Lecture 4: Optimization Background

Start Recording!

Usual Goal in ML

Empirical Risk Minimization:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)$$

2 perspectives:

Deterministic (a.k.a., Batch)

$$\min_{\theta \in \Theta} g(\theta)$$

Stochastic

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_n}[\ell(f_\theta(x), y)]$$

Our Algorithm (Batch Case) : Gradient Descent

Gradient Descent:

Descent Method!!!!

$$\theta_{t+1} = \theta_t - \eta \nabla g(\theta_t)$$

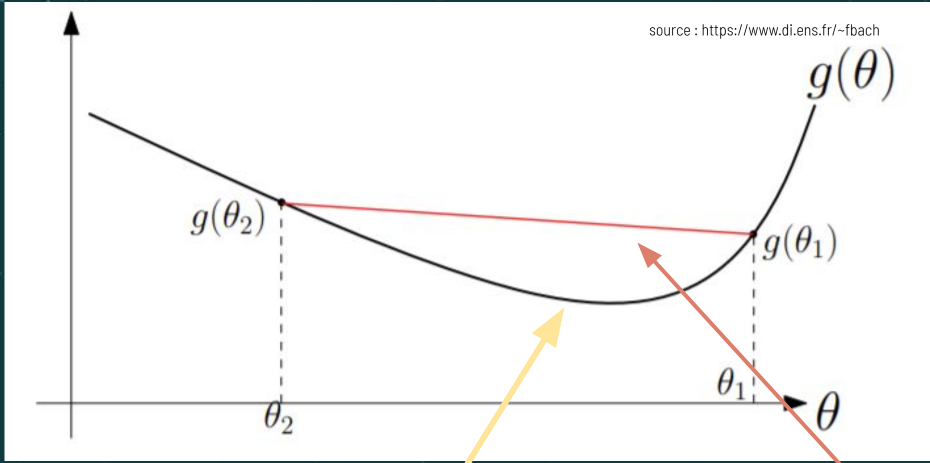
Step-Size (a.k.a learning rate)

Outline for Gradient Descent

- Standard assumptions
- Convergence in the convex case
- Convergence in the strongly-convex case

(we will cover non-convex case later)

Convexity



$$\forall \theta_1, \theta_2, \alpha \in [0,1],\ g(\alpha\theta_1 + (1-\alpha)\theta_2) \leq \alpha g(\theta_1) + (1-\alpha) g(\theta_2)$$

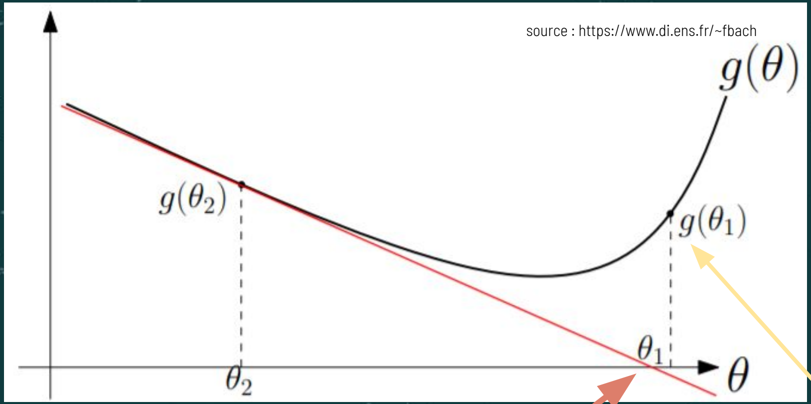
Convexity

Remark:

- Convexity is the most standard assumption
- Local minima are Global minima!
- We can prove convergence rates!
- We have convex duality. [Boyd and Vandenberghe (2004)]

$$\forall \theta_1, \theta_2, \alpha \in [0,1]\,,\ \overbrace{g(\alpha\theta_1 + (1-\alpha)\theta_2)} \leq \overbrace{\alpha g(\theta_1) + (1-\alpha) g(\theta_2)}$$

Convexity with Differentiable functions



$$\forall \theta_1, \theta_2, \quad g(\theta_2) + g'(\theta_2)^\top(\theta_1 - \theta_2) \leq g(\theta_1)$$

