Math 680 - HW3

A Gougeon

2018-04-01

1. Subgradients and Proximal Operators

1.

i)

To show $\partial f(x)$ is convex and closed, first define

$$\partial f(x) = \{z | z^T(y - x) \leq f(y) - f(x), \forall y \in dom(f)\},$$

that is, the set of all subgradients of f. Consider any two subgradients in $\partial f(x)$ (and $x = \lambda u + (1 - \lambda)v$),

$$\begin{aligned} f(u) &\geq f(x) + z^T(u - x) = f(x) + (1 - \lambda)z^T(u - v) \\ f(v) &\geq f(x) + z^T(v - x) = f(x) - \lambda z^T(u - v). \end{aligned}$$

This leads to

$$\begin{aligned} \lambda f(u) + (1 - \lambda)f(v) &\geq \lambda f(x) + \lambda(1 - \lambda)z^T(u - v) + (1 - \lambda)f(x) - (1 - \lambda)\lambda z^T(u - v) \\ &= f(x). \end{aligned}$$

Therefore, by definition, $\partial f(x)$ is convex. Furthermore, we note that the set $\partial f(x)$ is closed since it is an intersection of halfspaces.

ii)

For $x \neq 0$, f is differentiable and so the subgradient z,

$$z = \nabla f = \frac{x}{||x||_2}.$$

If $x = 0$, then by definition, we must have

$$\begin{aligned} f(y) = ||y||_2 &\geq f(x) + g^T(y - x) = g^T y, \forall y \\ \implies & ||y||_2 \geq g^T y \\ \implies & ||z||_2 \leq 1. \end{aligned}$$

We conclude that

$$\partial f(x) = \begin{cases} \frac{x}{||x||_2}, & \text{if } x \neq 0 \\ z : ||z||_2 \leq 1, & \text{if x=0} \end{cases} \tag{1}$$

as desired.

iii)

iv)

We wish to show
$$\partial f(x) = \{z : ||z||_q \text{ and } z^T x = ||x||_p\}$$
Let $z \in \partial f(x)$,
$$\implies f(y) \geq f(x) + z^T(y-x).$$
If $y = 0$,
$$0 = f(0) \geq f(x) - z^T x \implies ||x||_p \geq z^T x.$$
If $y = 2x$,
$$2z^T x = f(2x) \geq f(x) + z^T x \implies ||x||_p \leq z^T x.$$
We conclude that $||x||_p \leq z^T x$. It follows that
$$||y||_p \geq z^T y$$
for all y, and so $||z||_q \leq 1$.

2.

i)

$$\text{prox}_{h,t}(x) = \underset{z}{\text{argmin}} \frac{1}{2}||z-x||_2^2 + t(\frac{1}{2}z^T A z + b^T z + c)$$
Taking the derivative of the minimizing object with respect to z and setting equal to 0,
$$(z-x) + t(z^T A + b) = 0 \implies z = (I + tA)^{-1}(x - tb).$$
Therefore,
$$\text{prox}_{h,t}(x) = (I + tA)^{-1}(x - tb)$$

ii)

$$\text{prox}_{h,t}(x) = \underset{z}{\text{argmin}} \frac{1}{2}||z-x||_2^2 + t(-\sum_{i=1}^{n} \log z_i)$$
We consider the ith entry. Taking the derivative of the minimizing object with respect to z and setting equal to 0,
$$(z_i - x_i) - \frac{t}{z_i} = 0 \implies z_i = \frac{1}{2}(x_i - \sqrt{x_i^2 - 4t}).$$
Therefore,
$$\text{prox}_{h,t}(x_i) = \frac{1}{2}(x_i - \sqrt{x_i^2 - 4t})$$

iii)

