Chapter 1: Gradient Descent

Ernest K. Ryu

MATH 164: Optimization
University of California, Los Angeles
Department of Mathematics

Last edited: February 7, 2025

Gradient descent

Consider the optimization problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable.[1]

Gradient descent (GD) has the form

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

for $k = 0, 1, \ldots$, where $x_0 \in \mathbb{R}^n$ is a suitably chosen starting point and $\alpha_0, \alpha_1, \ldots \in \mathbb{R}$ is a positive step size sequence.

Under suitable conditions, we hope $x_k \overset{?}{\to} x_\star$ for some solution $x_\star$.

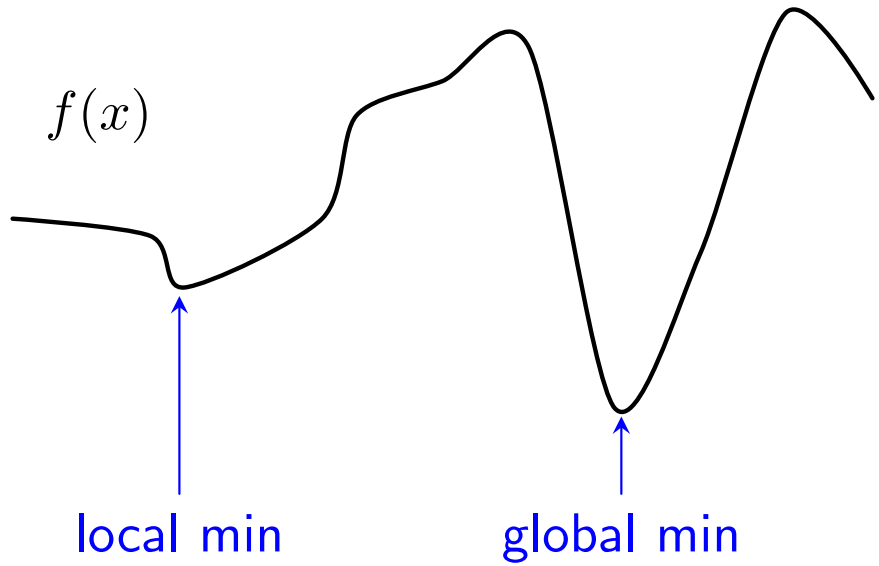
[1] If f is not differentiable, then *gradient* descent is not well defined, right?

Local vs. global minima

$x_\star$ is a *local minimum* if $f(x) \geq f(x_\star)$ within a small neighborhood.[2]

$x_\star$ is a *global minimum* if $f(x) \geq f(x_\star)$ for all $x \in \mathbb{R}^n$

In the worst case, finding the global minimum of an optimization problem is difficult. (The class of non-convex optimization problems is NP-hard.)



[2] if $\exists\, r > 0$ s.t. $\forall\, x$ s.t. $\|x - x_\star\| \leq r \Rightarrow f(x) \geq f(x_\star)$

What can we prove?

Without further assumptions, there is no hope of showing that GD finds the global minimum since GD can never "know" if it is stuck in a local minimum.

We cannot prove the function value converges to the global optimum. We instead prove $\nabla f(x_k) \to 0$. Roughly speaking, this is similar but weaker than proving that x_k converges to a local minimum.[3]

[3] Without further assumptions, we cannot show that x_k converges to a limit, and even x_k does converge to a limit, we cannot guarantee that that limit is not a saddle point or even a local maximum. Nevertheless, people commonly use the argument that x_k "usually" converges and that it is "unlikely" that the limit is a local maximum or a saddle point. More on this later.

$-\nabla f$ is steepest descent direction

From vector calculus, we know that ∇f is the steepest ascent direction, so $-\nabla f$ is the steepest descent direction. In other words,

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

is moving in the steepest descent direction, which is $-\nabla f(x_k)$ at the current position x_k, scaled by $\alpha_k > 0$.

Taylor expansion of f about x_k

$$f(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \mathcal{O}\big(\|x - x_k\|^2\big).$$

Plugging in x_{k+1}

$$f(x_{k+1}) = f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \mathcal{O}(\alpha_k^2).$$

For small (cautious) α_k, a GD step reduces function value.

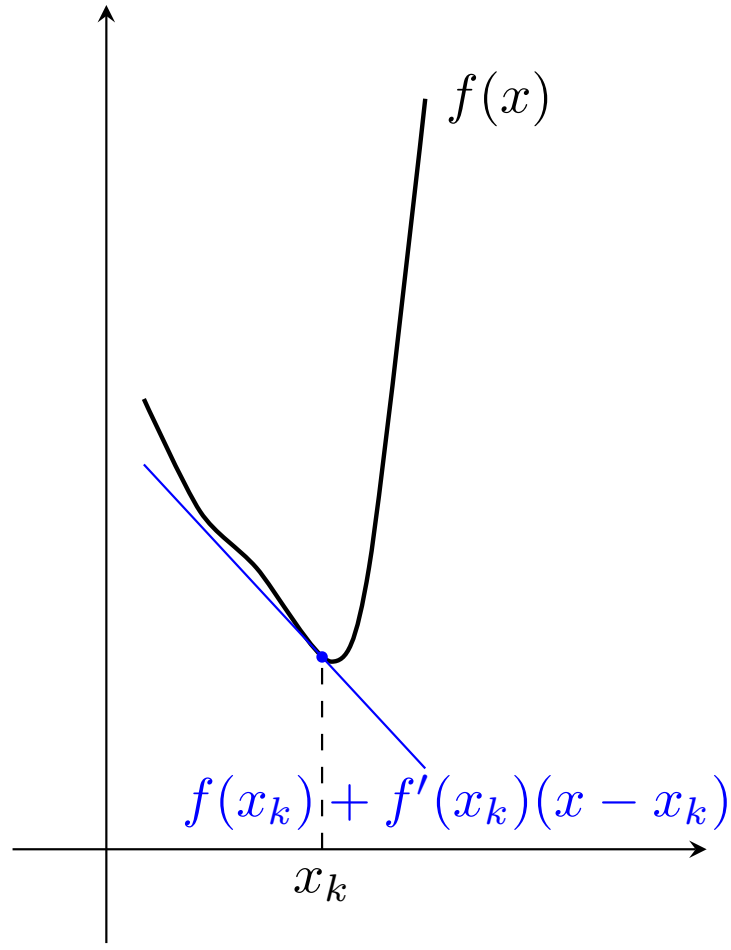
Is GD a "descent method"?