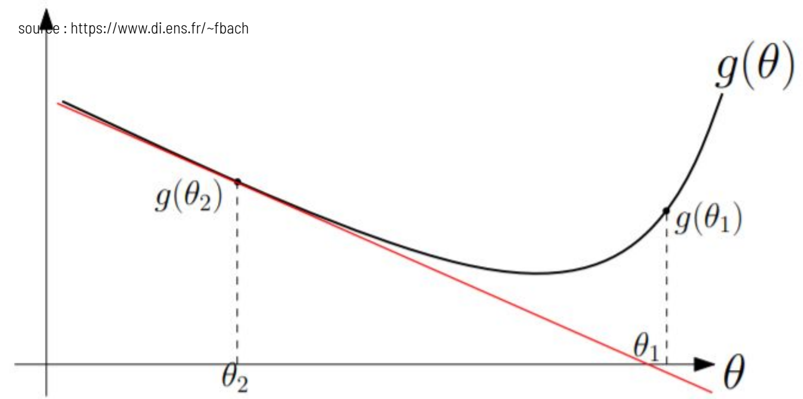
Convexity with Differentiable functions

Remark:

- Can extend this to non-differentiable function
- Any convex function is sub-differentiable

$$\forall\theta_1,\theta_2,\quad g(\theta_2)+g'(\theta_2)^\top(\theta_1-\theta_2)\le g(\theta_1)$$

Convexity with Twice differentiable functions



$$\forall \theta \, , \quad \nabla^2 g(\theta) \geq 0$$

Exercices:

- Show that for convex functions {local minima} = {global minima}
- Show that for differentiable functions def in Slides 6 and 7 are equivalent.
- Show that for twice differentiable functions def in Slides 6 and 7 are equivalent.

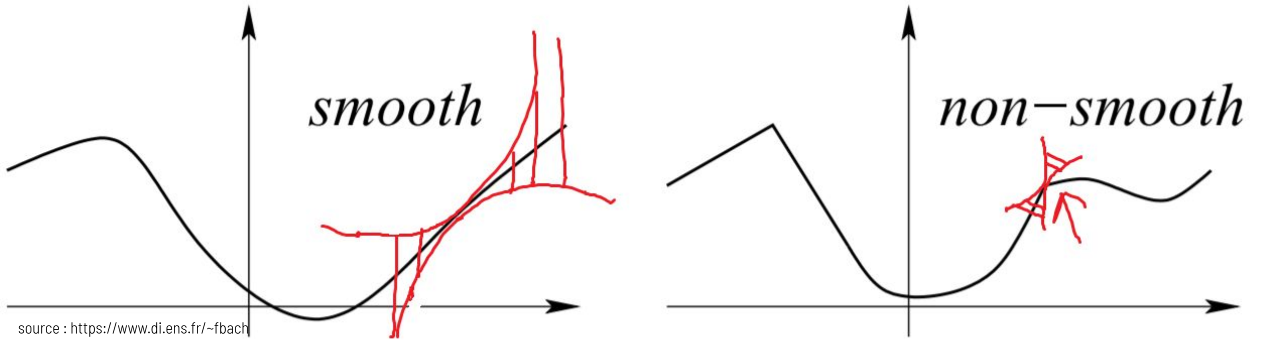
Smoothness and Strong Convexity

- A *smooth* function is a differentiable function with Lipschitz gradients:

$$\forall \theta_1, \theta_2, \quad \|\nabla g(\theta_1) - \nabla g(\theta_2)\|_2 \leq L \|\theta_1 - \theta_2\|_2$$

- If the function is twice differentiable we have:

$$\forall \theta \quad \nabla^2 g(\theta) \preceq L I_d$$



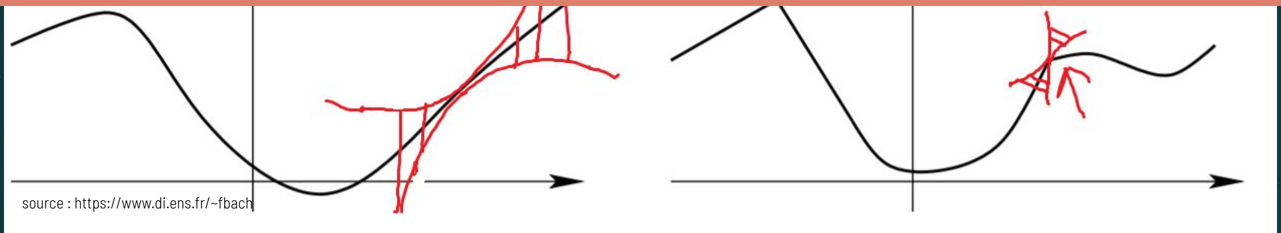
Smoothness and Strong Convexity

- Example:

$$g(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\theta^\top \varphi(x_i) - y_i)^2$$

$$\nabla^2 g(\theta) = \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)\varphi(x_i)^\top$$

