CSE 598 TEL - Homework 2

MJT

November 22, 2016

New instructions!

- **Everyone** must make an individual write-up and hand-in this time!
- 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 **Friday, November 18, at 11:00am**; no late homework accepted.
- Grading on this homework will be different: each part of each question will count for 1 point!
- Please consider using the provided LaTeX file as a template (apologies for its weird indentation).
- Policies on the webpage still apply (where not in conflict with the policies above).

1. **(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.

2. **(Baby representer theorem.)**

This question will establish the baby representer theorem from lecture.

Theorem 1 (Baby Representer Theorem)**.** *Let hinge loss $\ell(z) := \max\{0, 1+z\}$ be given, and define $\ell_i(v) := \ell(v_i)$. Let matrix $A \in \mathbb{R}^{m\times d}$, real scalar $\lambda > 0$ be given, and integer $n \leq m$ be given. Then*

$$\min\left\{\sum_{i=1}^{n} \ell_i(-Aw) + \frac{\lambda}{2}\|w\|_2^2 : w \in \mathbb{R}^d\right\} = \max\left\{\sum_{i=1}^{n} s_i - \frac{1}{2\lambda}\|A^\top s\|_2^2 : s \in [0,1]^n \times \{0\}^{m-n}\right\}.$$

Primal-dual optimal pairs $(\bar{w}, \bar{s})$ always exist. $\bar{w}$ is unique and has the form $\bar{w} = A^\top \bar{s}/\lambda$ for any optimal $\bar{s}$, and $\bar{s}$ is optimal iff $A^\top \bar{s} = \lambda \bar{w}$ and

$$\bar{s}_i \in \begin{cases} \{0\} & i > n, \\ \{0\} & i \leq n, (A\bar{w})_i > 1, \\ [0,1] & i \leq n, (A\bar{w})_i = 1, \\ \{1\} & i \leq n, (A\bar{w})_i < 1. \end{cases}$$

The proof proceeds in the following steps.

(a) Show that the primal optimum exists and is unique. (**Hint.** You may use without proof the following analysis fact: continuous functions attain minima over compact sets. To complete the proof from here, dig around in the convexity lectures for a proof that roughly says regularization implies bounded sublevel sets.)

(b) Show that given $g : \mathbb{R}^m \to \mathbb{R}$ defined as $g(v) := \sum_{i=1}^n f(v_i)$, then $g^*(s) = \sum_{i=1}^n f^*(s_i) + \sum_{i>n} \iota_{\{0\}}(s_i)$ and $s \in \partial g(v)$ iff $s_i \in \partial f(v_i)$ for $i \leq n$ and $s_i = 0$ for $i > n$.

(c) Show that if $f(x) = g(cx)$ for $c \neq 0$, then $f^*(s) = g^*(s/c)$.

(d) Show that if $f(x) = cg(x)$ for $c \neq 0$, then $f^*(s) = cg^*(s/c)$.

(e) Now prove the theorem via (in part) the preceding steps, the form of ℓ^* from an earlier problem, and some other convexity properties from lecture.

Solution.

3. **(Bregman projection.)**

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and λ-strongly-convex for some $\lambda > 0$. Define the *Bregman divergence* B_f as

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

Next, let a convex closed nonempty set $S \subseteq \mathbb{R}^d$ be given. Define the *Bregman projection* Π_S^f as

$$\Pi_S^f(y) := \operatorname*{arg\,min}_{x \in S} B_f(x, y).$$

Establish the following.

(a) Show $B_f(x, y) \geq \lambda \|x - y\|_2^2/2$. (This is effectively by definition.)

(b) Show $B_f(x, y) = B_{f^*}(\nabla f(y), \nabla f(x))$. (Note this is implicitly assuming f^* is differentiable; this is true, a consequence of strict convexity of f, but was not proved in class.)

