

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

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

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

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

defines a halfspace $\{z \mid b \geq a^T z\}$. It's easy to see that the complement $H_y^c = \{g \mid f(y) - f(x) < g^T(y - x)\}$ 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.,

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

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, at $x = 0$, we apply the definition of the subgradient

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

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 $g \in \partial f(x) \iff g = \arg \max_{\|z\|_q \leq 1} z^T x$. First, 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

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

Next, suppose $g = \arg \max_{\|z\|_q \leq 1} z^T x$. 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,

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

as desired.

Question 1.2

1.2.(i)

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

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

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

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

Setting this quantity to zero

$$0 = z - x + t A z + t b \implies z = (\mathbb{I} + t A)^{-1} (x - t b).$$

Therefore,

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

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 $i = 1, \dots, n$, we find the i^{th} component of the proximal operator to be

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

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 gradient of our proximal objective function and set it to zero,

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

To solve this we consider the map to polar coordinates $x \mapsto (r_x, \theta_x)$ where

$$\begin{aligned} r_x &= \|x\|_2 \\ \theta_x &= \tan^{-1} \left(\frac{x_1}{x_2} \right). \end{aligned}$$

Note that both terms of the above gradient $\frac{z}{\|z\|_2}$ and $x - z$ must have the same angle such that the angle of $\frac{z}{\|z\|_2}$ and z must be equation to either the positive or negative angle of x . This informs us that $z = ax$, for any $a \in \mathbb{R}$. Substituting this expression for z into our gradient yields

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

and so

$$a = \begin{cases} \frac{r_x - t}{r_x} & \text{if } r_x > t \\ 0 & \text{else.} \end{cases} \quad (1)$$

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

$$\text{prox}_{h,t}(x) = \begin{cases} x \frac{\|x\|_2 - t}{\|x\|_2} & \text{if } \|x\|_2 > t \\ 0 & \text{else,} \end{cases} \quad (2)$$

as desired.

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

2.(b)

We wish to show that, for $\forall x, y \in \mathbb{R}$, $u \in \partial f(x)$, and $v \in \partial f(y)$,

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

To see this, first note that if $u \in \partial f(x)$ then 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

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

so

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

Therefore, putting these two inequalities together,

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

as desired.

2.(d)

We wish to show

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

First, recall that

$$\text{prox}_t(x) = \arg \min_u \left\{ \frac{1}{2t} \|x - u\|_2^2 + h(u) \right\}.$$

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

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

as desired.

2.(e)

Section 3: Properties of Lasso

Question 3.1

First, note that the Lagrangian of the Lasso problem is

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

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

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

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

$$s_j = \text{sign}(\hat{\beta}_j).$$

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

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

and so, for $\hat{\beta}_j \equiv 0$, 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

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

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)

We wish to show that

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

is a subgradient of h evaluated at $x^{(i)}$. Note that h is convex, but not necessarily differentiable, and recall that from Question 2.(d) we had shown 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,

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

That is, s precisely satisfies the definition of a subgradient, $s \in \partial h(x^{(i)})$, as desired.

Question 4.(b)

We 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 our objective function f can be decomposed as

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

for g convex and differentiable, with ∇g being Lipschitz, and h convex. Therefore, f must also be convex. Now, 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,

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

On the other hand, s is a subgradient of h at $x^{(i)}$, $s \in \partial h(x^{(i)})$, so

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

Rearranging the above expressions, it follows that

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

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

Question 4.(c)

We now wish to show that the sequence $\{f(x^{(i)})\}$ is nonincreasing for $i = 0, \dots, k$. That is, we wish to show, for $i = 1, \dots, k$,

$$f(x^{(i)}) \leq f(x^{(i-1)}).$$

We 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

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

as desired.

Question 4.(d)

We will now derive the following inequality

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

for x^* the minimizer of f (where $f(x^*)$ is assumed to be finite). Using the previous inequality, and the fact that $f(x^{(i)}) \leq f(x^*)$,

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

Note 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

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

Furthermore, we have that $G_t(x^{(i-1)}) = \frac{1}{t}(x^{(i-1)} - x^{(i)})$. Hence,

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

as desired.

Question 4.(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,

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

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.

Question 4.(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, while

$$f(x - t\nabla f(x)) > f(x) - \frac{t}{2} \|\nabla f(x)\|_2^2,$$

shrink the step size $t = \beta t$. Now, in the context of this problem,

$$f(x - t\nabla f(x)) > f(x) - \frac{t}{2} \|\nabla f(x)\|_2^2$$

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

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

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

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

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

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

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

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

to be the mapping

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

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

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

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

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

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

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

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

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

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

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

where \tilde{S} is the group soft thresholding operator

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

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

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

Therefore,

$$\beta_j - \lambda t w_j \frac{\beta_j}{\|\beta_j\|_2} = \beta_j - \lambda t w_j s_j.$$

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

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

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

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

as desired.

