Name: Anna Huang

AMATH 515 Homework Set 2, Due Feb 13, 11:59 pm.

(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.

(a) Show that f_t is convex.

Answer to 1a: Since $f_t(y) = \min_x \frac{1}{2t}\|x - y\|^2 + f(x)$, then it must be less than or equal to $F(x, y) = \frac{1}{2t}\|x - y\|^2 + f(x)$ because $f_t(y)$ is the min of x.

$$f_t(y_i) <= F(x_i, y_i)$$

Using the strictly convex inequality and reassigning:

$$F(\lambda x_1 + (1 - \lambda)x_2, \lambda y_1 + (1 - \lambda)y_2) <= \lambda F(x_1, y_1) + (1 - \lambda)F(x_2, y_2)$$
$$F(x_3, y_3) <= \lambda F(x_1, y_1) + (1 - \lambda)F(x_2, y_2)$$

And once again, $f_t(y_3) <= F(x_3, y_3)$ since $f_t(y)$ is the minimizer. So to continue the inequality:

$$F(x_3, y_3) <= \lambda F(x_1, y_1) + (1 - \lambda)F(x_2, y_2)$$
$$f_t(y_3) <= F(x_3, y_3) <= \lambda F(x_1, y_1) + (1 - \lambda)F(x_2, y_2)$$
$$f_t(y_3) <= F(x_3, y_3) <= \lambda f(y_1) + (1 - \lambda)f(y_2)$$
$$f_t(y_3) <= \lambda f(y_1) + (1 - \lambda)f(y_2)$$

Now plugging y_3 back into $f_t(y_3)$:

$$f_t(\lambda y_1 + (1 - \lambda)y_2) <= \lambda f(y_1) + (1 - \lambda)f(y_2)$$

As a result, f_t is convex, as demonstrated above by the strictly convex inequality.

(b) Show that $\text{prox}_{tf}(y)$ is uniquely defined for any input y.

Answer to 1b: f_t is strictly convex, as demonstrated in 1a, but utilizing the assumption that f is continuous and bounded from below, then the sum of the function of the quadratic has to go to ∞ in any direction if you slice any level set. As a result, there is a unique minimizer x_1 for every y_i for the inequality to be true. Additionally, $f_t(y)$ is a sum of the convex function $f(x)$ and a simple quadratic $\frac{1}{2t}||x-y||^2$. It is guaranteed for every y_i, and there will be an unique x_i that solves it. This results in the following:

$$
\begin{aligned}
f_t(y) &= \min_x \frac{1}{2t}\|x-y\|^2 + f(x) \\
f_t(y_i) &= \min_{x_i} \frac{1}{2t}\|x_i-y_i\|^2 + f(x_i) \\
f_t(y_i) &= \frac{1}{2t}\|x_i-y_i\|^2 + f(x_i)
\end{aligned}
$$

For this to be true, then the $prox_{tf}(y)$ also exists and is unique.

$$
\begin{aligned}
\text{prox}_{tf}(y) &= \arg\min_x \frac{1}{2t}\|x-y\|^2 + f(x) \\
\text{prox}_{tf}(y_i) &= \arg\min_{x_i} \frac{1}{2t}\|x_i-y_i\|^2 + f(x_i) \\
\text{prox}_{tf}(y_i) &= \frac{1}{2t}\|x_i-y_i\|^2 + f(x_i)
\end{aligned}
$$

(c) Compute $\text{prox}_{tf}(y)$ and f_t, where $f(x) = \|x\|_1$.

Answer to 1c: The $\text{prox}_{tf}(y)$ is:

$$
\begin{aligned}
\text{prox}_{tf}(y) &= \arg\min_x \frac{1}{2t}\|x-y\|^2 + f(x) \\
\text{prox}_{tf}(y) &= \arg\min_{\hat{x}} \frac{1}{2t}\|\hat{x}-y\|^2 + \|x\|_1
\end{aligned}
$$

$$
\text{prox}_{tf}(y) = \begin{cases} y > t & \frac{1}{t} * (\hat{x} - y) + 1 \\ y < -t & \frac{1}{t} * (\hat{x} - y) - 1 \\ y \in [-t, t] & 0 \end{cases}
$$

Therefore the $\hat{x} = y - t$ for $\hat{x} > t$ and $\hat{x} = y + t$ for $\hat{x} < -t$. This gives the following:

$$
\hat{x} = \begin{cases} y > t & y - t \\ y < -t & y + t \\ y \in [-t, t] & 0 \end{cases}
$$

