Labs
Optimization for Machine Learning
Spring 2024

Saarland University
CISPA Helmholtz Center for Information Security
Sebastian Stich
TAs: Yuan Gao & Xiaowen Jiang
https://cms.cispa.saarland/optml24/

Problem Set 2, April 23, 2024 (Gradient Descent)

Convexity, Smoothness and Gradient descent

Exercise 1. (μ-strong convexity)
*A function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-**strongly convex** if f is differentiable and*

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

where $\mu > 0$ and the norm $||\mathbf{x}||$ is defined as $||\mathbf{x}||^2 := \mathbf{x}^T\mathbf{x}$.

- *Prove that a μ-strongly convex function has a unique minimizer $\mathbf{x}^\star \in \operatorname{argmin} f(\mathbf{x})$ and that it holds:*

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \leq \frac{1}{2\mu}||\nabla f(\mathbf{x})||^2, \quad \forall \mathbf{x} \in \mathbb{R}^d .$$

- *Assume that f is μ-strongly convex and L-smooth. Let us run gradient descent on f starting from $\mathbf{x}_0$. Recall that the sequence produced by GD satisfies:*

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

  where $\gamma \geq 0$ is a parameter. Prove that, for any $t \geq 0$, it holds:

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) \leq (1-\alpha)\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big),$$

  for a parameter α. For which α? What is the best γ? (You can use the result from the previous question.) (Hint: recall that when running GD on smooth functions, it holds that, $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \beta||\nabla f(\mathbf{x}_t)||^2$ for some $\beta > 0$ that depends on L and γ)

- *Following the previous question, please state the iteration complexity of gradient descent in big-$\mathcal{O}$ notation. Your expression can depend on the problem parameters γ,μ,L,$F_0 := f(\mathbf{x}_0) - f(\mathbf{x}^\star)$, $f(\mathbf{x}_0)$, $f(\mathbf{x}^*)$, $R_0^2 := ||\mathbf{x}_0 - \mathbf{x}^\star||^2$, and the target accuracy $\epsilon \geq 0$.*

Exercise 2. (ℓ_2-regularized least square)
Consider the objective function $f\colon \mathbb{R}^d \to \mathbb{R}$:

$$f(\mathbf{x}) = \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{a}_i^\top\mathbf{x} - b_i)^2 + \frac{\lambda}{2}||\mathbf{x}||_2^2,$$

where each $\mathbf{a}_i$ is a data vector with dimension d, each b_i is a label which is a scalar and $\lambda > 0$ is the regularization parameter.

- *Can you rewrite the objective function into a compact matrix form?*
- *What is the smoothness parameter L of f?*
- *Is $f(\mathbf{x})$ strongly convex? If yes, prove its strong convexity and write down the strongly convex parameter μ. Otherwise, give a reason why it is not necessarily strongly convex.*
- *What is the minimizer x^* of $f(\mathbf{x})$? Is it unique?*

Practical Implementation of Gradient Descent

Follow the Python notebook provided here:

colab.research.google.com/github/epfml/OptML_course/blob/master/labs/ex02/template/Lab 2 - Gradient Descent.ipynb

The notebook introduces the objective function $f(\mathbf{x})$ *defined in Exercise 2 with* $\lambda = 0$.