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

Without further assumptions, $-\nabla f(x_k)$ only provides directional information. How far should you go? How large should α_k be?

A step of GD need not result in descent, i.e., $f(x_{k+1}) > f(x_k)$ is possible.

Calculus only guarantees the accuracy of the Taylor expansion in an infinitesimal neighborhood.



Step size selection for GD

How do we choose the step size α_k and ensure convergence?

We consider 3 solutions:

- Make an assumption allowing us to choose α_k and ensures $f(x_k)$ will descend.
  - Estimate the L needed to choose α_k.
- Do a line search to ensure that $f(x_k)$ will descend.
- Drop the insistence that $f(x_k)$ must consistently go down.

Outline

GD for smooth non-convex functions

Consider the optimization problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is "L-smooth" (but not necessarily convex).

We consider GD with constant step size:

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

(So $\alpha = \alpha_0 = \alpha_1 = \cdots$.)

We will show the following.

Theorem.
Assume $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth and $\inf f > -\infty$. Let $\alpha \in (0, 2/L)$. Then, the GD iterates satisfy $\nabla f(x_k) \to 0$.

L-smoothness

For $L > 0$, we say $f : \mathbb{R}^n \to \mathbb{R}$ is *L-smooth* if f is differentiable and

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \qquad \forall\, x, y \in \mathbb{R}^n.$$

I.e., $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is L-Lipschitz continuous. We say f is *smooth* if it is L-smooth for some $L > 0$.

Interpretation 1: ∇f does not change too rapidly. This makes the first-order Taylor expansion reliable beyond an infinitesimal neighborhood. (Further quantified on next slide.)

If f twice-continuously differentiable, then L-smoothness is equivalent to

$$-L \leq \lambda_{\min}(\nabla^2 f(x)) \leq \lambda_{\max}(\nabla^2 f(x)) \leq L, \qquad \forall\, x \in \mathbb{R}^n.$$

Interpretation 2: The curvature f, quantified by $\nabla^2 f$, has lower and upper bounds $\pm L$.

The name "smoothness", as used in optimization, is somewhat confusing because in other areas of mathematics, "smoothness" often refers to infinite differentiability.

Smoothness $\Rightarrow$ first-order Taylor has small remainder

For GD to work with a fixed non-adaptive step size, we need assurance that the first-order Taylor expansion is a good approximation within a sufficiently large neighborhood. L-smoothness provides this assurance.

Lemma.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth. Then

$$\left|f(x+\delta) - \big(f(x) + \langle \nabla f(x), \delta\rangle\big)\right| \leq \frac{L}{2}\|\delta\|^2, \qquad \forall\, x, \delta \in \mathbb{R}^n.$$

Note

$$R_1(\delta; x) = f(x+\delta) - \big(f(x) + \langle \nabla f(x), \delta\rangle\big)$$

is the remainder between f and its first-order Taylor expansion about x. This lemma provides a quantitative bound $|R_1(\delta; x)| \leq \mathcal{O}(\|\delta\|^2)$.

L-smoothness lower and upper bounds

The claimed inequality

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

is equivalent to

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

We will only prove the upper bound $\leq$. The lower bound $\leq$ follows from the same reasoning with some sign changes. (Also, we only use $\leq$.)

Proof of the upper bound $\leq$. Define $g : \mathbb{R} \to \mathbb{R}$ by

$$g(t) = f(x + t\,\delta).$$

Then g is differentiable, and its derivative is

$$g'(t) = \langle \nabla f(x + t\,\delta), \delta \rangle.$$

Next, observe that g' is $(L\|\delta\|^2)$-Lipschitz continuous. Indeed,

$$\begin{aligned} |g'(t_1) - g'(t_0)| &= \big|\langle \nabla f(x + t_1\,\delta) - \nabla f(x + t_0\,\delta), \delta \rangle\big| \\ &\leq \big\|\nabla f(x + t_1\,\delta) - \nabla f(x + t_0\,\delta)\big\| \|\delta\| \leq L\|\delta\|^2 |t_1 - t_0|. \end{aligned}$$

Finally, we conclude that

$$\begin{aligned} f(x + \delta) = g(1) &= g(0) + \int_0^1 g'(t)\,dt \\ &\leq f(x) \; + \; \int_0^1 \big(g'(0) + L\|\delta\|^2 t\big)\,dt \\ &= f(x) + \langle \nabla f(x), \delta \rangle + \frac{L}{2}\|\delta\|^2. \end{aligned}$$

□

Summability lemma

Lemma.
Let $V_0, V_1, \ldots \in \mathbb{R}$ and $S_0, S_1, \ldots \in \mathbb{R}$ be nonnegative sequences satisfying

$$V_{k+1} \leq V_k - S_k$$

for $k = 0, 1, \ldots$. Then $S_k \to 0$.

