MATH 680: Assignment 3

Annik Gougeon, David Fleischer

Last Update: 02 May, 2018

Section 1: Subgradients and Proximal Operators

Question 1.1

1.1.(i)

Recall that a *subgradient* of f at point $x \in \mathbb{R}^n$ is defined as a vector $g \in \mathbb{R}^n$ satisfying the inequality

$$f(y) \geq f(x) + g^T(y - x), \quad \forall \, y.$$

The *subdifferential* of f at x is the set of all subgradients at x

$$\partial f(x) = \{ g \in \mathbb{R}^n \ : \ g \text{ is a subgradient of } f \text{ at } x \}.$$

Let $g_1, g_2 \in \partial f(x)$ be two subgradients of f at x so that

$$\begin{aligned} f(y) &\geq f(x) + g_1^T(y - x) \\ f(y) &\geq f(x) + g_2^T(y - x). \end{aligned}$$

Let $\lambda \in [0, 1]$ and consider the linear combination of the above two inequalities, yielding

$$\begin{aligned} \lambda f(y) + (1 - \lambda) f(y) &\geq \lambda \left[f(x) + g_1^T(y - x) \right] + (1 - \lambda) \left[f(x) + g_2^T(y - x) \right] \\ \iff f(y) &\geq f(x) + \left[\lambda g_1^T + (1 - \lambda) g_2^T \right] (y - x) \\ &= f(x) + \left[\lambda g_1 + (1 - \lambda) g_2 \right]^T (y - x). \end{aligned}$$

That is, vector $\lambda g_1 + (1 - \lambda) g_2$ is a valid subgradient of f at x since it satisfies the subgradient inequality. Therefore,

$$g_1, g_2 \in \partial f(x) \implies \lambda g_1 + (1 - \lambda) g_2 \in \partial f(x), \quad \lambda \in [0, 1]$$

which informs us that $\partial f(x)$ is indeed a convex set for all $x \in \text{dom}(f)$. To show that $\partial f(x)$ is a closed set we first note that for fixed $y \in \text{dom}(f)$ the set

$$H_y = \left\{ g \mid f(y) \geq f(x) + g^T(y - x) \right\} = \left\{ g \mid f(y) - f(x) \geq g^T(y - x) \right\}$$

defines a halfspace $\left\{ z \mid b \geq a^T z \right\}$. It's easy to see that the complement $H_y^c = \left\{ g \mid f(y) - f(x) < g^T(y - x) \right\}$ is an open set since, for $a_x < b_x$, $a_x, b_x \in \mathbb{R}$,

$$\forall \, x \in H_y^c, \ \exists \, (a_x, b_x) \subset H_y^c.$$

Therefore, each H_y must be a closed set. Next, note that we may express $\partial f(x)$ as the intersection of all halfspaces H_y over all $y \in \text{dom}(f)$, i.e.,

$$
\begin{aligned}
\partial f(x) &= \left\{ g \mid f(y) \geq f(x) - g^T(y-x),\ \forall\, y \in \mathrm{dom}(f) \right\} \\
&= \bigcap_{y \in \mathrm{dom}(f)} \left\{ g \mid f(y) \geq f(x) - g^T(y-x) \right\}.
\end{aligned}
$$

Recall that a (potentially uncountable) intersection of closed sets is closed. Therefore, $\partial f(x)$ is indeed a closed set, as desired.

1.1.(ii)

Note that f is differentiable for all $x \neq 0$. Therefore, the subgradient of f at x is simply the gradient given by

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

However, if $x = 0$, we apply the definition of the subgradient

$$
\begin{aligned}
\partial f(0) &= \left\{ z \mid f(y) \geq f(0) + z^T(y-0),\ \forall\, y \in \mathrm{dom}(f) \right\} \\
&= \left\{ z \mid \|y\|_2 \geq z^T y,\ \forall\, y \in \mathrm{dom}(f) \right\} \\
&= \left\{ z \mid 1 \geq \|z\|_2 \right\}.
\end{aligned}
$$

Thus,

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

as desired.

