

## Problem Set 2 — Solutions (Gradient Descent)

### Gradient Descent

**Exercise 5.** Consider the function  $\ell$  defined in (1.15). Prove that  $\ell$  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 (1.14). Using Lemma 1.18 (i) and (ii), it then follows that  $\ell$  is convex. We compute

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

The first summand is a *log-sum-exp* function and is convex (a proof goes via the Hessian [BV04, 3.1.5]). The second summand is a linear function and therefore also convex. Hence the sum is convex by Lemma 1.18 (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  $\ell$  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 (1.15) 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  $\ell$  has a global minimum if and only if all separators are trivial.

**Solution:**

**Left to right implication.** 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  $\ell$  and hence  $\ell$  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  $-\lambda y_0(\mathbf{w}'^\top \mathbf{x}_0) < 0$ .

**Right to left implication.** Now let us prove that if all separators are trivial, then  $\ell$  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  $V = \text{span}(x_1, \dots, x_n)$  be the linear span of the datapoints. We therefore have that  $V^\perp$  is exactly the set of trivial separators of  $P$ . Since every vector  $\mathbf{w} \in \mathbb{R}^d$  can be decomposed as  $\mathbf{w} = \mathbf{v} + \mathbf{u}$ , where  $\mathbf{v} \in V$ ,  $\mathbf{u} \in V^\perp$ , the previously proved property means that  $\ell(\mathbf{w}) = \ell(\mathbf{v} + \mathbf{u}) = \ell(\mathbf{v})$  and this means that

$$\inf_{\mathbf{w} \in \mathbb{R}^d} \ell(\mathbf{w}) = \inf_{\mathbf{w} \in \text{span}(x_1, \dots, x_n)} \ell(\mathbf{w}).$$

Thus without loss of generality, we can restrict ourselves to weight vectors in  $V$ . Now we define the following sublevel set  $A$ :

$$A = \{\mathbf{w} \in V : \ell(\mathbf{w}) \leq |P| \ln(2)\}.$$

$A$  is non-empty since  $\ell(0) = |P| \ln(2)$ . Hence if we show that  $A$  is bounded, then we can use Theorem 1.29 to finish the proof and show that  $\ell$  has a global minimum on  $V$  and hence on  $\mathbb{R}^d$ .

To see that  $A$  is indeed bounded we proceed by contradiction. Assume that  $A$  is not bounded, then we can construct a sequence of points  $(w_n)_{n \geq 0} \in A$  such that  $\|w_n\| \rightarrow \infty$ . Note that the sequence  $\frac{w_n}{\|w_n\|}$  lives in the compact (closed and bounded in finite dimension) set  $S = \{v \in V : \|v\| = 1\}$ , hence there exists a subsequence such that  $\frac{w_{\phi(n)}}{\|w_{\phi(n)}\|} \xrightarrow{n \rightarrow +\infty} d$  where  $d \in S$ . Since  $d \in \text{span}(x_1, \dots, x_n)$  and is not 0,  $d$  cannot be a trivial separator, by assumption it therefore cannot be a separator. Hence there must exist  $(\mathbf{x}_0, y_0) \in P$  such that  $y_0 d^\top \mathbf{x}_0 = -a < 0$ , where  $a > 0$ . Therefore  $y_0 \frac{1}{\|w_{\phi(n)}\|} w_{\phi(n)}^\top \mathbf{x}_0 \xrightarrow{n \rightarrow +\infty} -a$  and for  $n$  big enough,  $y_0 w_{\phi(n)}^\top \mathbf{x}_0 < -\frac{a}{2} \|w_{\phi(n)}\|$  and:

$$\begin{aligned}\ell(w_{\phi(n)}) &= \sum_{(\mathbf{x}, y) \in P} \ln \left( 1 + \exp \left( -y\mathbf{w}_{\phi(n)}^\top \mathbf{x} \right) \right) \\ &\geq \ln \left( 1 + \exp \left( -y_0 \mathbf{w}_{\phi(n)}^\top \mathbf{x}_0 \right) \right) \\ &\geq \ln \left( 1 + \exp \left( \frac{a}{2} \|w_{\phi(n)}\| \right) \right) \xrightarrow{n \rightarrow \infty} +\infty\end{aligned}$$

This is contradictory because we assumed  $\ell(w_n)$  bounded because  $w_n$  in  $V$ . Thus, the set  $V$  cannot be unbounded and this concludes the proof.

**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 1.22,

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

The second part follows from

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

**Exercise 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:** Note that we do not assume that  $Q$  is a positive matrix, hence  $f$  is not necessarily convex (note that the notion of smoothness is valid for any differentiable function). Straightforward computations lead to  $\nabla^2 f(x) = 2Q$  for all  $x \in \mathbb{R}^d$ . Hence, trivially,  $\|\nabla^2 f(x)\| \leq 2\|Q\|$  for all  $x \in \mathbb{R}^d$ . Hence using Theorem 1.9, we have that  $\nabla f$  is  $2\|Q\|$ -Lipschitz. To conclude we use implication (ii) $\Rightarrow$ (i) in Lemma 2.4 (note that, as explained in the lecture notes, we do not need any convexity assumption for this implication). This shows that  $f$  is smooth with parameter  $2\|Q\|$ .

## References

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. <https://web.stanford.edu/~boyd/cvxbook/>.