Key idea. S_k measures progress (decrease) made in iteration k. Since $V_k \geq 0$, V_k cannot decrease forever, so the progress (magnitude of S_k) must diminish to 0.

Proof. Sum the inequality from $i = 0$ to k

$$V_{k+1} + \sum_{i=0}^{k} S_i \leq V_0.$$

Let $k \to \infty$

$$\sum_{i=0}^{\infty} S_i \leq V_0 - \lim_{k\to\infty} V_k \leq V_0$$

Since $\sum_{i=0}^{\infty} S_i < \infty$, we conclude $S_i \to 0$. □

Convergence proof for smooth non-convex functions

Theorem.
Assume $f\colon \mathbb{R}^n \to \mathbb{R}$ *is* L*-smooth and* $\inf f > -\infty$*. Let* $\alpha \in (0, 2/L)$*. Then, the GD iterates satisfy* $\nabla f(x_k) \to 0$*.*

Proof. Use the Lipschitz gradient lemma with $x = x_k$ and $\delta = -\alpha \nabla f(x_k)$ to obtain

$$f(x_{k+1}) \leq f(x_k) - \alpha\big(1 - \tfrac{\alpha L}{2}\big)\|\nabla f(x_k)\|^2,$$

and

$$\overbrace{\big(f(x_{k+1}) - \inf_x f(x)\big)}^{\stackrel{\text{def}}{=} V_{k+1}} \leq \overbrace{\big(f(x_k) - \inf_x f(x)\big)}^{\stackrel{\text{def}}{=} V_k} - \overbrace{\underbrace{\alpha\big(1 - \tfrac{\alpha L}{2}\big)}_{\substack{>0 \\ \text{for } \alpha \in (0,2/L)}} \|\nabla f(x_k)\|^2}^{\stackrel{\text{def}}{=} S_k}.$$

By the summability lemma, we have $\|\nabla f(x_k)\|^2 \to 0$ and thus $\nabla f(x_k) \to 0$. □

GD experiments and curvature

GD with line search

Consider

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable but not necessarily smooth.

GD with *exact line search*

$$\begin{aligned} g_k &= \nabla f(x_k) \\ \alpha_k &\in \operatorname*{argmin}_{\alpha \in \mathbb{R}} f(x_k - \alpha g_k) \\ x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \end{aligned}$$

performs a one-dimensional search in the direction of the gradient.

Theorem.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Then GD with exact line search satisfies

$$f(x_k) \searrow f_\infty \in [-\infty, \infty).$$

Proof. By construction, we have $f(x_{k+1}) \leq f(x_k)$. A non-increasing sequence of real numbers converges to a value in $[-\infty, \infty)$. □

GD with inexact line search

Computing the exact line search is often expensive and unnecessary.
GD with *inexact line search*

$$
\begin{aligned}
g_k &= \nabla f(x_k) \\
\alpha_k &= \text{InexLineSearch}(f, x_k, g_k) \\
x_{k+1} &= x_k - \alpha_k \nabla f(x_k)
\end{aligned}
$$

InexLineSearch(f, x, g) :

$\alpha \leftarrow \beta$ // some initial constant > 0

if $g == 0$: return α

while $f(x - \alpha g) \geq f(x)$

  $\alpha \leftarrow \alpha/2$

return α

This inexact line search is also called a *backtracking line search*.

Theorem.

If f is differentiable, the line search terminates in finite steps.

Proof. Since f is differentiable,

$$f(x - \alpha g) = f(x) - \alpha \|g\|^2 + o(\alpha)$$

and there is a threshold $A > 0$ such that $f(x - \alpha g) < f(x)$ for $\alpha \in (0, A)$. The halving process of α eventually results in $f(x - \alpha g) < f(x)$ (by coincidence) or enters the interval $\alpha \in (0, A)$. $\square$

GD with inexact line search

The starting step size $\beta > 0$ is a parameter to be tuned.

With large β, we have to perform the backtracking loop many times, but we have the opportunity to take a long step.

With small β, the backtracking loop may terminate more quickly, but we won't take steps larger than β.

One can modify the algorithm to adaptively decrease or increase β based on the history of backtracking.

How to choose the starting point x_0

Most (if not all) optimization algorithms require a starting point x_0. It is optimal to choose x_0 to be close (or equal to) $x_\star$, but, of course, we don't know where $x_\star$ is.

If one has an estimate of $x_\star$ based on problem structure, should utilize it.

In convex optimization problems, we often have convergence to the global minimum regardless of x_0, so it is okay to choose $x_0 = 0$.

For non-convex optimization problems, the general prescription is to start with $x_0 =$ random noise.

In some non-convex optimization problems (such as training deep neural networks), one must not use $x_0 = 0$, and a well-tuned random initialization is crucial.

Outline

Convex optimization

The problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x)$$

is a *convex optimization* problem if $f : \mathbb{R}^n \to \mathbb{R}$ is convex, i.e., if

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y), \qquad \forall\, x, y \in \mathbb{R}^n,\, \theta \in [0,1].$$

Finding the global minimum of a convex function is tractable.

> *"In fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."*
> *— R. Tyrrell Rockafellar, in SIAM Review, 1993*

(In other areas of mathematics, linear things tend to be easier, while nonlinear things tend to be significantly harder, but not in optimization.)

$-\nabla f$ points toward $x_\star$

Why can GD find global minimizers of convex functions?