1.1.(iii)

Let $p, q > 0$ be conjugates so that $\frac{1}{p} + \frac{1}{q} = 1$. Then, we can express the p-norm through the q-norm via the relationship

$$
\|x\|_p = \max_{\|z\|_q \leq 1} z^T x.
$$

To prove Holder's inequality we define vectors z and w such that

$$
z = \frac{x}{\|x\|_p} \quad \text{and} \quad w = \frac{y}{\|y\|_q}.
$$

Hence, by Young's inequality,

$$
\sum_k |z_k w_k| \leq \sum_k \left(\frac{|z_k|^p}{p} + \frac{|w_k|^q}{q} \right).
$$

However, by construction we find that both z and w have unit length

$$
\|z\|_p^p = 1 \quad \text{and} \quad \|w\|_q^q = 1.
$$

Thus,

$$\sum_k |z_k w_k| = \sum_k \left(\frac{|z_k|^p}{p} + \frac{|w_k|^q}{q} \right) = \frac{1}{p} + \frac{1}{q} = 1$$

so

$$sum_k |z_k w_k| \leq 1.$$

That is,

$$\begin{aligned}
&\sum_k \left| \frac{x_k}{\|x\|_p} \cdot \frac{y_k}{\|y\|_q} \right| \leq 1 \\
&\iff \frac{1}{\|x\|_p \|y\|_q} \sum_k |x_k y_k| \leq 1 \\
&\iff x^T y \leq \|x^T y\|_1 \leq \|x\|_p \|y\|_q,
\end{aligned}$$

as desired.

1.1.(iv)

We wish to show that $g \in \partial f(x) \iff g = \underset{\|z\|_q \leq 1}{\arg\max}\, z^T x$. Let $g \in \partial f(x)$, then

$$f(y) \geq f(x) + g^T(y - x) \iff \|y\|_p \geq \|x\|_p + g^T(y - x)$$

Taking $y = 0$

$$0 \geq \|x\|_p - g^T x \iff g^T x \geq \|x\|_p.$$

Taking $y = 2x$

$$\|2x\|_p = 2\|x\|_p \geq \|x\|_p + g^T x \iff g^T x \leq \|x\|_p.$$

Applying both inequalities we find

$$g^T x = \|x\|_p \iff g^T x = \max_{\|z\|_q \leq 1} z^T x \iff g = \underset{\|z\|_q \leq 1}{\arg\max}\, z^T x.$$

Next, suppose $g = \underset{\|z\|_q \leq 1}{\arg\max}\, z^T x$. Then, $\|g\|_q \leq 1$ and

$$g^T x = \|x\|_p.$$

However, recall that $\partial f(x)$ is defined as the set of vectors z satisfying $\|z\|_q \leq 1$ and $z^T x = \|x\|_p$. Therefore,

$$g \in \partial f(x) = \left\{ z \mid \|z\|_q \leq 1 \text{ and } z^T x = \|x\|_p \right\},$$

as desired.

Question 1.2

NOTE TO SELF: Check http://www.siam.org/books/mo25/mo25_ch6.pdf, check Theorem 6.6 for 1.2.(iii)

1.2.(i)

If $h(z) = \frac{1}{2}z^T Az + b^T z + c$, $A \in \mathbb{S}^n_+$ then our proximal operator is the minimizier

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

Since the proximal objective is continuous with respect to z, we may simply take the gradient of our objective to obtain

$$\begin{aligned}\frac{\partial}{\partial z}\left[\frac{1}{2}(z-x)^T(z-x) + t\left(z^T Az + b^T z + c\right)\right] &= \frac{\partial}{\partial z}\left[\frac{1}{2}z^T z - z^T x + \frac{1}{2}x^T x + t\left(z^T Az + b^T z + c\right)\right] \\ &= z - x + tz^T A + tb\end{aligned}$$

Setting this quantity to zero

$$0 = z - x + tAz + tb \implies z = (\mathbb{I} + tA)^{-1}(x - tb).$$

Therefore,

$$\text{prox}_{h,t}(x) = (\mathbb{I} + tA)^{-1}(x - tb),$$

as desired.

