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 1 – Solutions (Convexity, Python Setup)

Convexity

Exercise (Recognizing convex functions).

Solution:

- A function $f : \mathbb{R} \to \mathbb{R}$ which is convex but not strictly convex could be $|\mathbf{x}|$.
- A function $f : \mathbb{R} \to \mathbb{R}$ which is strictly convex yet not bounded below could be $e^{\mathbf{x}} + \mathbf{x}$.
  Note that $f''(x) = e^{\mathbf{x}} > 0$ but $f(\mathbf{x}) \to -\infty$ as $\mathbf{x} \to -\infty$.

Exercise (Properties of convexity).

Solution:

- Consider two simple linear functions $f(\mathbf{x}) = 1 + \mathbf{x}$ and $g(\mathbf{x}) = 1 - \mathbf{x}$. Then $h(\mathbf{x}) = f(\mathbf{x})g(\mathbf{x}) = 1 - \mathbf{x}^2$ is concave instead of convex.
- Consider two simple quadratics $f(\mathbf{x}) = \mathbf{x}^2$ and $g(\mathbf{x}) = \mathbf{x}^2 - 1$. Then the composition $(f \circ g)(\mathbf{x}) = \mathbf{x}^4 - 2\mathbf{x}^2 + 1$ is not convex since $(f \circ g)''(0) = -4 < 0$.
- Again $f(\mathbf{x}) = 1 + \mathbf{x}$ and $g(\mathbf{x}) = 1 - \mathbf{x}$ are suffice to show that $h(\mathbf{x}) = \min(f(\mathbf{x}), g(\mathbf{x}))$ is not convex. However, $w(\mathbf{x}) = \max_i\{f_i(\mathbf{x})\}$ is always convex given that all $f_i(\mathbf{x})$ are convex. See also Exercise **??** from the lecture note.

Exercise (Recognizing simple quadratics).

Solution:

- f is convex iff $\mathbf{A} \succeq 0$
- f is strictly convex if $\mathbf{A} \succ 0$

Exercise (Norm).

Solution: Let $||\cdot||$ be a norm defined on $\mathbb{R}^d$. It satisfies the following three axioms:

1. $||a\mathbf{x}|| = |a|||\mathbf{x}||$ for all $a \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$
2. $||\mathbf{x} + \mathbf{y}|| \le ||\mathbf{x}|| + ||\mathbf{y}||$ for all $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^d$
3. $||\mathbf{x}|| \ge 0$ for all $\mathbf{x} \in \mathbb{R}^d$ and $||\mathbf{x}|| = 0$ iff $\mathbf{x} = \mathbf{0}$.

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\lambda \in [0, 1]$, we have:

$$\begin{aligned} ||(1-\lambda)\mathbf{x} + \lambda\mathbf{y}|| &\le ||(1-\lambda)\mathbf{x}|| + ||\lambda\mathbf{y}|| \\ &\le |1-\lambda|||\mathbf{x}|| + |\lambda|||\mathbf{y}|| \\ &\le (1-\lambda)||\mathbf{x}|| + \lambda||\mathbf{y}|| \end{aligned}$$

Therefore, $||\cdot||$ is convex. For any seminorm functions: the proof just generalizes the previous case. For $\lambda \in [0,1]$ we get:

$$\begin{aligned} f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) &\leq f(\lambda\mathbf{x})+f((1-\lambda)\mathbf{y}) \quad \text{(triangle inequality)} \\ &= |\lambda|f(\mathbf{x})+|(1-\lambda)|f(\mathbf{y}) \\ &= \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y}). \end{aligned}$$

Exercise (Prove Jensen's inequality). *(Lemma* **??***)*

Solution: For $m=1$, there is nothing to prove, and for $m=2$, the statement holds by convexity of f. For $m>2$, we proceed by induction. If $\lambda_m = 1$ (and hence all other λ_i are zero), the statement is trivial. Otherwise, let $\mathbf{x} = \sum_{i=1}^m \lambda_i \mathbf{x}_i$ and define

$$\mathbf{y} = \sum_{i=1}^{m-1} \frac{\lambda_i}{1-\lambda_m}\mathbf{x}_i.$$

Thus we have $\mathbf{x} = (1-\lambda_m)\mathbf{y}+\lambda_m\mathbf{x}_m$. Also observe that $\sum_{i=1}^{m-1}\frac{\lambda_i}{1-\lambda_m} = 1$. By convexity and Jensens's inequality that we inductively assume to hold for $m-1$ terms, we get

