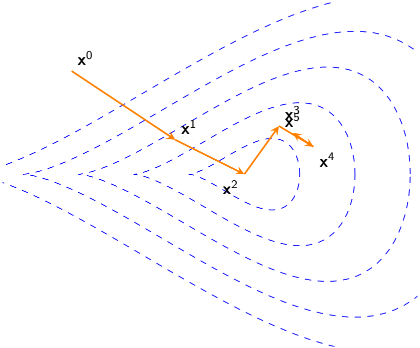
School of Electrical and Computer Engineering
Cornell Tech, Cornell University

September 15, 2025

Gradient descent (GD)

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

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

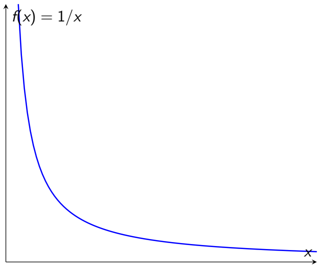


- traced to Augustin Louis Cauchy '1847 ...

Table of Contents

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:

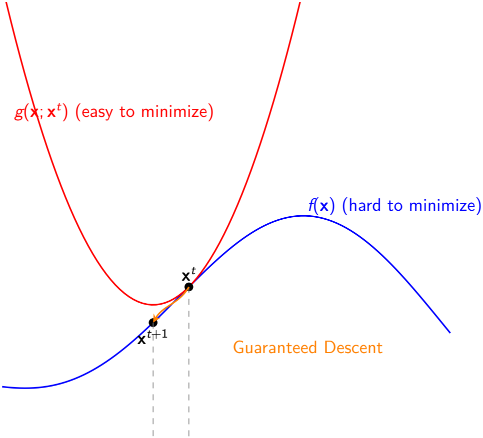
- can we ensure reduction of the objective value (i.e. $f(\mathbf{x}^{t+1}) < f(\mathbf{x}^t)$) without strong convexity?
- what stepsizes guarantee sufficient decrease?

Key idea: **majorization-minimization**

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

Majorization-Minimization principle

The idea is to replace a complex problem with a sequence of simpler ones.



Find a majorizing (quadratic) function of $f(\mathbf{x})$

From the L-smoothness assumption,

$$f(\mathbf{x}) \leq g(\mathbf{x}; \mathbf{x}^t) := f(\mathbf{x}^t) + \nabla f(\mathbf{x}^t)^\top (\mathbf{x} - \mathbf{x}^t) + \frac{L}{2} ||\mathbf{x} - \mathbf{x}^t||_2^2$$

Recall the **gradient descent** recursion

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

We replace $\mathbf{x}$ with $\mathbf{x}^{t+1}$

$$\begin{aligned} f(\mathbf{x}^{t+1}) &\leq f(\mathbf{x}^t) + \nabla f(\mathbf{x}^t)^\top (\mathbf{x}^{t+1} - \mathbf{x}^t) + \frac{L}{2} ||\mathbf{x}^{t+1} - \mathbf{x}^t||_2^2 \\ &= \underbrace{f(\mathbf{x}^t) - \eta_t ||\nabla f(\mathbf{x}^t)||_2^2 + \frac{\eta_t^2 L}{2} ||\nabla f(\mathbf{x}^t)||_2^2}_{\text{majorizing function of objective reduction due to smoothness}} \end{aligned}$$

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)

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

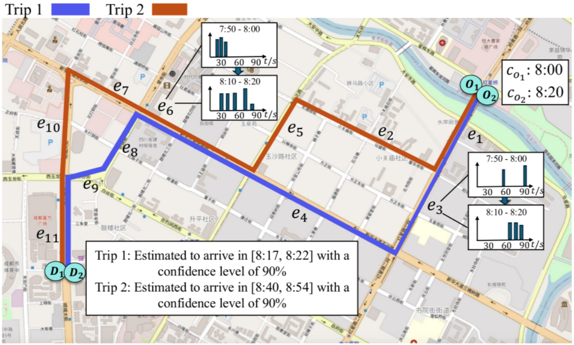
Objective improvement

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

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

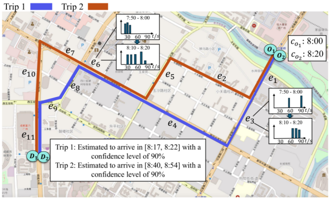
- for η_t sufficiently small, GD results in improvement in the objective
- does **NOT** rely on convexity!

Make connections to ETA



Miles per hour vs Improvement per iteration

How far from the destination?



