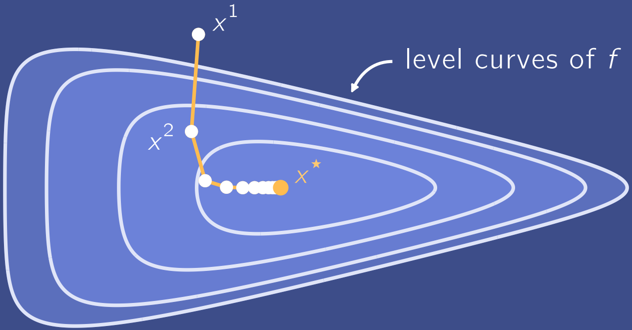
Gradient Descent

a classic algorithm seen through operators



Howard Heaton

Typal Academy

Motivation

"Turns out walking downhill is a powerful idea." – a convex philosopher

Problem Minimize a smooth and convex function f

Algorithm Given initial point $x^1 \in \mathbb{R}^n$ and step size $\alpha > 0$, update via

$$x^{k+1} = x^k - \alpha \nabla f(x^k)$$

Why it matters

Widespread Standard method across science and engineering

Fundamental Core part of many modern optimization algorithms

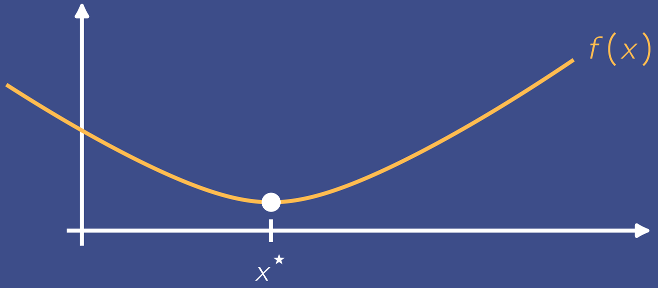
Problem Setting

Given

A convex and L-smooth function $f : \mathbb{R}^n \to \mathbb{R}$

Problem

Find the minimizer $x^\star$, *i.e.* solve $\min_{x \in \mathbb{R}^n} f(x)$



Definitions

Each operator is a function mapping $\mathbb{R}^n$ to $\mathbb{R}^n$.

An operator Q is nonexpansive if it is 1-Lipschitz, *i.e.*

$$\|Q(x) - Q(y)\| \leq \|x - y\|, \quad \text{for all } x \text{ and } y.$$

An operator T is averaged if there is $\alpha \in (0, 1)$ and nonexpansive Q such that

$$T(x) = (1 - \alpha)x + \alpha Q(x) = x + \alpha\big(Q(x) - x\big) \quad \text{for all } x \text{ and } y.$$

An operator F is L-contractive if there is $L \in [0, 1)$ such that

$$\|F(x) - F(y)\| \leq L\|x - y\|, \quad \text{for all } x \text{ and } y.$$

Optimality and Fixed Point Equivalence

Lemma 1 If f is convex and differentiable, $\alpha > 0$ and an operator T is defined by

$$T(x) = x - \alpha \nabla f(x),$$

then $x^\star$ is a minimizer of f if and only if $x^\star = T(x^\star)$.

Proof By the first-order optimality condition for f,

$$\begin{aligned}
x^\star \text{ minimizes } f \quad &\iff \quad 0 = \nabla f(x^\star) && \text{(optimality condition)} \\
&\iff \quad 0 = -\alpha \nabla f(x^\star) && \text{(multiply by } -\alpha) \\
&\iff \quad x^\star = T(x^\star). && \text{(add } x^\star \text{ and use definition of } T)
\end{aligned}$$

The result follows by transitivity of logical equivalences. ■

Cocercivity $\implies$ Nonexpansiveness

Lemma 2 If f is convex and L-smooth, then $Q(x) = x - \frac{2}{L}\nabla f(x)$ is nonexpansive.

Proof By the Baillon-Haddad theorem,[†] ∇f is $\frac{1}{L}$-cocoercive. This implies

$$\begin{aligned}
\|Q(x) - Q(y)\|^2 &= \|x-y\|^2 - \frac{4}{L}(x-y)^\top\big(\nabla f(x) - \nabla f(y)\big) + \frac{4}{L^2}\|\nabla f(x) - \nabla f(y)\|^2 \\
&= \|x-y\|^2 - \frac{4}{L}\underbrace{\left[(x-y)^\top\big(\nabla f(x) - \nabla f(y)\big) - \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2\right]}_{\geq 0} \\
&\leq \|x-y\|^2,
\end{aligned}$$

for all x and y. Taking square roots yields $\|Q(x) - Q(y)\| \leq \|x - y\|$. ■

[†] See prior lecture about this theorem for its details and proof.