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

Recall that

$$\partial f(x) = \begin{cases} \frac{x}{||x||_2}, & \text{if } x \neq 0 \\ z : ||z||_2 \leq 1, & \text{if x=0} \end{cases} \tag{2}$$

where f is differentiable everywhere except for one point. We begin by assuming that $z^* = \text{prox}_{h,t}(x) \neq 0$. Then, z has to satisfy

$$\frac{1}{t}(z^* - x) + \frac{z^*}{||z^*||_2} = 0. \tag{3}$$

It is now useful to consider polar coordinates, $x = (r_x, \theta_x)$ where $r_x = ||x||_2$ and $\theta_x = \tan^{-1}(\frac{x_1}{x_2})$. We notice that $\frac{z^*}{||z^*||_2}$ and $x - z^*$ must have the same angle, and the angle of $\frac{z^*}{||z^*||_2}$ and z^* must equal the angle of x or its negative. This leads to $z^* = ax$ for any $a \in \mathbb{R}$. Substituting this in (3), we get

$$\frac{a-1}{t} r_x + \text{sign}(a) = 0$$

and so

$$a = \begin{cases} \frac{r_x - t}{r_x}, & \text{if } r_x > t \\ 0, & \text{else} \end{cases} \tag{4}$$

and $z = ax^*$. Now, if $z = 0$, we see that $r_x \leq t$ and $\frac{1}{t}x \in \{||x||_2 \leq 1\}$. Therefore, we conclude that

$$\text{prox}_{h,t}(x) = \begin{cases} x\frac{||x||_2 - t}{||x||_2}, & \text{if } ||x||_2 > t \\ 0, & \text{else} \end{cases} \tag{5}$$

iv)

$$\text{prox}_{h,t}(x) = \underset{z}{\text{argmin}} \frac{1}{2}||z - x||_2^2 + t||z||_0$$

where $||z||_0 = |\{z_i : z_i \neq 0, i = 1, ...n\}|$

ugly - has jump discontinuities

2. Properties of Proximal Mappings and Subgradients

b)

Show that, for $\forall x, y \in \mathbb{R}, u \in \partial f(x), v \in \partial f(y)$

$$(x - y)^T(u - v) \geq 0.$$

If $u \in \partial f(x)$, then u is a subgradient of $f(x)$. Therefore, by definition,

$$f(y) \geq f(x) + u^T(y - x).$$

It follows that (result from Stanford's notes)

$$f(y) \leq f(x) \implies u^T(y-x) \leq 0$$

Similarly, if $v \in \partial f(y)$, then v is a subgradient of $f(y)$. Therefore,

$$f(x) \geq f(y) + v^T(x-y).$$

It follows that (result from Stanford's notes)

$$f(x) \leq f(y) \implies v^T(x-y) \leq 0$$

Therefore,

$$\begin{aligned} u^T(y-x) + v^T(x-y) &\leq 0 \\ \implies (x-y)^T(v-u) &\leq 0 \\ \implies (x-y)^T(u-v) &\geq 0 \end{aligned}$$

as desired.

d)

We wish to show

$$\text{prox}_t(x) = u \iff h(y) \geq h(u) + \frac{1}{t}(x-u)^T(y-u), \quad \forall y.$$

Recall that

$$\text{prox}_t(x) = \underset{u}{\text{argmin}} \frac{1}{2t}||x-u||_2^2 + h(u).$$

If h is closed and convex, then the proximal mapping exists and is unique for all x. That is, it is closed and bounded, and is strongly convex. It follows, from these optimality conditions, that

$$\begin{aligned} u = \text{prox}_t(x) &\implies x - u \in \partial h(u) \\ &\implies h(y) \geq h(u) + \frac{1}{t}(x-u)^T(y-u), \end{aligned}$$

as desired.

3. Properties of Lasso

1.

We begin by writting the Lasso problem in Lagrange form, that is,

$$\underset{\beta \in \mathbb{R}^p}{\text{argmin}} \frac{1}{2n}||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda||\beta||_1 \tag{6}$$

where $\mathbf{y}$ and $\mathbf{X}$ are centered. The solution to (6) satisfies the subgradient condition

$$-\frac{1}{n}\langle \mathbf{x}_j, \mathbf{y} - \mathbf{X}\hat{\beta}\rangle + \lambda s_j,$$

where $s_j \in \text{sign}(\hat{\beta}_j), j = 1, ..., p$. With this information, we find the solution $\hat{\beta}(\lambda_{max}) = 0$ as the subgradient condition

$$
\begin{aligned}
&-\frac{1}{n}\langle \mathbf{x}_j, \mathbf{y}\rangle + \lambda s_j \\
&\implies \lambda_{max} = \max_j |\frac{1}{n}\langle \mathbf{x}_j, \mathbf{y}\rangle|
\end{aligned}
$$

as desired.

