Labs
Optimization for Machine Learning
Spring 2019

EPFL
School of Computer and Communication Sciences
Martin Jaggi
github.com/epfml/OptML_course

Problem Set 2 — Solutions (Gradient Descent)

Gradient Descent

Exercise 5. *Consider the function ℓ defined in (1.10). Prove that ℓ is convex!*

Solution: It suffices to show that the function $-\ln z_j(\mathbf{y})$ is convex for all j, with z_j as in (1.9). Using Lemma 1.13 (i) and (ii), it then follows that ℓ is convex. We compute

$$-\ln z_j(\mathbf{y}) = \ln\left(e^{y_0} + \cdots e^{y_9}\right) - y_j .$$

The first summand is a *log-sum-exp* function and therefore convex (a proof goes via the Hessian). The second summand is a linear function and therefore also convex. Hence the sum is convex by Lemma 1.13 (i).

Exercise 6. *Consider the logistic regression problem with two classes. Given a training set P consisting of datapoint and label pairs $(\mathbf{x}, y)$ where $\mathbf{x} \in \mathbb{R}^d$ and $y \in \{-1, +1\}$, we define our loss ℓ for weight vector $\mathbf{w} \in \mathbb{R}^d$ to be*

$$\ell(\mathbf{w}) = \sum_{(\mathbf{x},y)\in P} -\ln\left(z(y\mathbf{w}^\top\mathbf{x})\right) ,$$

where $z(s) = 1/(1+\exp(-s))$. This loss function is in fact a simplification of (1.10) when we only have two classes.

We say that the weight vector $\mathbf{w}$ is a separator *for P if for all $(\mathbf{x}, y) \in P$,*

$$y(\mathbf{w}^\top\mathbf{x}) \geq 0 .$$

A separator is said to be trivial *if for all $(\mathbf{x}, y) \in P$,*

$$y(\mathbf{w}^\top\mathbf{x}) = 0 .$$

For example $\mathbf{w} = 0$ is a trivial separator. Depending on the data P, there may be other trivial separators.

Prove the following statement: the function ℓ has a global minimum if and only if all separators are trivial.

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:

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

To see why the last inequality is true, observe that $-y(\mathbf{w}'^\top \mathbf{x}) \le 0$ 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. We just showed 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 $\ell(0) = |P|\ln(2)$:

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

If we show that $\tilde{W}$ is bounded, we can appeal to Theorem 1.24 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

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

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.

Exercise 9. *Suppose that we have centered observations* $(\mathbf{x}_i, y_i)$ *such that* $\sum_{i=1}^n \mathbf{x}_i = \mathbf{0}, \sum_{i=1}^n y_i = 0$. *Let* $w_0^\star, \mathbf{w}^*$ *be the global minimum of the least squares objective*

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

Prove that $w_0^\star = 0$. *Also, suppose* $\mathbf{x}_i'$ *and* y_i' *are such that for all* i, $\mathbf{x}_i' = \mathbf{x}_i + \mathbf{q}$, $y_i' = y_i + r$. *Show that* $(w_0, \mathbf{w})$ *minimizes* f *if and only if* $(w_0 - \mathbf{w}^\top \mathbf{q} + r, \mathbf{w})$ *minimizes*

$$
f'(w_o, \mathbf{w}) = \sum_{i=1}^n (w_0 + \mathbf{w}^T \mathbf{x}_i' - y_i')^2.
$$

Solution: We compute

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

since the observations are centered. Also, by the first-order characterization of optimality as by Lemma 1.17,

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

The second part follows from

$$
\begin{aligned}
f'(w_o - \mathbf{w}^\top\mathbf{q} + r, \mathbf{w}) &= \sum_{i=1}^{n} (w_0 - \mathbf{w}^\top\mathbf{q} + r + \mathbf{w}^T\mathbf{x}_i' - y_i')^2 \\
&= \sum_{i=1}^{n} (w_0 - \mathbf{w}^\top\mathbf{q} + r + \mathbf{w}^T(\mathbf{x}_i + \mathbf{q}) - (y_i + r))^2 \\
&= \sum_{i=1}^{n} (w_0 + \mathbf{w}^T\mathbf{x}_i - y_i)^2 = f(w_0, \mathbf{w}).
\end{aligned}
$$

Exercise 11. *Prove Lemma 2.5! (Operations which preserve smoothness)*

Solution: For (i), we sum up the weighted smoothness conditions for all the f_i to obtain

$$
\sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) \leq \sum_{i=1}^{m} \lambda_i f_i(\mathbf{y}) + \sum_{i=1}^{m} \lambda_i \nabla f_i(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \sum_{i=1}^{m} \lambda_i \frac{L_i}{2} \|\mathbf{x} - \mathbf{y}\|^2.
$$

As the gradient is a linear operator, this equivalently reads as

$$
f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\sum_{i=1}^{m} \lambda_i L_i}{2} \|\mathbf{x} - \mathbf{y}\|^2,
$$

and the statement follows. For (ii), we apply smoothness of f at $\mathbf{x}' = A\mathbf{x} + \mathbf{b}$ and $\mathbf{y}' = A\mathbf{y} + \mathbf{b}$ to obtain

$$
f(A\mathbf{x} + \mathbf{b}) \leq f(A\mathbf{y} + \mathbf{b}) + \nabla f(A\mathbf{x} + \mathbf{b})^\top(A(\mathbf{y} - \mathbf{x})) + \frac{L}{2} \|A(\mathbf{x} - \mathbf{y})\|^2.
$$

As $\nabla(f \circ g)(\mathbf{x})^\top = \nabla f(A\mathbf{x} + \mathbf{b})^\top A$ (chain rule (Lemma 1.8), using that $Dg(\mathbf{x}) = A$, an easy consequence of Definition 1.7). This equivalently reads as

$$
(f \circ g)(\mathbf{x}) \leq (f \circ g)(\mathbf{y}) + \nabla(f \circ g)(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{L}{2} \|A(\mathbf{x} - \mathbf{y})\|^2.
$$

The statement now follows from $\|A(\mathbf{x} - \mathbf{y})\| \leq \|A\|\|\mathbf{x} - \mathbf{y}\|$.