AMATH 515 Homework #2 - Vinsensius
Due: Monday, February 13th, by 11:59 pm

Number 1

Let $x, y \in \mathbb{R}^n$, and consider a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$. We make the following definitions:

$$\text{prox}_{tf}(y) := \arg\min_x \frac{1}{2t}\|x - y\|^2 + f(x)$$
$$f_t(y) := \min_x \frac{1}{2t}\|x - y\|^2 + f(x).$$

Notice that $\text{prox}_{tf}(y)$ is the minimizer of an optimization problem; in particular it is a vector in $\mathbb{R}^n$, On the other hand $f_t(y)$ is a function from $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, just as f.

Suppose f is convex.

1. Show that f_t is convex.
2. Show that $\text{prox}_{tf}(y)$ is uniquely defined for any input y.
3. Compute prox_{tf} and f_t, where $f(x) = \|x\|_1$.
4. Compute prox_{tf} and f_t for $f = \delta_{\mathbb{B}_\infty}(x)$, where $\mathbb{B}_\infty = [-1, 1]^n$.

1. Given the definition of $f_t(y)$

$$f_t(y) := \min_x \frac{1}{2t}\|x - y\|^2 + f(x) = \min_x h(x, y)$$

we will denote the whole function as $h(x, y)$ for simplicity.

We want to prove that

$$f_t(\lambda y_1 + (1 - \lambda)y_2) \leq \lambda f_t(y_1) + (1 - \lambda)f_t(y_2)$$

To prove that f_t is convex, we will use the fact that the h(x,y) is (strongly) convex and has unique minimizer for any input of y from the next part.

Let x_1 be the unique minimizer for input y_1 and x_2 be unique minimizer for y_2.

This means that the following equations are true

$$f_t(y_1) = \frac{1}{2t}\|x_1 - y_1\|^2 + f(x_1) = h(x_1, y_1)$$
$$f_t(y_2) = \frac{1}{2t}\|x_2 - y_2\|^2 + f(x_2) = h(x_2, y_2)$$

because of the uniqueness of the minimizer.

Now, we can start proving the f_t is convex.

$$\begin{aligned}
f_t(\lambda y_1 + (1-\lambda)y_2) &= \min_x h(x, \lambda y_1 + (1-\lambda)y_2) \\
&= h(\lambda x_1 + (1-\lambda)x_2, \lambda y_1 + (1-\lambda)y_2) && (x_1 \text{ and } x_2 \text{ are uniquely chosen}) \\
&\leq h(\lambda x_1, \lambda y_1) + h((1-\lambda)x_2, (1-\lambda)y_2) && \text{(By convexity)} \\
&= \lambda h(x_1, y_1) + (1-\lambda)h(x_2, y_2) && \text{(convexity of f(x) and norm)} \\
&= \lambda f_t(y_1) + (1-\lambda)f_t(y_2)
\end{aligned}$$

Thus, f_t is convex.

2. In order to prove the uniqueness of minimizers we need to the function to be bounded from below and quadratically goes to infinity.

   Firstly, we will assume the function f is convex and is bounded from below by quadratic function.

   The function $\frac{1}{2t}\|x - y\|^2$ is strongly convex because it has quadratic lower bound. Thus, the whole function is strongly convex, because sum of convex and strongly convex is strongly convex function, so it has quadratic lower bound.

   When we apply optimality condition, we found that

$$\frac{1}{t}(y - z) \in \partial f(z)$$

   From the subdifferential, we show that a minimizer exists.

   Since the whole function is bounded below by quadratic and has a minimizer, the function goes to infinity in all direction.

   Thus, the level sets of the function are compact. This means that it has closed epigraph. Therefore, the function has unique minimizer for any input y.

3. We will be considering each coordinate case to find the prox_{tf}.

$$\text{prox}_{tf}(y_i) = \arg\min_{x_i} \frac{1}{2t}\|x_i - y_i\|^2 + |x_i|$$

Appply the optimality condition,

$$
\begin{aligned}
&0 \in \partial\left(\frac{1}{2t}|x_i - y_i|^2 + |x_i|\right) \\
&0 \in (x_i - y_i) + t\partial(|x_i|)
\end{aligned}
$$

We define $v \in \partial(|x_i|)$ So, the equation becomes

$$
\begin{aligned}
0 &= (x_i - y_i) + tv \\
x_i^* &= y_i - tv
\end{aligned}
$$

Where x_i^* is the prox_{tf} . The definition of $\partial(|x_i|)$ is

$$
v = \begin{cases} 1 & x_i^* > 0 \\ [-1, 1] & x_i^* = 0 \\ -1 & x_i^* < 0 \end{cases}
$$

We now can find the conditions of y_i given the subdifferential definition.

