

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

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

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

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

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:

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

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 $\ell(\mathbf{w} + \lambda \mathbf{w}') = \ell(\mathbf{w})$.

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

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 $w_0 = 0$ 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 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

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

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^*, \mathbf{w}^* be the global minimum of the least squares objective

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

Prove that $w_0^* = 0$. 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

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

Solution: We compute

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

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

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

The second part follows from

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

Exercise 11. Prove Lemma 2.3: The quadratic function $f(\mathbf{x}) = \mathbf{x}^\top Q \mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$ is smooth with parameter $2 \|Q\|$.

Solution: As the function $\mathbf{x} \mapsto \mathbf{b}^\top \mathbf{x} + c$ is affine and hence smooth with parameter 0, it suffices by Lemma 2.5 to restrict ourselves to the case $f(\mathbf{x}) := \mathbf{x}^\top Q \mathbf{x}$.

Because Q is symmetric, $\mathbf{x}^\top Q \mathbf{y} = \mathbf{y}^\top Q \mathbf{x}$ for any \mathbf{x} and \mathbf{y} . Thus, a simple calculation shows that

$$\begin{aligned}
f(\mathbf{y}) = \mathbf{y}^\top Q \mathbf{y} &= \mathbf{x}^\top Q \mathbf{x} + 2\mathbf{x}^\top Q(\mathbf{y} - \mathbf{x}) + (\mathbf{x} - \mathbf{y})^\top Q(\mathbf{x} - \mathbf{y}) \\
&= f(\mathbf{x}) + 2\mathbf{x}^\top Q(\mathbf{y} - \mathbf{x}) + (\mathbf{x} - \mathbf{y})^\top Q(\mathbf{x} - \mathbf{y}).
\end{aligned}$$

Cauchy-Schwarz for $(\mathbf{x} - \mathbf{y})^\top Q(\mathbf{x} - \mathbf{y}) \leq \|\mathbf{x} - \mathbf{y}\| \|Q(\mathbf{x} - \mathbf{y})\|$, and using and the definition of spectral norm for $\|Q(\mathbf{x} - \mathbf{y})\| \leq \|Q\| \|\mathbf{x} - \mathbf{y}\|$ we get

$$f(\mathbf{y}) \leq f(\mathbf{x}) + 2\mathbf{x}^\top Q(\mathbf{y} - \mathbf{x}) + \|Q\| \|\mathbf{x} - \mathbf{y}\|^2,$$

Because $\|\mathbf{x} - \mathbf{y}\|^2$ vanishes as $(\mathbf{x} - \mathbf{y})$ goes to 0, differentiability of f (Definition 1.7) implies that $\nabla f(\mathbf{x})^\top = 2\mathbf{x}^\top Q$, so we further get

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{2\|Q\|}{2} \|\mathbf{x} - \mathbf{y}\|^2,$$

That is, f is smooth with parameter $2\|Q\|$.