Bounded data implies smooth function (generalizes to other losses)



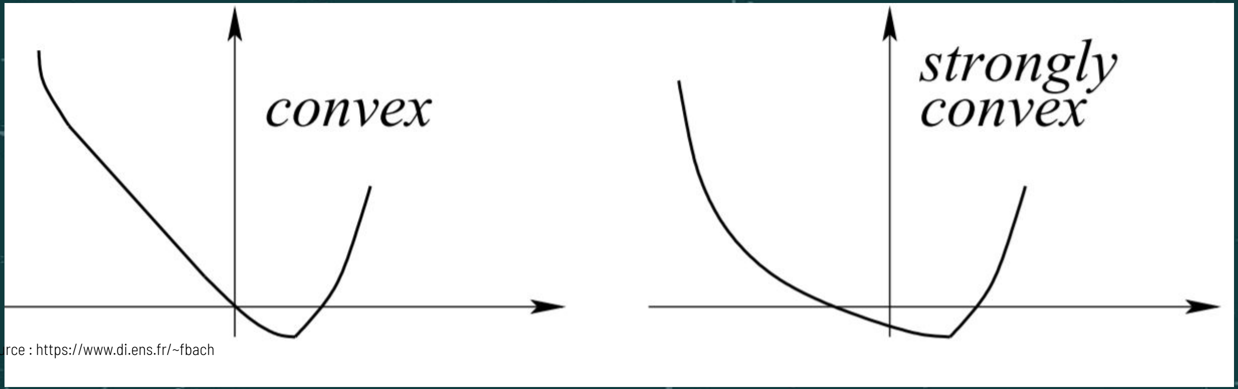
Smoothness and Strong Convexity

- A function is *strongly* convex if:

$$\forall \theta_1, \theta_2, \quad g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2) + \underbrace{\frac{\mu}{2}\|\theta_1 - \theta_2\|}_{\text{New term}} \leq g(\theta_1)$$

- If the function is twice differentiable we have:

$$\forall \theta \quad \nabla^2 g(\theta) \succeq \mu I_d$$



source : https://www.di.ens.fr/~fbach

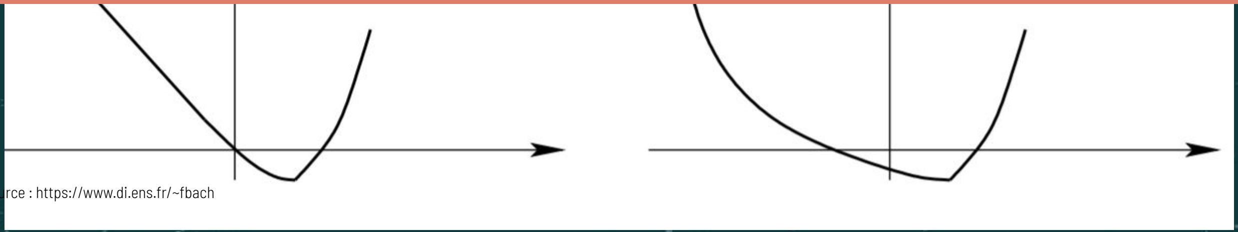
Smoothness and Strong Convexity

- Example:

$$g(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\theta^\top \varphi(x_i) - y_i)^2$$

