ML Theory — Homework 2

your NetID here

Version 3

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 28, at 3:00pm**; no late homework accepted.
- Please consider using the provided LaTeX file as a template.

1. **(Miscellaneous short questions.)**

   (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.

   Recall: the subgradient ∂h is defined as

$$\partial h(w) = \left\{ s \in \mathbb{R}^d \ : \ \forall v \in \mathbb{R}^d \centerdot h(v) \geq h(w) + \langle s, v - w\rangle \right\}.$$

   (b) Suppose $\Phi : \mathbb{R}^d \to \mathbb{R}$ is λ-strongly-convex (λ-sc) and differentiable, and define the *Bregman divergence*

$$D_\Phi(x, y) := \Phi(x) - \Big(\Phi(y) + \big\langle \nabla\Phi(y), x - y\big\rangle\Big).$$

   Prove that D_Φ is λ-sc in its first argument.

   (**Remark.** What about the second argument? Does a weaker property hold?)

   (c) Once again let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be λ-sc. Recall the definition of *Fenchel conjugate* $\Phi^*(s) := \sup_{x\in\mathbb{R}^d} \langle x, s\rangle - \Phi(s)$.

   The update rule of mirror descent may be written

$$w' := \arg\min_v \eta \left\langle \nabla f(w), v\right\rangle + D_\Phi(v, w).$$

   Prove this is equivalent to

$$w'' := \nabla\Phi^* \left(\nabla\Phi(w) - \eta\nabla f(w)\right).$$

   Hint: since Φ is strongly convex, then $(\nabla\Phi)^{-1}$ exists and is equal to $\nabla\Phi^*$ (you may use this without proof).

   (d) Suppose $Q \in \mathbb{R}^{d\times d}$ is symmetric positive definite, let $b \in \mathbb{R}^d$ be arbitary, and define $f(x) := \frac{1}{2}x^\top Qx + b^\top x$. Using direct computation (and not the preceding inverse gradient gradient fact), derive the Fenchel conjugate f^*, and prove it is correct.

   (e) Now suppose $Q \in \mathbb{R}^{d\times d}$ is merely symmetric positive *semi-definite* (it may fail to have an inverse), $b \in \mathbb{R}^d$ is again arbitrary, and define $f(x) := \frac{1}{2}x^\top Qx + b^\top x$. Derive the Fenchel conjugate f^*, and prove it is correct.

   (f) Freedman's inequality (Bernstein's inequality for martingales) implies: given martingale difference sequence $(Z_i)_{i=1}^n$ with $|Z_i| \leq b$ and $\sum_i \mathbb{E}(Z_i^2|Z_{<i}) \leq v$, then with probability at least $1 - \delta$,

$$\sum_i Z_i \leq \sqrt{2v\ln(1/\delta)} + \frac{b\ln(1/\delta)}{3}.$$

   Consider the setting of the theorem in Lecture 15, but additionally $\mathbb{E}(\|g_i - s_i\|^2 \mid w_{i-1}) \leq \sigma^2$, and that for any given w_{i-1} it is possible to obtain an arbitrary number of mutually conditionally independent stochastic gradients g_i with all stated properties.

   Use all these assumptions together with the above version of Freedman's inequality to provide a refinement of the theorem in Lecture 15.

   (g) Consider the setting of the previous part, but suppose a minibatch of size b is used (b conditionally independent stochastic gradients are averaged together for each step). State the optimal values of step size η and batch size b by optimizing the right hand side of the previous bound.

Solution.

(Your solution here.)

