ML Theory — Homework 2

your NetID here

Version 1

Instructions. (Same as homework 1.)

• Everyone must submit an individual write-up.

• You may discuss with up to 3 other people. State their NetIDs clearly on the first page. Outside of office hours, you should not discuss with anyone but these three.

• Homework is due Wednesday, November 29, at 3:30pm; no late homework accepted.

• Please consider using the provided LaTeX file as a template.
1. **Dual norms.**

Recall that for any norm \( \| \cdot \| \), there is also a dual norm
\[
\| s \|_* = \sup \{ \langle s, v \rangle : \| v \| \leq 1 \}.
\]

You may assume this is a valid norm without proof. For this problem, suppose vectors lie in \( \mathbb{R}^d \), but norm duality works beyond that.

**Note:** in all parts of this problem, assume a general norm and dual-norm pair! Do not assume \( l_2 \) norm!

(a) Prove \( |\langle s, v \rangle| \leq \| v \| \cdot \| s \|_* \) (a generalized Hölder inequality).

(b) Suppose \( f : \mathbb{R}^d \to \mathbb{R} \) has \( \beta \)-Lipschitz gradients wrt \( \| \cdot \| \), meaning
\[
\| \nabla f(x) - \nabla f(y) \|_* \leq \beta \| x - y \|.
\]
(Gradients live in dual space, get dual norm.) Prove
\[
\left| f(x + v) - f(x) - \langle \nabla f(x), v \rangle \right| \leq \frac{\beta}{2} \| v \|^2.
\]

(Major hint: repeat the integral calculation for \( \| \cdot \|_2 \) from lecture 12.)

(c) Suppose \( f \) is \( \beta \)-smooth wrt \( \| \cdot \| \) as above, and suppose the gradient descent iteration is replaced with the steps
\[
v := \arg \min \left\{ \langle \nabla f(w), v \rangle : \| v \| \leq 1 \right\}, \quad w' := w - v\|\nabla f(w)\|_*/\beta.
\]

Show that
\[
f(w') \leq f(w) - \|\nabla f(w)\|_*^2/(2\beta).
\]

(d) Suppose that \( f \) is \( \lambda \) strongly convex wrt \( \| \cdot \| \), meaning
\[
f(w + v) \geq f(w) + \langle \nabla f(w), v \rangle + \frac{\lambda}{2} \| v \|^2.
\]

Prove that a minimizer \( \bar{w} \) exists, is unique, and for any \( w \)
\[
f(\bar{w}) \geq f(w) - \frac{\|\nabla f(w)\|_*^2}{2\lambda}.
\]
(You may assume without proof that convex functions over \( \mathbb{R}^d \) are continuous.)

(e) Suppose that \( f \) is not only \( \beta \)-smooth wrt \( \| \cdot \| \) as above, but moreover it is \( \lambda \) strongly convex wrt \( \| \cdot \| \). Suppose \( (w_i)_{i \leq t} \) are given by the generalized gradient descent iteration in eq. (1). Show that
\[
f(w_t) - f(\bar{w}) \leq (f(w_0) - f(\bar{w})) \exp \left(-t\lambda/\beta \right),
\]
where \( \bar{w} \) is a unique minimizer (as established in the previous part).

**Solution.**

(Your solution here.)
2. Hölder’s inequality.

Let \((p, q)\) be conjugate exponents: \(p \geq 1, q \geq 1,\) and \(1/p + 1/q = 1\) (where \(p = \infty\) or \(q = \infty\) are allowed with the convention \(1/\infty = 0\)).

(a) Let convex function \(f : \mathbb{R} \to \mathbb{R} \cup \infty\) be given, and define \(g(v) := \sum_{i=1}^{d} f(v_i).\) Prove \(g^*(s) := \sum_{i=1}^{d} f^*(s_i),\) and \(s \in \partial g(v)\) iff \(s_i \in \partial f(v_i)\) for each \(i\).

(b) Given any \(p \in [1, \infty],\) define \(f_p : \mathbb{R}^d \to \mathbb{R}\) as
\[
f_p(v) := \begin{cases} \frac{1}{p}\|v\|_p^p = \frac{1}{p} \sum_{i=1}^{d} |v_i|^p & \text{when } p \in [1, \infty), \\ \iota_{[-1, +1]}^p(v) & \text{when } p = \infty. \end{cases}
\]

Show that for any conjugate exponents \((p, q)\), then \(f_p^* = f_q;\) moreover, \(s \in \partial f_p(v)\) iff \(s_i \in \partial g_p(v)\), where \(g_p : \mathbb{R} \to \mathbb{R} \cup \infty\) is the univariate version of \(f_p,\) and \(g_p\) has the following subgradient structure.

- If \(p = 1,\) then
  \[
  \partial g_p(v) = \begin{cases} \{-1\} & v < 0, \\ \{-1, +1\} & v = 0, \\ \{+1\} & v > 0. \end{cases}
  \]
- If \(p \in (1, \infty),\) then \(\partial g_p(v) = |v|^{p-1}\text{sgn}(v).\)
- If \(p = \infty,\) then
  \[
  \partial g_p(v) = \begin{cases} \emptyset & \text{when } v \notin [-1, +1], \\ (-\infty, 0] & v = -1, \\ \{0\} & v \in (-1, +1), \\ [0, \infty) & v = +1. \end{cases}
  \]

(c) Show \(f_p\) is convex.