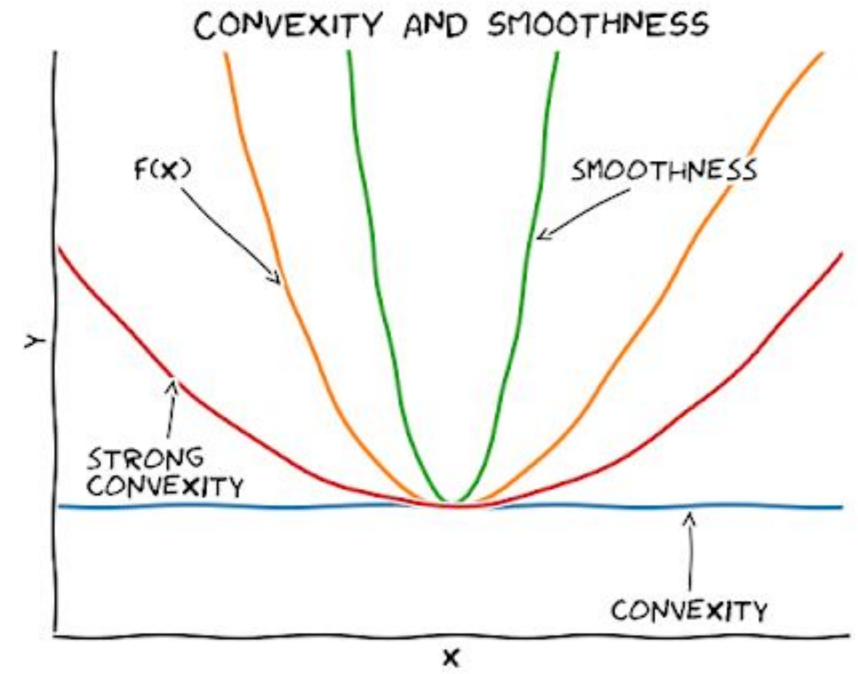
$$\nabla^2 g(\theta) = \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)\varphi(x_i)^\top$$

Invertible Covariance of the data implies strong convexity.



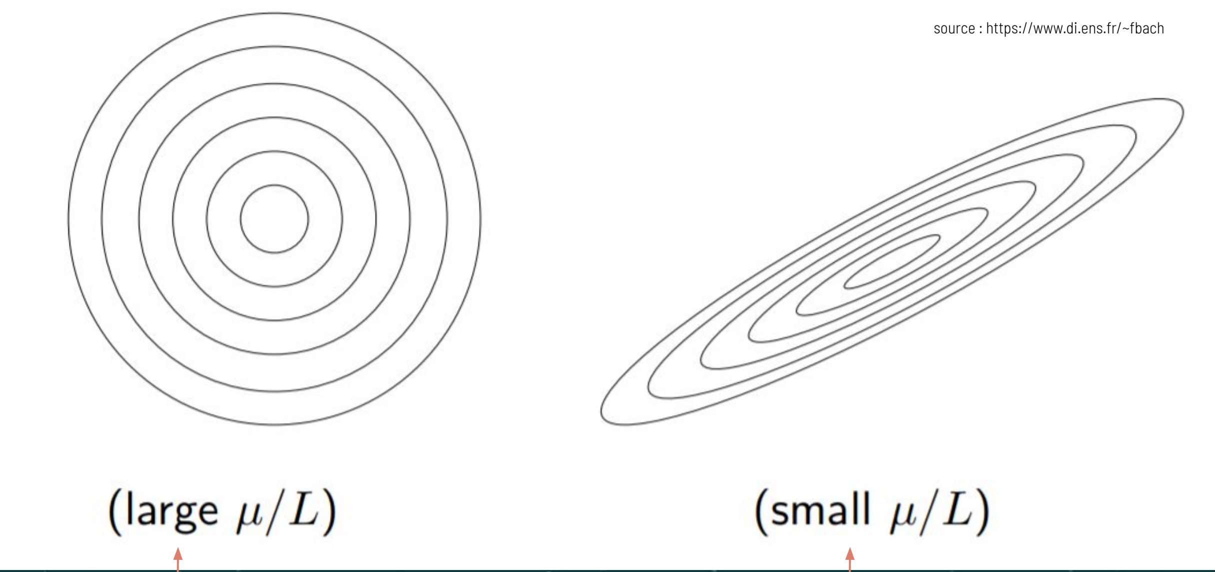
source : https://www.di.ens.fr/~fbach

Smoothness and Strong Convexity



Source: http://www.pokutta.com/

Smoothness and Strong Convexity



Large means close to 1 (easy problem)

Small means close to 0 (harder problem)

Back to Gradient Descent

Gradient Descent:

Descent Method!!!!