1.2.(ii)

Taking $h(z) = -\sum_{i=1}^n \log z_i$, $z \in \mathbb{R}^n_{++}$, we seek to solve the proximal operator

$$\text{prox}_{h,t}(x) = \arg\min_z \left\{\frac{1}{2}\|z - x\|_2^2 - t\sum_{i=1}^n \log z_i\right\}.$$

Noting that the objective is once again continuous (on $\mathbb{R}_{++}$), we take the gradient with respect to each z_i

$$\frac{\partial}{\partial z_i}\left[\frac{1}{2}\|z - x\|_2^2 - t\sum_{i=1}^n \log z_i\right] = z_i - x_i - \frac{t}{z_i}.$$

Setting this equal to zero yields

$$0 = z_i - x_i - \frac{t}{z_i} \iff z_i = \frac{1}{2}\left(x_i - \sqrt{x_i^2 - 4t}\right).$$

Thus, for $i = 1, ..., n$, we find the i^{th} component of the proximal operator to be

$$\left[\text{prox}_{h,t}(x)\right]_i = \frac{1}{2}\left(x_i - \sqrt{x_i^2 - 4t}\right),$$

as desired.

1.2.(iii)

Consider the proximal operator

$$\text{prox}_{h,t}(x) = \arg\min_z \left\{ \frac{1}{2}\|z - x\|_2^2 + t\|z\|_2 \right\}.$$

Recall that we had found the subgradient of $\|z\|_2$ to be

$$\partial h(z) = \begin{cases} \frac{z}{\|z\|_2} & \text{if } z \neq 0 \\ \{g \mid 1 \geq \|g\|_2\} & \text{if } z = 0. \end{cases}$$

Omitting the point $z = 0$ we take the derivative of our loss function and set it to zero,

$$0 = (z - x) + t\frac{z}{\|z\|_2}.$$

To solve this equality we consider the polar transform $x \mapsto (r_x, \theta_x)$ such that

$$r_x = \|x\|_2$$

and

$$\theta_x = \text{atan}\left(\frac{x_1}{x_2}\right).$$

1.2.(iv)

Finally, consider $h(z) = t\|z\|_0$ in the proximal operator

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

where $\|z\|_0$ denotes the sum of indicators

$$h(z) = \|z\|_0 = \sum_i \mathbb{I}_{\{z_i \neq 0\}}.$$

Note that,

$$t \cdot \mathbb{I}_{\{z_i \neq 0\}} = \begin{cases} t, & z_i \neq 0 \\ 0, & z_i = 0. \end{cases}$$

We can express this indicator as the sum $t \cdot \mathbb{I}(z_i) = t \cdot \mathbb{J}(z_i) + t$ for $\mathbb{J}$ given by

$$t \cdot \mathbb{J}(z_i) = \begin{cases} 0, & z_i \neq 0 \\ -t, & z_i = 0. \end{cases}$$

Section 2: Properties of Proximal Mappings and Subgradients

Question 2.1

Question 2.2

Question 2.3

Section 3: Properties of Lasso

Question 3.1

First, note that the Lagrangian of the Lasso problem is

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

for centered response vector $\mathbf{y} \in \mathbb{R}^n$ and centered design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. The solution $\widehat{\beta}_j$, $j = 1, ..., p$, to the above minimization problem must satisfy the subgradient condition

$$0 = -\frac{1}{n} \left\langle X_j, \mathbf{y} - X_j \widehat{\beta}_j \right\rangle + \lambda s_j,$$

where X_j denotes the j^{th} column/predictor of $\mathbf{X}$ and s_j is

$$s_j = \operatorname{sign}\left(\widehat{\beta}_j\right).$$

Therefore, for $\widehat{\beta}_j = 0$, $j = 1, ..., p$, we find that λ must satisfy

$$0 = -\frac{1}{n} \langle X_j, \mathbf{y} \rangle + \lambda s_j.$$

and so, for $\widehat{\beta}_j \equiv 0$, we find

$$\lambda = \left| \frac{1}{n} \langle X_j, \mathbf{y} \rangle \right|.$$

Hence, for all $\widehat{\beta}_j \equiv 0$ we must set

$$\lambda_{\max} = \max_j \left| \frac{1}{n} \langle X_j, \mathbf{y} \rangle \right|,$$

as desired.

