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)

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

3. Properties of Lasso

1.

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

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

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.