2. **(Dual norms.)**

Recall that for any norm $\|\cdot\|$, there is also a *dual norm*

$$\|s\|_* = \sup\left\{\langle s, v\rangle : \|v\| \le 1\right\}.$$

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

(b) Suppose $f : \mathbb{R}^d \to \mathbb{R}$ has β-Lipschitz gradients wrt $\|\cdot\|$, meaning

$$\|\nabla f(x) - \nabla f(y)\|_* \le \beta\|x - y\|.$$

(Gradients live in dual space, get dual norm.) Prove

$$\Big|f(x+v) - f(x) - \big\langle \nabla f(x), v\big\rangle\Big| \le \frac{\beta}{2}\|v\|^2.$$

(**Major hint:** repeat the integral calculation for $\|\cdot\|_2$ from lecture 11.)

(c) Suppose f is β-smooth wrt $\|\cdot\|$ as above, and suppose the gradient descent iteration is replaced with the steps

$$v := \arg\max\Big\{\big\langle \nabla f(w), v\big\rangle : \|v\| \le 1\Big\}, \qquad w' := w - v\|\nabla f(w)\|_*/\beta. \tag{1}$$

Show that

$$f(w') \le f(w) - \|\nabla f(w)\|_*^2/(2\beta).$$

(d) Suppose that f is λ strongly convex wrt $\|\cdot\|$, meaning

$$f(w+v) \ge f(w) + \big\langle \nabla f(w), v\big\rangle + \frac{\lambda}{2}\|v\|^2.$$

Prove that a minimizer $\bar{w}$ exists, is unique, and for any w

$$f(\bar{w}) \ge f(w) - \frac{\|\nabla f(w)\|_*^2}{2\lambda}.$$

(You may assume without proof that convex functions over $\mathbb{R}^d$ are continuous, and that continuous functions over $\mathbb{R}^d$ attain minima and maxima over closed bounded sets.)

(e) Suppose that f is not only β-smooth wrt $\|\cdot\|$ as above, but moreover it is λ strongly convex wrt $\|\cdot\|$. Suppose $(w_i)_{i\le t}$ are given by the generalized gradient descent iteration in eq. (1). Show that

$$f(w_t) - f(\bar{w}) \le \big(f(w_0) - f(\bar{w})\big)\exp\big(-t\lambda/\beta\big),$$

where $\bar{w}$ is a unique minimizer (as established in the previous part).

Solution.

(Your solution here.)

3. **(Frank-Wolfe.)**

Recall the Frank-Wolfe method from lecture 13 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

$$v_i := \arg\min_{v \in S} \left\langle \nabla f(w_{i-1}), v \right\rangle, \qquad w_i := (1 - \eta_i) w_{i-1} + \eta_i v_i.$$

Lastly, suppose f is convex and β-smooth.

(a) Suppose the lecture's step sizes are replaced with $\eta_i := 1/i$. 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 $\eta_i := 1/t$.

(b) **(Optional.)** Define

$$G(w) := \begin{cases} \infty & w \notin S, \\ \sup_{v \in S} \left\langle \nabla f(w), w - v \right\rangle & w \in S. \end{cases}$$

Prove $f(w) - \inf_{v \in S} f(v) \leq G(w)$ for all w.

Note: there are various ways to prove this with strong duality laws; you can for instance use the two omitted convexity lectures.

(c) Using the definition of G, the guarantee in the previous part, and steps from the proof of the Frank-Wolfe iteration guarantee: prove that for any i,

$$\eta_{i+1} G(w_i) \leq f(w_i) - f(w_{i+1}) + \frac{\beta \eta_{i+1}^2 D^2}{2}.$$

(d) In lecture, we've mentioned that in general we don't have a good way to stop convex programs. The Frank-Wolfe method, on the other hand, admits a nice stopping rule. Consider the following adjusted definition of the method.

i. Let $w_0 \in S$ and $\epsilon > 0$ be given.

ii. For $i \in \{1, 2, \ldots\}$:

A. $v_i := \arg\min_{v \in S} \left\langle \nabla f(w_{i-1}), v \right\rangle$.

B. **Return** w_{i-1} if $\left\langle \nabla f(w_{i-1}), w_{i-1} - v_i \right\rangle \leq \epsilon$.

C. $w_i := (1 - \eta_i) w_{i-1} + \eta_i v_i$ where $\eta_i := 2/(i+1)$.

Prove the method terminates with output w_{t-1} where

$$t \leq \frac{128 \beta D^2}{\epsilon} \qquad \text{and} \qquad f(w_{t-1}) - \inf_{v \in S} f(v) \leq G(w_{t-1}) \leq \epsilon.$$

Note: the '128' should give you some wiggle room.

