Math 273A Notes:
Chapter1

Ernest K. Ryu

October 29, 2025

1 Gradient descent-type methods

Consider the optimization problem

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

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

Gradient descent (GD) has the form

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

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

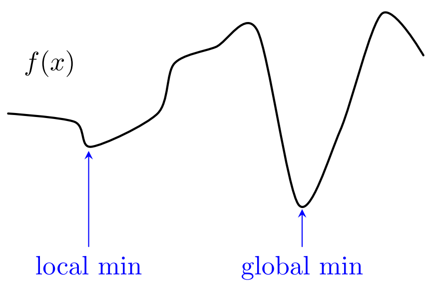
Under suitable conditions, we hope $x_k \overset{?}{\to} x_\star$ for some solution $x_\star$.
$x_\star$ is a *local minimum* if $f(x) \ge f(x_\star)$ within a small neighborhood.[2]

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

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

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

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



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]

$-\nabla f$ 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 the function value.

Is GD a "descent method"?

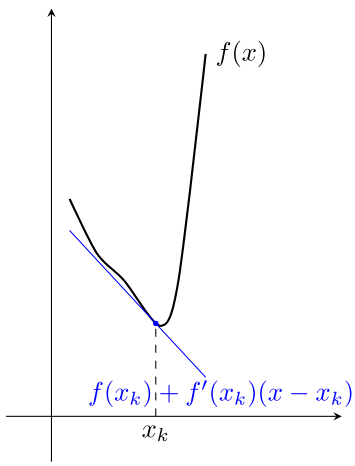
$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

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

[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.

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.

1.1 GD for smooth non-convex functions

Consider the optimization problem

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

where $f\colon \mathbb{R}^n \to \mathbb{R}$ 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 1. *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 $\nabla f(x_k) \to 0$.*

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

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

I.e., $\nabla f\colon \mathbb{R}^n \to \mathbb{R}^n$ is L-Lipschitz continuous. We say f is *smooth* if it is L-smooth for some $L > 0$. [4]

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.)

Lemma 1. *If f twice-continuously differentiable, then L-smoothness is equivalent to*

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

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

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

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

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 $|R_1(\delta; x)| \le \mathcal{O}(\|\delta\|^2)$.

Proof of the upper bound $\le$. 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,

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

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

Finally, we conclude that

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

□

Summability lemma

Lemma 3. *Let $V_0, V_1, \ldots \in \mathbb{R}$ and $S_0, S_1, \ldots \in \mathbb{R}$ be nonnegative sequences satisfying*

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

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

Key idea. S_k measures progress (decrease) made in iteration k. Since $V_k \ge 0$, 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

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

Let $k \to \infty$

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

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

Convergence proof for smooth non-convex functions

Theorem 2. *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

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

and

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

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

GD with line search Consider

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

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

GD with *exact line search*

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

performs a one-dimensional search in the direction of the gradient. (XXX we need to assume the iterations exist.)

Theorem 3. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be differentiable. Then GD with exact line search satisfies*