Gradient Descent Convergence

Theorem 1 If $\alpha \in (0, 2/L)$ and f is convex, L-smooth and has a minimizer, then $T(x) = x - \alpha \nabla f(x)$ is averaged and has a fixed point, which is a minimizer of f.

The following corollary applies the result of Krasnosel'skiĭ and Mann.

Corollary 1 The gradient descent iteration $x^{k+1} = T(x^k)$ generates a sequence $\{x^k\}$ converging to a minimizer of f with $\|\nabla f(x^k)\|^2 = \mathcal{O}(1/k)$.

Gradient Descent Proofs

Theorem 1 Proof Set $\theta = \alpha L/2$ so that $\theta \in (0, 1)$ and, by Lemma 2,

$$T(x) = x - \frac{2\theta}{L}\nabla f(x) = \theta x + (1-\theta)Q(x) \quad \Longrightarrow \quad T \text{ is averaged.}$$

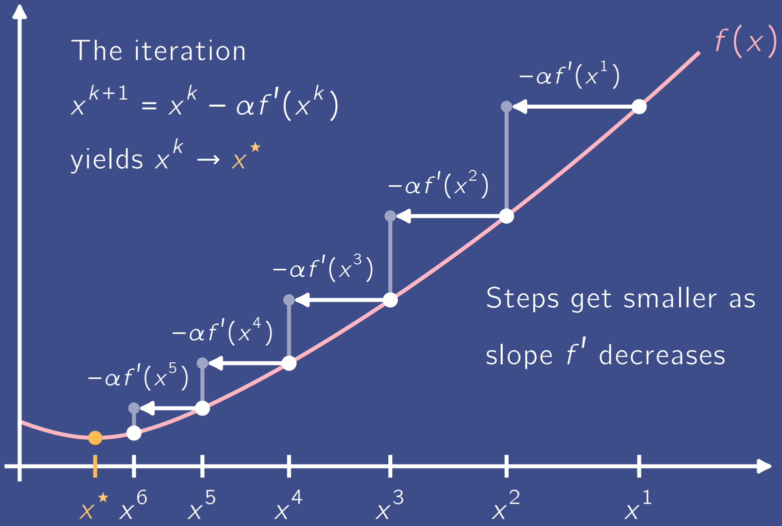
Since f has a minimizer $x^\star$, Lemma 1 asserts $x^\star$ is a fixed point of T. ■

Corollary 1 Proof Because T is averaged with a fixed point, the Krasnosel'skiĭ-Mann iteration[†] $x^{k+1} = T(x^k)$ converges to a fixed point, *i.e.* minimizer of f. Moreover,

$$\alpha^2 \|\nabla f(x^k)\|^2 = \|x^{k+1} - x^k\|^2 = \mathcal{O}\left(\frac{1}{k}\right).$$ ■

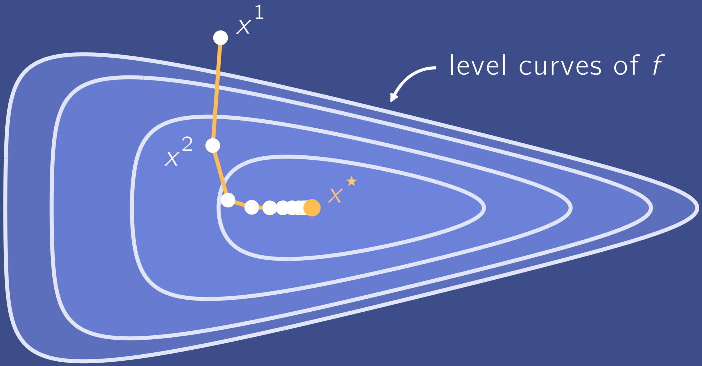
[†] See prior lecture on Krasnosel'skiĭ-Mann iteration for convergence result and proof.

Gradient Descent in 1D



Gradient Descent in 2D

The iteration $x^{k+1} = x^k - \alpha \nabla f(x^k)$ ensures $\{x^k\}$ converges to the minimizer $x^\star$

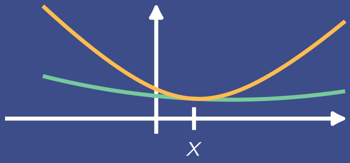


$$f(x^\star) = \min_x f(x) = \min_x \left\{ \ln\left(1 + e^{x_1 + 2x_2}\right) + \ln\left(1 + e^{x_1 - 2x_2}\right) + \ln\left(1 + e^{-x_1}\right) \right\}$$

Strong Convexity

Recall a function f is μ-strongly convex provided

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x) + \frac{\mu}{2}\|y - x\|^2, \quad \text{for all } x \text{ and } y.$$



Baillon and Haddad's result may be strengthened when f is strongly convex.

