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 — Solutions (Gradient Descent)

Convexity, Smoothness and Gradient descent

Exercise (μ-strong convexity).

Solution:

- We first show that $f(\mathbf{x})$ is strictly convex. From the definition of μ-strong convexity, we have:

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

  Therefore $f(\mathbf{x})$ admits at most one global minimum. It suffices to show that $f(\mathbf{x})$ admits at least one global minimum. For any $\mathbf{x} \in \mathbb{R}^d$, we define the ball: $B = \{\mathbf{y} \in \mathbb{R}^d : ||\mathbf{y}-\mathbf{x}|| \leq r\}$ where $r = \frac{4||\nabla f(\mathbf{x})||}{\mu}$, it holds $\forall y \in \mathbb{R}^d, y \notin B$:

$$\begin{aligned}
f(\mathbf{y}) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}||\mathbf{y}-\mathbf{x}||^2 \\
&\geq f(\mathbf{x}) - ||\nabla f(\mathbf{x})||||\mathbf{y}-\mathbf{x}|| + \frac{\mu}{2}||\mathbf{y}-\mathbf{x}||^2 \\
&= f(\mathbf{x}) + \frac{\mu}{2}(||\mathbf{y}-\mathbf{x}||^2 - \frac{1}{2}r||\mathbf{y}-\mathbf{x}||) \\
&\geq f(\mathbf{x}) + \frac{\mu}{4}r^2 \, .
\end{aligned}$$

  The above inequality shows that for $\forall y \in \mathbb{R}^d, y \notin B$, $f(y) \geq f(\mathbf{x})$. On the other hand, since B is closed and bounded, and $f(\mathbf{x})$ is continuous, from Weierstrass theorem, f attains its minimum in B, that is $\forall \mathbf{z} \in B$, there exists $\mathbf{x}^*$ such that $f(\mathbf{z}) \geq f(\mathbf{x}^*)$. Since $\mathbf{x} \in B$, we conclude that $\mathbf{x}^*$ is a minimizer of f in $\mathbb{R}^d$. Since $f(\mathbf{x})$ is strictly convex, $\mathbf{x}^*$ is the unique minimizer of f in $\mathbb{R}^d$.
  To prove the inequality, $\forall \mathbf{x} \in \mathbb{R}^d$, we let $g(\mathbf{y}) := f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}||\mathbf{y}-\mathbf{x}||^2$. Since $g(\mathbf{y})$ is strongly convex, we can explicitly compute its minimum by finding its first-order critical point:

$$\nabla g(\mathbf{y}^\star) = \nabla f(\mathbf{x}) + \mu(\mathbf{y}^* - \mathbf{x}) = 0 \Rightarrow \mathbf{y}^\star = \mathbf{x} - \mu^{-1}\nabla f(\mathbf{x}) \, .$$

  Plugging $\mathbf{y}^*$ into $g(\mathbf{y})$, we obtain that: $\min g(\mathbf{y}) = f(\mathbf{x}) - \frac{1}{2\mu}||\nabla f(\mathbf{x})||^2$. By the definition of μ-strong convexity, we have $\forall \mathbf{y}, \mathbf{x} \in \mathbb{R}^d$:

$$f(\mathbf{y}) \geq g(\mathbf{y}) \geq \min g(\mathbf{y}) = f(\mathbf{x}) - \frac{1}{2\mu}||\nabla f(\mathbf{x})||^2 \, .$$

  Setting $\mathbf{y}$ to be $\mathbf{x}^*$ and rearranging give the result.

- According to the hint, we need to lower bound $||\nabla f(\mathbf{x}_t)||^2$. Using the definition of L-smoothness, we get:

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

  Plugging the update rule of GD into this inequality, we obtain:

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

Let $\beta := \gamma - \frac{L\gamma^2}{2}$. Combining this inequality with the one provided in the first question, we get:

$$2\mu\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \leq ||\nabla f(\mathbf{x}_t)||^2 \leq \frac{1}{\beta}\big(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})\big) = \frac{1}{\beta}\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) - \frac{1}{\beta}\big(f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star)\big) \ .$$

Rearranging, we get, for any $t \geq 0$:

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \leq (1 - 2\mu\beta)\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \ .$$

We thus have

$$\alpha = 2\mu\beta = 2\mu(\gamma - \frac{L\gamma^2}{2}) \ .$$

Maximizing α w.r.t γ, we obtain the best choice for γ, which is:

$$\gamma = \frac{1}{L} \qquad \text{and} \qquad \alpha = \frac{\mu}{L} \ .$$

- From question 2, we have for any $t \geq 0$:

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \leq (1 - \frac{\mu}{L})\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \ .$$

Recursively applying this inequality, for any $T \geq 0$, we have:

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq (1 - \frac{\mu}{L})^T\big(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\big) = (1 - \frac{\mu}{L})^T F_0 \ .$$

To reach ϵ-accuracy, i.e. $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \epsilon$, we can let:

$$(1 - \frac{\mu}{L})^T F_0 \leq \epsilon \ .$$

This implies: $T \geq \frac{\ln(\frac{F_0}{\epsilon})}{\ln(\frac{1}{1-\frac{\mu}{L}})}$. Note that $\frac{1}{\ln(\frac{1}{1-\frac{\mu}{L}})} = \frac{1}{-\ln(1-\frac{\mu}{L})} \leq \frac{1}{\frac{\mu}{L}} = \frac{L}{\mu}$. Therefore, it suffices to have:

$$T \geq \frac{L}{\mu}\ln(\frac{F_0}{\epsilon}) \ .$$

Since we work on the upper bound for $f(\mathbf{x}_T) - f(\mathbf{x}^\star)$, the iteration complexity is thus $\mathcal{O}\big(\frac{L}{\mu}\ln(\frac{F_0}{\epsilon})\big)$.

Exercise (ℓ_2-regularized least square).

Solution:

- $f(\mathbf{x})$ can also be expressed as: $f(\mathbf{x}) = \frac{1}{2n}||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2 + \frac{\lambda}{2}||\mathbf{x}||_2^2$ for a $n \times d$ data matrix A (with rows $\mathbf{a}_i^T \in \mathbb{R}^{1\times d}$, $i = 1, \ldots, n$) and $n \times 1$ vector $\mathbf{b}$ (with rows b_i, $i = 1, \ldots, n$)
- The Hessian of f can computed as: $\nabla f^2(\mathbf{x}) = \frac{1}{n}\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}$. Therefore the smoothness parameter $L \geq \frac{1}{n}\lambda_{\max} + \lambda$ where $\lambda_{\max}$ is the largest eigenvalue of the matrix $\mathbf{A}^T\mathbf{A}$.
- $f(\mathbf{x})$ is strongly convex since $\frac{1}{n}\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I} \succ 0$ and hence $\nabla f^2(x) \succ 0$. The parameter μ is then $\frac{1}{n}\lambda_{\min}(\mathbf{A}^T\mathbf{A}) + \lambda$.
- Since $f(\mathbf{x})$ is strongly convex, there exists a unique global minimizer $\mathbf{x}^*$ which satisfies: $\nabla f(\mathbf{x}^*) = 0$. After computation, we can explicitly derive $x^* = (\frac{1}{n}\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}(\frac{1}{n}\mathbf{A}^T\mathbf{A}b)$.