Homework 5

Convex Optimization 10-725

Due Friday November 15 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.)

Choose to solve **one of Q3 or Q4**.
That is, you will either submit Q1+Q2+Q3 or Q1+Q2+Q4.

Note that the programming questions in this assignment, Q1(e), Q3(g), Q4(f), will be **peer-graded**. Instructions to follow about how this will be done.

Total: 63 points

1 Coordinatewise optima of smooth + separable convex functions (23 points)

Let $f(x) = g(x) + \sum_{i=1}^n h_i(x_i)$, where g and h_i, $i = 1, \ldots, n$ are convex, and g is differentiable. Let x be a point that is a *coordinatewise minimizer*, i.e.,

$$f(x + ve_i) \geq f(x), \quad \text{for all } v \text{ and } i = 1, \ldots, n.$$

(Here e_i is denotes the ith standard basis vector, which is all 0s except for a 1 in the ith component, for $i = 1, \ldots, n$.) We will show that x must be a global minimizer of f, by fixing an arbitrary y and establishing that $f(y) \geq f(x)$.

(a, 2 pts) Using convexity of g, show that

$$f(y) - f(x) \geq \sum_{i=1}^n \underbrace{\big(\nabla_i g(x)(y_i - x_i) + h_i(y_i) - h_i(x_i)\big)}_{a_i},$$

where $\nabla_i g$ denotes the ith component of the vector-valued gradient map ∇g.

(b, 4 pts) Using the fact that x is a coordinatewise minimizer, and subgradient optimality, show that $a_i \geq 0$, $i = 1, \ldots, n$, and thus $f(y) \geq f(x)$.

(c, 5 pts) Now let $g(x) = \frac{1}{2}x^T Q x - b^T x$ for $Q \succ 0$, and $h(x) = \lambda \|x\|_1$. Write out the updates for one cycle of coordinate descent:

$$x_i^{(k)} = \underset{x_i}{\operatorname{argmin}} \ f\big(x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_i, x_{i+1}^{(k-1)}, \ldots, x_n^{(k-1)}\big), \quad i = 1, \ldots, n,$$

as expicitly as possible. Write out the updates for one cycle of coordinate proximal gradient descent:

$$x_i^{(k)} = \mathrm{prox}_{h_i,t_{ki}}\Big(x_i^{(k-1)} - t_{ki}\cdot\nabla_i g\big(x_1^{(k)},\ldots,x_i^{(k-1)},\ldots,x_n^{(k-1)}\big)\Big), \quad i=1,\ldots,n,$$

again as explicitly as possible. Show that these two are equal under certain choices of step sizes t_{ki}, $i = 1,\ldots,n$.

(d, 2 pts) Argue that the step sizes, which make coordinate proximal gradient descent equivalent to coordinate descent, can be viewed as the result of *exact* step size optimization, i.e., exact line search,

$$t_{ki} = \underset{t\geq 0}{\mathrm{argmin}}\ f\big(x_1^{(k)},\ldots,x_{i-1}^{(k)},x_i^{(k)}(t),x_{i+1}^{(k-1)},\ldots,x_n^{(k-1)}\big), \quad i=1,\ldots,n,$$

where $x_i^{(k)}(t)$ denotes the ith update from coordinate proximal gradient descent with step size t.

(e, 10 pts) Design and conduct an experiment to empirically investigate, for the given class of functions $f = g + h$ (quadratic plus ℓ_1), the use of exact steps sizes in coordinate proximal gradient descent. That is, we know (from parts (c) and (d)) that coordinate descent is the same as using exact line search at each step of coordinate proximal gradient descent; how much does this help over fixed step sizes, or backtracking line search?

Some general tips: be completely explicit about all your experimental design choices; think in particular about the problem conditioning; use figures rather than tables to report what you find; aggregate results over multiple simulation instances. Your simulation will be graded on the following criteria (3 points each):

- is the setup clearly explained?
- are the results clearly explained?
- are the conclusions justified?

As usual, append all code in an appendix (1 point for readable/organized code).

2 Conjugates, duality, and proximal mappings (18 points)

Let f, g be closed and convex functions, and f^*, g^* denote their conjugates.