Condition 1: $||\nabla f(\mathbf{x})||_2^2 \geq c(f(\mathbf{x}) - f(\underbrace{\mathbf{x}^*}_{\text{minimizer}})),$ for all $\mathbf{x}$.

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

Which condition describes a "sharper" or "steeper" minimum?

Condition 1: Steep curvature

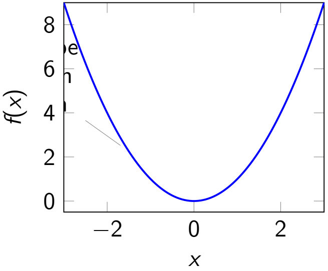
- **Steep curvature:** $\|\nabla f(\mathbf{x})\|_2^2 \geq c\big(f(\mathbf{x}) - f(\mathbf{x}^*)\big)$

- **Interpretation:** The squared gradient is at least **linearly** proportional to the optimality gap.

- **Analogy:** This describes a **sharp, V-shaped valley** or a quadratic bowl. The slope is always significant as long as you are not at the minimum.

- **Result:** Guarantees fast (linear) convergence. All strongly convex functions satisfy this.

Example: $f(x) = x^2$



Strong convexity $\implies$ Steep curvature

For a μ-strongly convex function f, the following holds for all $\mathbf{x}, \mathbf{y}$:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 \tag{1}$$

Let's find the value of $\mathbf{y}$ that minimizes it:

$$\nabla f(\mathbf{x}) + \mu(\mathbf{y} - \mathbf{x}) = 0 \implies \mathbf{y}^* = \mathbf{x} - \frac{1}{\mu} \nabla f(\mathbf{x})$$

Strong convexity $\implies$ Steep curvature

For a μ-strongly convex function f, the following holds for all $\mathbf{x}, \mathbf{y}$:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \tag{1}$$

Let's find the value of $\mathbf{y}$ that minimizes it:

$$\nabla f(\mathbf{x}) + \mu(\mathbf{y} - \mathbf{x}) = 0 \implies \mathbf{y}^* = \mathbf{x} - \frac{1}{\mu}\nabla f(\mathbf{x})$$

$f(\mathbf{x}^*)$ must be not smaller than the minimum value of the RHS of (1).

$$\begin{aligned} f(\mathbf{x}^*) &\geq \min_{\mathbf{y}} \left[f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \right] \\ &= f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \left(-\frac{1}{\mu}\nabla f(\mathbf{x}) \right) + \frac{\mu}{2} \left\| -\frac{1}{\mu}\nabla f(\mathbf{x}) \right\|_2^2 = f(\mathbf{x}) - \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|_2^2 \end{aligned}$$

Rearranging leads to the steep curvature condition with constant $c = 2\mu$.

Condition 2: Flat curvature

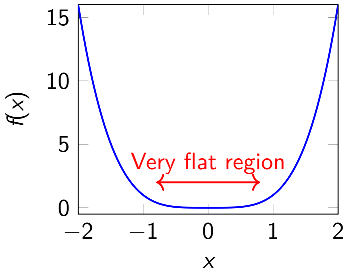
- **Flat curvature:** $\|\nabla f(\mathbf{x})\|_2^2 \geq c\big(f(\mathbf{x}) - f(\mathbf{x}^*)\big)^2$

- **Interpretation:** The squared gradient is proportional to the **square** of the optimality gap.

- **Analogy:** This describes a **flat-bottomed canyon**. The slope can become extremely gentle near the minimum, even if the function value is not yet optimal.

- **Result:** Can lead to very slow (sublinear) convergence.

Example: $f(x) = x^4$



Convexity $\implies$ Flat curvature

The convexity states that for any $\mathbf{x}^t$ and the minimizer $\mathbf{x}^*$:

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

Rearranging this gives us a lower bound on the optimality gap:

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \leq \nabla f(\mathbf{x}^t)^\top (\mathbf{x}^t - \mathbf{x}^*) \tag{2}$$

Convexity $\implies$ Flat curvature

The convexity states that for any $\mathbf{x}^t$ and the minimizer $\mathbf{x}^*$:

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

Rearranging this gives us a lower bound on the optimality gap:

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \leq \nabla f(\mathbf{x}^t)^\top (\mathbf{x}^t - \mathbf{x}^*) \tag{2}$$

We can bound the right-hand side of (2) using Cauchy-Schwarz:

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

Convexity $\implies$ Flat curvature

The convexity states that for any $\mathbf{x}^t$ and the minimizer $\mathbf{x}^*$:

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

Rearranging this gives us a lower bound on the optimality gap:

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \leq \nabla f(\mathbf{x}^t)^\top (\mathbf{x}^t - \mathbf{x}^*) \tag{2}$$

We can bound the right-hand side of (2) using Cauchy-Schwarz:

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

Rearranging the terms gives us a lower bound on the gradient norm:

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

Now, we assume $\|\mathbf{x}^t - \mathbf{x}^*\|_2 \leq \|\mathbf{x}^0 - \mathbf{x}^*\|_2$ for all $t \geq 0$. This is a reasonable assumption for GD on convex functions (prove later).