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

Proof. By construction, we have $f(x_{k+1}) \le f(x_k)$. 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) :
α ← β   // some initial constant > 0
if g == 0 : return α
while f(x − αg) ≥ f(x)
   α ← α/2
return α
```

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

Theorem 4. *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 $\alpha \in (0, A)$. □

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.

2 Smooth convex GD

Convex optimization The problem

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

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

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

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.)

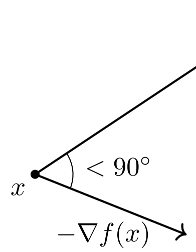
$-\nabla f$ points toward $x_\star$ Why can GD find global minimizers of convex functions?

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 5. *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 $\nabla f(x_k) \neq 0$.

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

Lemma 4. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ 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. $\square$

Theorem 6. *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, $\nabla f(x_k) \to 0$.

The proof of $x_k \to x_\star$ for some $x_\star \in \operatorname{argmin} f$ requires a somewhat delicate analysis argument. 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_∞. I.e., there is a convergent subsequence $x_{k_j} \to x_\infty$, where x_∞ is the limit point.

(ii) Accumulation point x_∞ satisfies $\nabla f(x_\infty) = 0$, as $\nabla f(x_k) \to 0$ and ∇f is continuous, i.e., $x_\infty \in \operatorname{argmin} f$. (We now know that the limit point x_∞ is a solution.)

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

$\square$

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

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

2.1 Convergence <u>rate</u> of GD for smooth convex functions

Theorem 7. *Let $f \colon \mathbb{R}^n \to \mathbb{R}$ 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

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

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

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

where the first inequality follows from the L-smoothness lemma

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

and the second inequality follows from the cocoercivity inequality

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

□

Theorem 8. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be L-smooth and μ-strongly convex. Consider GD with $\alpha_k = 1/L$. Then, for $k = 0, 1, \ldots,$*

$$\|x_k - x_\star\|^2 \le \big(1 - \tfrac{\mu}{L}\big)^k\|x_0 - x_\star\|^2.$$

Proof. From L-smoothness, we have

$$\begin{aligned}
&f(x_k) \ge f(x_\star) + \tfrac{1}{2L}\|\nabla f(x_k)\|^2 \\
&f(x_\star) \ge f(x_k) + \langle \nabla f(x_k), x_\star - x_k\rangle + \tfrac{1}{2L}\|\nabla f(x_k)\|^2.
\end{aligned}$$

Adding the two gives us

$$\langle \nabla f(x_k), x - x_k\rangle \ge \tfrac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Likewise from μ-strong convexity, we have

$$f(x_k) \geq f(x_\star) + \tfrac{\mu}{2}\|x_k - x_\star\|^2$$
$$f(x_\star) \geq f(x_k) + \langle \nabla f(x_k), x_\star - x_k \rangle + \tfrac{\mu}{2}\|x_k - x_\star\|^2.$$

Adding the two gives us

$$\langle \nabla f(x_k), x_k - x_\star \rangle \geq \mu \|x_k - x_\star\|^2.$$

Then, we have

$$\begin{aligned}
\|x_{k+1} - x_\star\|^2 &= \|x_k - x_\star\|^2 - \tfrac{2}{L}\langle \nabla f(x_k), x_k - x_\star \rangle + \tfrac{1}{L^2}\|\nabla f(x_k)\|^2 \\
&\leq \|x_k - x_\star\|^2 - \tfrac{1}{L}\langle \nabla f(x_k), x_k - x_\star \rangle \\
&\leq \big(1 - \tfrac{\mu}{L}\big)\|x_k - x_\star\|^2
\end{aligned}$$

for $k = 0, 1, \ldots$. Finally, by a recursive argument, we have