Question 3.2

3.2.(a)

Suppose solutions $\widehat{\beta}$, $\widehat{\gamma}$ have common optimum c^* such that

$$\mathbf{X}\widehat{\beta} \neq \mathbf{X}\widehat{\gamma}.$$

Recall that the squared-loss function $f(a) = \|y - a\|_2^2$ is strictly convex, and that the ℓ_1 norm is convex, implying that the lasso minimization problem must also be strictly convex. Therefore, the solution set $\mathcal{B}$ to the lasso problem must also be convex. Thus, by convexity of $\mathcal{B}$,

$$\alpha\widehat{\beta} + (1-\alpha)\widehat{\gamma} \in \mathcal{B}$$

for $0 < \alpha < 1$. It follows that

$$\begin{aligned}\frac{1}{2}\|\mathbf{y} - \mathbf{X}\left[\alpha\widehat{\beta} + (1-\alpha)\widehat{\gamma}\right]\|_2^2 + \lambda\|\alpha\widehat{\beta} + (1-\alpha)\widehat{\gamma}\|_1 &< \alpha\left(\frac{1}{2}\|\mathbf{y} - \mathbf{X}\widehat{\beta}\|_2^2 + \lambda\|\widehat{\beta}\|_1\right) + (1-\alpha)\left(\frac{1}{2}\|\mathbf{y} - \mathbf{X}\widehat{\gamma}\|_2^2 + \lambda\|\widehat{\gamma}\|_1\right) \\ &= \alpha c^* + (1-\alpha)c^* \\ &= c^*.\end{aligned}$$

This implies that the solution of $\alpha\widehat{\beta} + (1-\alpha)\widehat{\gamma}$ attains a new optima $c^{\text{new}} < c^*$, which is a contradiction. Therefore, we must conclude

$$\mathbf{X}\widehat{\beta} = \mathbf{X}\widehat{\gamma},$$

as desired.

3.2.(b)

The statement $\|\widehat{\beta}\|_1 = \|\widehat{\gamma}\|_1$, for $\lambda > 0$, is directly implied by the above proof. Specifically, since $\mathbf{X}\widehat{\beta} = \mathbf{X}\widehat{\gamma}$, we must have that both solutions must have the same squared residuals

$$\|\mathbf{y} - \mathbf{X}\widehat{\beta}\|_2^2 = \|\mathbf{y} - \mathbf{X}\widehat{\gamma}\|_2^2,$$

and since both Lagrangian loss functions attain the same optimum c^* we find that the penalty terms must also be equal

$$\lambda\|\widehat{\beta}\|_1 = \lambda\|\widehat{\gamma}\|_1,$$

as desired.

Section 4: Convergence Rates for Proximal Gradient Descent

Question 4.(a)

Question 4.(b)

Question 4.(c)

Question 4.(d)

Question 4.(e)

Question 4.(f)

Section 5: Proximal Gradient Descent for Group Lasso

Question 5.(a)

Consider design matrix $X \in \mathbb{R}^{n\times(p+1)}$ split in J *groups* such that we may express as

$$X = \left[\mathbf{1}\ X_{(1)}\ X_{(2)}\ \cdots\ X_{(J)}\right],$$

where $\mathbf{1} = [1, ..., 1] \in \mathbb{R}^n$ and $X_{(j)} \in \mathbb{R}^{n\times p_j}$ for $\sum_j^J p_j = p$. The *group lasso* problem seeks to estimate grouped coefficients $\beta = \left[\beta_{(0)}, \beta_{(1)}, ..., \beta_{(J)}\right]$ through the minimization problem

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^{p+1}} \left\{g(\beta) + h(\beta)\right\},$$

such that g is a convex and differentiable loss function, and the group-lasso-specific h is defined as

$$h(\beta) = \lambda \sum_{j=1}^{J} w_j \left\|\beta_{(j)}\right\|_2,$$

for tuning parameter $\lambda > 0$ and weights $w_j > 0$.

5.(a).1

Recall that for convex, differentiable g and convex h, we define the proximal operator of the minimization problem

$$\min_\beta f(\beta) = \min_x \left\{g(\beta) + h(\beta)\right\}$$

to be the mapping

$$\text{prox}_{h,t}(\beta) = \arg\min_\beta \left\{\frac{1}{2}\left\|\beta - z\right\|_2^2 + t\cdot h(z)\right\}.$$

Therefore, to find the proximal operator for the group lasso problem we seek to solve

$$\text{prox}_{h,t}(\beta) = \arg\min_\beta \left\{\frac{1}{2}\left\|\beta - z\right\|_2^2 + \lambda t \sum_{j=1}^{J} w_j \left\|z_{(j)}\right\|_2\right\}.$$