Linear convergence under steep curvature

From the per-iteration objective improvement

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

Linear convergence under steep curvature

From the per-iteration objective improvement

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

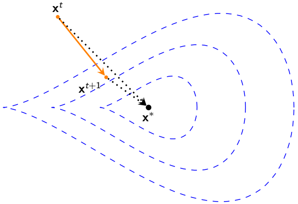
where (i) follows from Fact 7, and (ii) comes from the so-called Polyak-Lojasiewicz (PL) condition (implied by strong convexity)

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

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 $\eta \le 1/L$, it is easily seen that

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

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

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 $f(\cdot)$.

Proof of Theorem 10 (cont.)

From Fact 7,

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

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

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

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

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

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.

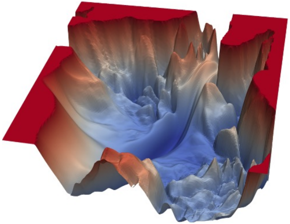
Table of Contents

Nonconvex problems are everywhere

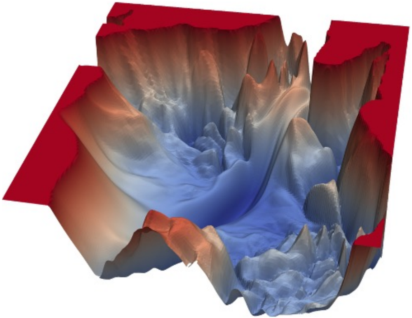
Many empirical risk minimization tasks are nonconvex

$$\min_{\mathbf{x}} \; f(\mathbf{x}; \text{data})$$

- low-rank matrix completion
- blind deconvolution
- dictionary learning
- mixture models
- **learning deep neural nets**
- ...



Challenges



- there may be bumps and local minima everywhere
  - e.g. 1-layer neural net (Auer, Herbster, Warmuth '96; Vu '98)
- no algorithm can solve nonconvex problems efficiently in all cases

Typical convergence guarantees

We cannot hope for efficient global convergence to global minima in general, but we may have

- convergence to stationary points (i.e. $\nabla f(\mathbf{x}) = 0$)
- convergence to local minima
- local convergence to global minima (i.e. when initialized suitably)

Making gradients small

Suppose we are content with any (approximate) stationary point ...

This means that our goal is merely to find a point $\mathbf{x}$ with

$$||\nabla f(\mathbf{x})||_2 \leq \epsilon \quad \text{(called } \epsilon\text{-}\textbf{approximate stationary point}\text{)}$$

Question: can GD achieve this goal? If so, how fast?

Making gradients small

Theorem 11 Let f be L-smooth and $\eta_k \equiv \eta = 1/L$. Assume t is even.

- In general, GD obeys

$$\min_{0 \le k < t} \|\nabla f(\mathbf{x}^k)\|_2 \le \sqrt{\frac{2L(f(\mathbf{x}^0) - f(\mathbf{x}^*))}{t}}$$

- If $f(\cdot)$ is convex, then GD obeys

$$\|\nabla f(\mathbf{x}^t)\|_2 \le \frac{4L\|\mathbf{x}^0 - \mathbf{x}^*\|_2}{t}$$

- Does not imply GD converges to stationary points; it only says that $\exists$ approximate stationary point in the GD trajectory

Proof of Theorem 11

From Fact 7, we know

$$\frac{1}{2L}||\nabla f(\mathbf{x}^k)||_2^2 \leq f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}), \quad \text{for all} k$$

This leads to a telescopic sum when summed over $k = t_0$ to $k = t - 1$:

$$\begin{aligned} \frac{1}{2L}\sum_{k=t_0}^{t-1} ||\nabla f(\mathbf{x}^k)||_2^2 &\leq \sum_{k=t_0}^{t-1}(f(\mathbf{x}^k) - f(\mathbf{x}^{k+1})) = f(\mathbf{x}^{t_0}) - f(\mathbf{x}^t) \\ &\leq f(\mathbf{x}^{t_0}) - f(\mathbf{x}^*) \end{aligned}$$

$$\implies \min_{t_0 \leq k < t} ||\nabla f(\mathbf{x}^k)||_2 \leq \sqrt{\frac{2L(f(\mathbf{x}^{t_0}) - f(\mathbf{x}^*))}{t - t_0}} \qquad (11)$$

Proof of Theorem 11 (cont.)

