Homework 5
Convex Optimization 10-725
Due Friday, November 30 at 11:59pm

Submit your work as a single PDF on Gradescope. Make sure to prepare your solution to each problem on a separate page. (Gradescope will ask you select the pages which contain the solution to each problem.)

Total: 65 points

1 Exponential families and convexity (24 points)

In this problem, we’ll study convexity (and concavity) in exponential families and generalized linear models. Consider an exponential family density (or probability mass) function over \( y \in D \subseteq \mathbb{R}^n \), of the form

\[
f(y; \theta) = \exp \left( y^T \theta - b(\theta) \right) f_0(y).
\] (1)

Note \( \theta \in \mathbb{R}^n \) is called the natural parameter in this family.

1. (6 pts) Prove that \( b : C \to \mathbb{R} \) is a convex function, where \( C = \text{dom}(b) \). Hint: use the fact that \( f(y; \theta) \) is a density (or probability mass) function to derive an expression for \( b(\theta) \).

2. (2 pts) Assume that \( \theta_i = x_i^T \beta \) for each \( i = 1, \ldots, n \), where \( x_i \in \mathbb{R}^p \) are predictor measurements (considered fixed, i.e., nonrandom) and \( \beta \in \mathbb{R}^p \) is a coefficient vector. Prove that the domain of \( \beta \), \( B = \{ \beta : (x_1^T \beta, \ldots, x_n^T \beta) \in C \} \), is a convex set.

3. (3 pts) Write down the log likelihood function \( \ell(\beta; Y) \) for a random vector \( Y \in \mathbb{R}^n \) drawn from the distribution in (1). Prove that maximizing this log likelihood over \( \beta \in B \) is a concave maximization problem, i.e., a convex optimization problem.

Note: taking \( \theta_i = x_i^T \beta, i = 1, \ldots, n \) as we’ve done is the same as considering a generalized linear model with canonical link function. What you’ve just shown: maximum likelihood in any generalized linear model (with canonical link) is a convex optimization problem.

4. (4 pts) Argue that when \( b(\theta) = \|\theta\|_2^2/2 \), the maximum log likelihood problem is the same as linear regression, and that when \( b(\theta) = \sum_{i=1}^n \log(1 + \exp(\theta_i)) \), it is the same as logistic regression.

5. (9 pts) Argue whether or not each of the following regularized maximum likelihood problems is a convex optimization problem, as written. Your justifications can be one line (or less). Below, \( \lambda, t, k \geq 0 \) are all constants.

(a) \( \max_{\beta \in B} \ell(\beta) - \lambda \|\beta\|_1 \)
(b) \( \max_{\beta \in B} \ell(\beta) \) subject to \( \beta_1 \geq 0, \ldots, \beta_p \geq 0 \)
(c) \( \max_{\beta \in B} \ell(\beta) \) subject to \( \beta^T Q \beta = t \), for a matrix \( Q \succeq 0 \)
(d) max_{β ∈ B} ℓ(β) subject to ∥β∥_2 ≤ t
(e) max_{β ∈ B} ℓ(β) − λ log \sum_{i \neq j} \exp(β_i − β_j)
(f) max_{β ∈ B} ℓ(β) subject to max_{i=1,...,p-1} |β_i − β_{i+1}| ≤ t
(g) max_{β ∈ B} ℓ(β) subject to max_{α:∥α∥_0≤k} ∥β − α∥_2 ≤ t
(h) max_{β ∈ B} ℓ(β) subject to β_1A_1 + ... + β_pA_p ≥ 0, for symmetric matrices A_1, ... A_p

2 Subgradients, conjugates, and duality (24 points)

Let f be a closed and convex function, and let f^* its conjugate. Recall that for a linear map A, the problem
\[ \min_x f(x) + g(Ax) \] (2)
has a dual problem
\[ \max_y -f^*(-A^Ty) - g^*(y). \] (3)

Suppose that g is convex and has a known proximal operator
\[ \text{prox}_{g,t}(x) = \text{argmin}_z \frac{1}{2t} \|x - z\|^2_2 + g(z). \]

Note that this does not necessarily mean that we know the proximal operator for h(x) = g(Ax). Therefore we cannot easily apply proximal gradient descent to the primal problem (2). However, as you will show in the next few parts, knowing the proximal mapping of g does lead to the proximal mapping of g^*, which leads to an algorithm on the dual problem (3).