2.

a)

Suppose there are two lasso solutions $\hat{\beta}$ and $\hat{\gamma}$ with common optimal value c^* and $X\hat{\beta} \neq X\hat{\gamma}$. We note that $f(a) = ||y - a||_2^2$ is strictly convex, and that the ℓ_1 norm is convex. This implies that lasso is strictly convex. Therefore, the solution set is convex, and so $\alpha\hat{\beta} + (1-\alpha)\hat{\gamma}$ is also a solution for some $0 < \alpha < 1$. It follows that

$$\frac{1}{2}||y - X[\alpha\hat{\beta} + (1-\alpha)\hat{\gamma}]||_2^2 + \lambda||\alpha\hat{\beta} + (1-\alpha)\hat{\gamma}|| < \alpha c^* + (1-\alpha)c^* = c^*$$

where the "$<$" comes from the strict convexity of lasso. This signifies that $\alpha\hat{\beta} + (1-\alpha)\hat{\gamma}$ attains a $c^{new} < c^*$, which is a contradiction. We conclude that $X\hat{\beta} = X\hat{\gamma}$, as desired.

b)

This statement is implied by a). Both solutions have the same fitted values,

$$\frac{1}{2}||y - X\hat{\beta}||_2^2 = \frac{1}{2}||y - X\hat{\gamma}||_2^2.$$

They also attain the same optimal value, c^*. This implies that

$$\lambda||\hat{\beta}|| = \lambda||\hat{\gamma}||$$

as desired.

4) Convergence Rate for Proximal Gradient Descent

a)

We wish to show that

$$s = G_t(x^{(i-1)}) - \nabla g(x^{(i-1)})$$

is a subgradient of h evaluated at $x^{(i)}$. We recall that h is convex, but not necessarily differentiable. From Question 2d), we showed that

$$\text{prox}_t(x) = u \iff h(y) \geq h(u) + \frac{1}{t}(x-u)^T(y-u), \quad \forall y.$$

In this case,

$$
\begin{aligned}
x^{(i)} &= \mathrm{prox}_{t,h}(x^{(i-1)} - t\nabla g(x^{(i-1)})) \\
\iff h(y) &\geq h(x^{(i)}) + \frac{1}{t}(x^{(i-1)} - x^{(i)} - t\nabla g(x^{(i-1)}))^T(y - x^{(i)}) \\
&= h(x^{(i)}) + (G_t(x^{(i-1)}) - \nabla g(x^{(i-1)}))^T(y - x^{(i)}) \\
&= h(x^{(i)}) + s^T(y - x^{(i)})
\end{aligned}
$$

This implies that $s \in \partial h(x^{(i)})$, as desired.

b)

We now wish to derive the following inequality

$$f(x^{(i)}) \leq f(z) + G_t(x^{(i-1)})^T(x^{(i-1)} - z) - \frac{t}{2}||G_t(x^{(i-1)})||_2^2.$$

Recall that $f(x)$ is our objective function and can be written as

$$f(x) = g(x) + h(x),$$

where g is convex, differentiable, with ∇g being Lipschitz, and h is convex. Therefore, f must also be convex. By (A4),

$$g(x^{(i)}) \leq g(x^{(i-1)}) - t\nabla g(x^{(i-1)})^T G_t(x^{(i-1)}) + \frac{t}{2}||G_t(x^{(i-1)})||_2^2.$$

Furthermore, since ∇g is Lipschitz,

$$
\begin{aligned}
||\nabla g(x^{(i-1)}) - g(x^{(i)})|| &\leq L||x^{(i-1)} - x^{(i)}|| \\
&= \frac{1}{t}||x^{(i-1)} - x^{(i)}|| \\
&= ||G_t(x^{(i-1)})||.
\end{aligned}
$$