For a general $f(\cdot)$, taking $t_0 = 0$ immediately establishes the claim.

If $f(\cdot)$ is convex, invoke Theorem 10 to obtain

$$f(\mathbf{x}^{t_0}) - f(\mathbf{x}^*) \leq \frac{2L\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{t_0}$$

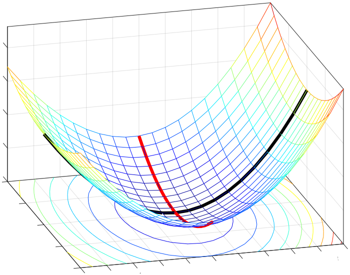
Taking $t_0 = t/2$ and combining it with (11) give

$$\min_{t_0 \leq k < t} \|\nabla f(\mathbf{x}^k)\|_2 \leq \frac{2L}{\sqrt{t(t-t_0)}} \|\mathbf{x}^0 - \mathbf{x}^*\|_2 = \frac{4L\|\mathbf{x}^0 - \mathbf{x}^*\|_2}{t}$$

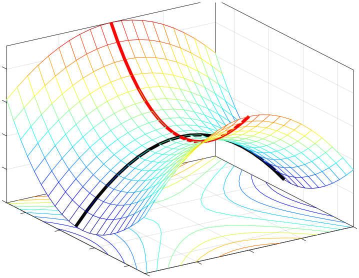
In view of Lemma 9 (smooth and convex),
$\min_{t_0 \leq k < t} \|\nabla f(\mathbf{x}^k)\|_2 = \|\nabla f(\mathbf{x}^t)\|_2$, thus concluding the proof.

Escaping saddles

There are at least two kinds of points with vanishing gradients



global and local minimum



saddle point

Escaping saddle points

Saddle points look like "unstable" critical points; can we hope to at least avoid saddle points?

GD cannnot always escape saddles

- e.g. if $\underbrace{\mathbf{x}^0 \text{ happens to be a saddle}}_{\text{can often be prevented by random initialization}}$, then GD gets trapped (since $\nabla f(\mathbf{x}^0) = 0$)

Fortunately, under mild conditions, **randomly initialized** GD converges to local (sometimes even global) minimum almost surely (Lee et al.)!

Example

Consider a simple **nonconvex** quadratic minimization problem

$$\min_{\mathbf{x}} \ f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x}$$

- $\mathbf{A} = \mathbf{u}_1\mathbf{u}_1^\top - \mathbf{u}_2\mathbf{u}_2^\top$, where $||\mathbf{u}_1||_2 = ||\mathbf{u}_2||_2 = 1$ and $\mathbf{u}_1^\top \mathbf{u}_2 = 0$

This problem has (at least) a saddle point: $\mathbf{x} = \mathbf{0}$ (why?)

- if $\mathbf{x}^0 = \mathbf{0}$, then GD gets stuck at $\mathbf{0}$ (i.e. $\mathbf{x}^t \equiv \mathbf{0}$)
- what if we initialize GD randomly? can we hope to avoid saddles?

Example (cont.)

Fact 12 If $\mathbf{x}^0 \sim \mathcal{N}(0, \mathbf{I})$, then with prob. approaching 1, GD with $\eta < 1$ obeys

$$\|\mathbf{x}^t\|_2 \to \infty \quad \text{as } t \to \infty$$

- Interestingly, GD (almost) never gets trapped in the saddle 0!

Example (cont.)