Hint: use the previous part, and also the iteration guarantee from lecture. Divide the iterate sequence into two halves, and reason about each half differently.

Solution.

(Your solution here.)

4. **(Cross entropy.)**

Let $f : \mathbb{R}^d \to \mathbb{R}^k$ 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* $\sigma_{\text{sm}} : \mathbb{R}^k \to \mathbb{R}^k$ as the final activation; the softmax is defined per-coordinate as $\sigma_{\text{sm}}(v)_i := e^{v_i} / \sum_j e^{v_j}$. Composing this with the cross entropy loss yields the *modified cross entropy loss*

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

(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 $\tilde{y} = 2y - 3 \in \{-1, +1\}$,

$$\ell(Wx, y) = \ell_{\log}(\left\langle W^\top v, -x\tilde{y} \right\rangle).$$

(Include a rigorous derivation!)

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

Solution.

(Your solution here.)

5. **(Max of random variables; moment generating functions.)**

An important object in the study of random variables is the moment generating function (MGF), $M_X(t)$, defined as $M_X(t) := \mathbb{E}(\exp(tX))$. ($M_X$ will in general fail to be finite for all $t \geq 0$, but in this question it is finite for all $t \geq 0$.)

Given a family $(X_i, \ldots, X_d)$ of i.i.d. random variables drawn according to some distribution, this question will investigate the behavior of the random variable $Z := \|(X_1, \ldots, X_d)\|_\infty = \max_i |X_i|$.

(a) Prove the following inequality, which will be convenient in the remainder of the question: for any $t > 0$,

$$\mathbb{E}(Z) \leq \frac{1}{t} \ln \Big(d \cdot \mathbb{E} \left(\exp(tX_1) + \exp(-tX_1) \right) \Big).$$

Note. You will want to use *Jensen's inequality*, namely $\mathbb{E}\left(\ln(f(X))\right) \leq \ln\left(\mathbb{E}f(X)\right)$.

(b) **(Optional.)** Suppose X_1 distributed according to a Gumbel distribution with scale parameter σ, whereby $\mathbb{E}(\exp(sX_1)) = \Gamma(1 - s\sigma)$ for all $s \in \mathbb{R}$, where Γ denotes the gamma function. Prove that

$$\mathbb{E}(Z) \leq 2\sigma \ln(d\sqrt{\pi}).$$

Hint: the inequality from the first part holds for all t... can you find a particularly nice choice of t?

(c) Prove that Gaussian distribution is *subgaussian*: in particular, if X_1 is Gaussian with mean 0 and variance σ^2, then $\mathbb{E}(\exp(tX_1)) = \exp(t^2\sigma^2/2)$ for every $t \in \mathbb{R}$.

(d) Prove that if X_1 is *subgaussian with variance proxy* σ^2, meaning $\mathbb{E}(\exp(tX_1)) \leq \exp(t^2\sigma^2/2)$ for every $t \in \mathbb{R}$, then

$$\mathbb{E}(Z) \leq \sigma\sqrt{2\ln(2d)}.$$

(Together with the preceding part, this implies the bound for X_1 a Gaussian with mean 0 and variance σ^2.)

(e) Was it necessary to assume $(X_1, \ldots, X_d)$ were i.i.d.? Answer this question however you like.

When the dust has settled, I urge you to ponder the power of this modest little technique of replacing max with $\ln \sum \exp$.

Solution.

(Your solution here.)