$$\|x_k - x_\star\|^2 \leq \big(1 - \tfrac{\mu}{L}\big)^k \|x_0 - x_\star\|^2.$$

□

Using the Polyak–Łojasiewicz inequality, we can obtain a rate on f.

Theorem 9. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be L-smooth, convex, and μ-strongly convex. Consider gradient descent with constant stepsize $\alpha_k = 1/L$. Then, for $k = 0, 1, \ldots,$*

$$f(x_k) - f(x_\star) \leq \big(1 - \tfrac{\mu}{L}\big)^k (f(x_0) - f(x_\star))$$

Proof.

$$\begin{aligned}
f(x_{k+1}) - f(x_\star) &= f(x_k - \alpha \nabla f(x_k)) - f(x_\star) \\
&\leq f(x_k) - \alpha \langle \nabla f(x_k), \nabla f(x_k)) \rangle + \frac{\alpha^2 L}{2}\|\nabla f(x_k)\|^2 - f(x_\star) \\
&= f(x_k) - f(x_\star) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \\
&\leq (1 - 1/\kappa)(f(x_k) - f(x_\star)),
\end{aligned}$$

where the first inequality follow from L-smoothness and the third follows from PL. □

2.2 Projected gradient method

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 $f\colon \mathbb{R}^n \to \mathbb{R}$ 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

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

Then, Π_C is the thresholding operator

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

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

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

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.

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

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

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

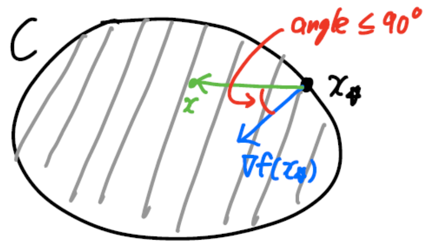
Then, $x_\star$ being a solution is defined as

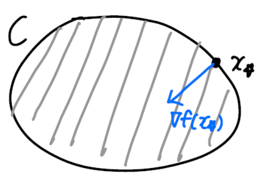
$$\langle g, x \rangle \geq \langle g, x_\star \rangle, \qquad \forall\, x \in C.$$

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

Theorem 10. *Let $C \subset \mathbb{R}^n$ be nonempty closed convex and $f\colon \mathbb{R}^n \to \mathbb{R}$ 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

Proof. ($\Rightarrow$) 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

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

($\Leftarrow$) 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. $\square$

Optimality $\Leftrightarrow$ stationarity

Theorem 11. *Let $C \subset \mathbb{R}^n$ be nonempty closed convex and $f\colon \mathbb{R}^n \to \mathbb{R}$ 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

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

$\square$

Lemma 5. *Let $C \subseteq \mathbb{R}^n$ be nonempty closed covnex. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be μ-strongly convex. Then,*

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

has a unique solution.

Theorem 12. *Let $C \subseteq \mathbb{R}^n$ be nonempty closed covnex. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be L-smooth, and μ-strongly convex. Let $x_\star = \operatorname{argmin}_{x \in C} f(x)$. Consider projcted gradient descent with constant stepsize $\alpha_k = 1/L$. Then, for $k = 0, 1, \dots,$*

$$\|x_k - x_\star\|^2 \leq \left(1 - \tfrac{\mu}{L}\right)^k \|x_0 - x_\star\|^2.$$

Proof.

$$\begin{aligned} \|x_{k+1} - x_\star\|^2 &= \|\Pi_C(x_k - \alpha \nabla f(x_k)) - \Pi_C(x_\star - \alpha \nabla f(x_\star))\|^2 \\ &\leq \|(x_k - \alpha \nabla f(x_k)) - (x_\star - \alpha \nabla f(x_\star))\|^2 \\ &= \|x_k - x_\star\|^2 - 2\alpha \langle \nabla f(x_k) - \nabla f(x_\star), x_k - x_\star \rangle + \alpha^2 \|\nabla f(x_k) - \nabla f(x_\star)\|^2 \\ &\leq \|x_k - x_\star\|^2 - 2\alpha \langle \nabla f(x_k) - \nabla f(x_\star), x_k - x_\star \rangle + \alpha^2 L \langle \nabla f(x_k) - \nabla f(x_\star), x_k - x_\star \rangle \\ &= \|x_k - x_\star\|^2 - \frac{1}{L} \langle \nabla f(x_k) - \nabla f(x_\star), x_k - x_\star \rangle \\ &\leq \left(1 - \tfrac{\mu}{L}\right) \|x_k - x_\star\|^2 \end{aligned}$$