(a, 2 pts) For a matrix $A \in \mathbb{R}^{m\times n}$, prove that the dual problem of

$$\min_x\ f(x) + g(Ax) \tag{1}$$

is

$$\max_y\ -f^*(-A^Ty) - g^*(y). \tag{2}$$

(b, 3 pts) Assume that f is strictly convex. Prove that this implies f^* is differentiable, and that

$$\nabla f^*(y) = \underset{z}{\mathrm{argmin}}\ f(z) - y^Tz.$$

Hint: use the fact that $x \in \partial f^*(y) \iff y \in \partial f(x)$, as you established if Q2(d) of Homework 3.

From now on, assume that f is strictly convex and smooth, and g is not smooth, but we know its proximal operator

$$\mathrm{prox}_{g,t}(x) = \underset{z}{\mathrm{argmin}}\ \frac{1}{2t}\|x-z\|_2^2 + g(z).$$

We 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 (1). However, as you will show in the next few parts, knowing the the proximal mapping of g *does* lead to the proximal mapping of g^*, which leads to an algorithm on the dual problem (2).

(c, 4 pts) Prove first that

$$\text{prox}_{g,1}(x) + \text{prox}_{g^*,1}(x) = x,$$

for all x. This is sometimes called *Moreau's theorem.* Note the specification $t = 1$ in the above. Hint: again, use $x \in \partial g^*(y) \iff y \in \partial g(x)$.

(d, 4 pts) Verify that for $t > 0$, we have $(tg)^*(x) = tg^*(x/t)$. Use this, and part (c), 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 (c) to the function tg. Then note that $\text{prox}_{g,t}(x) = \text{prox}_{tg,1}(x)$, and the same for g^*.

(e, 2 pts) Now write down a proximal gradient ascent algorithm for the dual problem (2). Use parts (b) and (d) of this question to express all quantities in terms of f and g. That is, your proximal gradient ascent updates should not have any appearences of ∇f^* and $\text{prox}_{g^*,t}(\cdot)$.

(f, 3 pts) Write down the steps of ADMM applied to problem (1), after substituting $g(z)$ for $g(Ax)$ in the criterion, and introducing the inequality constrained $Ax = z$. Compare these to the steps of the dual proximal gradient algorithm from part (e). How are they different? Briefly explain any advantages/disadvantages you see to using each method.

3 Coordinate descent for the graphical lasso (22 points)

Let $X \in \mathbb{R}^{n\times p}$ be a data matrix whose rows are independent observations from $N(0, \Sigma)$. Normality theory tells us that for $x \sim N(0, \Sigma)$, $\Sigma^{-1}_{ij} = 0$ implies that the variables x_i and x_j are conditionally independent given all the other variables $\{x_k\}_{k\notin\{i,j\}}$. So, if we believe that many pairs of features recorded in X are conditionally independent given the other features (which is often a reasonable belief if p is large), then we want an estimate of Σ such that Σ^{-1} is sparse. This goal can be achieved by solving the *graphical lasso* problem

$$\min_{\Theta} \; -\log\det\Theta + \text{tr}(S\Theta) + \lambda\|\Theta\|_1. \tag{3}$$

Here the domain of the minimization problem is $\mathbb{S}^p_{++}$ (the space of symmetric $p \times p$ positive definite matrices), $S = X^TX/n$ is the samples covariance matrix, and $\|\Theta\|_1 = \sum_{i,j=1}^p |\Theta_{ij}|$. Note that the solution $\hat{\Theta}$ in the above serves as our estimate for Σ^{-1}.

(a, 3 pts) Prove that the subgradient optimality condition for the graphical lasso problem (3) is

$$-\Theta^{-1} + S + \lambda\Gamma = 0, \tag{4}$$

where $\Gamma_{ij} \in \partial|\Theta_{ij}|$ for each i, j. Let $W = \Theta^{-1}$. Verify that the above implies that $W_{ii} = S_{ii} + \lambda$ for each i.