Lemma 3 If f is μ-strongly convex and L-smooth, then

$$(\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \frac{1}{L + \mu}\left[\|\nabla f(x) - \nabla f(y)\|^2 + L\mu\|x - y\|^2\right],$$

for all x and y.

Smooth + Strongly Convex $\implies$ Contractive

Theorem 2 If f is L-smooth and μ-strongly convex and $\alpha \in (0, 2/L)$, then the operator $T(x) = x - \alpha \nabla f(x)$ is θ-contractive with $\theta = \max\{|1 - \alpha\mu|, |1 - \alpha L|\}$.

The following corollary is a direct application of Banach's fixed point theorem.

Corollary 2 For each $x^1 \in \mathbb{R}^n$, the iteration $x^{k+1} = T(x^k)$ generates a sequence $\{x^k\}$ converging to the unique minimizer of f with $\|\nabla f(x^k)\| = \mathcal{O}(\theta^k)$.

→ Gradient descent converges linearly for smooth and strongly convex f.

Contractive Gradient Descent Operator Proof

Let $x, y \in \mathbb{R}^n$ and $\alpha \in (0, 2/L)$ be given. It suffices to show

$$\|T(x) - T(y)\| \leq \max\{|1 - \alpha\mu|,\ |1 - \alpha L|\}\|x - y\|.$$

Squaring and expanding the left hand side yields

$$\begin{aligned}\|T(x) - T(y)\|^2 = \|x - y\|^2 &+ \alpha^2\|\nabla f(x) - \nabla f(y)\|^2 \\ &- 2\alpha(\nabla f(x) - \nabla f(y)) \cdot (x - y).\end{aligned}$$

Lemma 3 may be applied to deduce

$$\|T(x) - T(y)\|^2 \leq \left(1 - \frac{2\alpha L\mu}{L + \mu}\right)\|x - y\|^2 + \left(\alpha^2 - \frac{2\alpha}{L + \mu}\right)\|\nabla f(x) - \nabla f(y)\|^2.$$

Contractive Gradient Descent Operator Proof

By the μ-strong convexity and L-smoothness of f,

$$-\|\nabla f(x) - \nabla f(y)\| \le -\mu\|x - y\| \quad \text{and} \quad \|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|.$$

Thus, if $0 < \alpha \le \frac{2}{L+\mu}$, then[†]

$$\left(\alpha^2 - \frac{2\alpha}{L+\mu}\right)\|\nabla f(x) - \nabla f(y)\|^2 \le \mu^2\left(\alpha^2 - \frac{2\alpha}{L+\mu}\right)\|x - y\|^2.$$

Similarly, if $\frac{2}{L+\mu} < \alpha < \frac{2}{L}$, then

$$\left(\alpha^2 - \frac{2\alpha}{L+\mu}\right)\|\nabla f(x) - \nabla f(y)\|^2 \le L^2\left(\alpha^2 - \frac{2\alpha}{L+\mu}\right)\|x - y\|^2.$$

Note the expression is negative when $0 < \alpha \le \frac{2}{L+\mu}$.

Contractive Gradient Descent Operator Proof

Combining the bounds and simplifying reveals

$$\|T(x) - T(y)\|^2 \leq \begin{cases} (1-\alpha\mu)^2\|x-y\|^2 & \text{if } 0 \leq \alpha \leq \frac{2}{L+\mu} \\ (1-\alpha L)^2\|x-y\|^2 & \text{if } \frac{2}{L+\mu} < \alpha \leq \frac{2}{L}. \end{cases}$$

Since $L \geq \mu$ and $\alpha > 0$,

$$\alpha \leq \frac{2}{L+\mu} \quad \Longleftrightarrow \quad (1-\alpha L)^2 \leq (1-\alpha\mu)^2.$$

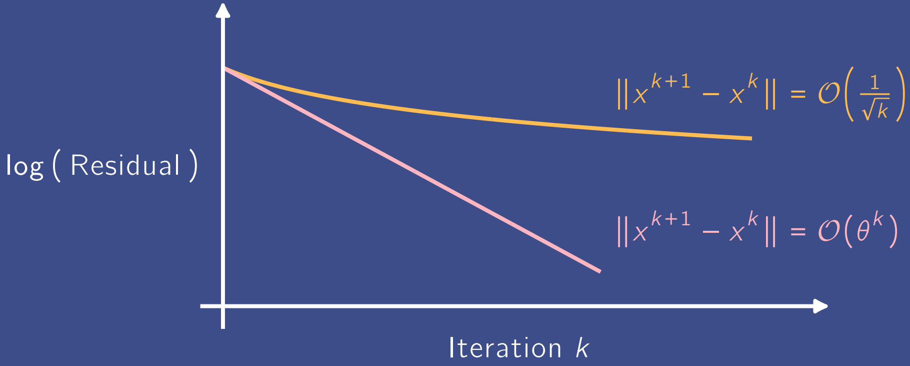
This implies the bound on $\|T(x) - T(y)\|^2$ in each case for α is the maximum of $(1-\alpha\mu)^2$ and $(1-\alpha L)^2$. Thus, these cases may combined to obtain

$$\|T(x) - T(y)\|^2 \leq \max\left\{(1-\alpha\mu)^2,\ (1-\alpha L)^2\right\}\|x-y\|^2.$$