This results in $\max(\min(y+t,0), y-t)$. The f_t is the following when plugging $\hat{x}$ back into f_t:

$$f_t = \begin{cases} y > t & \frac{1}{2t}\|-t\|^2 + f(y-t) \\ y < -t & \frac{1}{2t}\|+t\|^2 + f(y+t) \\ y \in [-t,t] & \frac{||y||^2}{2} \end{cases}$$

$$f_t = \begin{cases} y > t & \frac{1}{2t}\|-t\|^2 + \|y-t\|_1 \\ y < -t & \frac{1}{2t}\|+t\|^2 + \|y+t\|_1 \\ y \in [-t,t] & \frac{||y||^2}{2} \end{cases}$$

(d) Compute prox_{tf} and f_t for $f = \delta_{\mathbb{B}_\infty}(x)$, where $\mathbb{B}_\infty = [-1,1]^n$.
Answer to 1d: The prox_{tf} for the B_∞ is reduce to each coordinate (x_i, y_i):

$$\text{prox}_{tf} = \begin{cases} -1 \leq y_i \leq 1, & y_i \\ y_i > 1, & -1 \\ y_i < -1, & 1 \end{cases}$$

$$\text{prox}_{tf} = \max[(\min(1, x_i), -1]$$

Since -1 and 1 are in the set of the indicator function, then $f(-1) = \delta_{\mathbb{B}_\infty}(-1) = 0$ and $f(1) = \delta_{\mathbb{B}_\infty}(1) = 0$. The f_t of the B_∞ is the following:

$$f_t = \begin{cases} -1 \leq y_i \leq 1, & 0 \\ y_i > 1, & \frac{1}{2t}(-1-y_i)^2 \\ y_i > 1, & \frac{1}{2t}(1-y_i)^2 \end{cases}$$

(2) More prox identities.

(a) 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}.

Answer to 2a: The $\text{prox}_{tg}(y)$ below:

$$\begin{aligned}
f_t(y) &= \min_x \frac{1}{2t}\|x-y\|^2 + f(x) \\
g_t(y) &= = \min_x \frac{1}{2t}\|x-y\|^2 + \frac{1}{2s}||x-x_0||^2 + f(x) \\
\text{prox}_{tf}(y) &= \arg\min_x \frac{1}{2t}\|x-y\|^2 + f(x) \\
\text{prox}_{tg}(y) &= \arg\min_x \frac{1}{2t}\|x-y\|^2 + f(x) + \frac{1}{2s}||x-x_0||^2
\end{aligned}$$

First we set $z = \text{prox}_{tg}(y)$ and given the optimality conditions:

$$\begin{aligned}
0 \in \frac{1}{t}(z-y) &\in \partial g(x) \\
\frac{1}{t}(y-z) &\in \partial g(x)
\end{aligned}$$

This gives the following with $g(x)$ to find the $\partial g(x)$:

$$\begin{aligned}
g(x) &= \frac{1}{2s}||x-x_0||^2 + f(x) \\
\partial g(x) &= \frac{1}{s}(x-x_0) + \partial f(x)
\end{aligned}$$

Since $\frac{1}{t}(y-z) \in \partial g(x)$, this results:

$$\begin{aligned}
\frac{1}{t}(y-z) &= \frac{1}{s}(x-x_0) + \partial f(x) \\
\frac{1}{t}(y-z) + \frac{1}{s}(x_0-x) &= \partial f(x)
\end{aligned}$$

Since $x = z$:

$$\frac{1}{t}(y - z) + \frac{1}{s}(x_0 - z) = \partial f(z)$$
$$s(y - z) + t(x_0 - z) = st\partial f(z)$$
$$sy - sz + tx_0 - tz = st\partial f(z)$$
$$sy + tx_0 - sz - tz = st\partial f(z)$$
$$sy + tx_0 - z(s + t) = st\partial f(z)$$
$$\frac{sy + tx_0}{s + t} - z = \frac{st}{s + t}\partial f(z)$$
$$\frac{s + t}{st}(\frac{sy + tx_0}{s + t} - z) = \partial f(z)$$
$$\frac{s + t}{st}(\frac{sy + tx_0}{s + t} - z) \in \partial f(z)$$

As compared to the optimality conditions: $\frac{1}{t}(y - z) \in \partial f(x)$. We set $\hat{y}$ and $\hat{t}$ as the following:

$$\hat{y} = \frac{sy + tx_0}{s + t}$$
$$\hat{t} = \frac{st}{s + t}$$

This will result in:

$$z = \text{prox}_{\hat{t}\hat{f}}(\frac{sy + tx_0}{s + t})$$
$$z = \text{prox}_{\hat{t}\hat{f}}(\hat{y})$$
$$z = \text{prox}_{tg}(y) = \text{prox}_{\hat{t}\hat{f}}(\hat{y})$$