- If $x_i^* > 0$,

$$
y_i - t > 0 \implies y_i > t
$$

- If $x_i^* = 0$,

$$
\begin{aligned}
y_i - t[-1, 1] = 0 \\
y_i = [-t, t] \implies |y_i| \leq t
\end{aligned}
$$

- If $x_i^* < 0$,

$$
y_i + t > 0 \implies y_i < -t
$$

Substituting conditions for y_i and v we get the conditions for x_i^*

$$
\begin{aligned}
x_i^* &= \begin{cases} y_i - t & y_i > t \\ 0 & |y_i| \leq t \\ y_i + t & y_i < -t \end{cases} \\
&= \text{sign}(y_i) \max(|y_i| - t, 0)
\end{aligned}
$$

To find the f_t, we will substitute x^* back into the equation. Since x^* is the vector of points that minimizes the whole function, the min operator can be dropped

$$\begin{aligned} f_t(y) &= \min_x \frac{1}{2t}\|x-y\|^2 + \|x\|_1 \\ &= \frac{1}{2t}\|x^*-y\|^2 + \|x^*\|_1 \\ &= \frac{1}{2t}\sum_{i=1}^{\infty}(x_i^*-y_i)^2 + |x_i^*| \end{aligned}$$

4. Firstly we calculate the $\text{prox}_{t\delta_{\mathbb{B}_\infty}}$

$$\begin{aligned} \text{prox}_{t\delta_{\mathbb{B}_\infty}} &= \arg\min_x \frac{1}{2t}\|x-y\|^2 + \delta_{\mathbb{B}_\infty}(x) \\ &= \arg\min_{x\in\mathbb{B}_\infty} \frac{1}{2t}\|x-y\|^2 \\ &= \text{proj}_{\mathbb{B}_\infty}(y) \end{aligned}$$

From lecture in class the projection on to the $\mathbb{B}_\infty$ is

$$\text{proj}_{\mathbb{B}_\infty}(y_i) = x_i^* = \max\left(\min\left(1, y_i\right), 0\right)$$

To find the f_t, we will substitute x^* back into the equation. Since x^* is the vector of points that minimizes the whole function, the min operator can be dropped

$$\begin{aligned} f_t(y) &= \min_x \frac{1}{2t}\|x-y\|^2 + \delta_{\mathbb{B}_\infty}(x) \\ &= \frac{1}{2t}\|x^*-y\|^2 \\ &= \frac{1}{2t}\sum_{i=1}^{\infty}(\max\left(\min\left(1, y_i\right), 0\right) - y_i)^2 \end{aligned}$$

Number 2

More prox identities.

1. Suppose f is convex and let $g_s(x) = f(x) + \frac{1}{2s}\|x - x_0\|^2$. Find the formula for prox_{tg} in terms of prox_{tf}.

2. Let $f(x) = \|x\|_2$. Write $\text{prox}_{tf}(y)$ in closed form.

3. Let $f(x) = \frac{1}{2}\|x\|_2^2$. Write $\text{prox}_{tf}(y)$ in closed form.

4. Let $f(x) = \frac{1}{2}\|Cx\|^2$. Write $\text{prox}_{tf}(y)$ in closed form.

1. We apply optimality condition on $\text{prox}_{tf}(y)$, let $\epsilon = \text{prox}_{tf}(y)$

$$\begin{aligned} 0 &\in \frac{1}{t}(\epsilon - y) + \partial f(\epsilon) \\ \frac{1}{t}(y - \epsilon) &\in \partial f(\epsilon) \end{aligned}$$

   So our goal to make a similar equation form as above when we applied optimality condition on $\text{prox}_{tg}(y)$. Let $z = \text{prox}_{tg}(y)$.

$$\begin{aligned} 0 &\in \frac{1}{t}(z - y) + \frac{1}{s}(z - x_0) + \partial f(z) \\ \frac{1}{t}(y + z) + \frac{1}{s}(x_0 - z) &\in \partial f(z) \\ \frac{t+s}{ts}\left(\frac{ys + x_0 t}{t+s} - z\right) &\in \partial f(z) \\ \frac{1}{\frac{ts}{t+s}}\left(\frac{ys + x_0 t}{t+s} - z\right) &\in \partial f(z) \end{aligned}$$

   So, we can write $\text{prox}_{tg}(y)$ as

$$z = \text{prox}_{tg}(y) = \text{prox}_{\left(\frac{ts}{t+s}f\right)}\left(\frac{ys + x_0 t}{t+s}\right)$$