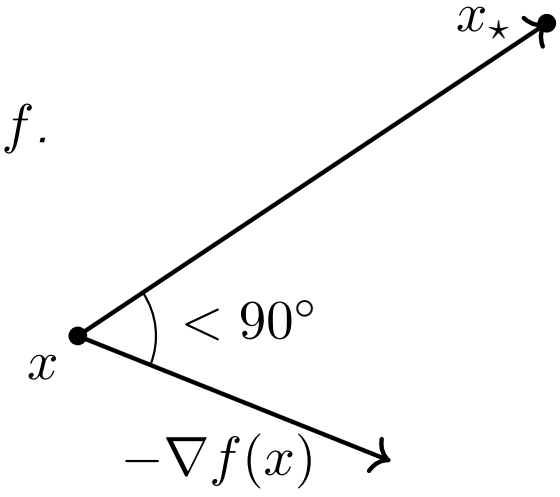
Reason 1. Moving in the $-\nabla f$ direction reduces the function value, taking you to a local minimum, which is a global minimum by convexity.

Reason 2. The $-\nabla f$ direction points toward global minimizers. (This is the more fundamental reason.)

Theorem.

Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be differentiable and convex.
Assume f has a minimizer and let $x_\star \in \operatorname{argmin} f$.
Let $x \in \mathbb{R}^n$ such that $\nabla f(x) \neq 0$. Then,

$$\langle x_\star - x, -\nabla f(x) \rangle > 0.$$



$-\nabla f$ points toward $x_\star$

Theorem.
Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be differentiable and convex. Assume f has a minimizer and let $x_\star \in \operatorname{argmin} f$. Let $x \in \mathbb{R}^n$ such that $\nabla f(x) \neq 0$. Then,

$$\langle x_\star - x, -\nabla f(x)\rangle > 0.$$

Proof. Note that x is not a local or global minimizer since $\nabla f(x) \neq 0$. So, $f(x) - f(x_\star) > 0$. By the convexity inequality, we conclude

$$\langle x_\star - x, -\nabla f(x)\rangle \geq f(x) - f(x_\star) > 0. \qquad \square$$

Consequence: For small α_k, a GD step reduces the distance to a solution.

$$\|\underbrace{x_k - \alpha_k \nabla f(x_k)}_{=x_{k+1}} - x_\star\|^2 = \|x_k - x_\star\|^2 - 2\alpha_k \underbrace{\langle x_k - x_\star, \nabla f(x_k)\rangle}_{>0} + \alpha_k^2 \|\nabla f(x_k)\|^2$$
$$< \|x_k - x_\star\|^2$$

for sufficiently small $\alpha_k > 0$, if $\nabla f(x_k) \neq 0$.

Convergence of GD for smooth convex functions

We quickly establish an inequality we need for the subsequent proof.

Lemma.
Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth and convex. Let $x_\star \in \operatorname{argmin} f$ be a minimizer. Then

$$\langle \nabla f(x), x - x_\star \rangle \geq \frac{1}{L} \|\nabla f(x)\|^2$$

Proof. Note, $\nabla f(x_\star) = 0$. By the cocoercivity inequality, we have

$$f(x_\star) \geq f(x) + \langle \nabla f(x), x_\star - x \rangle + \frac{1}{2L} \|\nabla f(x)\|^2$$

and

$$f(x) \geq f(x_\star) + \frac{1}{2L} \|\nabla f(x)\|^2.$$

Adding these two inequalities yield the stated result. □

Convergence of GD for smooth convex functions

Theorem.
Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be L-smooth and convex. Assume f has a minimizer. Then GD with constant stepsize α satisfying $\alpha \in (0, 2/L)$ converges in the sense of $x_k \to x_\star$ for some $x_\star \in \operatorname{argmin} f$.

Proof. Let $\tilde{x}_\star \in \operatorname{argmin} f$. Using the cocoercivity inequality,

$$\begin{aligned}
\|x_{k+1} - \tilde{x}_\star\|^2 &= \|x_k - \tilde{x}_\star - \alpha \nabla f(x_k)\|^2 \\
&= \|x_k - \tilde{x}_\star\|^2 - 2\alpha \langle \nabla f(x_k), x_k - \tilde{x}_\star \rangle + \alpha^2 \|\nabla f(x_k)\|^2 \\
&\leq \|x_k - \tilde{x}_\star\|^2 - \tfrac{2\alpha}{L} \|\nabla f(x_k)\|^2 + \alpha^2 \|\nabla f(x_k)\|^2 \\
&= \|x_k - \tilde{x}_\star\|^2 - \underbrace{\alpha\left(\tfrac{2}{L} - \alpha\right)}_{>0} \|\nabla f(x_k)\|^2.
\end{aligned}$$

By the summability lemma, $\nabla f(x_k) \to 0$.

Convergence of GD for smooth convex functions

The proof of $x_k \to x_\star$ for some $x_\star \in \operatorname{argmin} f$ is analysis-heavy, and it somewhat exceeds the scope of this class. Nevertheless, we show it for the sake of completeness.

By,

$$\|x_{k+1} - \tilde{x}_\star\|^2 \leq \|x_k - \tilde{x}_\star\|^2 \tag{1}$$

$\|x_k - \tilde{x}_\star\|^2$ is a decreasing sequence and thus has a limit, but the limit is not necessarily 0 (especially if the minimizer is not unique). We argue that $x_k \to x_\star$ for some $x_\star \in \operatorname{argmin} f$ with the steps: (i) x_k has an accumulation point (ii) this accumulation point is a minimizer (iii) this is the only accumulation point.