On the other hand, $s \in \partial h(x^{(i)})$,

$$
\begin{aligned}
&h(x^{(i-1)}) \geq h(x^{(i)}) + s^T(x^{(i-1)} - x^{(i)}) \\
&\iff h(x^{(i)}) < h(x^{(i-1)}) + G_t(x^{(i-1)}) - \nabla g(x^{(i-1)})^T(x^{(i)} - x^{(i-1)}).
\end{aligned}
$$

With some rewritting and some rearranging, it follows that

$$
\begin{aligned}
&f(x^{(i)}) = g(x^{(i)}) + h(x^{(i)}) \\
&\qquad \leq h(x^{(i-1)}) + g(x^{(i-1)}) + G_t(x^{(i-1)}) - \nabla g(x^{(i-1)})^T(x^{(i)} - x^{(i-1)}) - t\nabla g(x^{(i-1)})G_t(x^{(i-1)}) + \frac{t}{2}||G_t(x^{(i-1)})||_2^2 \\
&\qquad \leq f(z) + G_t(x^{(i-1)})^T(x^{(i-1)} - z) - \frac{t}{2}||G_t(x^{(i-1)})||_2^2,
\end{aligned}
$$

for some $z \in \mathbb{R}^n$, as desired.

c)

We now wish to show that the sequence $\{f(x^{(i)})\}$ is nonincreasing for $i = 0, ..., k$. That is,

$$f(x^{(i)}) \leq f(x^{(i-1)}), \quad i = 1, ..., k.$$

Recall the inequality from the previous question,

$$f(x^{(i)}) \leq f(z) + G_t(x^{(i-1)})^T(x^{(i-1)} - z) - \frac{t}{2}||G_t(x^{(i-1)})||_2^2, \quad z \in \mathbb{R}^n.$$

If we let $z = x^{(i-1)}$, we see that

$$\begin{aligned} f(x^{(i)}) &\leq f(x^{(i-1)}) + G_t(x^{(i-1)})^T(x^{(i-1)} - x^{(i-1)}) - \frac{t}{2}||G_t(x^{(i-1)})||_2^2 \\ &= f(x^{(i-1)}) - \frac{t}{2}||G_t(x^{(i-1)})||_2^2. \end{aligned}$$

Note that $\frac{t}{2}||G_t(x^{(i-1)})||_2^2$ will always be positive unless $G_t(x^{(i-1)}) = 0$. This implies that

$$f(x^{(i)}) \leq f(x^{(i-1)}), \quad i = 1, ..., k$$

and so the sequence of objection function evaluations is nonincreasing.

d)

We will now derive the following inequality

$$f(x^{(i)}) - f(x^*) \leq \frac{1}{2t}(||x^{(i-1)} - x^*||_2^2 - ||x^{(i)} - x^*||_2^2.$$

Using the inequality derived above, and the fact that $f(x^{(i)}) \leq f(x^*)$,

$$f(x^{(i)}) \leq f(x^*) + G_t(x^{(i-1)})^T(x^{(i-1)} - x^*) - \frac{t}{2}||G_t(x^{(i-1)})||_2^2$$

$$f(x^{(i)}) - f(x^*) \leq \frac{1}{2t}\{2t \cdot G_t(x^{(i-1)})^T(x^{(i-1)} - x^*) - t^2||G_t(x^{(i-1)})||_2^2$$

$$f(x^{(i)}) - f(x^*) \leq \frac{1}{2t}\{2t \cdot G_t(x^{(i-1)})^T(x^{(i-1)} - x^*) - t^2||G_t(x^{(i-1)})||_2^2 - ||x^{(i-1)} - x^*||_2^2 + ||x^{(i-1)} - x^*||_2^2\}.$$

Notice that

$$||x^{(i-1)} - x^* - tG_t(x^{(i-1)})||_2^2 = ||x^{(i-1)} - x^*||_2^2 - 2t \cdot G_t(x^{(i-1)})^T(x^{(i-1)} - x^*) + t^2||G_t(x^{(i-1)})||_2^2.$$

