AMATH 515 Problem Set 2

Tyler Chen

Problem 1

(a) Show that a $\mathcal{C}^1$-smooth function f is α-strongly convex if and only if the function $g(x) = f(x) - \frac{\alpha}{2}\|x\|^2$ is convex.

(b) Suppose that f is α-strongly convex. Show that any minimizer of f is unique.

(c) Suppose that f is α-strongly convex with minimizer x^*. Show that

$$f(x) \geq f(x^*) + \frac{\alpha}{2}\|x - x^*\|^2.$$

Solution

(a) Observe that,

$$\nabla g(x) = \nabla f(x) - \alpha x$$

Then we have the following equivalence,

$$\begin{aligned}
& g(x) \text{ is convex} \\
\iff \quad & \langle \nabla g(y) - \nabla g(x), y - x\rangle \geq 0 \\
\iff \quad & \langle \nabla f(y) - \nabla f(x) - (\alpha y - \alpha x), y - x\rangle \geq 0 \\
\iff \quad & \langle \nabla f(y) - \nabla f(x), y - x\rangle - \alpha \langle y - x, y - x\rangle \geq 0 \\
\iff \quad & \langle \nabla f(y) - \nabla f(x), y - x\rangle - \alpha \|y - x\|^2 \geq 0 \\
\iff \quad & \langle \nabla f(y) - \nabla f(x), y - x\rangle \geq \alpha \|y - x\|^2 \\
\iff \quad & f(x) \text{ is } \alpha\text{-strictly convex}
\end{aligned}$$

□

(b) Suppose x^* is a minimizer of f. Then $\nabla f(x^*) = 0$ so,

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^*\rangle + \frac{\alpha}{2}\|x - x^*\|^2 = f(x^*) + \frac{\alpha}{2}\|x - x^*\|^2$$

Therefore, for any $\alpha > 0$, $f(x) = f(x^*)$ if and only if $x = x^*$. That is, the minimizer x^* is unique. □

(c) This was shown as an intermediate step in the previous result. □

Problem 2

Recall that

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

Suppose f is convex.

(a) Prove that f_t is convex.

(b) Prove that prox_{tf} is a single-valued mapping.

(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) Assume f is convex. Recall that for a fixed x, $\|x-y\|^2$ is a convex function of y. Therefore, for all x,

$$\|x - (\lambda y + (1-\lambda)z)\|^2 \leq \lambda \|x-y\|^2 + (1-\lambda)\|x-z\|^2$$

Observe that for any x_1 and x_2, by the convexity of $\|\cdot\|^2$,

$$\begin{aligned}\|\lambda x_1 + (1-\lambda)x_2 - (\lambda y + (1-\lambda)z)\|^2 &= \|\lambda(x_1-y) + (1-\lambda)(x_2-z)\|^2 \\ &\leq \lambda\|x_1-y\|^2 + (1-\lambda)\|x_2-z\|^2\end{aligned}$$

Similarly,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

Then,

$$\begin{aligned}f_t(\lambda y + (1-\lambda)z) &= \min_x \left[\frac{1}{2t}\|x - (\lambda y + (1-\lambda)z)\|^2 + f(x)\right] \\ &\leq \lambda \min_{x_1}\left[\frac{1}{2t}\|x_1-y\|^2 + f(x)\right] + (1-\lambda)\min_{x_2}\left[\frac{1}{2t}\|x_2-z\|^2 + f(x)\right] \\ &= \lambda f_t(y) + (1-\lambda)f_t(z)\end{aligned}$$

where the inequality holds by observing that the convex combination of whichever x_1 and x_2 are used in the middle line provided an upper bound for the top line.

(b) Fix y and suppose $x_1 \neq x_2$ satisfy,

$$\frac{1}{2t}\|x_1-y\|^2 + f(x_1) = \frac{1}{2t}\|x_2-y\|^2 + f(x_2)$$

Recall that $\|x-y\|^2$ as a function of x is strictly convex. Therefore, the point $x = (x_1+x_2)/2$ satisfies,

$$\begin{aligned}\frac{1}{2t}\|x-y\|^2 + f(x) &< \frac{1}{2}\left(\frac{1}{2t}\|x_1-y\|^2 + f(x_1)\right) + \frac{1}{2}\left(\frac{1}{2t}\|x_2-y\|^2 + f(x_2)\right)\\ &= \frac{1}{2t}\|x_1-y\|^2 + f(x_1)\\ &= \frac{1}{2t}\|x_2-y\|^2 + f(x_2)\end{aligned}$$

Therefore neither of x_1 and x_1 can be an output to prox_{tf}. □

(c) By definition,

$$\begin{aligned}\text{prox}_{tf}(y) &= \arg\min_x \left[\frac{1}{2t}\|x-y\|^2 + \lambda\|x\|_1\right]\\ &= \arg\min_x \left[\sum_{i=1}^n \frac{1}{2t}(x_i-y_i)^2 + \lambda|x_i|\right]\end{aligned}$$

Now observe that the components are decoupled so that we only need to compute,

$$\arg\min_{x_i}\left[\frac{1}{2t}(x_i-y_i)^2 + \lambda|x_i|\right]$$

As a function of x_i, the argument above, which we well denote $h_i(x)$, is a quadratic centered at y_i plus an absolute value centered at 0.

We consider three cases, $y_i \in (-\infty, -\lambda t)$, $y_i \in [-\lambda t, \lambda t]$, and $y_i \in (\lambda t, \infty)$. For each of theses cases we consider $x_i < 0$, $x_i > 0$ and $x_i = 0$.