Step-Size (a.k.a learning rate)

$$\theta_{t+1} = \theta_t - \eta \nabla g(\theta_t)$$

Today two main results (smooth function):

- Convergence in the strongly convex case (faster)
- Convergence in the convex case (slower)

Descent Method

Lemma on smooth function:

$$\forall \theta, \theta', \quad g(\theta') \leq g(\theta) + \nabla g(\theta)^\top (\theta - \theta') + \frac{L}{2} \|\theta - \theta'\|_2^2$$

Descent Lemma:

$$\forall t \geq 0, \eta \leq 1/L \,, \quad g(\theta_{t+1}) \leq g(\theta_t) - \underbrace{\frac{\eta}{2} \|\nabla g(\theta_t)\|_2^2}_{\text{Descent!!!!}}$$

Exercice: Prove these two lemmas.

Strongly convex case

Minimum of the function

Lemma for Strongly convex functions

$$\forall \theta, \quad \|\nabla g(\theta)\|_2^2 \geq 2\mu(g(\theta) - g^*)$$

Convergence Result:

$$g(\theta_t) - g^* \leq (1 - \frac{\mu}{L})^t (g(\theta_0) - g^*)$$

Exercice: prove this!

Convex case

Point where the min is achieved

Le

$g(\theta_t$ ... $g(\theta_t)\|_2^2$

$\|_2^2)$

Remark:

- Many variations (Different step-sizes, Projections ....)
- We want the step-size to be as big as possible (bigger means faster)
- Many proofs use similar ideas!!!!

Co

$g(\theta_t) - g(\theta^*) \leq \frac{\ldots}{2\eta(t+1)}$

Exercice: prove this!

Convex case

Point where the min is achieved

Lemmas for convex functions

$$g(\theta_t) - g^* \leq \nabla g(\theta_t)^\top (\theta_t - \theta^*) + \eta \|\nabla g(\theta_t)\|_2^2$$

$$g(\theta_t) - g(\theta^*) \leq \frac{1}{2\eta}(\|\theta_t - \theta^*\|_2^2 - \|\theta_{t+1} - \theta^*\|_2^2)$$

Convergence Result (for the right step-size):

$$g(\theta_t) - g(\theta^*) \leq \frac{\|\theta_0 - \theta^*\|_2^2}{2\eta(t+1)}$$

Exercice: prove this!

Convex case

Point where the min is achieved

Le

$g(\theta_t$ … $g(\theta_t)\|_2^2$

… $\|_2^2)$

Remark:

- Many variations (Different step-sizes, Projections ….)
- We want the step-size to be as big as possible (bigger means faster)
- Many proofs use similar ideas!!!!

Co

$$g(\theta_t) - g(\theta^*) \leq \frac{\ldots}{2\eta(t+1)}$$

Exercice: prove this!

Summary

Condition number: The quantity of interest for convergence speed.

Strongly convex case: (Linear rate)

$$g(\theta_t) - g^* \leq (1 - \frac{\mu}{L})^t (g(\theta_0) - g^*)$$

Convex case: (Linear rate)

$$g(\theta_t) - g(\theta^*) \leq \frac{L\|\theta_0 - \theta^*\|_2^2}{2(t+1)}$$

Why care about rate?

Summary

Condition number: The quantity of interest for convergence speed.

Why care about rate?

$$\mathbb{E}_{(x,y)\sim p_{data}}[\ell(f(x),y)] \quad \text{v.s.} \quad \min_\theta \frac{1}{n}\sum_{i=1}^{n}\ell(f_\theta(x_i),y_i)$$

Three kind of errors: $\mathcal{E}_{app} + \mathcal{E}_{est} + \mathcal{E}_{opt}$

- Approximation error (Bias)
- Estimation error (Variance)
- Optimization error (useless to be too small)

[Bottou and Bousquet (2008)] – In machine learning, no need to optimize below estimation error

First Last Algorithm: Steepest Descent

$$\theta_{t+1} = \theta_t + \eta d \qquad d := \arg\min_{\|d\| \leq 1} \nabla g(\theta_t)^\top d$$