Therefore, we can write

$$f(x^{(i)}) - f(x^*) \leq \{\ ||x^{(i-1)} - x^*||_2^2 - ||x^{(i-1)} - tG_t(x^{(i-1)}) - x^*||_2^2\}.$$

Furthermore, by definition, we have that $G_t(x^{(i-1)}) = \frac{1}{t}(x^{(i-1)} - x^{(i)})$ and so

$$\begin{aligned} f(x^{(i)}) - f(x^*) &\leq \{\ ||x^{(i-1)} - x^*||_2^2 - ||x^{(i-1)} - t(\frac{1}{t}(x^{(i-1)} - x^{(i)})) - x^*||_2^2\} \\ &= ||x^{(i-1)} - x^*||_2^2 - ||x^{(i)} - x^*||_2^2, \end{aligned}$$

as desired.

e)

We will now show
$$f(x^{(k)}) - f(x^*) \leq \frac{1}{2kt}||x^{(0)} - x^*||_2^2.$$
We begin with the result above, summing over all k iterations,

$$\begin{aligned}
\sum_{i=1}^{k} f(x^{(i)}) - f(x^*) &\leq \sum_{i=1}^{k} \frac{1}{2t}(||x^{(i-1)} - x^*||_2^2 - ||x^{(i)} - x^*||_2^2) \\
&= \frac{1}{2t}(||x^{(0)} - x^*||_2^2 - ||x^{(k)} - x^*||_2^2) \\
&\leq \frac{1}{2t}(||x^{(0)} - x^*||_2^2).
\end{aligned}$$

Since the sequence of objection function evaluations is nonincreasing,

$$\begin{aligned}
f(x^{(k)}) - f(x^*) &\leq \frac{1}{k}\sum_{i=1}^{k} f(x^{(i)}) - f(x^*) \\
&\leq \frac{||x^{(0)} - x^*||_2^2}{2kt}
\end{aligned}$$

as desired.

f)

The method of selecting the step size according to backtracking line search consists of fixing some $0 < \beta < 1$ and starting with $t = 1$. Then, at each iteration,
$$f(x - t\nabla f(x)) > f(x) - \frac{t}{2}||\nabla f(x)||_2^2,$$
update $t = \beta t$. Now, in the context of this problem,
$$f(x - t\nabla f(x)) > f(x) - \frac{t}{2}||\nabla f(x)||_2^2,$$

6) Practice with KKT Conditions and Duality

We begin with the usual least squares problem,
$$\min_{\beta \in \mathbb{R}^p} ||y - X\beta||_2^2.$$
We begin by noting the primal of this problem is
$$\underset{v \in \mathbb{R}^n}{min} \frac{1}{2}||v||_2^2 \quad \text{subject to } y = X\beta + v.$$
We can now write the Lagrangian,
$$L(v, \beta, \lambda) = \frac{1}{2}||v||_2^2 + \lambda(y - X\beta - v).$$

It follows that the first order necessary conditions are

$$\frac{\partial L}{\partial u} = v - \lambda = 0$$

$$\frac{\partial L}{\partial \beta} = -X\lambda = 0$$

$$\frac{\partial L}{\partial \lambda} = y - X\beta - v = 0.$$

From the first partial derivative, we can see that we can simplify the Lagrangian since

$$v = \lambda \quad \text{and} \quad v^T v = v^T \lambda.$$

We can now rewrite the Lagrangian as

$$\begin{aligned} L(v, \beta, \lambda) &= \frac{1}{2}||v||_2^2 + \lambda(y - X\beta - v) \\ &= \frac{1}{2}||v||_2^2 + \lambda y - \lambda u \\ &= \frac{1}{2}||v||_2^2 - v^T y - ||v||_2^2 \\ &= ||y - v||_2^2. \end{aligned}$$

Therefore, we conclude that the dual of the model can be written as

$$\underset{v \in \mathbb{R}^n}{min} ||y - v||_2^2 \quad \text{subject to } X^T v = 0,$$

as desired.