Optimization for Machine Learning

Lecture 4: SGD & Coordinate Descent

Sebastian Stich

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

May 7, 2024

Quiz Week 4

Which of these statements are true?

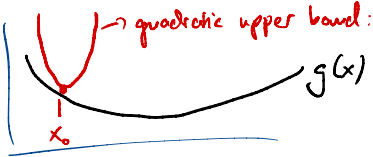
1. Let f be a convex function. Then $f(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}\|^2$ is μ-strongly convex. $\mu > 0$
2. Let f be a convex function. Then $f(\mathbf{x}) + \frac{L}{2}\|\mathbf{x}\|^2$ is L-smooth. $\nabla^2(\quad) = \nabla^2 f(x) + L \cdot 1 \overset{?}{\leq} L \cdot 1$
3. Let f be a L-smooth function. Then $f(\mathbf{x}) + \frac{M}{2}\|\mathbf{x}\|^2$ is $(L+M)$-smooth. ✓
4. Let f be a L-smooth function. Then $f(\mathbf{x}) - \frac{M}{2}\|\mathbf{x}\|^2$ is $(L-M)$-smooth. ✗
5. Let f be a μ-strongly convex function. Then $f(\mathbf{x}) - \frac{\nu}{2}\|\mathbf{x}\|^2$ is $(\mu-\nu)$-strongly convex. μ? ν? ≥ 0

Quiz Week 4

1. if $\nabla^2 f(x)$ exists

$$\nabla^2\left(f(x) + \frac{\mu}{2}\|x\|^2\right) = \underbrace{\nabla^2 f(x)}_{\geq 0} + \underbrace{\mu \cdot I}_{>0}$$

2. • for simplicity assume $\nabla^2 f(x)$ exists



- g ... L-smooth if $\|\nabla^2 g(x)\| \leq L$

$$\nabla(\quad) = \nabla g(x_0) + L\cdot(x-x_0)$$

$$\nabla(\quad) \qquad \|\nabla^2 g(x)\| \leq \|L \cdot 1\|$$

in general "no"
(except a few special cases)

Stochastic Gradient Descent

Case 2: **Bounded Variance**

Bounded Variance Assumption

$\mathbb{E}\, g(x) = \nabla f(x)$

For a gradient oracle $\mathbf{g}: \mathbb{R}^d \to \mathbb{R}^d$, assume that there exists a constant $\sigma \geq 0$, such that:

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|^2\right] \leq \sigma^2$$

for all $\mathbf{x} \in \mathbb{R}^d$.

This assumption can be generalized:

Assume that there exists constants $\sigma \geq 0$, $M \geqslant 0$ such that:

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|^2\right] \leq M\|\nabla f(\mathbf{x})\|^2 + \sigma^2$$

for all $\mathbf{x} \in \mathbb{R}^d$.

Exercise sheet 3:

$\leq M'\cdot(f(x) - f(x^*)) + \sigma^2$

Bounded variance: $\mathcal{O}(1/\varepsilon^2)$ steps

Theorem (Lecture-4).1

Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be L-smooth and assume $f(\mathbf{x}) \geq f^\star$. Define $F_0 = f(\mathbf{x}_0) - f^\star$. Choosing the constant stepsize

$$\gamma := \min\left\{\frac{1}{2L(1+M)}, \frac{\sqrt{F_0}}{\sqrt{TL\sigma^2}}\right\}$$

stochastic gradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2 = \mathcal{O}\left(\frac{\sqrt{LF_0}\sigma}{\sqrt{T}} + \frac{LF_0(1+M)}{T}\right).$$

$M=0,\ \sigma^2=0 \Rightarrow \mathcal{O}\left(\frac{LF_0}{T}\right)$

- Recovers the gradient descent convergence rate (Theorem (lecture-2).5)
- Note the daifferent impact of M and σ^2!