5.(i).(b)

We load our data

```
X <- as.matrix(read.csv("../data/birthwt/X.csv"))
y <- as.matrix(read.csv("../data/birthwt/y.csv"))

yc <- scale(y, scale = F)
Xc <- scale(X, scale = F)
ybar <- attributes(yc)$`scaled:center`
Xbar <- attributes(Xc)$`scaled:center`
```

and define some useful functions

```
norm_p <- function(v, p) {
  sum(abs(v)^p)^(1/p)
}
grad_g <- function(X, y, b) {
  -crossprod(X, y - X %*% b)
}
Stilde_groupj <- function(beta_groupj, lambda, t_step, w_groupj) {
  beta_groupj_norm2 <- norm_p(beta_groupj, 2)

  beta_groupj/beta_groupj_norm2 *
    max(beta_groupj_norm2 - lambda * t_step * w_groupj, 0)
}
```

Next, we set some parameters and define the group structure, as well as initialize our solution $\beta^{(0)} = \mathbf{0}$

```
fstar <- 84.5952
lambda <- 4
t_step <- 0.002
max_steps <- 1e3
group_idx <- list()
group_idx[[1]] <- 1:3 # age1, age2, age3
group_idx[[2]] <- 4:6 # lwt1, lwt2, lwt3
group_idx[[3]] <- 7:8 # white, black
group_idx[[4]] <- 9 # smoke
group_idx[[5]] <- 10:11 # ptl1, ptl2m
group_idx[[6]] <- 12 # ht
group_idx[[7]] <- 13 # ui
```

```
group_idx[[8]] <- 14:16 # ftv1, ftv2, ftv3m
n_groups <- length(group_idx)

w <- sapply(group_idx, function(groupj) sqrt(length(groupj)))
```

First, we compute the traditional proximal gradient descent algorithm

```
beta_init <- rep(0, ncol(Xc))
beta <- matrix(nrow = max_steps, ncol = length(beta_init))
beta[1, ] <- beta_init - t_step * grad_g(Xc, yc, beta_init)

for (k in 2:max_steps) {
  # update step
  beta[k, ] <- beta[k - 1,] - t_step * grad_g(Xc, yc, beta[k - 1,])

  # proximal step
  for (j in 1:n_groups) {
    beta[k, group_idx[[j]]] <-
      Stilde_groupj(beta[k, group_idx[[j]]], lambda, t_step, w[j])
  }
}

beta_prox_sol <- beta[max_steps,] # extract solution

f <- apply(beta, 1, function(b) {
  h <- lambda * sum(w * sapply(group_idx, function(groupj) norm_p(b[groupj], 2)))
  crossprod(yc - Xc %*% b) + h
})
```

Next, we implement the accelerated proximal algorithm

```
beta_init_m1 <- rep(0, ncol(Xc))
beta_init_00 <- rep(0, ncol(Xc))

beta <- matrix(nrow = max_steps + 2, ncol = ncol(Xc))
beta[1, ] <- beta_init_m1
beta[2, ] <- beta_init_00
#beta[1, ] <- beta_init - t_step * grad_g(Xc, yc, beta_init)

for (k in 3:nrow(beta)) {
  # momentum step
  v <- beta[k - 1,] + (k - 4)/(k - 1) * (beta[k - 1,] - beta[k - 2,])
  # update step
  beta[k, ] <- v - t_step * grad_g(Xc, yc, beta[k - 1,])

  # proximal step
  for (j in 1:n_groups) {
    beta[k, group_idx[[j]]] <-
      Stilde_groupj(beta[k, group_idx[[j]]], lambda, t_step, w[j])
  }
}

f_acc <- apply(beta, 1, function(b) {
  h <- lambda * sum(w * sapply(group_idx, function(groupj) norm_p(b[groupj], 2)))
  crossprod(yc - Xc %*% b) + h
})
```

```

})

acc_min_idx <- which(f_acc == min(f_acc))
beta_acc_prox_sol <- beta[acc_min_idx,] # extract solution

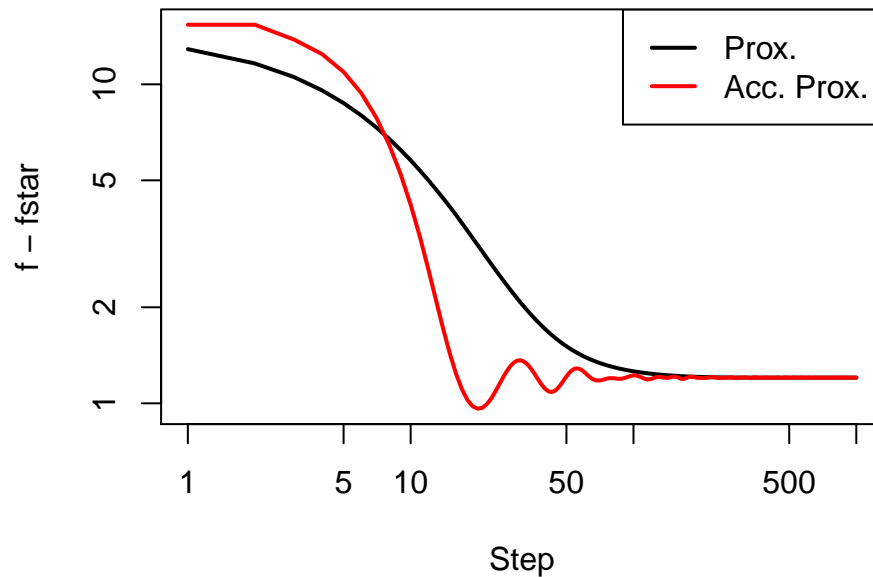
```

