

Optimization for Data Science

Instructor

Mateo Díaz (mateodd@uchicago.edu)

Office: Suzuki Center, 304.

Office Hours: M 4:40 - 5:40 pm.

TA: ??

Resources

- Website: mateodd25.github.io/opt-ds
- Canvas
- Gradescope
- Piazza

Grading scheme

Your grade will be determined by four components:

- Homework (Approximately 3)
- Midterm (In-class)
- Final project (Groups 3-6 people, document, talk)
- Participation.

Let G_H , G_M , G_F , and G_P be the grades for each component (normalized to be in $[0, 1]$). Then, your final grade is the solution to the following problem.

$$\max_{(H, M, F, P)} G_H \cdot H + G_M \cdot M + G_F \cdot F + G_P \cdot P$$

$$\text{s.t.} \quad H + M + F + P = 100$$

$$10 \leq H \leq 25,$$

$$30 \leq M,$$

$$M \leq F,$$

$$M + F \leq 80,$$

$$0 \leq P \leq 10.$$

Textbook

No textbook, the syllabus includes suggested books.

AI Policy

- Midterm

No AI

- Homework

Use it as a TA, write solutions by yourself (oral defense).

- Final project Use AI as you wish.

Why taking this class?

Traditional optimization theory deals with
 $\min_{x \in C} f(x)$.

Historically the objectives have changed.

▷ (1940 - 1960) Linear Programming becomes tractable thanks to Dantzig's simplex method and the introduction of computers.

▷ (1960 - 2000) Convex optimization becomes more prevalent thanks to work on gradient-based methods, interior point methods.

▷ (2000 - 2010) Well-matured software led to the wide adoption of these methods in applications (compressed sensing, matrix completion, control).

▷ (2010 -) Deep learning illustrated that nonconvex optimization is extremely practical. Why does it work?

convex? $\times$ exact gradients/Hessians? $\times$
smooth? $\times$

This is what we will try to understand. Unlike convex optimization, this is not as well understood. This is a moving target.

A typical statistical learning problem. We will often be interested in a pair of objectives of the form

$f(w) = \mathbb{E}_{z \sim D} \ell(w, z)$ and $f_n(w) = \frac{1}{n} \sum_{i=1}^n \ell(w, z_i)$

Population objective Empirical objective $z_i \sim D$

Our true goal might be to solve $\inf_w f(w)$, but we don't know about D , we can merely sample from it to obtain $z_1, \dots, z_n \stackrel{iid}{\sim} D$.

A popular approach is to instead

solve empirical risk minimization (ERM)
 $\inf_w f_n(w)$.

Example (linear least squares).

Suppose I want to estimate $w^* \in \mathbb{R}^d$ from measurements of the form

$$y = \underbrace{\langle x, w^* \rangle}_{\sim N(0, I)} + \underbrace{\epsilon}_{\sim N(0, \sigma^2 I)}.$$

A standard objective for this is

$$f(w) = \frac{1}{2} \mathbb{E}_{x,y} (\langle x, w \rangle - y)^2 \text{ and } f_n(w) = \frac{1}{2n} \sum_{i=1}^n (\langle x_i, w \rangle - y_i)^2.$$

Everything is ^{strongly} convex, smooth and nice when $n \gg d$.
↓

Example (LASSO).

Suppose now that we are told that w^* is sparse, i.e., $\#\{i \mid w_i^* \neq 0\} \ll d$. Further, we don't have many samples. Then a natural objective (Tibshirani '96) is

$$f_n(w) = \frac{1}{2n} \sum_{i=1}^n (\langle x_i, w \rangle - y_i)^2 + \lambda \|w\|_1$$

$\sum |w_i|$ promotes sparsity.

The ℓ_1 regularizer makes the objective non-

smooth, and when $d > n$, we no longer have strong convexity. $\rightarrow$

Example (Statistical learning)

Finally suppose that we want to learn a function $F: \mathbb{R}^m \rightarrow \mathbb{R}^k$ using samples

$$y = \underset{\substack{\uparrow \\ \text{unknown}}}{F}(x) + \underset{\substack{\uparrow \\ N(0, \sigma^2 I)}}{\varepsilon}.$$

Since we don't know how F looks like we decide to use a neural network to parameterized $w \in \mathbb{R}^d$ to approximate it:

$$f_n(w) = \frac{1}{2n} \sum_{i=1}^n (H(x_i; w) - y_i)^2.$$

$\uparrow$
NN

Unlike before $f_n(\cdot)$ might be highly nonconvex, nonsmooth, and $H(x; \cdot)$ might be overparameterized. Yikes! $\rightarrow$

What's a reasonable goal?

Before we even start to design algorithms we should step back and think about this question. There are multiple natural options.