Consider now partitioning W as

$$W = \begin{pmatrix} W_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix},$$

where $W_{11} \in \mathbb{R}^{(p-1)\times(p-1)}$, $w_{12} \in \mathbb{R}^{p-1}, w_{21}^T \in \mathbb{R}^{p-1}$, and $w_{22} \in \mathbb{R}$. Consider partitioning the other matrices Θ, S, and Γ in the same manner.

(b, 2 pts) Using the fact that $W\Theta = I$ and the subgradient optimality condiiton from part (a), show that

$$w_{12} = -W_{11}\theta_{12}/\theta_{22}$$

and therefore

$$W_{11}\frac{\theta_{12}}{\theta_{22}} + s_{12} + \lambda\gamma_{12} = 0.$$

(c, 3 pts) Let now $\beta = \theta_{12}/\theta_{22} \in \mathbb{R}^{p-1}$. Write a lasso problem (with β as the optimization variable) such that $\beta = \theta_{12}/\theta_{22}$ is the solution. Note: in this lasso problem, you may directly write out the quadratic and linear terms in the loss (i.e., this will be easier than directly specifying the form of the least squares loss).

Observe that, once the lasso problem from part (c) is solved, it is easy to recover $w_{12} = -W_{11}\beta$ and $w_{21} = w_{12}^T$. Furthermore, θ_{12}, θ_{21}^T, and θ_{22} are directly obtained by solving

$$\begin{pmatrix} W_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & 1 \end{pmatrix}.$$

By rearranging appropriately the entries of W, Θ, S, and Γ, one can then iterate this procedure to solve for the entire matrix W.

(d, 3 pts) Describe an algorithm to estimate Θ based on iterative blockwise minimization using the results of parts (a), (b), and (c).

(e, 1 pts) Explain why, strictly speaking, such an algorithm is in fact *not* a blockwise coordinate descent algorithm for the graphical lasso problem as formulated in (3).

(f, 6 pts) We will now show that the coordinate descent algorithm suggested by parts (a), (b), and (c), which optimizes over the matrix W and is known as the *glasso* algorithm[1], is in fact a proper coordinate descent algorithm for the *dual* of the graphical lasso problem[2]. Prove that the dual of (3) is (equivalent to)

$$\min_{\tilde{\Gamma}} \; -\log\det(\tilde{\Gamma} + S) - p \quad \text{subject to} \quad \|\tilde{\Gamma}\|_\infty \le \lambda. \tag{5}$$

Write down the KKT conditions for the dual problem (5). Show that the subgradient optimality condition for the primal graphical lasso problem (3) can be retrieved from the KKT conditions for the dual problem (possibly after appropriate changes of variable). Finally, briefly clarify why the glasso algorithm that you derived in part (d) is a proper coordinate descent for (5).

(g, 4 pts) Produce an empirical example to verify that the glasso algorithm (which you will implement) is not a descent algorithm on the primal graphical lasso problem (3), but is indeed a descent algorithm on its equivalent dual (5). Explain clearly your setup and results. As always, append your code.

[1] Friedman et al. (2007), "Sparse inverse covariance estimation with the graphical lasso".

[2] Mazumder and Hastie (2013), "The graphical lasso: New insights and alternatives".

4 Coordinate descent and Dykstra (22 points)

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, consider the regularized least squares program

$$\min_{w \in \mathbb{R}^p} \frac{1}{2}\|y - Xw\|_2^2 + \sum_{i=1}^d h_i(w_i), \tag{6}$$

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. Consider coordinate descent, which repeats the following updates:

$$w_i^{(k)} = \underset{w_i \in \mathbb{R}^{p_i}}{\operatorname{argmin}} \ \frac{1}{2}\Bigg\| y - \sum_{j<i} X_j w_j^{(k)} - \sum_{j>i} X_j w_j^{(k-1)} - X_i w_i \Bigg\|_2^2 + h_i(w_i), \quad i = 1, \ldots, d, \tag{7}$$

for $k = 1, 2, 3, \ldots$. Assume 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.