2. We can decompose vector x in the direction of y and perpendicular to y.

$$x = p\frac{y}{\|y\|} + z$$

where $\langle y, z\rangle = 0$. Then the objective function becomes

$$\frac{1}{2t}\|x-y\|^2 + \|x\|_2 = \frac{1}{2t}\|z\|^2 + \frac{1}{2t}\left(p - \|y\|\right)^2 + \sqrt{p^2 + \|z\|^2}$$

Note that the expression is minimized when $z = 0$, so the problem reduced to 1-D problem

$$\min_p \frac{1}{2t}\left(p - \|y\|\right)^2 + |p|$$

We apply optimality condition

$$0 \in \frac{1}{t}(p - \|y\|) + \partial(|p|)$$
$$\frac{1}{t}(p - \|y\|) + v = 0$$
$$p^* = \|y\| - tv$$

Because of the the value of $\|\cdot\| \geq 0$, so v is restricted to only 1. So, the final result will be

$$p^* = \|y\| - t$$

When $\|y\| < t$, we want the p value to be 0 such that x does not go too far from y in the opposite direction, so the expression is

$$p^* = \max(\|y\| - t, 0)$$

Thus, the $\text{prox}_{tf}(y)$ is

$$x = \max(\|y\| - t, 0)\frac{y}{\|y\|}$$

3. Since f(x) is differentiable, we can find the gradient of the function that we minimizing and set it to 0 to find the $\text{prox}_{tf}(y)$.

$$\nabla\left\{\frac{1}{2t}\|x-y\|^2 + \frac{1}{2}\|x\|_2^2\right\} = 0$$
$$\frac{1}{t}(x-y) + x = 0$$
$$x^* = \frac{y}{1+t}$$

Where x^* is the $\text{prox}_{tf}(y)$.

4. Since f(x) is differentiable, we can find the gradient of the function that we minimizing and set it to 0 to find the $\text{prox}_{tf}(y)$.

$$\begin{aligned}\nabla\left\{\frac{1}{2t}\|x-y\|^2+\frac{1}{2}\|Cx\|^2\right\}&=0\\ \frac{1}{t}(x-y)+C^TCx&=0\\ x^*&=\left[I+tC^TC\right]^{-1}y\end{aligned}$$

Where x^* is the $\text{prox}_{tf}(y)$. Since C^TC and I are HPD, their linear combination is invertible.

Number 3

Complete three generic solvers we learned from the class in `solvers.py`, including,

- proximal gradient descent,
- accelerated gradient descent.
- accelerated proximal gradient descent.

The answer are in `solvers.py`.

Number 4

Compressive sensing, consider the sparse regression problem,

$$\min_x \quad \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1$$

where $A \in \mathbb{R}^{m\times n}$ and $m < n$. When x is sparse, it is possible to recover using the ℓ_1 regularizer. We choose $\lambda = \|A^\top b\|_\infty/10$.

(a) By treating $f(x) = \frac{1}{2}\|Ax - b\|^2$ and $g(x) = \lambda\|x\|_1$, complete the function w.r.t. to f and g.

(b) Apply the proximal gradient algorithm. Do you recover the signal?

(c) Apply accelerated proximal gradient, is it faster than method of (b)?

(a) Completed in the code

(b) Yes we recover the signal

(c) it is faster because it converged with more iteration than proximal gradient descent.

Number 5

Logistic regression on MNIST data, recall the logistic regression problem,

$$\min_x \quad \sum_{i=1}^{m} \{\ln(1+\exp(\langle a_i, x\rangle)) - b_i\langle a_i, x\rangle\} + \frac{\lambda}{2}\|x\|^2.$$

We will use logistic regression to classify the "0" and "1" images from MNIST. In this example, a_i is our vectorized image, and b_i is the corresponding label. We want to obtain an classifier, so that for a new image a_{new}, we can predict

$$\begin{cases} a_{\text{new}} \text{ is a } 0, & \text{if } \langle a_{\text{new}}, x\rangle \leq 0 \\ a_{\text{new}} \text{ is a } 1, & \text{if } \langle a_{\text{new}}, x\rangle > 0 \end{cases}.$$

(a) Complete the function, gradient and Hessian for the logistic regression.

(b) Apply gradient, accelerate gradient and Newton's method to solve the problem. Which one is the fastest and which one is the slowest?

(c) What is your accuracy of the classification for the test data.

1. Done in the code
2. The fastest method is Newton's method because it converged with least iteration while the other 2 methods reach maximum number of iterations.
3. The accuracy is 100%.