Proceeding in the typical manner, we find the subgradient of the corresponding objective function to our proximal operator (with respect to group component (j))

$$\begin{aligned}\partial_{(j)}\left\{\frac{1}{2}\|\beta-z\|_2^2+\lambda t\sum_{j=1}^{J}w_j\left\|z_{(j)}\right\|_2\right\}&=\beta_{(j)}-z_{(j)}+\lambda t\cdot\partial_{(j)}\left\{\sum_{j=1}^{J}w_j\left\|z_{(j)}\right\|_2\right\}\\&=\beta_{(j)}-z_{(j)}+\lambda tw_j\cdot\partial_{(j)}\left\|z_{(j)}\right\|_2.\end{aligned}$$

From question 1.1.(ii) we find the final subgradient to be

$$\partial_{(j)}\left\|z_{(j)}\right\|_2=\begin{cases}\frac{z_{(j)}}{\left\|z_{(j)}\right\|_2} & \text{if } z_{(j)}\neq\mathbf{0}\\ \{v\,:\,\|v\|_2\leq 1\} & \text{if } z_{(j)}=\mathbf{0}.\end{cases}$$

Therefore, if $z_{(j)}\neq\mathbf{0}$ we find the subgradient to be

$$\partial_{(j)}\left\{\frac{1}{2}\|\beta-z\|_2^2+\lambda t\sum_{j=1}^{J}w_j\left\|z_{(j)}\right\|_2\right\}=\beta_{(j)}-z_{(j)}+\lambda tw_j\frac{z_{(j)}}{\left\|z_{(j)}\right\|_2}.$$

We obtain the proximal operator by setting this quantity to zero, yielding optimum

$$\begin{aligned}0&=\beta_{(j)}-z_{(j)}+\lambda tw_j\frac{z_{(j)}}{\left\|z_{(j)}\right\|_2}\\\iff z_{(j)}&=\left[\widetilde{S}_{\lambda t}\left(\beta\right)\right]_{(j)},\end{aligned}$$

where $\widetilde{S}$ is the group soft thresholding operator

$$\left[\widetilde{S}_{\lambda t}\left(\beta\right)\right]_{(j)}=\begin{cases}\beta_{(j)}-\lambda tw_j\frac{\beta_{(j)}}{\left\|\beta_{(j)}\right\|_2} & \text{if } \left\|\beta_{(j)}\right\|_2>\lambda t\\ \mathbf{0} & \text{otherwise.}\end{cases}$$

Note that in the case where $J=p$ we find $\beta_{(j)}=\beta_j\in\mathbb{R}$, so

$$\frac{\beta_{(j)}}{\left\|\beta_{(j)}\right\|_2}=\frac{\beta_j}{\|\beta_j\|_2}=\frac{\beta_j}{|\beta_j|}=\operatorname{sign}\left(\beta_j\right)=:s_j$$

Therefore,

$$\beta_j-\lambda tw_j\frac{\beta_j}{\|\beta_j\|_2}=\beta_j-\lambda tw_js_j.$$

So, if we set $w_j\equiv 1$ for all j, we obtain

$$\left[\widetilde{S}_{\lambda t}\left(\beta\right)\right]_j=\begin{cases}\beta_j-\lambda ts_j & \text{if } \beta_j>\lambda t\\ 0 & \text{otherwise,}\end{cases}$$

which is precisely the proximal operator for the (ungrouped) lasso problem.

Question 5.(i)

5.(i).(a)

For $g(\beta) = \|y - X\beta\|_2^2$ we find the gradient

$$\begin{aligned}
\nabla g(\beta) &= \nabla \left(y - X\beta\right)^T \left(y - X\beta\right) \\
&= \nabla \left[y^T y - 2\beta^T X^T y + \beta^T X^T X\beta\right] \\
&= -X^T y + X^T X\beta,
\end{aligned}$$

as desired.

5.(i).(b)

5.(i).(a)

5.(i).(c)

5.(i).(d)

Question 5.3

5.3.(i).(a)

$$\begin{aligned}
\nabla g(\beta) &= \nabla \left(\sum_{i=1}^{n} -y_i X_i\beta + \log\left(1 + \exp\left\{X_i\beta\right\}\right)\right) \\
&= \sum_{i=1}^{n} -y_i X_i + \frac{X_i \exp\left\{X_i\beta\right\}}{1 + \exp\left\{X_i\beta\right\}}
\end{aligned}$$

5.3.(i).(b)

5.3.(i).(c)

Section 6: Practice with KKT Conditions and Duality