- If the norm is the L_2 norm then we recover gradient descent.
- Exercice: what do we get if we use the L_∞ norm???
- Remark: proof not trivial. A more natural extension is a penalized version:

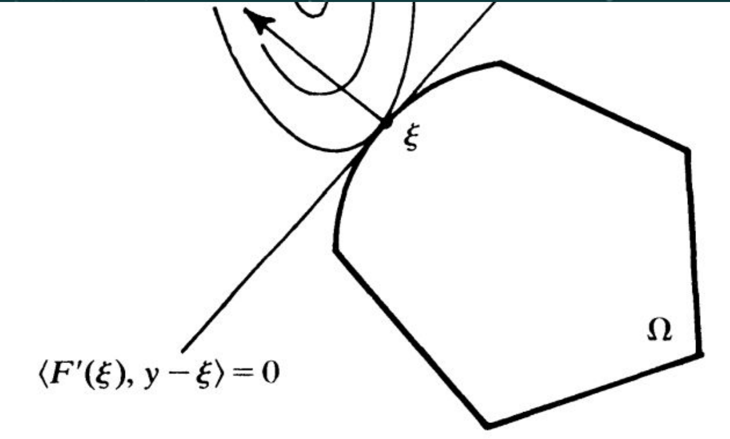
$$\theta_{t+1} := \arg\min_{\theta \in \mathbb{R}^d} \nabla g(\theta_t)^\top (\theta - \theta_t) + \frac{1}{2\eta} \|\theta - \theta_t\|^2$$

Arbitrary norm!!!

Second Last Algorithm: Projected Gradient Descent

- Gradient Descent + Projection step.

$$\theta_{t+1} = P_{\Theta}[\theta_t - \eta \nabla g(\theta_t)]$$



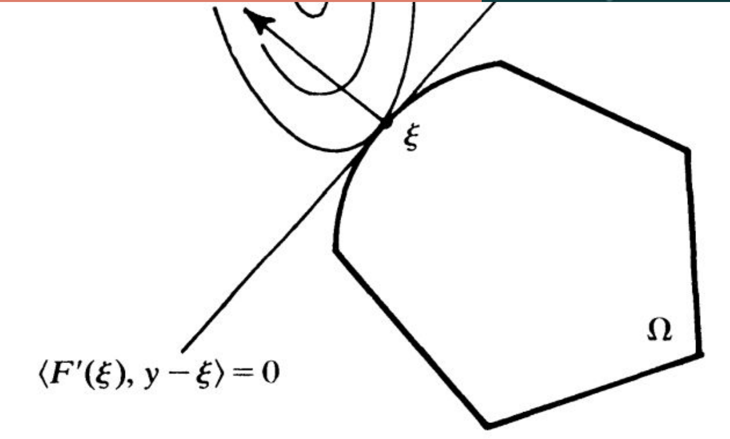
- Steepest-descent version:

$$\theta_{t+1} := \arg\min_{\substack{\theta\in\mathbb{R}^d \\ \Theta}} \nabla g(\theta_t)^\top(\theta - \theta_t) + \frac{1}{2\eta}\|\theta - \theta_t\|^2$$

Second Last Algorithm: Projected Gradient Descent

Remark:

- Different notion of optimality
- Extending the proof is quite Straightforward. (projection is contractive)
- Rich literature on lower-bound And faster algorithms.



$$\theta_{t+1} := \arg\min_{\theta \in \cancel{\mathbb{R}^d}\ \Theta} \nabla g(\theta_t)^\top (\theta - \theta_t) + \frac{1}{2\eta}\|\theta - \theta_t\|^2$$

Conclusion

- Many different proofs
- Standard Assumption to get convergence Rates:
  - Lipschitz gradient
  - Convexity

  } Not Valid in Deep-Learning

- Many variants:
  - Projected Gradient Descent
  - Steepest Descent

  } Application: Adversarial Examples (next course)

For more: check the books in the next slides.

References:

Nesterov, Yurii. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2003.

Bubeck, Sébastien. "Convex optimization: Algorithms and complexity." *arXiv preprint arXiv:1405.4980* (2014).

Bottou, Léon, and Olivier Bousquet. "The tradeoffs of large scale learning." *Advances in neural information processing systems*. 2008.

Boyd, Stephen, Stephen P. Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.