(c) Show Π_S^f is a well-defined function, meaning $\min_{x \in S} B_f(x, y)$ is attained uniquely for every $y \in \mathbb{R}^d$. (**Hint.** You may assume f is continuous (which is true for every convex function which is finite over $\mathbb{R}^d$), and then use then use the attainment hint from the previous problem.)

(d) Let $y \in \mathbb{R}^d$ be given, and set $x := \Pi_S^f(y)$. Show that for any $z \in S$,

$$\big\langle \nabla f(y) - \nabla f(x), z - x \big\rangle \leq 0.$$

(e) Let $y \notin S$ be given, and set $x := \Pi_S^f(y)$. Define hyperplane

$$H_y := \Big\{ v \in \mathbb{R}^d : \big\langle \nabla f(y) - \nabla f(x), v - x \big\rangle = 0 \Big\}.$$

Show that $x = \Pi_{H_y}^f(y)$.

(f) Let $y \in \mathbb{R}^d$ and $z \in S$ be given. Then

$$B_f(z, y) \geq B_f(z, \Pi_S^f(y)) + B_f(\Pi_S^f(y), y) \geq B_f(z, \Pi_S^f(y)).$$

Additionally show the first inequality is an equality when S is affine.

(g) **(Optional.)** What breaks if we weaken f to only strict convexity or just convexity?

(h) **(Optional.)** Interpret the above properties when $f(x) := \|x\|_2^2/2$. Personally, I think about this in terms of pictures, with some right and obtuse angles.

Solution.

4. **(Least squares.)**

Consider random variables (X, Y) with $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$, and suppose each is bounded with probability 1. (Boundedness is not necessary, but helps with some technicalities.) Define objective function

$$f(w) := \frac{1}{2}\mathbb{E}\left((X^\top w - Y)^2\right).$$

Set $M := \mathbb{E}(XX^\top)$, and henceforth let $M = U\Sigma V^\top$ denote the SVD of M; as further notation, take $U \in \mathbb{R}^{d\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$, $V \in \mathbb{R}^{r\times d}$, where r is the rank of M (assumed to be positive). Define $S := \operatorname{span}(\{u_1, \dots, u_r\})$, and recall pseudoinverse $A^\dagger := V\Sigma^{-1}U^\top$.

(a) Establish the *normal equations*: $\bar{w} \in \mathbb{R}^d$ minimizes f iff

$$M\bar{w} = \mathbb{E}(XX^\top)\bar{w} = \mathbb{E}(XY).$$

(Hint: first order conditions!)

(b) Recall the definition $B_f(w, v) := f(w) - \left(f(v) + \left\langle \nabla f(v), w - v\right\rangle\right)$ from the previous problem. First show for any $v, w \in \mathbb{R}^d$ that

$$B_f(w, v) = B_f(v, w) = \|w - v\|_M^2/2$$

where $\|\bar{w}-v\|_M^2 = \left\langle \bar{w} - v, M(\bar{w} - v)\right\rangle$. Additionally, if $\bar{w}$ is optimal, show $B_f(v, \bar{w}) = f(v)-f(\bar{w}) = \|\bar{w} - v\|_M^2/2$.

(c) Prove M is s-psd and $U = V$. (**Easy mode:** assume the SVD is unique.)

(d) Prove $X \in S$ with probability 1.

(e) Prove that $\bar{w} := M^\dagger\mathbb{E}(XY)$ minimizes f.

(f) Suppose $Y := X^\top\tilde{w} + \varepsilon$, where $\tilde{w} \in \mathbb{R}^d$ is a fixed deterministic vector, and ε is independent of X and satisfies $\mathbb{E}(\varepsilon) = 0$. Then $\bar{w}$ is optimal iff satisfies $M\bar{w} = M\tilde{w}$.

(g) Prove $\bar{w}$ is unique iff M is full rank.

(h) **(Optional.)** In the context of the preceding two problems, M full rank means $\bar{w} = \tilde{w}$. Write something about the consequences of this on the whole "recovery vs prediction" discussion we had in the clustering lecture.

Solution.