(i) Inequality (1) tells us $\{x_k\}_k$ lie within $\{x \mid \|x - \tilde{x}^\star\| \leq \|x_0 - \tilde{x}^\star\|\}$, a compact set, so $\{x_k\}_k$ has an accumulation point $x_\star$.

(ii) Accumulation point $x_\star$ satisfies $\nabla f(x_\star) = 0$, as $\nabla f(x_k) \to$ and ∇f is continuous, i.e., $x_\star \in \operatorname{argmin} f$.

(iii) Apply (1) to this accumulation point $x_\star \in \operatorname{argmin} f$ (i.e., plug in $\tilde{x}_\star = x_\star$) to conclude $\|x_k - x_\star\|$ monotonically decreases to 0, i.e., the entire sequence converges to $x_\star$.

□

Convergence of GD for smooth convex functions

Note, $x_k \to x_\star$ immediately implies $f(x_k) \to f(x_\star)$ and $\nabla f(x_k) \to 0$. (L-smoothness implies f and ∇f are continuous.)

As we show next, we can establish a rate (speed) guarantee on $f(x_k) \to f(x_\star)$. Namely, we will show

$$f(x_k) - f(x_\star) \leq \mathcal{O}(1/k).$$

It is also possible to establish a rate guarantee on $\nabla f(x_k) \to 0$. It can be shown that

$$\|\nabla f(x_k)\| \leq \mathcal{O}(1/k).$$

Convergence rate of GD for smooth convex functions

Theorem.
Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth and convex. Assume f has a minimizer $x_\star$. Consider gradient descent with constant stepsize $\alpha = 1/L$. Then, for $k = 1, 2, \ldots,$

$$f(x_k) - f(x_\star) \leq \frac{L}{2k}\|x_0 - x_\star\|^2.$$

Outline of proof. This proof technique is called an *energy function* analysis, *potential function* analysis, or *Lyapunov analysis*. The key insight is to define an appropriate dissipative (non-increasing) quantity.

The main challenge is in identifying the right energy function, which in some cases is highly non-obvious. (The "energy functions" are often unrelated to any notion of physical energy.)

Proof. Define the energy function

$$\mathcal{E}_k = k\big(f(x_k) - f(x_\star)\big) + \frac{L}{2}\|x_k - x_\star\|^2$$

for $k = 0, 1, \ldots$. If the energy is dissipative, then we conclude

$$k(f(x_k) - f(x_\star)) \leq \mathcal{E}_k \leq \cdots \leq \mathcal{E}_0 = \frac{L}{2}\|x_0 - x_\star\|^2.$$

It remains to show $\mathcal{E}_{k+1} \leq \mathcal{E}_k$ for $k = 0, 1, \ldots$. We have

$$\begin{aligned}
\mathcal{E}_{k+1} - \mathcal{E}_k &= (k+1)\big(f(x_{k+1}) - f(x_\star)\big) - k\big(f(x_k) - f(x_\star)\big) \\
&\qquad - \alpha L\langle \nabla f(x_k), x_k - x_\star\rangle + \frac{\alpha^2 L}{2}\|\nabla f(x_k)\|^2 \\
&\leq f(x_k) - f(x_\star) - \frac{k+1}{2L}\|\nabla f(x_k)\|^2 - \langle \nabla f(x_k), x_k - x_\star\rangle + \frac{1}{2L}\|\nabla f(x_k)\|^2 \\
&\leq -\frac{1}{2L}\|\nabla f(x_k)\|^2 - \frac{k+1}{2L}\|\nabla f(x_k)\|^2 + \frac{1}{2L}\|\nabla f(x_k)\|^2 = -\frac{k}{2L}\|\nabla f(x_k)\|^2 \leq 0,
\end{aligned}$$

where the first inequality follows from the L-smoothness lemma

$$(k+1)f(x_{k+1}) = (k+1)f\big(x_k - \tfrac{1}{L}\nabla f(x_k)\big) \leq (k+1)f(x_k) - \frac{(k+1)}{2L}\|\nabla f(x_k)\|^2$$

and the second inequality follows from the cocoercivity inequality

$$f(x_k) - f(x_\star) - \langle \nabla f(x_k), x_k - x_\star\rangle \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2.$$

□

Outline

Projected gradient descent

Constrained optimization problem

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{\text{minimize}} & f(x), \\ \text{subject to} & x \in C, \end{array}$$

where $C \subset \mathbb{R}^n$ is a nonempty closed convex set and $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Assume the constraint set C is computationally easy to project onto.

Projected gradient descent has the form

$$x_{k+1} = \Pi_C\big(x_k - \alpha \nabla f(x_k)\big)$$

for $k = 0, 1, \ldots,$ where $x_0 \in \mathbb{R}^n$ is a suitably chosen starting point and $\alpha \in \mathbb{R}$ is a positive step size.

In other words, projected GD alternates gradient descent steps and projections onto C.

Example: Projection onto ℓ_∞-ball

Consider the ℓ_∞-ball

$$C = \{x \in \mathbb{R}^n \mid \|x\|_\infty \leq 1\} = \{x \in \mathbb{R}^n \mid |x_i| \leq 1, \text{ for } i = 1, \ldots, n\}.$$

Then, Π_C is the thresholding operator

$$\big(\Pi_C(x)\big)_i = \Pi_{[-1,1]}(x_i) = \begin{cases} -1 & \text{if } x_i < -1 \\ x_i & \text{if } -1 \leq x_i \leq 1 \\ +1 & \text{if } 1 < x_i \end{cases}$$

applied element-wise for $i = 1, \ldots, n$.

Since projected GD uses Π_C every iteration, it is important that computing Π_C is inexpensive.

