Next week: only zoom!

Optimization for Machine Learning

Lecture 2: Gradient Descent

Sebastian Stich

CISPA – https://cms.cispa.saarland/optml24/

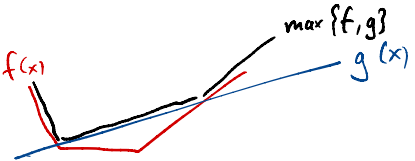
April 23, 2024

Quiz Week 1


Let $f\colon \mathbb{R} \to \mathbb{R}$ and $g\colon \mathbb{R} \to \mathbb{R}$ be two convex functions. Which of the following combinations of f and g are convex:

1. $f(\mathbf{x}) + g(\mathbf{x})$ ✓
2. $f(\mathbf{x}) \cdot g(\mathbf{x})$
3. $\max\{f(\mathbf{x}), g(\mathbf{x})\}$ ✓
4. $\min\{f(\mathbf{x}), g(\mathbf{x})\}$
5. $f(g(\mathbf{x}))$
6. $e^{f(\mathbf{x})}$ ✓

$f(x) = -x$, $g(x) = x$, $f(x) \cdot g(x) = -x^2$



$f(x) = -x$, $g(x) = x^2$, $f(g(x)) = -x^2$

$\Rightarrow$ same picture

$\frac{d^2}{dx} e^{f(x)}$ $\quad \frac{d}{dx} e^{f(x)} = e^{f(x)} \cdot f'(x)$ $\quad$ etc ...

Chapter 3

Gradient Descent

The Algorithm

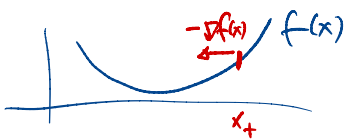
$\min_{x \in \mathbb{R}^d} f(x)$

Given: Objective function $f : \mathbb{R}^d \to \mathbb{R}$.

Iterative Algorithm: choose $\mathbf{x}_0 \in \mathbb{R}^d$.

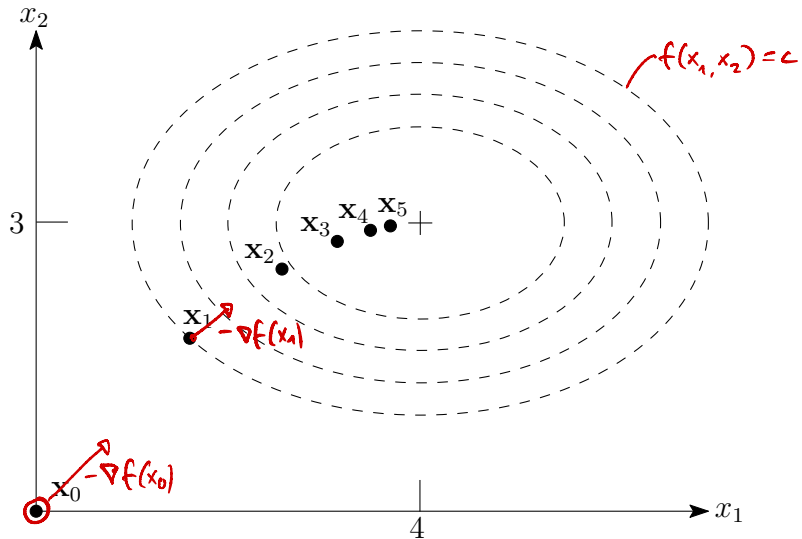
$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t),$$

for **timesteps** $t = 0, 1, \ldots,$ and **stepsize** $\gamma \geq 0$.



"function value gets smaller"

Example



$$f(x_1, x_2) := 2(x_1 - 4)^2 + 3(x_2 - 3)^2, \mathbf{x}_0 := (0, 0), \gamma := 0.1$$

What does it mean to 'solve' an optimization problem?

We need to define approximate solutions:

- With respect to $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$: $\Rightarrow \|x_+ - x^*\| \leq \varepsilon$

- With respect to $\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial f}{\partial x_d}(\mathbf{x})\right)$: $\|\nabla f(x_+)\| \leq \varepsilon$

How difficult is it to solve an optimization problem?

- Example 1:

- Example 2:

Summary: ⇒ "it depends"

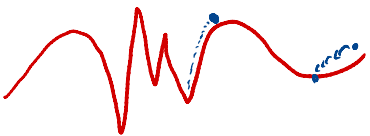
Example: Lipschitz functions

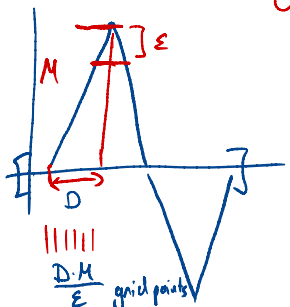
A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is M-Lipschitz, if

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq M \|\mathbf{x} - \mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

Problem: $\quad \text{minimize} \quad f(\mathbf{x}) \quad \text{with } \mathbf{x} \in [0,1]^d$

A strategy to solve this problem:





Option 0: sampling $\geq \left(\frac{M}{\varepsilon}\right)^d$ times

Option 1: Grid with $\left(\frac{M}{\varepsilon}\right)^d$ query points

$\Rightarrow$ this is optimal!