(d) Suppose \((p, q)\) are conjugate exponents. Show \(|\langle u, v \rangle| \leq \|u\|_p \|v\|_q,\) with equality iff \(v \in \partial f_p(u).\)

\textbf{Hint:} Fenchel-Young; consider cases \(u = 0, \|u\|_p = 1, \ldots\)

\textbf{Remark:} this is Hölder’s inequality.

(e) Suppose \((p, q)\) are conjugate exponents. Prove \((\| \cdot \|_p)_* = \| \cdot \|_q.\)

\textbf{Solution.}

\textit{(Your solution here.)}
3. Convexity odds and ends.

(a) Let $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a convex loss, and fix any distribution on $(x, y)$; consider our familiar setting of risk minimization for linear functions, meaning $f(w) := \mathbb{E}\ell(\langle w, -xy \rangle)$. Show that given a random draw $(x, y)$ and any $g \in \partial\ell(\langle w, -xy \rangle)$, then $\mathbb{E}(-xyg) \in \partial f(w)$.

Remark: this problem justifies the choice of stochastic gradient descent used in practice.

(b) Define $f(x) := x^\top Qx/2$, where $Q \in \mathbb{R}^{d \times d}$ a symmetric positive definite matrix. Derive an explicit form for $f^*$, and then provide a rigorous proof. (Hint: try to guess the answer first.)

(c) Define $f(x) := x^\top Qx/2$ once again, but now $Q \in \mathbb{R}^{d \times d}$ is merely symmetric positive semi-definite. Derive a (new) explicit form for $f^*$, and provide a rigorous proof. (Hint: there are many ways here, but again try to guess the answer, and in times of great need never forget your friend S-V-D.)

Solution.

(Your solution here.)

Let \( f : \mathbb{R}^d \rightarrow \mathbb{R}^k \) denote the function computed by a neural network; note the output space has \( k \) dimensions for \( k \) classes.

The standard loss is the cross entropy loss: given an example \((x, y)\) with \( x \in \mathbb{R}^d \) and \( y \in \{1, \ldots, k\} \), the loss is

\[
- \ln(f(x)_y);
\]

similarly, the risk can be defined.

Networks usually have the softmax \( \sigma_{sm} : \mathbb{R}^k \rightarrow \mathbb{R}^k \) as the final activation; the softmax is defined per-coordinate as

\[
\sigma_{sm}(v)_i := \frac{e^{v_i}}{\sum_j e^{v_j}}.
\]

Composing this with the cross entropy loss yields the modified cross entropy loss

\[
\ell(f(x), y) := - \ln(\sigma_{sm}(f(x))_y).
\]

(a) Prove \( g(v) := \ln \sum_i \exp(v_i) \) is convex.

(b) For any linear operator \( A \) and convex function \( g, g \circ A \) is convex.

(c) Let data \(((x_i, y_i))_{i=1}^n\) be given. Show that the modified cross-entropy risk

\[
\mathcal{R}_\ell(W) := \frac{1}{n} \sum_{i=1}^n \ell(Wx_i, y_i)
\]

is convex in \( W \in \mathbb{R}^{k \times d} \).

(Note: if you’re not comfortable with matrix variables, just unroll it into a vector and appropriately re-define \( Wx_i \), etc.)

(d) Define the logistic loss \( \ell_{log}(z) := \ln(1 + \exp(z)) \), and let matrix \( W \in \mathbb{R}^{k \times d} \) be given. Find a vector \( v \in \mathbb{R}^2 \) so that for any \( x \in \mathbb{R}^d, y \in \{1, 2\} \), and \( \tilde{y} = 2y - 3 \in \{-1, +1\} \),

\[
\ell(Wx, y) = \ell_{log}(\langle W^T v, -x\tilde{y} \rangle).
\]

(Include a rigorous derivation!)

Remark: this shows that logistic loss is equivalent to binary cross-entropy.

Solution.

(Your solution here.)
5. Conjugates of key losses.

Derive and then provide rigorous proofs for the conjugates of the following loss functions.

(a) Squared loss: \( \ell(z) := (1 + z)^2 / 2 \).
(b) Hinge loss: \( \ell(z) := \max\{0, 1 + z\} \).
(c) Logistic loss: \( \ell(z) := \ln(1 + \exp(z)) \).
(d) Exponential loss: \( \ell(z) := \exp(z) \).
(e) Impagliazzo/Zhang loss:

\[
\ell(z) := \begin{cases} 
0 & \text{when } z < -1, \\
(1 + z)^2 / 2 & \text{when } z \in [-1, +1], \\
2z & \text{when } z > 1.
\end{cases}
\]

Solution.

(Your solution here.)
6. Frank-Wolfe.
Recall the Frank-Wolfe method from lecture 15 and its associated notation: there is a bounded closed convex constraint set $S$, it has diameter $D := \sup_{x, y \in S} \|x - y\|$, and the iterates are defined via $w_0 \in S$ (arbitrary) and thereafter
\[
v_i := \arg\min_{v \in S} \langle \nabla f(w_{i-1}), v \rangle, \quad w_i := (1 - \eta_i)w_{i-1} + \eta_i v_i.
\]
Lastly, suppose $f$ is $\beta$-smooth.

(a) Suppose the lecture’s step sizes are replaced with $\eta_i := 1/(i + 1)$. Show that for every $t \geq 1$ and $z \in S$,
\[
f(w_t) - f(z) \leq \frac{\beta D^2 (1 + \ln(t))}{2t}.
\]
Remark: notice that something goes wrong if you instead pick $\eta_i := 1/(t + 1)$.

(b) Coming soon!!!

(c) Coming soon!!!

Solution.
(Your solution here.)