$\square$

2.2.1 Sublinear convergence results

The following results can be show with more work, but we shall move on.

Theorem 13. *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 $x_\star \in \operatorname{argmin}_{x\in C} f(x)$.*

Theorem 14. *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. Consider projected gradient descent with constant stepsize $\alpha = 1/L$. Then, for $k = 1, 2, \dots,$*

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

2.3 Subgradient methods

Consider the optimization problem

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

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

The subgradient method is

$$x_{k+1} \in x_k + \alpha_k \partial f(x_k)$$

Here, the inclusion notation is really a shorthand for

$$\begin{aligned} g_k &\in \partial f(x_k) \\ x_{k+1} &= x_k + \alpha_k \partial f(x_k) \end{aligned}$$

Lemma 6 (Quasi-summability Lemma)**.** *Let $\{V_k\}_{k\in\mathbb{N}}$, $\{S_k\}_{k\in\mathbb{N}}$, $\{U_k\}_{k\in\mathbb{N}}$ be sequences of nonnegative real numbers satisfying the inequality*

$$V_{k+1} \le V_k - S_k + U_k$$

for $k = 0, 1, \dots$ and

$$\sum_{k=0}^{\infty} U_k < \infty.$$

Then,

$$\sum_{k=0}^{\infty} S_k < \infty, \qquad V_k \to V_\infty \in \mathbb{R}.$$

Proof. By summation, we have

$$V_K \le V_0 - \sum_{k=0}^{K} S_k + \sum_{k=0}^{K} U_k.$$

Reorganizing, we get

$$\sum_{k=0}^{K} S_k \le V_0 - V_k + \sum_{k=0}^{K} U_k \le V_0 + \sum_{k=0}^{\infty} U_k.$$

By letting $K \to \infty$, we have

$$\sum_{k=0}^{\infty} S_k < \infty.$$

Next, we define

$$\tilde{V}_k = V_k - \sum_{j=0}^{k-1} U_j.$$

Then

$$\tilde{V}_{k+1} \le \tilde{V}_k - S_k$$

and the nonincreasing sequence $\{\tilde{V}_k\}_{k\in\mathbb{N}}$ (lower bounded by $-\sum_{k=0}^{\infty} U_k$) has a limit. So V_k has a limit. □

Theorem 15. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be convex. Assume $\|\partial f(x)\| \le G$. Let α_k be a sequence of positive scalars such that*

$$\sum_k \alpha_k = \infty, \qquad \sum_k \alpha_k^2 < \infty$$

Then

$$\begin{aligned} g_k &\in \partial f(x_k) \\ x_{k+1} &= x_k - \alpha_k g_k \end{aligned}$$

converges in the sense of $x_k \to x_\infty \in \operatorname{argmin} f$.

Proof. Let $\tilde{x}_\star \in \operatorname{argmin} f$.

$$\begin{aligned} \|x_{k+1} - \tilde{x}_\star\|^2 &= \|x_k - \tilde{x}_\star - \alpha_k g_k\|^2 \\ &= \|x_k - \tilde{x}_\star\|^2 - 2\alpha_k \langle g_k, x_k - \tilde{x}_\star\rangle + \alpha_k^2 \|g_k\|^2 \\ &\le \|x_k - \tilde{x}_\star\|^2 - 2\alpha_k \big(f(x_k) - f_\star\big) + \alpha_k^2 G^2 \end{aligned}$$

By the quasi-summability lemma, $\|x_k - \tilde{x}_\star\|$ is bounded. we have

$$\sum_k^{\infty} \alpha_k \big(f(x_k) - f_\star\big) < \infty.$$

Since α_k is not summable this implies

$$\liminf_{k\to\infty} \big(f(x_k) - f_\star\big) = 0.$$

Pick a convergent subsequence x_{k_j} such that $f(x_{k_j}) \to f_\star$. Since $\{x_k\}_k$ is bounded, we can choose a further convergent subsequence to ensure $x_{k_j} \to x_\infty$. Since f is continuous, $f(x_\infty) = f_\star$ and thus $x_\infty \in \operatorname{argmin} f$. Finally, since $\|x_k - x_\infty\|^2$ has a limit, we know that the limit is 0, i.e., the entire sequence converges to x_∞. □

Lemma 7 (Jensen's inequality)**.** *Let $X \in \mathbb{R}^n$ be a random variable such that $\mathbb{E}[X] \in \mathbb{R}^n$ is well defined, and let $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ be convex. Then,*

$$\varphi\big(\mathbb{E}[X]\big) \le \mathbb{E}[\varphi(X)].$$

Proof. Let $g \in \partial\varphi(\mathbb{E}[X])$. Then,

$$\varphi(X) \ge \varphi(\mathbb{E}[X]) + \langle g, X - \mathbb{E}[X]\rangle.$$

Taking expectations on both sides completes the proof. □