Proof I

$\|x_{t+1} - x^*\|^2$ ← f is convex

$f(x_{t+1}) \leq$ ← f is smooth

$$f(x_t) + \nabla f(x_t)^T \cdot (x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|^2$$

$-\gamma \cdot g_t$ $\quad -\gamma g_t$

SGD: $x_{t+1} = x_t - \gamma \cdot g_t$

$$f(x_t) + \nabla f(x_t)^T \cdot (-\gamma g_t) + \frac{L}{2}\gamma^2 \|g_t\|^2$$

$E[\quad]$ $\quad$ note: $E g_t = \nabla f(x_t)$

Proof I

Assumption $\mathbb{E}\|g_t - \nabla f(x)\|^2 \le M \cdot \|\nabla f(x)\|^2 + \sigma^2$

We use the quadratic upper bound:

$$\begin{aligned}
\mathbb{E} f(\mathbf{x}_{t+1}) &\le f(\mathbf{x}_t) - \gamma \nabla f(\mathbf{x}_t)^\top \nabla f(\mathbf{x}_t) + \frac{\gamma^2 L}{2} \mathbb{E}\|\mathbf{g}_t\|^2 \\
&\le f(\mathbf{x}_t) - \gamma \|\nabla f(\mathbf{x}_t)\|^2 + \frac{\gamma^2 L}{2}\left((1+M)\|\nabla f(\mathbf{x}_t)\|^2 + \sigma^2\right) \\
&\le f(\mathbf{x}_t) - \frac{\gamma}{2}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\gamma^2 L \sigma^2}{2}
\end{aligned}$$

$= \mathbb{E}\|g_t - \nabla f(x_t)\|^2 + \|\nabla f(x_t)\|^2$

where we used that $\gamma \le \frac{1}{2L(1+M)}$.

$\|\nabla f(x_t)\|^2 \cdot \left(-\gamma + \frac{\gamma^2 L}{2}(1+M)\right)$

$\left(-\gamma + \gamma \cdot \frac{L(1+M)}{2} \cdot \frac{1}{2L(1+M)}\right)$

Proof II

Re-arrange:

$$2\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2 \leq \frac{\mathbb{E}f(\mathbf{x}_t) - \mathbb{E}f(\mathbf{x}_{t+1})}{\gamma} + \frac{\gamma L\sigma^2}{2}$$

And sum from $t = 0, \ldots, T-1$:

$$\frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2 \leq \frac{f(\mathbf{x}_0) - f(\mathbf{x}_T)}{\gamma T} + \frac{\gamma L\sigma^2}{2} \leq \frac{F_0}{\gamma T} + \gamma L\sigma^2 .$$

Pick the optimal stepsize:

$$\gamma = \min\left\{\frac{1}{2L(1+M)}, \operatorname*{argmin}_{\gamma}\left(\frac{F_0}{\gamma T} + \gamma L\sigma^2\right)\right\} = \min\left\{\frac{1}{2L(1+M)}, \frac{\sqrt{F_0}}{\sqrt{TL\sigma^2}}\right\} .$$

from the proof

best possible

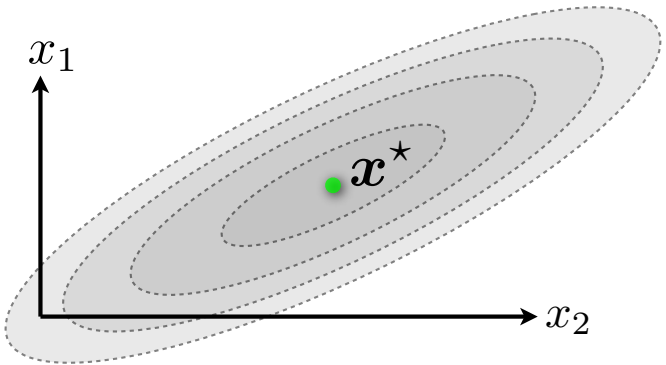
$$-\frac{F_0}{\gamma^2 T} + L\sigma^2 \doteq 0$$