Proof of Fact 12: Observe that

$$\mathbf{I} - \eta\mathbf{A} = \mathbf{I}_{\perp} + (1-\eta)\mathbf{u}_1\mathbf{u}_1^{\top} + (1+\eta)\mathbf{u}_2\mathbf{u}_2^{\top}$$

where $\mathbf{I}_{\perp} := \mathbf{I} - \mathbf{u}_1\mathbf{u}_1^{\top} - \mathbf{u}_2\mathbf{u}_2^{\top}$. It can be easily verified that

$$(\mathbf{I} - \eta\mathbf{A})^t = \mathbf{I}_{\perp} + (1-\eta)^t\mathbf{u}_1\mathbf{u}_1^{\top} + (1+\eta)^t\mathbf{u}_2\mathbf{u}_2^{\top}$$

$$\begin{aligned} \Longrightarrow \quad \mathbf{x}^t &= (\mathbf{I} - \eta\mathbf{A})\mathbf{x}^{t-1} = \cdots = (\mathbf{I} - \eta\mathbf{A})^t\mathbf{x}^0 \\ &= \mathbf{I}_{\perp}\mathbf{x}^0 + \underbrace{(1-\eta)^t(\mathbf{u}_1^{\top}\mathbf{x}^0)}_{=:\alpha_t}\mathbf{u}_1 + \underbrace{(1+\eta)^t(\mathbf{u}_2^{\top}\mathbf{x}^0)}_{=:\beta_t}\mathbf{u}_2 \end{aligned}$$

Clearly, $\alpha_t \to 0$ as $\underbrace{t \to \infty, \text{ and } |\beta_t| \to \infty}_{\text{and hence } \|\mathbf{x}^t\|_2 \to \infty}$ $\underbrace{\text{as long as } \beta_0 \neq 0}_{\text{happens with prob. 1}}$

Table of Contents

Convex and smooth problems (cont'd)

Nonconvex and smooth problems

Special cases of linear convergence*

Is strong convexity necessary for linear convergence?

So far, linear convergence under strong convexity and smoothness.

Strong convexity requirement can often be relaxed

- local strong convexity
- Polyak-Lojasiewicz condition

Example: logistic regression

Suppose we obtain m independent binary samples

$$y_i = \begin{cases} 1, & \text{with prob. } \frac{1}{1+\exp(-\mathbf{a}_i^\top \mathbf{x}^\natural)} \\ -1, & \text{with prob. } \frac{1}{1+\exp(\mathbf{a}_i^\top \mathbf{x}^\natural)} \end{cases}$$

where $\{\mathbf{a}_i\}$: known design vectors; $\mathbf{x}^\natural \in \mathbb{R}^n$: unknown parameters

Example: logistic regression

The maximum likelihood estimate (MLE) is given by (after a little manipulation)

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} \log\left(1+\exp(-y_i\mathbf{a}_i^\top\mathbf{x})\right)$$

- $\nabla^2 f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m \underbrace{\frac{\exp(-y_i\mathbf{a}_i^\top\mathbf{x})}{(1+\exp(-y_i\mathbf{a}_i^\top\mathbf{x}))^2}}_{\to 0 \text{ if } \mathbf{x}\to\infty} \mathbf{a}_i\mathbf{a}_i^\top \xrightarrow{\mathbf{x}\to\infty} \mathbf{0} \implies f$ is

  0-strongly convex

- Does it mean we no longer have linear convergence?

Local strong convexity

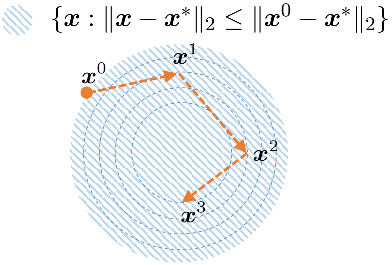
Theorem (GD for locally strongly convex and smooth functions)

Let f be **locally** μ-strongly convex and L-smooth such that

$$\mu \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}, \quad \text{for all} \mathbf{x} \in \mathcal{B}_0$$

where $\mathcal{B}_0 := \{\mathbf{x} : ||\mathbf{x} - \mathbf{x}^*||_2 \leq ||\mathbf{x}^0 - \mathbf{x}^*||_2\}$ and $\mathbf{x}^*$ is the minimizer. **Then Theorem 2.1 continues to hold**.

Local strong convexity



- Suppose $\mathbf{x}^t \in \mathcal{B}_0$. Then repeating our previous analysis yields

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

- This also means $\mathbf{x}^{t+1} \in \mathcal{B}_0$, so the above bound continues to hold for the next iteration ...

Local strong convexity

Back to the logistic regression example, the local strong convexity parameter is given by

$$\inf_{\mathbf{x}:\|\mathbf{x}-\mathbf{x}^*\|_2 \le \|\mathbf{x}^0-\mathbf{x}^*\|_2} \lambda_{\min}\left(\frac{1}{m}\sum_{i=1}^{m}\frac{\exp(-y_i\mathbf{a}_i^\top\mathbf{x})}{(1+\exp(-y_i\mathbf{a}_i^\top\mathbf{x}))^2}\mathbf{a}_i\mathbf{a}_i^\top\right) \tag{6}$$

which is often strictly bounded away from 0, thus enabling linear convergence.