1. (5 pts) Show that
\[ y ∈ \partial f(x) ⇐⇒ x ∈ \partial f^*(y). \]
   Hint: show that y ∈ \partial f(x) ⇒ x ∈ \partial f^*(y) by using the rule for subgradients of a maximum of functions. Then apply what you know about f'' for closed, convex f to show the converse.

2. (4 pts) Assume henceforth that f is strictly convex. Show that this implies f^* is differentiable, and that
\[ \nabla f^*(y) = \text{argmin}_x f(x) - y^T x. \]
   Hint: use part 1.

3. (5 pts) Prove that
\[ \text{prox}_{g,1}(x) + \text{prox}_{g^*,1}(x) = x, \]
for all x. This is called Moreau’s theorem. Note the specification t = 1 in the above. Hint: again use part 1.

4. (5 pts) Verify that for t > 0, we have (tg)^*(x) = tg^*(x/t). Use this, and part 3, to prove that for any t > 0,
\[ \text{prox}_{g,t}(x) + t \cdot \text{prox}_{g^*,1/t}(x/t) = x, \]
for all x. Hint: apply part 3 to the function tg. Then note \text{prox}_{g,t}(x) = \text{prox}_{tg,1}(x), and the same for g^*.

5. (5 pts) Lastly, write down a proximal gradient descent algorithm for the dual problem (3). Use parts 2 and 4 of this question to express all quantities in terms of f and g. That is, your proximal gradient descent updates should not have any appearances of \nabla f^* or \text{prox}_{g^*,1}(·).
Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, consider the regularized least squares program

$$
\min_{w \in \mathbb{R}^p} \frac{1}{2} \|y - Xw\|^2 + \sum_{i=1}^d h_i(w_i), \tag{4}
$$

where $w = (w_1, \ldots, w_d)$ is a block decomposition with $w_i \in \mathbb{R}^{p_i}$, $i = 1, \ldots, d$, and where $h_i$, $i = 1, \ldots, d$ are convex functions. Let $X_i \in \mathbb{R}^{n \times p_i}$, $i = 1, \ldots, d$ be a corresponding block decomposition of the columns of $X$, and $g(w) = \|y - Xw\|^2/2$.

1. (4 pts) Consider coordinate descent, which repeats the following updates:

$$
w_i^{(k)} = \arg\min_{w_i \in \mathbb{R}^{p_i}} \frac{1}{2} \|y - \sum_{j<i} X_j w_j^{(k)} - \sum_{j>i} X_j w_j^{(k-1)} - X_i w_i\|^2 + h_i(w_i), \quad i = 1, \ldots, d, \tag{5}
$$

for $k = 1, 2, 3, \ldots$. Consider also coordinate proximal gradient descent, which repeats:

$$
w_i^{(k)} = \prox_{h_i, t_i g_i} \left( w_i^{(k-1)} - t_k \nabla g_i(w_1^{(k)}, \ldots, w_{i-1}^{(k)}, w_i^{(k-1)}, \ldots, w_d^{(k-1)}) \right), \quad i = 1, \ldots, d, \tag{6}
$$

for $k = 1, 2, 3, \ldots$. Assume we initialize these algorithms at the same point. Show that when each $p_i = 1$ (all coordinate blocks are of size 1), under appropriate step sizes for coordinate proximal gradient descent, these two methods are exactly the same. (Assume each $X_i \neq 0$.)

2. (2 pts) When at least one $p_i > 1$, give an example to show that these two methods are not the same, for any choice of step sizes in coordinate proximal gradient descent.

3. (3 pts) Assume henceforth that $h_i$, $i = 1, \ldots, d$ are each support functions

$$
h_i(v) = \max_{u \in D_i} \langle u, v \rangle, \quad i = 1, \ldots, d.
$$

where $D_i \subseteq \mathbb{R}^{p_i}$, $i = 1, \ldots, d$ are closed, convex sets. Show that the dual of (4) is what is sometimes called the best approximation problem

$$
\min_{u \in \mathbb{R}^n} \|y - u\|^2_2 \quad \text{subject to} \quad u \in C_1 \cap \cdots \cap C_d. \tag{7}
$$

where each $C_i = (X_i^T)^{-1}(D_i) \subseteq \mathbb{R}^n$, the inverse image of $D_i$ under the linear map $X_i^T$. Show also that the relationship between the primal and dual solutions $w, u$ is

$$
u = y - Xw \tag{8}
$$

4. (2 pts) Assume that each $X_i$ has full column rank. Show that, for each $i$ and any $a \in \mathbb{R}^n$,

$$
w_i^* = \arg\min_{w_i \in \mathbb{R}^{p_i}} \frac{1}{2} \|a - X_i w_i\|^2 + h_i(w_i) \quad \iff \quad X_i w_i^* = a - P_{C_i}(a).
$$