(a, 3 pts) Show that the dual of (6) is what is sometimes called the *best approximation problem*

$$\min_{u \in \mathbb{R}^n} \|y - u\|_2^2 \quad \text{subject to} \quad u \in C_1 \cap \cdots \cap C_d. \tag{8}$$

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

$$u = y - Xw. \tag{9}$$

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

$$w_i^* = \underset{w_i \in \mathbb{R}^{p_i}}{\operatorname{argmin}} \ \frac{1}{2}\|a - X_i w_i\|_2^2 + h_i(w_i) \quad \iff \quad X_i w_i^* = a - P_{C_i}(a).$$

Hint: write $X_i w_i^*$ in terms of a proximal operator then use Moreau's theorem in Q2(c).

Dykstra's algorithm for problem (8) can be described as follows. We set $u_d^{(0)} = y$, $z_1^{(0)} = \cdots = z_d^{(0)} = 0$, and then repeat:

$$\begin{aligned} u_0^{(k)} &= u_d^{(k-1)}, \\ u_i^{(k)} &= P_{C_i}(u_{i-1}^{(k)} + z_i^{(k-1)}), \\ z_i^{(k)} &= u_{i-1}^{(k)} + z_i^{(k-1)} - u_i^{(k)}, \end{aligned} \left.\vphantom{\begin{aligned} a \\ b \\ c \end{aligned}}\right\} \quad \text{for } i = 1, \ldots, d, \tag{10}$$

for $k = 1, 2, 3, \ldots$. As $k \to \infty$, the iterate $u_0^{(k)}$ in (10) will approach the solution in (8).

(c, 6 pts) Assuming we initialize $w^{(0)} = 0$, show that coordinate descent (7) for problem (6) and Dykstra's algorithm (10) for problem (8) are in fact completely equivalent, and satisfy

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

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

Now let $\gamma_1, \ldots, \gamma_d > 0$ be arbitrary weights with $\sum_{i=1}^d \gamma_i = 1$. Consider the problem

$$\min_{u=(u_1,\ldots,u_d)\in\mathbb{R}^{nd}} \; \sum_{i=1}^d \gamma_i \|y - u_i\|_2^2 \quad \text{subject to} \quad u \in C_0 \cap (C_1 \times \cdots \times C_d), \tag{11}$$

where $C_0 = \{(u_1, \ldots, u_d) \in \mathbb{R}^{nd} : u_1 = \cdots = u_d\}$. Observe that this is equivalent to (8), and is sometimes called the *product-space reformulation* of (8), or the *consensus form* of (8).

(d, 3 pts) Rescale (11) 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:

$$\begin{aligned} u_0^{(k)} &= \sum_{i=1}^d \gamma_i u_i^{(k-1)}, \\ \left.\begin{aligned} u_i^{(k)} &= P_{C_i}(u_0^{(k)} + z_i^{(k-1)}), \\ z_i^{(k)} &= u_0^{(k)} + z_i^{(k-1)} - u_i^{(k)}, \end{aligned}\right\} &\quad \text{for } i = 1, \ldots, d, \end{aligned} \tag{12}$$

for $k = 1, 2, 3, \ldots$. Importantly, the steps enclosed in curly brace above can all be performed in parallel, so that (12) is a parallel version of Dykstra's algorithm (10) for problem (8).

(e, 4 pts) Prove that the iterations (12) can be rewritten in equivalent form as

$$w_i^{(k)} = \operatorname*{argmin}_{w_i \in \mathbb{R}^{p_i}} \; \frac{1}{2}\Big\| y - Xw^{(k-1)} + X_i w_i^{(k-1)}/\gamma_i - X_i w_i/\gamma_i \Big\|_2^2 + h_i(w_i/\gamma_i), \quad i = 1, \ldots, d, \tag{13}$$

for $k = 1, 2, 3, \ldots$. Importantly, the updates above can all be performed in parallel, so that (13) is a parallel version of coordinate descent (7) for problem (6). Hint: use an inductive argument and the result in part (b), similar to your proof in part (c).

(f, 4 pts) Produce an empirical example to verify that coordinate descent and Dykstra on (6) and (8), respectively (which you will both implement), are equivalent. Explain clearly your setup and results. As always, append your code.