(b) Let $f(x) = \|x\|_2$. Write $\text{prox}_{tf}(y)$ in closed form.
Answer to 2b: First setting $x = my$ where m is a placeholder. The $\text{prox}_{tf}(y)$ below:

$$\begin{aligned}
\text{prox}_{tf}(y) &= \arg\min_x \frac{1}{2t}\|x-y\|^2 + f(x) \\
\text{prox}_{tf}(y) &= \arg\min_x \frac{1}{2t}\|x-y\|^2 + \|x\|_2 \\
\text{prox}_{tf}(y) &= \arg\min_x \frac{1}{2t}\|my-y\|^2 + \|my\|_2 \\
\text{prox}_{tf}(y) &= \arg\min_x \frac{1}{2t}(m-1)^2\|y\|^2 + m\|y\|_2 \\
\nabla_m &= \frac{1}{t} * (m-1)\|y\|^2 + \|y\|_2 = 0 \\
m &= 1 - \frac{t}{\|y\|} \\
\frac{x}{y} &= 1 - \frac{t}{\|y\|} \\
x &= y - \frac{yt}{\|y\|}
\end{aligned}$$

However, $y \neq 0$ or else the expression does not exist, so it would be:

$$\text{prox}_{tf}(y) = \max(y - \frac{yt}{\|y\|}, 0)$$

(c) Let $f(x) = \frac{1}{2}\|x\|_2^2$. Write $\text{prox}_{tf}(y)$ in closed form.
Answer to 2c: The $\text{prox}_{tf}(y)$ below:

$$\begin{aligned}
\text{prox}_{tf}(y) &= \arg\min_x \frac{1}{2t}\|x-y\|^2 + f(x) \\
\text{prox}_{tf}(y) &= \arg\min_x \frac{1}{2t}\|x-y\|^2 + \frac{1}{2}\|x\|_2^2 \\
\nabla &= \frac{1}{t} * (x-y) + x = 0 \\
0 &= \frac{x}{t} - \frac{y}{t} + x \\
0 &= x - y + x * t \\
y &= x(1+t) \\
x &= \frac{y}{1+t} \\
\text{prox}_{tf}(y) &= \frac{y}{1+t}
\end{aligned}$$

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

Answer to 2d:

$$\begin{aligned}
\text{prox}_{tf}(y) &= \arg\min_x \frac{1}{2t}\|x-y\|^2 + f(x) \\
\text{prox}_{tf}(y) &= \arg\min_x \frac{1}{2t}\|x-y\|^2 + \frac{1}{2}\|Cx\|^2 \\
\nabla &= \frac{1}{t} * (x-y) + C^T Cx = 0 \\
0 &= \frac{x}{t} - \frac{y}{t} + C^T Cx \\
0 &= x - y + C^T Cxt \\
y &= x(1 + C^T Ct) \\
x &= \frac{y}{1 + C^T Ct} \\
\text{prox}_{tf}(y) &= \frac{y}{1 + C^T Ct}
\end{aligned}$$

Assuming C is a semi-positive definite matrix, then $C^T * C$ will always be positive semi-definite when taking the gradient.

Coding Assignment

Please download `515Hw2_Coding.ipynb` `solvers.py` and `mnist01.npy` to complete the coding problem (3), (4) and (5).

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

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

(4) Compressive sensing, consider the sparse regression problem,

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

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.
Answer to 4a: Complete

(b) Apply the proximal gradient algorithm. Do you recover the signal?
Answer to 4b: Yes, the signal is recovered.

(c) Apply accelerated proximal gradient, is it faster than method of (b)?
Answer to 4c: Yes it is faster than (b). Proximal Gradient Descent took 0.2260749340057373 seconds and Accelerated Proximal Gradient Descent took 0.1549389362335205 seconds.

(5) Logistic regression on MNIST data, recall the logistic regression problem,

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

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

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

(a) Complete the function, gradient and Hessian for the logistic regression.
Answer to 4a: Complete

(b) Apply gradient, accelerate gradient and Newton's method to solve the problem. Which one is the fastest and which one is the slowest?
Answer to 5b: Gradient Descent took 0.670673131942749 seconds, Accelerate Gradient took 0.7892789840698242 seconds, and Newton's method took 0.6741549968719482 seconds. The fastest method is Gradient Descent and the slowest method is Accelerate Gradient Descent.

(c) What is your accuracy of the classification for the test data.
Answer to 5c: The accuracy of the classification is 1.000.