Taking square roots yields the result. ■

Convergence Rate Plots

For L-smooth and convex or L smooth and μ-strongly convex f, we can bound the rate of convergence of the residual $\|x^{k+1} - x^k\| = \alpha\|\nabla f(x^k)\|$ to zero.



Takeaways

- Gradient descent is a fixed point iteration for $T(x) = x - \alpha \nabla f(x)$
- When f is convex and L-smooth, T is averaged
- When f is μ-strongly convex and L-smooth, T is contractive
- When T is averaged or contractive, fixed-point convergence results apply

References

- Baillon, Haddad. *Quelques proprieétés des opérateurs angle-bornés et n-cycliquement monotones*, 1977.
- Beck. *First-Order Methods in Optimization*, 2017.
- Bishop. *Prove that gradient descent is a contraction for strongly convex smooth functions (StackExchange post).* 2021.
- Vandenberghe. *Gradient method (lecture slides).* 2022.

Appendix – Baillon-Haddad-type Inequality

Lemma 3 If f is μ-strongly convex and L-smooth, then

$$(\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \frac{1}{L + \mu}\left[\| \nabla f(x) - \nabla f(y) \|^2 + L\mu \|x - y\|^2 \right],$$

for all x and y.

Proof We proceed in two steps.

1. If f is μ-strongly convex and L-smooth, then $g(x) = f(x) - \mu \|x\|^2/2$ is convex and $(L - \mu)$-smooth.
2. If g is convex and L-smooth, then the inequality holds by the result of Baillon and Haddad.

Step 1: Perturbed function is convex and $(L-\mu)$-smooth

Since f is μ-strongly convex, g is convex. Let $x, y \in \mathbb{R}^n$ be given. Setting $\Delta = \nabla f(x) - \nabla f(y)$ and $\delta = x - y$, observe $\nabla g(x) = \nabla f(x) - \mu x$ and

$$
\begin{aligned}
\|\nabla g(x) - \nabla g(y)\|^2 &= \|\Delta\|^2 - 2\mu\langle\Delta, \delta\rangle + \mu^2\|\delta\|^2 && \text{(expand terms)} \\
&\leq (L - 2\mu)\langle\Delta, \delta\rangle + \mu^2\|\delta\|^2 && \text{(Baillon-Haddad)} \\
&\leq (L - 2\mu)\|\Delta\|\|\delta\| + \mu^2\|\delta\|^2 && \text{(Cauchy-Schwarz)} \\
&\leq (L - 2\mu)L\|\delta\|^2 + \mu^2\|\delta\|^2 && \text{(L-smoothness of f)} \\
&= (L-\mu)^2\|\delta\|^2. && \text{(simplify)}
\end{aligned}
$$

Taking square roots yields

$$
\|\nabla g(x) - \nabla g(y)\| \leq (L-\mu)\|\delta\| = (L-\mu)\|x-y\|.
$$

Step 2: Apply Baillon-Haddad Theorem

If $\mu = L$, then f is quadratic, *i.e.* there is $c \in \mathbb{R}^n$ such that

$$f(x) = \frac{\mu}{2}\|x\|^2 + c \cdot x.$$

In this case, the result follows upon direct substitution. Now suppose $L > \mu$. As g is convex and $(L-\mu)$-smooth, Baillon and Haddad's theorem asserts

$$(\nabla g(x) - \nabla g(y)) \cdot (x - y) \geq \frac{1}{L-\mu}\|\nabla g(x) - \nabla g(y)\|^2.$$

Adding $\mu\|x-y\|^2 + 2\mu(\nabla f(x) - \nabla f(y)) \cdot (x-y)/(L-\mu)$ to each side yields

$$\frac{L+\mu}{L-\mu}\Big(\nabla f(x) - \nabla f(y)\Big) \cdot (x-y) \geq \frac{1}{L-\mu}\left[\|\nabla f(x) - \nabla f(y)\|^2 + L\mu\|x-y\|^2\right]$$

Multiplying by $(L-\mu)/(L+\mu)$ gives the desired result. ■