Homework 2
Due: February 17, 2021

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.

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

(c) Compute prox_{tf} and f_t, where $f(x) = \|x\|_1$.

(d) Compute prox_{tf} and f_t for $f = \delta_{\mathbb{B}_\infty}(x)$, where $\mathbb{B}_\infty = [-1, 1]^n$.

Solution.

(a) Let $h(y) = \frac{1}{2t}\|x-y\|^2 + f(x)$. Since $x - y$ is affine and $\|\cdot\|^2$ is convex, $\frac{1}{2t}\|x-y\|^2$ is convex. Since f is convex, their sum $h(y)$ must also be convex. Then

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

Now consider

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

which holds for any x_1 in $h(y_1)$ and x_2 in $h(y_2)$. Consider $x_1 = \text{prox}_{tf}(y_1)$ and $x_2 = \text{prox}_{tf}(y_2)$, then $h(x_1, y_1) = f_t(y_1)$ and $h(x_2, y_2) = f_t(y_2)$. Thus we can rewrite the inequality as

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

This proves f_t is convex. □

(b) Since $\lim_{|y|\to\infty} h(y) = \infty$, $h(y)$ is coercive and at least one minimizer exists. We note that $\frac{1}{2t}\|x-y\|^2$ is strictly convex since the Hessian is $2\mathbf{I}$. Then $h(y)$ must be strictly convex since it is the sum of a strictly convex function and a convex function. Then the minimizer must be unique, i.e. $\text{prox}_{tf}(y)$ is uniquely defined. □

(c) By the definition of $\text{prox}_{tf}(y)$

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

Since the equations are decoupled, we only need to consider the i-th one,

$$\arg\min_{x_i} \frac{1}{2t}\|x_i - y_i\|^2 + |x_i|$$

The optimality condition is

$$0 \in \frac{x-y}{t} + \partial\|x\|_1$$

If $x_i < 0$, then $x_i = y_i + t$ and $y_i < -t$. If $x_i > 0$, then $x_i = y_i - t$ and $y_i > t$. If $x_i = 0$, then $|y_i| \leq t$. Thus

$$\big(\text{prox}_{tf}(y)\big)_i = \begin{cases} y_i + t & y_i < -t \\ y_i - t & y_i > t \\ 0 & |y_i| \leq t \end{cases}$$

$$(f_t(y))_i = \begin{cases} t/2 + |y_i + t| & y_i < -t \\ t/2 + |y_i - t| & y_i > t \\ y_i^2/(2t) & |y_i| \leq t \end{cases}$$

(d) By the definition of $\text{prox}_{tf}(y)$

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

Thus $f_t(y) = \frac{1}{2t}\|\max(\min(y,1),-1) - y\|^2$.

2. More prox identities.

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

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

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

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

Solution.

(a) By the definition of $\text{prox}_{tg}(y)$

$$\text{prox}_{tg}(y) = \arg\min_x \frac{1}{2t}\|x - y\|^2 + f(x) + \frac{1}{2s}\|x - x_0\|^2$$

The optimality condition is

$$0 \in \frac{1}{t}(x - y) + \partial f(x) + \frac{1}{s}(x - x_0)$$

Rearranging

$$\begin{aligned}\frac{1}{t}(y - x) + \frac{1}{s}(x_0 - x) &\in \partial f(x)\\ \frac{s+t}{st}(\frac{sy + tx_0}{s+t} - x) &\in \partial f(x)\end{aligned}$$

Let $\lambda = \frac{st}{s+t}$ and $z = \frac{sy+tx_0}{s+t}$, then

$$\text{prox}_{tg}(y) = \text{prox}_{\lambda f}(z)$$

Let $\text{prox}_{tg}(y) = x^*$. By the definition of $g_t(y)$

$$\begin{aligned}g_t(y) &= \min_x \frac{1}{2t}\|x - y\|^2 + f(x) + \frac{1}{2s}\|x - x_0\|^2\\ &= \frac{1}{2t}\|x^* - y\|^2 + f(x^*) + \frac{1}{2s}\|x^* - x_0\|^2\end{aligned}$$

Similarly for $f_\lambda(z)$

$$f_\lambda(z) = \frac{1}{2\lambda}\|x^* - z\|^2 + f(x^*)$$

Rewriting in terms of t and y

$$f_\lambda(z) = \frac{\|s(x^* - y) + t(x^* - x_0)\|^2}{2st(s+t)} + f(x^*)$$

Computing the difference between $g_t(y)$ and $f_\lambda(z)$, we find

$$g_t(y) - f_\lambda(z) = \frac{1}{2(s+t)}\|y - x_0\|^2$$
$$g_t(y) = f_\lambda(z) + \frac{1}{2(s+t)}\|y - x_0\|^2$$

(b) The optimality condition is

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

Solving yields

$$x(1 + \frac{t}{\|x\|}) = y$$

Since $t > 0$, $x = cy$ for some $c > 0$. Then

$$y = cy(1 + \frac{t}{c\|y\|})$$
$$c = 1 - \frac{t}{\|y\|}$$

Thus $\text{prox}_{tf}(y) = (1 - \frac{t}{\|y\|})y$.

(c) The optimality condition is

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

Solving yields

$$\text{prox}_{tf}(y) = x = \frac{y}{t+1}$$

(d) The optimality condition is

$$\nabla(\frac{1}{2t}\|x-y\|^2 + \frac{1}{2}\|Cx\|^2) = \frac{1}{t}(x-y) + C^TCx = 0$$

Solving yields

$$\text{prox}_{tf}(y) = x = (tC^TC + I)^{-1}y$$

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,

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

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)?

(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.