Theorem 16. *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a G-Lipschitz continuous convex function. Assume f has a minimizer $x_\star$. Let $x_0 \in \mathbb{R}^d$ be a starting point and write $R = \|x_0 - x_\star\|_2$. Let $K > 0$ be the total iteration count. Then, subgradient descent with the constant stepsize*

$$\alpha_k = \alpha = \frac{R}{G\sqrt{K+1}}$$

exhibits the rate

$$\min_{0\le k\le K} f(x_k) - f(x_\star) \le \frac{GR}{\sqrt{K+1}}$$

and

$$f((\bar{x}_K) - f(x_\star) \le \frac{GR}{\sqrt{K+1}},$$

where

$$(\bar{x}_K = \frac{1}{K+1}\sum_{k=0}^{K} x_k.$$

Proof. for $k = 0, 1, 2, \dots,$

$$\begin{aligned}
\|x_{k+1} - x_\star\|_2^2 &= \|x_k - \alpha g_k - x_\star\|_2^2 \\
&= \|x_k - x_\star\|_2^2 - 2\alpha\langle g_k, x_k - x_\star\rangle + \alpha^2\|g_k\|_2^2 \\
&\le \|x_k - x_\star\|_2^2 - 2\alpha(f(x_k) - f(x_\star)) + \alpha^2 G^2.
\end{aligned}$$

Therefore,

$$2\alpha(f(x_k) - f(x_\star)) \le \|x_k - x_\star\|_2^2 - \|x_{k+1} - x_\star\|_2^2 + \alpha^2 G^2.$$

With a telescoping sum argument, we get

$$\begin{aligned}
2\alpha\sum_{k=0}^{K}(f(x_k) - f(x_\star)) &\le \|x_0 - x_\star\|_2^2 - \|x_{K+1} - x_\star\|_2^2 + \sum_{k=0}^{K}\alpha^2 G^2 \\
&\le R^2 + (K+1)\alpha^2 G^2,
\end{aligned}$$

and

$$\frac{1}{K+1}\sum_{k=0}^{K} f(x_k) - f(x_\star) \le \frac{R^2 + \alpha^2 G^2(K+1)}{2\alpha(K+1)} = \frac{GR}{\sqrt{K+1}}.$$

Therefore,

$$\begin{aligned}\min_{0\le k\le K} f(x_k)-f(x_\star) &= \frac{1}{K+1}\sum_{k=0}^{K}\min_{0\le k\le K} f(x_k)-f(x_\star)\\ &\le \frac{1}{K+1}\sum_{k=0}^{K} f(x_k)-f(x_\star) \le \frac{GR}{\sqrt{K+1}}.\end{aligned}$$

Likewise, using Jensen's inequality, we conclude

$$\begin{aligned}f(\bar{x}_K)-f(x_\star) &\le \mathop{\mathbb{E}}_{k\sim\mathrm{Uniform}(\{0,1,\dots,K\})}\big[f(x_k)-f(x_\star)\big]\\ &= \frac{1}{K+1}\sum_{k=0}^{K} f(x_k)-f(x_\star) \le \frac{GR}{\sqrt{K+1}}.\end{aligned}$$

□

Note that $\bar{x}_K$ can be computed with the following online averaging formula (without having to store all of the $x_0, x_1, \dots, x_K$):

$$\bar{x}_K = \frac{1}{K+1}x_K + \frac{K}{K+1}\bar{x}_{K-1}.$$

Theorem 17. *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a G-Lipschitz continuous convex function. Assume f has a minimizer $x_\star$. Let $x_0 \in \mathbb{R}^d$ be a starting point and write $R = \|x_0 - x_\star\|_2$. Then, subgradient descent with positive stepsizes $\{\alpha_k\}_{k=0}^K$ satisfies*

$$f(\bar{x}_K)-f(x_\star) \le \frac{R^2 + G^2\sum_{k=0}^K \alpha_k^2}{2\sum_{k=0}^K \alpha_k},$$

where

$$\bar{x}_K = \frac{\sum_{k=0}^K \alpha_k x_k}{\sum_{k=0}^K \alpha_k}.$$

Proof. For $k = 0, 1, \dots, K$,

$$\|x_{k+1}-x_\star\|_2^2 = \|x_k - \alpha_k g_k - x_\star\|_2^2 \le \|x_k - x_\star\|_2^2 - 2\alpha_k\langle g_k, x_k - x_\star\rangle + \alpha_k^2\|g_k\|_2^2.$$

By convexity, $\langle g_k, x_k - x_\star\rangle \ge f(x_k) - f(x_\star)$ for $g_k \in \partial f(x_k)$, and by G-Lipschitzness $\|g_k\|_2 \le G$. Hence

$$2\alpha_k\big(f(x_k)-f(x_\star)\big) \le \|x_k - x_\star\|_2^2 - \|x_{k+1}-x_\star\|_2^2 + \alpha_k^2 G^2.$$

Summing from $k=0$ to K gives

$$2\sum_{k=0}^K \alpha_k\big(f(x_k)-f(x_\star)\big) \le \|x_0-x_\star\|_2^2 - \|x_{K+1}-x_\star\|_2^2 + G^2\sum_{k=0}^K\alpha_k^2 \le R^2 + G^2\sum_{k=0}^K \alpha_k^2.$$

Let $A_K = \sum_{k=0}^{K} \alpha_k$. By convexity of f,

$$f(\bar{x}^K) \leq \frac{1}{A_K} \sum_{k=0}^{K} \alpha_k f(x_k), \qquad \bar{x}^K = \frac{1}{A_K} \sum_{k=0}^{K} \alpha_k x_k.$$

Therefore

$$\begin{aligned} 2A_K\big(f(\bar{x}^K) - f(x_\star)\big) &\leq \mathop{\mathbb{E}}_{k} \left[f(x_k) - f(x_\star)\right] \\ &= 2\sum_{k=0}^{K} \alpha_k \big(f(x_k) - f(x_\star)\big) \leq R^2 + G^2 \sum_{k=0}^{K} \alpha_k^2, \end{aligned}$$

with the distribution on the random index k defined as

$$\mathbb{P}(k = j) = \frac{\alpha_j}{\sum_{k=0}^{i} \alpha_i}.$$

This yields the claim. □

Corollary 1. *With stepsize* $\alpha_k = C/\sqrt{K}$*, we get the rate*

$$f(\bar{x}_k) - f(x_\star) \leq \mathcal{O}(\log k/\sqrt{k})$$