$$\begin{aligned} f(\mathbf{x}) &= f((1-\lambda_m)\mathbf{y}+\lambda_m\mathbf{x}_m) \\ &\leq (1-\lambda_m) f\left(\sum_{i=1}^{m-1}\frac{\lambda_i}{1-\lambda_m}\mathbf{x}_i\right)+\lambda_m f(\mathbf{x}_m) \\ &\leq (1-\lambda_m)\left(\sum_{i=1}^{m-1}\frac{\lambda_i}{1-\lambda_m}f(\mathbf{x}_i)\right)+\lambda_m f(\mathbf{x}_m) = \sum_{i=1}^{m}\lambda_i f(\mathbf{x}_i). \end{aligned}$$

For any random variable $\mathbf{X}$, it holds that $f(\mathbb{E}[\mathbf{X}]) \leq \mathbb{E}[f(\mathbf{X})]$.

Exercise Least square (Exercise ?? from lecture note).

Solution: We compute

$$\frac{\partial f(w_0,\mathbf{w})}{\partial w_0} = 2\sum_{i=1}^{n}(w_0+(\mathbf{w}^\star)^\top\mathbf{x}_i - y_i) = 2\sum_{i=1}^{n} w_0 = 2nw_0.$$

since the observations are centered. Also, by the first-order characterization of optimality as by Lemma **??**,

$$0 = \frac{\partial f(w_0,\mathbf{w})}{\partial w_0}\Big|_{w_0=w_0^\star,\mathbf{w}=\mathbf{w}^\star} = 2nw_0^\star.$$

The second part follows from

$$\begin{aligned} f'(w_o-\mathbf{w}^\top\mathbf{q}+r,\mathbf{w}) &= \sum_{i=1}^{n}(w_0-\mathbf{w}^\top\mathbf{q}+r+\mathbf{w}^\top\mathbf{x}_i'-y_i')^2 \\ &= \sum_{i=1}^{n}(w_0-\mathbf{w}^\top\mathbf{q}+r+\mathbf{w}^\top(\mathbf{x}_i+\mathbf{q})-(y_i+r))^2 \\ &= \sum_{i=1}^{n}(w_0+\mathbf{w}^\top\mathbf{x}_i-y_i)^2 = f(w_0,\mathbf{w}). \end{aligned}$$

Exercise (Logistic regression). *(Exercise* **??** *from lecture note)*

Solution:

First we show that if $\mathbf{w}'$ is a nontrivial separator, then for every $\mathbf{w}$, $\ell(\mathbf{w}+\lambda\mathbf{w}') < \ell(\mathbf{w})$ for all $\lambda > 0$. So if there exists a nontrivial separator, we can always decrease the value of ℓ and hence ℓ cannot have a global minimum.

Fix some $\mathbf{w}\in\mathbb{R}^d$, some number $\lambda>0$ and some nontrivial separator $\mathbf{w}'$. By definition of a nontrivial separator, there exists some $(\mathbf{x}_0, y_0) \in P$ such that $y_0(\mathbf{w}'^\top \mathbf{x}_0) > 0$ and $(\mathbf{w}'^\top \mathbf{x})y \geq 0$ for all $(\mathbf{x}, y) \in P$. We get:

$$\begin{aligned}
&\ell(\mathbf{w} + \lambda \mathbf{w}') = \\
&= \sum_{(\mathbf{x},y)\in P} \ln\left(1 + \exp\left(-y\left(\mathbf{w} + \lambda \mathbf{w}'\right)^\top \mathbf{x}\right)\right) \\
&= \sum_{(\mathbf{x},y)\in P} \ln\left(1 + \exp\left(-y\mathbf{w}^\top \mathbf{x} - \lambda y \mathbf{w}'^\top \mathbf{x}\right)\right) \\
&= \sum_{(\mathbf{x},y)\in P} \ln\left(1 + \exp\left(-y\mathbf{w}^\top \mathbf{x}\right)\exp\left(-\lambda y \mathbf{w}'^\top \mathbf{x}\right)\right) \\
&< \sum_{(\mathbf{x},y)\in P} \ln\left(1 + \exp\left(-y\mathbf{w}^\top \mathbf{x}\right)\right) = \ell(\mathbf{w})\,.
\end{aligned}$$

To see why the last inequality is true, observe that $-y(\mathbf{w}'^\top \mathbf{x}) \leq 0$ and that both $\exp$ and $\ln$ are increasing functions. The inequality is strict for $\lambda > 0$ because there exists a term in the summation such that $-\lambda y_0(\mathbf{w}'^\top \mathbf{x}_0) < 0$.