(so this is a hard problem in general)

Performance of Numerical Methods

- Given a problem class $\mathcal{P}$ (Example: M-Lipschitz)
  - (and the definition of an approximate solution) (Example: $\|\nabla f\| < \varepsilon$)
- and a method $\mathcal{M}$ → (Example: gradient descent)
  - with oracle access to the problem instance $p \in \mathcal{P}$ → $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$, etc.
- the performance of $\mathcal{M}$ on $\mathcal{P}$ is the amount of computational effort required to solve $\mathcal{P}$.

(Want to know: how difficult it is to "solve ~~every~~ problem in $\mathcal{P}$" = the hardest)

Computational effort can be measured as:

- analytic complexity (oracle calls)
- arithmetic complexity (additions, multiplications)

Gradient Descent on Smooth Functions

Smooth functions

"Not too curved"

Definition (Lecture-2).1

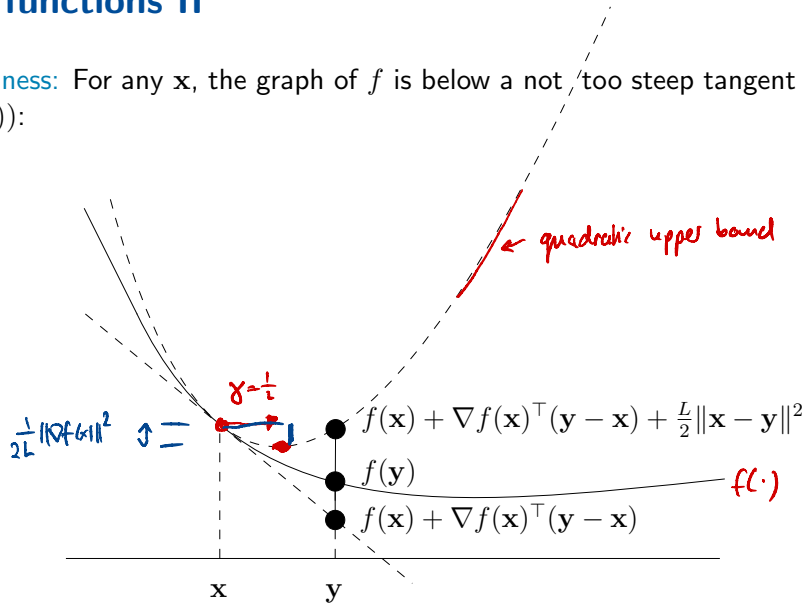
Let f: $\mathbf{dom}(f) \to \mathbb{R}$ be differentiable, $X \subseteq \mathbf{dom}(f)$, $L \in \mathbb{R}_+$. f is called smooth (with parameter L) over X if

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

f smooth $:\Leftrightarrow$ f smooth over $\mathbb{R}^d$.

Smooth functions II

Smoothness: For any $\mathbf{x}$, the graph of f is below a not too steep tangent paraboloid at $(\mathbf{x}, f(\mathbf{x}))$:



Smooth functions III

- In general: quadratic functions are smooth (**Exercise 19**).
- Operations that preserve smoothness (the same that preserve convexity):

Lemma (Lecture-2).2 (Exercise 22)

(i) *Let $f_1, f_2, \ldots, f_m$ be functions that are smooth with parameters $L_1, L_2, \ldots, L_m$, and let $\lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{R}_+$. Then the function $f := \sum_{i=1}^m \lambda_i f_i$ is smooth with parameter $\sum_{i=1}^m \lambda_i L_i$.*

(ii) *Let f be smooth with parameter L, and let $g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, for $A \in \mathbb{R}^{d \times m}$ and $\mathbf{b} \in \mathbb{R}^d$. Then the function $f \circ g$ is smooth with parameter $L\|A\|^2$, where is $\|A\|$ is the* **spectral norm** *of A (Definition 2.2).*

Smooth: Summary

$f(x) = \frac{1}{2} 10^{10} x^2$ ✓

$f(x) = \frac{1}{2} x^4$ ✗

- Lipschitz continuity of ∇f

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

- Quadratic upper bound:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$$

- For twice differentiable functions:

$$\left\|\nabla^2 f(\mathbf{x})\right\| \leq L$$

Sufficient decrease

Lemma (Lecture-2).3

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and smooth with parameter L. With stepsize

$$\gamma := \frac{1}{L},$$

value of the stepsize

gradient descent satisfies

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2, \quad t \geq 0.$$

decrease! "sufficient decrease"

Remark (Lecture-2).4

More specifically, this already holds if f is smooth with parameter L over the line segment connecting $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$.

Sufficient decrease II

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

Proof.