Chapter 7

Coordinate Descent

Coordinate Descent

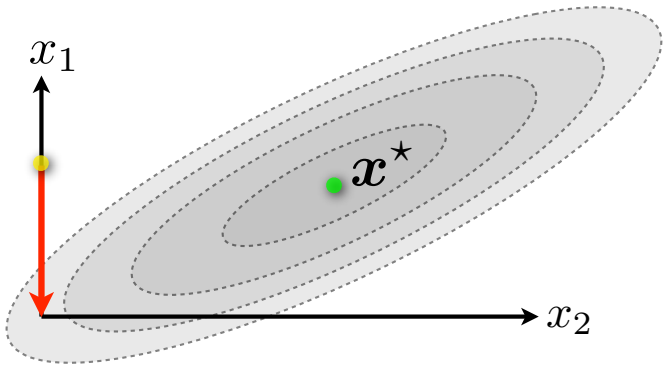
Goal: Find $\mathbf{x}^\star \in \mathbb{R}^d$ minimizing $f(\mathbf{x})$. (Example: $d = 2$)



Idea: Update one coordinate at a time, while keeping others fixed.

Coordinate Descent

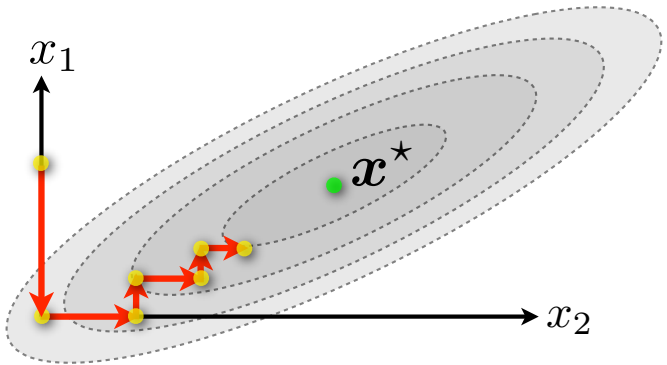
Goal: Find $\mathbf{x}^\star \in \mathbb{R}^d$ minimizing $f(\mathbf{x})$.



Idea: Update one coordinate at a time, while keeping others fixed.

Coordinate Descent

Goal: Find $\mathbf{x}^\star \in \mathbb{R}^d$ minimizing $f(\mathbf{x})$.



Idea: Update one coordinate at a time, while keeping others fixed.

Coordinate Descent

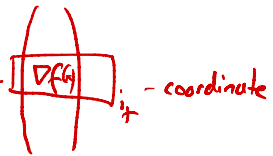
Modify only one coordinate per step:

$$\text{select } i_t \in [d]$$
$$\mathbf{x}_{t+1} := \mathbf{x}_t + \gamma \mathbf{e}_{i_t}$$

Two main variants:

- Gradient-based step-size:

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \tfrac{1}{L} \nabla_{i_t} f(\mathbf{x}_t)\, \mathbf{e}_{i_t}$$



- Exact coordinate minimization: solve the single-variable minimization $\operatorname{argmin}_{\gamma \in \mathbb{R}} f(\mathbf{x}_t + \gamma \mathbf{e}_{i_t})$ in closed form.

Randomized Coordinate Descent

select $i_t \in [d]$ uniformly at random

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \frac{1}{L} \nabla_{i_t} f(\mathbf{x}_t)\, \mathbf{e}_{i_t}$$

- Faster convergence than gradient descent
  (if coordinate step is significantly cheaper than full gradient step)
- is state-of-the-art for generalized linear models $f(\mathbf{x}) := g(A\mathbf{x}) + \sum_i h_i(x_i)$.