(It's also nice for humans if the code for Π_C is easy to implement.)

Example: ℓ_∞-constrained logistic regression

Consider the ℓ_∞-constrained logistic regression problem

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & \displaystyle\sum_{i=1}^{N} \log\left(1+\exp(v_i^\intercal x)\right) \\ \text{subject to} & \|x\|_\infty \leq 1 \end{array}$$

for some $v_1, \ldots, v_N \in \mathbb{R}$.

Projected GD is

$$x_{k+1} = \Pi\Big(x_k - \alpha \sum_{i=1}^{N} \frac{1}{1+\exp(-v_i^\intercal x_k)} v_i\Big),$$

where Π is the element-wise projection onto $[-1, 1]$.

This is quite simple to implement.

Optimality condition for constrained optimization

Recall that in unconstrained optimization, $\nabla f(x) = 0$ is a necessary condition for x to be a solution. This is called an *optimality condition*. We have an analogous optimality condition for constrained optimization.

Theorem.

Let $C \subset \mathbb{R}^n$ be a nonempty closed convex set and $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. If $x_\star \in \operatorname{argmin}_{x \in C} f(x)$, then

$$\langle \nabla f(x_\star), x - x_\star \rangle \geq 0, \qquad \forall\, x \in C.$$

Motivation. Imagine we are minimizing a linear objective subject to a constraint:

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{\text{minimize}} & \langle g, x \rangle \\ \text{subject to} & x \in C. \end{array}$$

Then, $x_\star$ being a solution is defined as

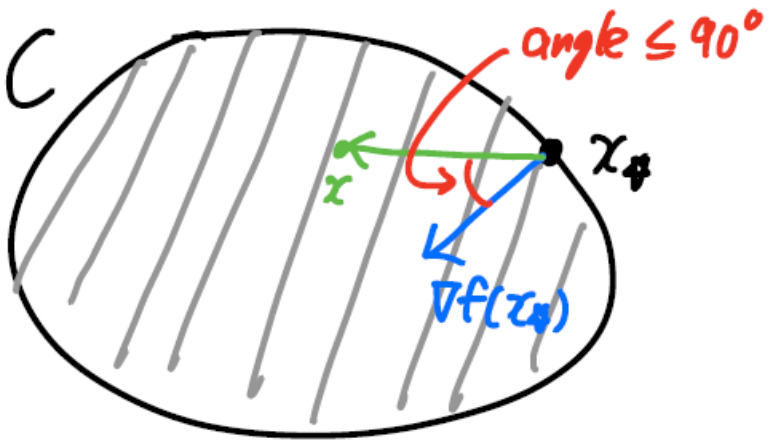
$$\langle g, x \rangle \geq \langle g, x_\star \rangle, \qquad \forall\, x \in C.$$

When f is not linear, we expect something similar within a neighborhood.

Optimality condition for constrained optimization

Theorem.
Let $C \subset \mathbb{R}^n$ be nonempty closed convex and $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. If $x_\star \in \operatorname{argmin}_{x \in C} f(x)$, then

$$\langle \nabla f(x_\star), x - x_\star \rangle \geq 0, \qquad \forall\, x \in C.$$



Optimality condition for constrained optimization

Theorem.
Let $C \subset \mathbb{R}^n$ be nonempty closed convex and $f\colon \mathbb{R}^n \to \mathbb{R}$ be differentiable. If $x_\star \in \operatorname{argmin}_{x\in C} f(x)$, then

$$\langle \nabla f(x_\star), x - x_\star \rangle \geq 0, \qquad \forall\, x \in C.$$

Proof. Let $x \in C$. If $x = x_\star$, there is nothing to prove, so assume $x \neq x_\star$. Then,

$$f(x_\star) \leq f\Big(\underbrace{x_\star + \theta(x - x_\star)}_{=(1-\theta)x_\star + \theta x \in C} \Big) \qquad \forall\, \theta \in (0, 1].$$

and

$$0 \leq \lim_{\theta \to 0} \frac{f(x_\star + \theta(x - x_\star)) - f(x_\star)}{\theta} = \langle \nabla f(x_\star), x - x_\star \rangle.$$

□

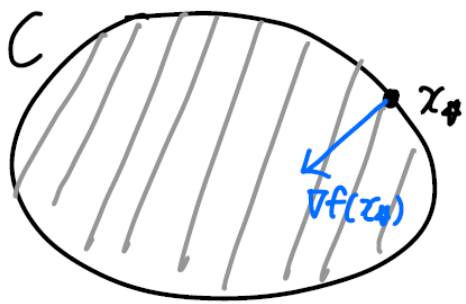
Optimality condition for constrained optimization

For unconstrained convex optimization, $\nabla f(x) = 0$ is a necessary and sufficient condition for optimality. The same pattern holds for constrained convex optimization.

Theorem.