▷ small objective gap

Maybe we want to find $\hat{w}$ such that

$$f(\hat{w}) - f(w^*) \leq \delta.$$

Let's reason about finite samples

$$\begin{aligned} f(\hat{w}) - f(w^*) &\leq [f(\hat{w}) - f_n(\hat{w})] + [f_n(\hat{w}) - f_n(w^*)] \\ &\quad + [f_n(w^*) - f(w^*)] \\ &\leq \underbrace{f_n(\hat{w}) - \inf_w f_n(w)}_{\text{optimization error}} \\ &\quad + 2 \underbrace{\sup_w |f_n(w) - f(w)|}_{\text{generalization / statistical error}}. \end{aligned}$$

▷ Distance to a minimizer

Let $w^* \in R(w)$, maybe we would like to find a point $\hat{w}$ s.t.

$$\|\hat{w} - w^*\| < \delta \leftarrow \text{For some small } \delta > 0.$$

Again, we don't even get to see $f(\cdot)$ so there might be limit to how small can δ be with n samples.

Warning: There might be many minimizers in overparameterized objective, which one is reasonable?

▷ Critical point

A more reasonable goal is to achieve what most local algorithms can guarantee

↪ $\|\nabla f_n(\hat{w})\| \leq \delta. \leftarrow \delta\text{-critical.}$

Or some other notion of nonsmooth criticality.

It ain't much, but it is often the only thing we can ensure.

Topics

We will cover a few broad topics.

▷ Smooth optimization

- ↳ Foundations
- ↳ PL inequality & overparameterization
- ↳ Landscape analysis

▷ Stochastic optimization

- ↳ SGD
- ↳ Variance Reduction

↪ Adaptativity, preconditioning, momentum.

▷ Nonsmooth optimization

- ↳ Weak convexity
- ↳ Proximal methods

▷ Minimax problems (Time permitting)

Smooth optimization

Let us start simple and assume, f is a smooth function, with L -Lipschitz gradients

$$\|\nabla f(w_1) - \nabla f(w_2)\| \leq L \|w_1 - w_2\|.$$

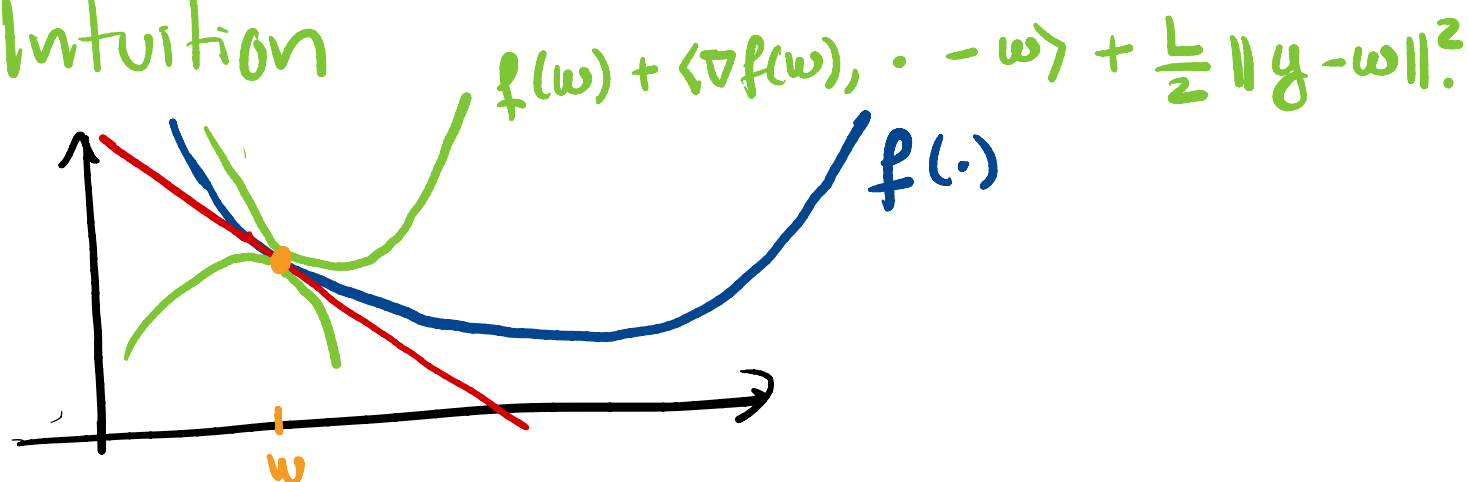
Let's think of f as a black box (we don't know its exact structure). We can merely query $w \mapsto (f(w), \nabla f(w))$. Our goal is to find a δ -critical point.

Theorem (Taylor): Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be an L -smooth function, then

$$|f(y) - f(w) - \langle \nabla f(w), y - w \rangle| \leq \frac{L}{2} \|y - w\|^2 \rightarrow$$

Armed with this result we can derive an incredibly useful lemma.