- For example, when $\mathbf{x}^* = 0$ and $\mathbf{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_n)$, one often has (6) $\ge c_0$ for some universal constant $c_0 > 0$ with high prob if $m/n > 2$ (Sur et al. '17).

Example: over-parametrized linear regression

- m data samples $\{\mathbf{a}_i \in \mathbb{R}^n, y_i \in \mathbb{R}\}_{1 \le i \le m}$
- linear regression: find a linear model that best fits the data

$$\min_{\mathbf{x} \in \mathbb{R}^n} \; f(\mathbf{x}) \triangleq \frac{1}{2} \sum_{i=1}^{m} (\mathbf{a}_i^\top \mathbf{x} - y_i)^2$$

Over-parameterization: model dimension $>$ sample size (i.e. $n > m$)

— a regime of particular importance in deep learning

Example: over-parametrized linear regression

While this is a convex problem, it is not strongly convex, since

$$\nabla^2 f(\mathbf{x}) = \sum_{i=1}^{m} \mathbf{a}_i \mathbf{a}_i^\top \text{ is rank-deficient if } n > m$$

But for most "non-degenerate" cases, one has $f(\mathbf{x}^*) = 0$ (why?) and the PL condition is met, and hence GD converges linearly

Example: over-parametrized linear regression

Fact 6 Suppose that $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_m]^\top \in \mathbb{R}^{m \times n}$ has rank m, and that $\eta_t \equiv \eta = \frac{1}{\lambda_{\max}(\mathbf{A}\mathbf{A}^\top)}$. Then GD obeys

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \leq \left(1 - \frac{\lambda_{\min}(\mathbf{A}\mathbf{A}^\top)}{\lambda_{\max}(\mathbf{A}\mathbf{A}^\top)}\right)^t (f(\mathbf{x}^0) - f(\mathbf{x}^*)), \quad \text{for all } t$$

- very mild assumption on $\{\mathbf{a}_i\}$
- no assumption on $\{y_i\}$

Example: over-parametrized linear regression

Fact 6 Suppose that $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_m]^\top \in \mathbb{R}^{m \times n}$ has rank m, and that $\eta_t \equiv \eta = \frac{1}{\lambda_{\max}(\mathbf{A}\mathbf{A}^\top)}$. Then GD obeys

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \leq \left(1 - \frac{\lambda_{\min}(\mathbf{A}\mathbf{A}^\top)}{\lambda_{\max}(\mathbf{A}\mathbf{A}^\top)}\right)^t (f(\mathbf{x}^0) - f(\mathbf{x}^*)), \quad \text{for all } t$$

- **(aside)** while there are many global minima for this over-parametrized problem, GD has **implicit bias**
  - GD converges to a global min closest to initialization $\mathbf{x}^0$!

Proof of Fact 6

Everything boils down to showing the PL condition

$$\|\nabla f(\mathbf{x})\|_2^2 \geq 2\lambda_{\min}(\mathbf{A}\mathbf{A}^\top) f(\mathbf{x}) \tag{9}$$

If this holds, then the claim follows immediately from Theorem 5 and the fact $f(\mathbf{x}^*) = 0$.

To prove (9), let $\mathbf{y} = [y_i]_{1 \leq i \leq m}$, and observe $\nabla f(\mathbf{x}) = \mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{y})$. Then

$$\begin{aligned} \|\nabla f(\mathbf{x})\|_2^2 &= (\mathbf{A}\mathbf{x} - \mathbf{y})^\top \mathbf{A}\mathbf{A}^\top (\mathbf{A}\mathbf{x} - \mathbf{y}) \\ &\geq \lambda_{\min}(\mathbf{A}\mathbf{A}^\top) \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 \\ &= 2\lambda_{\min}(\mathbf{A}\mathbf{A}^\top) f(\mathbf{x}), \end{aligned}$$

which satisfies the PL condition (9) with $\mu = \lambda_{\min}(\mathbf{A}\mathbf{A}^\top)$.

Recap and fine-tuning

- What we have talked about today?
  - $\Rightarrow$ How GD performs in convex and smooth problems?
  - $\Rightarrow$ Without convexity, where it converges to? How fast?

Welcome anonymous survey!