  Regression, classification (with different regularizers)

A data matrix

Example: Regression $\min_x \|Ax - b\|^2$

Convergence Analysis

Assume coordinate-wise smoothness:

$$f(\mathbf{x} + \gamma \mathbf{e}_i) \leq f(\mathbf{x}) + \gamma \nabla_i f(\mathbf{x}) + \frac{L}{2}\gamma^2 \qquad \forall \mathbf{x} \in \mathbb{R}^d,\ \forall \gamma \in \mathbb{R},\ \forall i$$

Is equivalent to coordinate-wise Lipschitz gradient:

$$|\nabla_i f(\mathbf{x} + \gamma \mathbf{e}_i) - \nabla_i f(\mathbf{x})| \leq L|\gamma|, \quad \forall \mathbf{x} \in \mathbb{R}^d,\ \forall \gamma \in \mathbb{R},\ \forall i.$$

- Additionally assume strong convexity (or the PL condition)

Convergence Analysis: Linear Rate

Theorem (Lecture-4).2

Let f be coordinate-wise smooth with constant L, and strongly convex with parameter $\mu > 0$. Then, coordinate descent with a step-size of $1/L$,

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \tfrac{1}{L} \nabla_{i_t} f(\mathbf{x}_t)\, \mathbf{e}_{i_t} .$$

when choosing the active coordinate i_t uniformly at random, has an expected linear convergence rate of

$$\mathbb{E}\big[f(\mathbf{x}_t) - f^\star\big] \leq \Big(1 - \frac{\mu}{dL}\Big)^t [f(\mathbf{x}_0) - f^\star].$$

Convergence Proof

Proof.
Plugging the update rule, into the smoothness condition, we have

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

Take expectation with respect to i_t:

$$\begin{aligned}
\mathbb{E}\left[f(\mathbf{x}_{t+1})\right] &\leq f(\mathbf{x}_t) - \frac{1}{2L}\mathbb{E}\left[|\nabla_{i_t} f(\mathbf{x}_t)|^2\right] \\
&= f(\mathbf{x}_t) - \frac{1}{2L}\frac{1}{d}\sum_i |\nabla_i f(\mathbf{x}_t)|^2 \\
&= f(\mathbf{x}_t) - \frac{1}{2dL}\|\nabla f(\mathbf{x}_t)\|^2.
\end{aligned}$$

recall exercise sheet 2

[**Lemma:** strongly convex f satisfy **PL**: $\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \geq \mu(f(\mathbf{x}) - f^\star)\ \forall \mathbf{x}$]
Subtracting $f^\star$ from both sides, we therefore obtain

$$\mathbb{E}[f(\mathbf{x}_{t+1}) - f^\star] \leq \left(1 - \frac{\mu}{dL}\right)[f(\mathbf{x}_t) - f^\star].$$

□

Coordinatewise Smoothness - Example

Consider a convex quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x}$ for a positive semidefinite (symmetric) matrix $A \in \mathbb{R}^{d\times d}$.

- Recall: the smoothness constant of f is $\|A\|$
- individual smoothness constants $L_i = (A)_{ii}$ for each coordinate i

$$f(\mathbf{x}+\gamma\mathbf{e}_i) = f(\mathbf{x}) + \gamma\nabla_i f(\mathbf{x}) + \tfrac{1}{2}\gamma^2(A)_{ii}$$

Lemma (Lecture-4).3

It holds $\max_i(A)_{ii} \le \|A\| \le d\cdot\max_i(A)_{ii}$ *and both inequalities are tight.*

Proof.

- $(A)_{ii} = \mathbf{e}_i^\top A\mathbf{e}_i \le \max_{\|\mathbf{u}\|=1}\mathbf{u}^\top A\mathbf{u} = \|A\|$, and consider $A = I_d$.
- $\|A\| = \lambda_{\max}(A) \le \sum_{i=1}^d (A)_{ii} \le d\max_i(A)_{ii}$, and consider $A = \mathbf{1}^\top\mathbf{1}$.

□

Importance Sampling [Nes12]

Uniformly random selection is not always the best!