Intuition



Descent lemma: Let f be L -smooth

and $\eta \in (0, 2/L)$. Then

$$f(w - \eta \nabla f(w)) \leq f(w) - \underbrace{\eta \left(1 - \frac{L\eta}{2}\right)}_{\frac{1}{2L} \text{ when } \eta = \frac{1}{L}} \|\nabla f(w)\|^2.$$

Proof: This follows from Taylor's with $y = w - \eta \nabla f(w)$, since

$$\begin{aligned} f(y) - f(w) &+ \eta \|\nabla f(w)\|^2 \\ &= f(y) - f(w) + \langle \nabla f(w), y - w \rangle \\ &\leq \frac{L}{2} \|y - w\|^2 \\ &= \frac{L\eta^2}{2} \|\nabla f(w)\|^2. \end{aligned}$$

□

Theorem: Suppose that f is L -smooth and $\inf f > -\infty$. Consider the iterates

$$(\odot) \quad w_{k+1} \leftarrow w_k - \frac{1}{L} \nabla f(w_k).$$

Then,

$$\min_{\substack{k \in [T-1] \\ k=0,1,\dots,T-1}} \|\nabla f(w_k)\| \leq \sqrt{\frac{2L(f(w_0) - \inf f)}{T}}.$$

Proof: Applying the descent lemma to each index and adding all inequalities

ties up yields

$$\begin{aligned}\frac{1}{2L} \sum_{k=0}^{T-1} \|\nabla f(w_k)\| &\leq \sum_{k=0}^{T-1} f(w_k) - f(w_{k+1}) \\ &= f(w_0) - f(w_T) \\ &\leq f(w_0) - \inf f.\end{aligned}$$

Using that the minimum is smaller than the average gives

$$\begin{aligned}\min_{k \in [T-1]} \|\nabla f(w_k)\|^2 &\leq \frac{1}{2L} \sum_{k=0}^{T-1} \|\nabla f(w_k)\|^2 \\ &\leq f(w_0) - \inf f.\end{aligned}$$

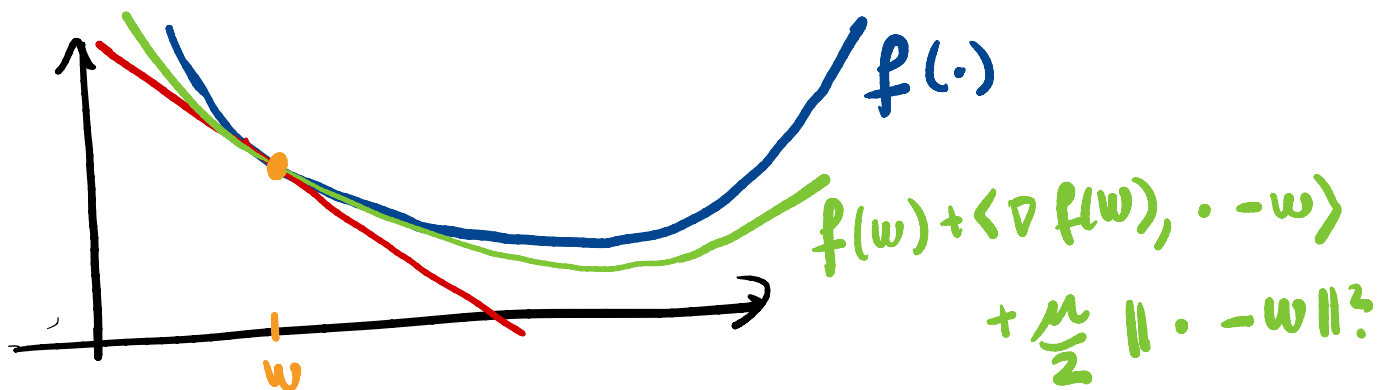
Completing the proof. $\square$

A natural next question is when is it that we can guarantee something about the distance to a solution. For that we need additional regularity assumptions.

Definition: We say that function f is μ -strongly convex if for all $w, y \in \mathbb{R}^d$,

$$f(y) \geq f(w) + \langle \nabla f(w), y - w \rangle + \frac{\mu}{2} \|y - w\|^2.$$

Intuition



There are very strong consequences of this assumption.

Proposition (Quadratic growth): Suppose that f is μ -strongly convex. Then, f has a unique minimizer w^* , and for all $w \in \mathbb{R}^d$

$$\frac{\mu}{2} \|w - w^*\|^2 \leq f(w) - f(w^*).$$

+

Theorem: Suppose that f is L -smooth and μ -strongly convex. Then, the iterates of (:) satisfy

$$f(w_k) - f(w^*) \leq \left(1 - \frac{\mu}{L}\right)^k (f(w_0) - f(w^*)).$$

+

A simple computation yields that if we want to reach

$$f(w_T) - f(w^*) \leq \delta,$$

then it suffices to take

(why?) $T \geq \frac{L}{\mu} \log \left(\frac{f(x_0) - f(w^*)}{\delta} \right).$

Further quadratic growth ensures

$$\|w_T - w^*\|^2 \leq 2\delta/\mu.$$