Finally, we visualize the results

```

plot(f - fstar, ylim = range(c(f, f_acc) - fstar),
     xlab = "Step", ylab = "f - fstar",
     log = 'xy', type = 'l', lwd = 2)
lines(f_acc - fstar, col = 'red', lwd = 2)
legend("topright", legend = c("Prox.", "Acc. Prox."),
      lwd = 2, seg.len = 1.5, col = c("black", "red"))

```



5.(i).(c)

We now display the estimated coefficients of both the proximal and accelerated proximal algorithms

```

round(beta_prox_sol, 4)

## [1] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.2449 -0.0616
## [9] -0.2443 -0.1166 0.0110 -0.0962 -0.3786 0.0000 0.0000 0.0000

round(beta_acc_prox_sol, 4)

## [1] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.2680 -0.0626
## [9] -0.2720 -0.1378 0.0130 -0.0648 -0.4036 0.0000 0.0000 0.0000

```

In both algorithms we see that predictors (7 = white, 8 = black), (9 = smoke), (10 = pt11, 11 = pt12), (12 = ht), and (13 = ui) are selected, corresponding to groups 3, 4, 5, 6, 7.

5.(i).(d)

Using the same framework we now compute the lasso with $\lambda = 0.35$


```

##### LASSO #####
lambda <- 0.35
t_step <- 0.002
max_steps <- 1e4
group_idx <- list()
for (i in 1:ncol(Xc))
  group_idx[[i]] <- i
n_groups <- length(group_idx)

w <- sapply(group_idx, function(groupj) sqrt(length(groupj)))
beta_init <- rep(0, ncol(Xc))

beta_lasso <- matrix(nrow = max_steps, ncol = length(beta_init))
beta_lasso[1, ] <- beta_init - t_step * grad_g(Xc, yc, beta_init)

for (k in 2:max_steps) {
  # update step
  beta_lasso[k,] <- beta_lasso[k - 1,] - t_step * grad_g(Xc, yc, beta_lasso[k - 1,])

  # proximal step
  for (j in 1:n_groups) {
    beta_lasso[k, group_idx[[j]]] <-
      Stilde_groupj(beta_lasso[k, group_idx[[j]]], lambda, t_step, w[j])
  }
}

f <- apply(beta_lasso, 1, function(b) {
  h <- lambda * sum(w * sapply(group_idx, function(groupj) norm_p(b[groupj], 2)))
  crossprod(yc - Xc %*% b) + h
})

```

Comparing the lasso results to the proximal and accelerated proximal results

```
round(beta_prox_sol, 4)
```

```
## [1] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.2449 -0.0616
## [9] -0.2443 -0.1166 0.0110 -0.0962 -0.3786 0.0000 0.0000 0.0000
```

```
round(beta_acc_prox_sol, 4)
```

```
## [1] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.2680 -0.0626
## [9] -0.2720 -0.1378 0.0130 -0.0648 -0.4036 0.0000 0.0000 0.0000
```

```
round(beta_lasso[max_steps,], 4)
```

```
## [1] 0.0000 1.2000 0.5571 1.4376 0.0000 0.9979 0.3063 -0.1146
## [9] -0.2846 -0.3004 0.1363 -0.5059 -0.4722 0.0827 0.0027 -0.1292
```

We find that the lasso solution does not apply groupwise sparsity, instead setting some predictors to zero in the same group as nonzero predictors.

5.3.(i).(a)

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

5.3.(i).(b)

5.3.(i).(c)

Section 6: Practice with KKT Conditions and Duality

We begin with the usual least squares problem,

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2.$$

Note that the corresponding primal problem is given by

$$\min_{v \in \mathbb{R}^n} \frac{1}{2} \|v\|_2^2 \quad \text{subject to } y = X\beta + v.$$

In this form we see that the Lagrangian is the function

$$L(v, \beta, \lambda) = \frac{1}{2} \|v\|_2^2 + \lambda(y - X\beta - v).$$

It follows that the first order necessary conditions are

$$\begin{aligned}0 &= \frac{\partial L}{\partial v} = v - \lambda \cdot \mathbf{1} \\ 0 &= \frac{\partial L}{\partial \beta} = -X\lambda \\ 0 &= \frac{\partial L}{\partial \lambda} = y - X\beta - v.\end{aligned}$$

Note that from these first order conditions we find

$$v = \lambda \cdot \mathbf{1} \quad \text{and} \quad v^T v = v^T \lambda,$$

permitting us to simplify 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 problem is given by

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

as desired.