Let $C \subset \mathbb{R}^n$ be nonempty closed convex and $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and convex. Then, $x_\star \in \operatorname{argmin}_{x \in C} f(x)$ if and only if

$$\langle \nabla f(x_\star), x - x_\star \rangle \geq 0, \qquad \forall\, x \in C.$$



for convex functions, moving in direction ∇f (infinitesimal or not) will certainly increase f

Optimality condition for constrained optimization

Theorem.
Let $C \subset \mathbb{R}^n$ be nonempty closed convex and $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and convex. Then, $x_\star \in \operatorname{argmin}_{x \in C} f(x)$ if and only if

$$\langle \nabla f(x_\star), x - x_\star \rangle \geq 0, \qquad \forall\, x \in C.$$

Proof. It remains to show the direction ($\Leftarrow$) under the assumption of convexity. Assume

$$\langle \nabla f(x_\star), x - x_\star \rangle \geq 0, \qquad \forall\, x \in C.$$

By the convexity inequality,

$$\begin{aligned} f(x) &\geq f(x_\star) + \langle \nabla f(x_\star), x - x_\star \rangle \\ &\geq f(x_\star), \end{aligned}$$

and we conclude $x_\star$ is a global minimizer. □

Optimality $\Leftrightarrow$ stationarity

Theorem.
Let $C \subset \mathbb{R}^n$ be nonempty closed convex and $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and convex. Let $\alpha > 0$. Then, $x_\star \in \operatorname{argmin}_{x \in C} f(x)$ if and only if

$$x_\star = \Pi_C(x_\star - \alpha \nabla f(x_\star)).$$

I.e., projected GD stops moving if and only if you are at a solution.

Proof. By the optimality condition, $x_\star$ is a solution if and only if

$$\langle \nabla f(x_\star), x - x_\star \rangle \geq 0, \qquad \forall\, x \in C.$$

This holds if and only if

$$\langle x - x_\star, x_\star - \alpha \nabla f(x_\star) - x_\star \rangle \leq 0, \qquad \forall\, x \in C.$$

By the projection theorem, this holds if and only if

$$x_\star = \Pi_C(x_\star - \alpha \nabla f(x_\star)).$$

□

G-mapping

Let $\alpha > 0$. Let $C \subset \mathbb{R}^n$ be nonempty closed convex and $f\colon \mathbb{R}^n \to \mathbb{R}$ be differentiable. Define $G_\alpha\colon \mathbb{R}^n \to \mathbb{R}$ such that

$$\Pi_C(x - \alpha \nabla f(x)) = x - \alpha G_\alpha(x).$$

In other words, let

$$G_\alpha(x) = \frac{1}{\alpha}\big(x - \Pi_C(x - \alpha \nabla f(x))\big).$$

With this notation, we can express projected GD as

$$x_{k+1} = x_k - \alpha G_\alpha(x_k).$$

We will call G_α the *G-mapping*. In other references, this is called the "gradient mapping," but I dislike this terminology because G_α is not a gradient, although it is a generalization of the gradient.

Note, if $C = \mathbb{R}^n$, then $\Pi_C(x) = x$ and $G_\alpha = \nabla f$.

Descent lemma

First, an intermediate inequality, a consequence of the projection theorem.

Lemma.
Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth and convex. Let $C \subset \mathbb{R}^n$ be nonempty closed convex. Let $\alpha > 0$ and $x_+ = x - \alpha G_\alpha(x)$. Then

$$\langle \nabla f(x), y - x_+ \rangle \geq \langle G_\alpha(x), y - x_+ \rangle.$$

for any $x \in \mathbb{R}^n$ and $y \in C$.

Proof. By the projection theorem,

$$\langle y - x_+, x - \alpha \nabla f(x) - \underbrace{(x - \alpha G_\alpha(x))}_{=x_+} \rangle \leq 0.$$

Reorganizing the terms, we get

$$\langle y - x_+, \nabla f(x) - G_\alpha(x) \rangle \geq 0.$$

Further reorganizing, we get the stated result. □

Descent lemma

Next, we establish our main descent lemma used for the convergence proof of projected GD.

Lemma.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth and convex. Let $C \subset \mathbb{R}^n$ be nonempty closed convex. If $\alpha \in (0, 1/L]$, then

$$f(y) \geq f(\underbrace{x - \alpha G_\alpha(x)}_{=x_+}) + \langle G_\alpha(x), y - x \rangle + \frac{\alpha}{2}\|G_\alpha(x)\|^2$$

for any $x \in \mathbb{R}^n$ and $y \in C$.

This lemma resembles the cocoercivity inequality, but it is not a strict generalization. (When $C = \mathbb{R}^n$ and $G_\alpha(x) = \nabla f(x)$, the resulting inequality is weaker than the cocoercivity inequality.)

$$f(y) \geq f(x_+) + \langle G_\alpha(x), y - x\rangle + \frac{\alpha}{2}\|G_\alpha(x)\|^2$$

Proof. By the L-smoothness lemma, convexity of f, and consequence of the projection theorem, we have

$$\begin{aligned}
f(x_+) &\leq f(x) + \langle \nabla f(x), x_+ - x\rangle + \frac{L}{2}\|x_+ - x\|^2 \\
&\leq f(x) + \langle \nabla f(x), x_+ - x\rangle + \frac{1}{2\alpha}\|x_+ - x\|^2 \\
&= f(x) + \langle \nabla f(x), y - x\rangle + \langle \nabla f(x), x_+ - y\rangle + \frac{\alpha}{2}\|G_\alpha(x)\|^2 \\
&\leq f(y) - \langle \nabla f(x), y - x_+\rangle + \frac{\alpha}{2}\|G_\alpha(x)\|^2 \\
&\leq f(y) - \langle G_\alpha(x), y - x_+\rangle + \frac{\alpha}{2}\|G_\alpha(x)\|^2 \\
&= f(y) - \langle G_\alpha(x), y - x\rangle - \langle G_\alpha(x), \underbrace{x - x_+}_{=\alpha G_\alpha(x)}\rangle + \frac{\alpha}{2}\|G_\alpha(x)\|^2 \\
&= f(y) - \langle G_\alpha(x), y - x\rangle - \frac{\alpha}{2}\|G_\alpha(x)\|^2.
\end{aligned}$$

Reorganizing, we get the stated result. □

Bounty!

In my view, the proofs for the cocoercivity inequality and this descent lemma are opaque.