Use smoothness and definition of gradient descent $(\mathbf{x}_{t+1} - \mathbf{x}_t = -\nabla f(\mathbf{x}_t)/L)$:

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2. \end{aligned}$$

□

Smooth functions: $\mathcal{O}(1/\varepsilon)$ steps

Theorem (Lecture-2).5

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable smooth with parameter L and suppose $f^\star \leq \min f(\mathbf{x})$. With the stepsize

$$\gamma := \frac{1}{L},$$

gradient descent yields

$$\min_{t \in \{0,\ldots,T-1\}} \|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2L(f(\mathbf{x}_0) - f^\star)}{T}, \quad T > 0.$$

Proof

Consider the sufficient decrease condition:

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

Equivalently:

$$\|\nabla f(\mathbf{x}_t)\|^2 \leq 2L(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})) .$$

By summing these equations over $t = 0, \ldots, T-1$, and dividing by T:

$$\min_{t \in \{0,\ldots,T\}} \|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2L}{T}(f(\mathbf{x}_0) - f(\mathbf{x}_T)) \leq \frac{2L}{T}(f(\mathbf{x}_0) - f^\star) .$$

"telescopic sum"

Discussion

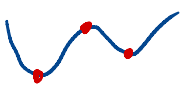
- $\min_{t \in \{0,\dots,T-1\}} \|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2L(f(\mathbf{x}_0) - f^\star)}{T} \quad \Leftrightarrow \quad T \in \mathcal{O}\left(\frac{L(f(\mathbf{x}_0) - f^\star)}{\epsilon}\right)$

  "convergence rate" — "how many steps" / "complexity estimate"

- $\min_{t \in \{0,\dots,T-1\}} \|\nabla f(\mathbf{x}_t)\|^2$ vs. $\|\nabla f(\mathbf{x}_T)\|^2$

  ↑ practice!

- $\min_{t \in \{0,\dots,T-1\}} \|\nabla f(\mathbf{x}_t)\|^2 \to 0$ does not imply convergence to a global (or local!) minima!



Gradient Descent on Smooth Convex Functions

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

Theorem (Lecture-2).6

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that f is smooth with parameter L. Choosing stepsize

$$\gamma := \frac{1}{L},$$

gradient descent yields

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

last iterate

Proof I

Consider $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2$ and $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$.

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 &= \left\|\mathbf{x}_t - \mathbf{x}^\star - \frac{1}{L}\nabla f(\mathbf{x}_t)\right\|^2 \\ &= \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \frac{2}{L}\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star)) + \frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2\end{aligned}$$

From the first-order characterization of convexity ($f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})$):

$$\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star) \geq f(\mathbf{x}_t) - f(\mathbf{x}^\star)$$

And from the sufficient decrease lemma ($f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$):

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

Proof II

Putting everything together:

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{L}{2}\left(\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right) + f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})$$

By summing up over $t = 0, \dots T$

$$\sum_{t=0}^{T} f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 - \frac{L}{2}\|\mathbf{x}_T - \mathbf{x}^\star\|^2 + f(\mathbf{x}_0) - f(\mathbf{x}_T)$$

≥ 0 $\quad$ $\geq f^\star$

Using $f(\mathbf{x}_T) \geq f(\mathbf{x}^\star)$ and rewriting:

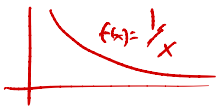
$$f(\mathbf{x}_T) - f^\star \leq \frac{1}{T}\left(\sum_{t=1}^{T} f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$$

Where we also used that the last iterate is the best (sufficient decrease)!

Discussion

- Can we also prove convergence $\|\mathbf{x}_t - \mathbf{x}^\star\|^2 \to 0$?

many $x^* \in \operatorname{argmin} f(x)$



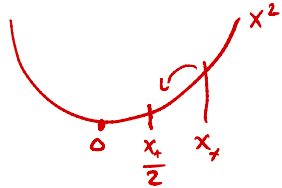
- We used the stepsize $\gamma = \frac{1}{L}$. What can we do when we do not know L?

"stepsize fuuig"

(see also **Exercise 23**)

- What is the benefit of Theorem (Lecture-2).6, if we already knew from Theorem (Lecture-2).5 that the gradient norm converges?

Can Gradient Descent Converge faster?



- Consider $f(x) := x^2$. Stepsize $\gamma := \frac{1}{4}$

$$x_{t+1} = x_t - \frac{1}{4}\nabla f(x_t) = x_t - \frac{x_t}{2} = \frac{x_t}{2},$$

so $f(x_t) = f\left(\frac{x_0}{2^t}\right) = \frac{1}{2^{2t}}x_0^2$.

- Exponential in t !

Note that f is smooth and strongly convex **(see Exercise sheet 2)!**

Lecture 2 Recap

- We have seen two convergence criteria: suboptimality gap and distance to the optimum.
- We have seen a key proof technique: telescoping.
- We have seen (template) convergence proofs for gradient descent on smooth functions, and on convex functions.

Discussion

Given: data $A = \begin{bmatrix} | & | & | & | & | & | \end{bmatrix}$

1) how does f depend on A?
2) how is y related to variance?

$$f(x) = \|Ax - b\|^2$$

$$\gamma = \frac{1}{\text{"smoothness"}}$$

Discussion

Discussion