- individual smoothness constants L_i for each coordinate i

$$f(\mathbf{x}+\gamma\mathbf{e}_i) \leq f(\mathbf{x}) + \gamma\nabla_i f(\mathbf{x}) + \tfrac{L_i}{2}\gamma^2$$

Coordinate descent using this modified selection probabilities $P[i_t = i] = \frac{L_i}{\sum_i L_i}$, and using a step-size of $1/L_{i_t}$ converges (Exercise 47) with the faster rate of

$$\mathbb{E}[f(\mathbf{x}_t) - f^\star] \leq \left(1 - \frac{\mu}{d\bar{L}}\right)^t [f(\mathbf{x}_0) - f^\star],$$

where $\bar{L} = \frac{1}{d}\sum_{i=1}^d L_i$.

Often: $\bar{L} \ll L = \max_i L_i$!

Discussion

- **Greedy Coordinate Descent:**

$$i_t := \operatorname*{argmax}_{i \in [d]} |\nabla_i f(\mathbf{x}_t)| \,.$$

  - often difficult to implement efficiently,
  - has the same convergence rate as random CD!
  - (this can slightly be improved, see [NSL+15])

- $g(x) = d\,\nabla_{i_t} f(x)$
- $\mathbb{E}\,g(x) = \frac{d}{d} \cdot \sum_{i=1}^{d} \nabla_i f(x) = \frac{d}{d} \cdot \nabla f(x)$

- **CD can be seen as SGD!**
  - the bounded variance assumption holds with $\sigma^2 = 0$, $M = d - 1$

$$\mathbb{E}[d\nabla_i f(\mathbf{x})] = \nabla f(\mathbf{x}) \qquad \mathbb{E}\,\|d\nabla_i f(\mathbf{x})\|^2 \le d\,\|\nabla f(\mathbf{x})\|^2$$

  - an important example where $\sigma^2 = 0$ (linear convergence)

$\mathbb{E}\,\|d\cdot\nabla_i f(x) - \nabla f(x)\|^2 = \mathbb{E}\,\|d\nabla_i f(x)\|^2 - \|\nabla f(x)\|^2 \le d\cdot\|\nabla f(x)\|^2 - 1\cdot\|\nabla f(x)\|^2$

definition: $\frac{1}{d}\sum_{i=1}^{d} \|d\cdot\nabla_i f(x)\|^2 = d\cdot\|\nabla f(x)\|^2$

Non-smooth objectives

Have proved everything for smooth f. What about non-smooth?

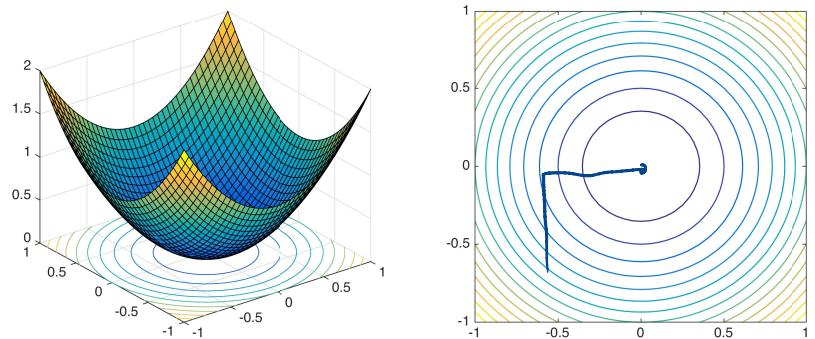


Figure: A smooth function: $f(\mathbf{x}) := \|\mathbf{x}\|^2$.

figure by Alp Yurtsever & Volkan Cevher, EPFL

Non-smooth objectives

For general non-smooth f, coordinate descent fails: gets permanently stuck:

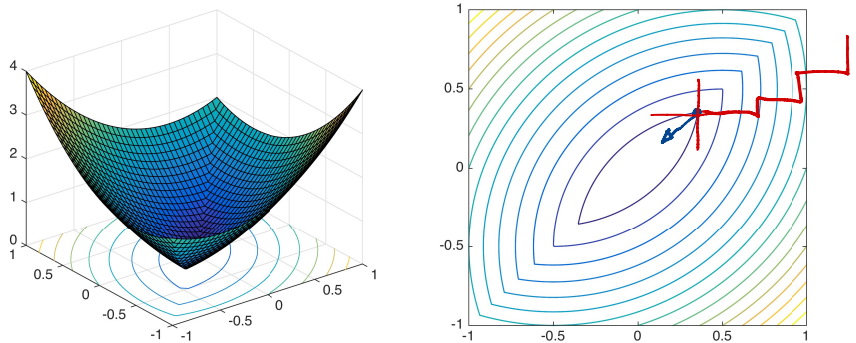


Figure: A non-smooth function: $f(\mathbf{x}) := \|\mathbf{x}\|^2 + |x_1 - x_2|$.

figure by Alp Yurtsever & Volkan Cevher, EPFL

Non-smooth separable objectives

What if the non-smooth part is separable over the coordinates?

$$f(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x}) \quad \text{with } h(\mathbf{x}) = \sum_i h_i(x_i)\,,$$

- global convergence!

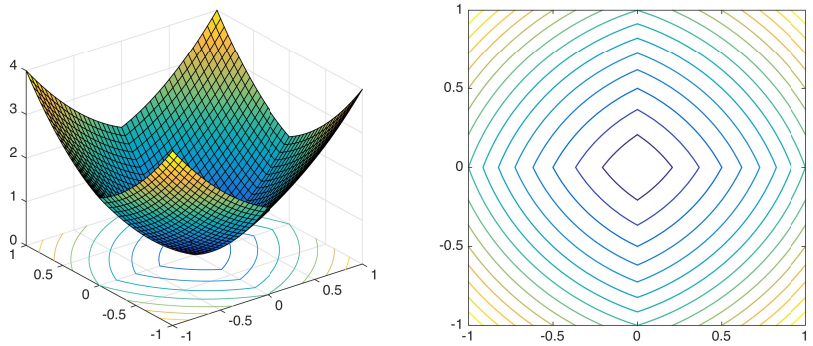


Figure: A non-smooth but separable function: $f(\mathbf{x}) := \|\mathbf{x}\|^2 + \|\mathbf{x}\|_1$.

figure by Alp Yurtsever & Volkan Cevher, EPFL

Lecture 4 Recap

- Bounded Variance Assumption
  - a key assumption that allows an unified view on gradient descent ($\sigma^2 = 0$), coordinate descent ($\sigma^2 = 0$, $M > 0$) and SGD.
  - on the technical side: we understand where the formulas for the stepsizes are coming from, and how to derive the (theoretically) optimal stepsize
- Coordinate Descent
  - low per-iteration cost—if coordinates of the gradients can be computed efficiently! This is sometimes the case for ML problems, for example $f(\mathbf{x}) := \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2$, see exercise sheet.
  - new coordinate-wise smoothness allows to explain the potential speedup
  - examples where coordinate descents works well, and where it does not work

Bibliography I

Yurii Nesterov.

Efficiency of coordinate descent methods on huge-scale optimization problems.

SIAM Journal on Optimization, 22(2):341–362, 2012.

Julie Nutini, Mark W Schmidt, Issam H Laradji, Michael P Friedlander, and Hoyt A Koepke.

Coordinate Descent Converges Faster with the Gauss-Southwell Rule Than Random Selection.

In *ICML - Proceedings of the 32nd International Conference on Machine Learning*, pages 1632–1641, 2015.

Discussion

$$\frac{F_0}{\gamma \cdot T} + \gamma \sigma^2 \cdot L$$ want $= 0$ minimize in γ!

$$\gamma = \sqrt{\frac{F_0}{LT\sigma^2}} \Rightarrow \sqrt{\frac{L\sigma^2 F}{T}}$$

from quadratic upper bound: $\gamma \leq \frac{2}{L}$ $\Rightarrow$ $\frac{F_0 L}{T}$

Discussion

Discussion