If you can find a a substantively simpler or intuitive proof for these inequalities, I will add $+20$ points (out of 100 points) on the final exam.

Descent lemma

Plugging $\alpha = 1/L$, $y = x_{k+1}$, and $x = x_k$ into the lemma to get

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|G_\alpha(x_k)\|^2.$$

This is a guarantee on the improvement from x_k to x_{k+1}; the improvement will be proportional to the squared magnitude of the movement.

By L-smoothness, the mapping

$$x_k \mapsto x_k - \alpha\nabla f(x_k)$$

will reduce the function value, but the projection step

$$x_k - \alpha\nabla f(x_k) \mapsto \Pi_C(x_k - \alpha\nabla f(x_k)) = x_{k+1}$$

can and often will increase the function value. The descent property above assures us that the decrease and increase add up to a decrease.

Convergence of projected GD

Theorem.
Let $C \subset \mathbb{R}^n$ be nonempty closed convex and $f\colon \mathbb{R}^n \to \mathbb{R}$ be L-smooth and convex. Assume $\operatorname{argmin}_{x\in C} f(x)$ has a solution. Then projected GD with constant stepsize α satisfying $\alpha \in (0, 1/L]$ converges in the sense of $x_k \to x_\star$ for some $x_\star \in \operatorname{argmin}_{x\in C} f(x)$.[4]

Proof. Let $\tilde{x}_\star \in \operatorname{argmin} f$ and $f_\star = f(\tilde{x}_\star)$. Using the descent lemma,

$$
\begin{aligned}
&\|x_{k+1} - \tilde{x}_\star\|^2 \\
&= \|x_k - \alpha G_\alpha(x_k) - \tilde{x}_\star\|^2 \\
&= \|x_k - \tilde{x}_\star\|^2 - 2\alpha\langle G_\alpha(x_k), x_k - \tilde{x}_\star\rangle + \alpha^2\|G_\alpha(x_k)\|^2 \\
&\le \|x_k - \tilde{x}_\star\|^2 - 2\alpha\big(f(x_{k+1}) - f_\star\big) - \tfrac{\alpha}{L}\|G_\alpha(x_k)\|^2 + \alpha^2\|G_\alpha(x_k)\|^2 \\
&= \|x_k - \tilde{x}_\star\|^2 - 2\alpha\big(f(x_{k+1}) - f_\star\big) - \underbrace{\alpha\big(\tfrac{1}{L} - \alpha\big)}_{\ge 0}\|G_\alpha(x_k)\|^2 \\
&\le \|x_k - \tilde{x}_\star\|^2 - 2\alpha\big(f(x_{k+1}) - f_\star\big)
\end{aligned}
$$

By the summability lemma, $f(x_k) \to f_\star$. With a subsequence argument, we can show $x_k \to x_\star$. □

[4] It is possible to show convergence for $\alpha \in (0, 2/L)$ with more work.

Convergence rate of projected GD

Theorem.
Let $C \subset \mathbb{R}^n$ be nonempty closed convex and $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth and convex. Assume $\operatorname{argmin}_{x \in C} f(x)$ has a solution. Consider projected gradient descent with constant stepsize $\alpha = 1/L$. Then, for $k = 1, 2, \ldots,$

$$f(x_k) - f(x_\star) \leq \frac{L}{2k} \|x_0 - x_\star\|^2.$$

Proof. Define the energy function

$$\mathcal{E}_k = k\big(f(x_k) - f(x_\star)\big) + \frac{L}{2}\|x_k - x_\star\|^2$$

for $k = 0, 1, \ldots$. If the energy is dissipative, then we conclude

$$k(f(x_k) - f(x_\star)) \leq \mathcal{E}_k \leq \cdots \leq \mathcal{E}_0 = \frac{L}{2}\|x_0 - x_\star\|^2.$$

Convergence rate of projected GD

$$\mathcal{E}_k = k\big(f(x_k) - f(x_\star)\big) + \frac{L}{2}\|x_k - x_\star\|^2$$

It remains to show $\mathcal{E}_{k+1} \le \mathcal{E}_k$ for $k = 0, 1, \ldots$. We have

$$\begin{aligned}
\mathcal{E}_{k+1} - \mathcal{E}_k &= (k+1)\big(f(x_{k+1}) - f(x_\star)\big) - k\big(f(x_k) - f(x_\star)\big) \\
&\qquad - \alpha L\langle G_\alpha(x_k), x_k - x_\star\rangle + \frac{\alpha^2 L}{2}\|G_\alpha(x_k)\|^2 \\
&\le f(x_{k+1}) - f(x_\star) - \frac{k}{2L}\|G_\alpha(x_k)\|^2 - \langle G_\alpha(x_k), x_k - x_\star\rangle + \frac{1}{2L}\|G_\alpha(x_k)\|^2 \\
&\le -\frac{1}{2L}\|G_\alpha(x_k)\|^2 - \frac{k}{2L}\|G_\alpha(x_k)\|^2 + \frac{1}{2L}\|G_\alpha(x_k)\|^2 = -\frac{k}{2L}\|G_\alpha(x_k)\|^2 \le 0,
\end{aligned}$$

where the first inequality follows from the descent lemma

$$kf(x_{k+1}) \le kf(x_k) - \frac{k}{2L}\|G_\alpha(x_k)\|^2$$

and the second inequality also follows from the descent lemma

$$f(x_{k+1}) - f(x_\star) - \langle G_\alpha(x_k), x_k - x_\star\rangle \le -\frac{1}{2L}\|G_\alpha(x_k)\|^2.$$

□