Now let us prove that if all separators are trivial, then ℓ has a global minimum. Note that a separator $\mathbf{w}' \neq 0$ is trivial only if $\mathbf{w}'$ is orthogonal to all datapoints $\mathbf{x}$. For any such trivial separator $\mathbf{w}'$, $\mathbf{w} \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$, the loss value $\ell(\mathbf{w} + \lambda \mathbf{w}') = \ell(\mathbf{w})$.

$$\begin{aligned}
&\ell(\mathbf{w} + \lambda \mathbf{w}') = \\
&= \sum_{(\mathbf{x},y)\in P} \ln\left(1 + \exp\left(-y\left(\mathbf{w} + \lambda \mathbf{w}'\right)^\top \mathbf{x}\right)\right) \\
&= \sum_{(\mathbf{x},y)\in P} \ln\left(1 + \exp\left(-y\mathbf{w}^\top \mathbf{x} - \lambda y \mathbf{w}'^\top \mathbf{x}\right)\right) \\
&= \sum_{(\mathbf{x},y)\in P} \ln\left(1 + \exp\left(-y\mathbf{w}^\top \mathbf{x}\right)\right) = \ell(\mathbf{w})\,.
\end{aligned}$$

Let W' be the set of all such trivial separators of P and $(W')^\perp$ its orthogonal complement. Since every vector $\mathbf{w} \in \mathbb{R}^d$ can be decomposed as $\mathbf{w} = \mathbf{u} + \mathbf{v}$, where $\mathbf{u} \in W'$, $\mathbf{v} \in (W')^\perp$, the previously proved property means that $\ell(\mathbf{w}) = \ell(\mathbf{u} + \mathbf{v}) = \ell(\mathbf{v})$ and this means that

$$\inf_{\mathbf{w}\in\mathbb{R}^d} \ell(\mathbf{w}) = \inf_{\mathbf{w}\perp W'} \ell(\mathbf{w})\,.$$

Thus without loss of generality, we can restrict ourselves to weight vectors $\mathbf{w} \perp W'$. Now define the sublevel set of $\mathbf{w}_0 = 0$ with $\ell(0) = |P|\ln(2)$:

$$\tilde{W} = \{\mathbf{w} \perp W' : \ell(\mathbf{w}) \leq |P|\ln(2)\}\,.$$

If we show that $\tilde{W}$ is bounded, we can appeal to Theorem **??** to finish the proof that ℓ has a global minimum.

To see that $\tilde{W}$ is indeed bounded, consider any fixed $\mathbf{w} \in \tilde{W}$. Since $\mathbf{w}$ is not a separator, there exists $(\mathbf{x}_0, y_0) \in P$ such that $y_0 \mathbf{w}^\top \mathbf{x}_0 < 0$. Then

$$\begin{aligned}
&\lim_{\lambda\to\infty} \ell(\lambda \mathbf{w}) = \\
&= \lim_{\lambda\to\infty} \sum_{(\mathbf{x},y)\in P} \ln\left(1 + \exp\left(-y\left(\lambda \mathbf{w}\right)^\top \mathbf{x}\right)\right) \\
&\geq \lim_{\lambda\to\infty} \ln\left(1 + \exp\left(-\lambda y_0 \mathbf{w}^\top \mathbf{x}_0\right)\right) = \infty\,.
\end{aligned}$$

The last equality is true since $-y_0 \mathbf{w}^\top \mathbf{x}_0 > 0$. This shows that for a large enough λ, $\ell(\lambda \mathbf{w}) > |P|\ln(2)$ and so $\lambda \mathbf{w} \notin \tilde{W}$. Thus, the set $\tilde{W}$ cannot be unbounded.

Getting Started with Python

Many exercises in this course use Python notebooks. We recommend running these notebooks in the cloud using Google Colab. This way, you do not have to install anything, and you can even get a free GPU. If you prefer to work locally, follow the `python_setup_tutorial.md` provided on our GitHub repository.

The first practical exercise is a primer on NumPy, a scientific computing library for Python. You can open the corresponding notebook in Colab with this link:

colab.research.google.com/github/epfml/OptML_course/blob/master/labs/ex00/npprimer.ipynb

For computational efficiency, avoid `for`-loops in favor of NumPy's built-in commands. These commands are vectorized and thoroughly optimized and bring the performance of numerical Python code (like for Matlab) closer to lower-level languages like C.