Hint: write $X_i w_i^*$ in terms of a proximal operator then use Moreau’s theorem in Q2 part 3.

5. (6 pts) Dykstra’s algorithm for problem (7) can be described as follows. We initialize $u_0 = y$, $z_1 = \cdots = z_d = 0$, and then repeat:

$$
\begin{align*}
&u_0^{(k)} = u_d^{(k-1)}, \\
&u_i^{(k)} = P_{C_i} \left( u_i^{(k-1)} + z_i^{(k-1)} \right), \\
&z_i^{(k)} = u_i^{(k)} + z_i^{(k-1)} - u_i^{(k)},
\end{align*}
$$

for $i = 1, \ldots, d$, \quad \tag{9}

\[3\]
for $k = 1, 2, 3, \ldots$. As $k \to \infty$, the iterate $u_0^{(k)}$ in (9) will approach the solution in (7).

Assuming that we initialize $w(0) = 0$, show that coordinate descent (5) for problem (4) and Dykstra’s algorithm (9) for problem (7) are in fact completely equivalent, and satisfy

$$z_i^{(k)} = X_i w_i^{(k)} \quad \text{and} \quad u_i^{(k)} = y - \sum_{j \leq i} X_j w_j^{(k)} - \sum_{j > i} X_j w_j^{(k-1)}, \quad \text{for } i = 1, \ldots, d,$$

at all iterations $k = 1, 2, 3, \ldots$. Hint: use an inductive argument, and the result in part 4.

6. (Bonus, 3 pts) Let $\gamma_1, \ldots, \gamma_d > 0$ be arbitrary weights with $\sum_{i=1}^d \gamma_i = 1$. Consider the problem

$$\min_{u=(u_1, \ldots, u_d) \in \mathbb{R}^{nd}} \sum_{i=1}^d \gamma_i \|y - u_i\|_2^2 \quad \text{subject to} \quad u \in C_0 \cap (C_1 \times \cdots \times C_d),$$

(10)

where $C_0 = \{(u_1, \ldots, u_d) \in \mathbb{R}^{nd} : u_1 = \cdots = u_d\}$. Observe that this is equivalent to (7), and is sometimes called the product-space reformulation of (7), or the consensus form of (7).

Rescale (10) to turn the loss into an unweighted squared loss, then apply Dykstra’s algorithm to the resulting best approximation problem. Show that the resulting algorithm repeats:

$$u_0^{(k)} = \sum_{i=1}^d \gamma_i u_i^{(k-1)},$$

$$u_i^{(k)} = P_{C_i}(u_0^{(k)} + z_i^{(k-1)}),$$

$$z_i^{(k)} = u_0^{(k)} + z_i^{(k-1)} - u_i^{(k)},$$

for $i = 1, \ldots, d$,

(11)

for $k = 1, 2, 3, \ldots$. Importantly, the steps enclosed in curly brace above can all be performed in parallel, so that (11) is a parallel version of Dykstra’s algorithm (9) for problem (7).

7. (Bonus, 4 pts) Prove that the iterations (11) can be rewritten in equivalent form as

$$w_i^{(k)} = \arg\min_{w_i \in \mathbb{R}^{p_i}} \frac{1}{2} \|y - X w_i^{(k-1)} + X_i w_i^{(k-1)}/\gamma_i - X_i w_i/\gamma_i\|_2^2 + h_i(w_i/\gamma_i), \quad i = 1, \ldots, d,$$

(12)

for $k = 1, 2, 3, \ldots$. Importantly, the updates above can all be performed in parallel, so that (12) is a parallel version of coordinate descent (5) for problem (4). Hint: use an inductive argument and the result in part 4, similar to your proof in part 5.