First, suppose $y_i \in (-\infty, -\lambda t)$. If $x_i \leq 0$ then setting the derivative to zero we find $x_i = y_i + \lambda t$ minimizes h_i and that $h_i(y_i+\lambda t) = \lambda^2 t/2 + |y_i + \lambda t|$. If $x_i > 0$ then the derivative is nonzero so h_i is not minimized.

Similarly, suppose $y_i \in (\lambda t, \infty)$. If $x_i \leq 0$ then setting the derivative to zero we find $x_i = y_i - \lambda t$ minimizes h_i and $h_i(y_i - \lambda t) = \lambda^2 t/2 + |y_i - \lambda t|$. If $x_i < 0$ then the derivative is nonzero so h_i is not minimized.

Finally, suppose $y_i \in [-\lambda t, \lambda t]$. If $x_i = 0$ then the subgradient is $y_i/t + [-\lambda, \lambda]$, so zero is in the subgradient exactly when $|y_i| < \lambda t$. In this case, $h_i(0) = y_i^2/2t$. If $x_i \neq 0$ then zero is not contained in the subgradient so h_i is not minimized.

Therefore,

$$\left[\text{prox}_{tf}(y)\right]_i = \arg\min_{x_i}\left[\frac{1}{2t}(x_i-y_i)^2 + \lambda|x_i|\right] = \begin{cases} y_i + \lambda t, & y_i \in (-\infty, -\lambda t)\\ 0, & y_i \in [-\lambda t, \lambda t]\\ y_i - \lambda t & y_i \in (\lambda t, \infty)\end{cases}$$

Similarly,

$$f_t = \sum_i [f_t]_t$$

where

$$[f_t]_i = \min_{x_i} \left[\frac{1}{2t}(x_i - y_i)^2 + \lambda |x_i| \right] = \begin{cases} \frac{\lambda^2 t}{2} + |y_i + \lambda t|, & y_i \in (-\infty, -\lambda t) \\ \frac{y_i^2}{2t}, & y_i \in [-\lambda t, \lambda t] \\ \frac{\lambda^2 t}{2} + |y_i - \lambda t| & y_i \in (\lambda t, \infty) \end{cases}$$

(d) Note that for a fixed y, $\|x - y\|$ is the distance from x to y, which is minimized at y. Then, the projection of y to some set minimizes this quantity over that set. Therefore,

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

where min and max are taken pointwise.

Then,

$$f_t = \frac{1}{2t} \left[\text{dist}_{\mathbb{B}_\infty}(y) \right]^2 = \frac{1}{2t} \sum_i [f_t]_i$$

where,

$$[f_t]_i = \begin{cases} (y_i + 1)^2 & y_i \in (-\infty, -1) \\ 0 & y_i \in [-1, 1] \\ (y_i - 1)^2 & y_i \in (1, \infty) \end{cases}$$

Problem 3

More prox identities.

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

(b) The elastic net penalty is used to detect groups of correlated predictors:

$$g(x) = \beta\|x\|_1 + (1-\beta)\frac{1}{2}\|x\|^2, \quad \beta \in (0,1).$$

Write down the formula for prox_{tg} and g_t.

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

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

Solution

(a) By definition,

$$\begin{aligned}\text{prox}_{tg}(y) &= \arg\min_x \left[\frac{1}{2t}\|x-y\|^2 + g(x)\right] \\ &= \arg\min_x \left[\frac{1}{2t}\|x-y\|^2 + f(x) + \frac{1}{2}\|x-x_0\|^2\right]\end{aligned}$$

Observe that by completing the square,

$$\|x-y\|^2 + t\|x-x_0\|^2 = (1+t)\left\|x - \frac{y+tx_0}{1+t}\right\|^2 + \frac{t}{1+t}\|y-x_0\|^2$$

Thus, define,

$$s = \frac{t}{1+t}, \qquad z = \frac{y+tx_0}{1+t}$$

Then, observing that the last terms when completing the square to not affect prox_{tg} as they depend only on y and x_0 but not x, we find that,

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

Similarly, by definition,

$$\begin{aligned}g_t(y) &= \min_x \left[\frac{1}{2t}\|x-y\|^2 + g(x)\right] \\ &= \min_x \left[\frac{1}{2t}\|x-y\|^2 + f(x) + \frac{1}{2}\|x-x_0\|^2\right]\end{aligned}$$

So, adding the difference term from $f_s(z)$ and $g_t(y)$ we find,

$$g_t(y) = f_s(z) + \frac{\|y-x\|^2}{2(1+t)}$$

(b) Observe that by completing the square,

$$\frac{1}{2t}\|x-y\|^2+\frac{1-\beta}{2}\|x\|^2=\frac{1+t(1-\beta)}{2t}\left\|x-\frac{y}{1+t(1-\beta)}\right\|^2 + \frac{1}{2t}\left(1-\frac{1}{1+t(1-\beta)}\right)\|y\|^2$$

Thus, define,

$$s=\frac{t\beta}{1+t(1-\beta)}, \qquad z=\frac{y}{1+t(1-\beta)}$$

Then, with $f=\|x\|_1$ as in 2(c),

$$\text{prox}_{tg}(y)=\text{prox}_{sf}(z), \qquad g_t(y)=\beta f_s(z)+\frac{1}{2t}\left(1-\frac{1}{1+t(1-\beta)}\right)\|y\|^2$$

(c) Observe that,

$$\nabla\left[\frac{1}{2t}\|x-y\|^2+\frac{1}{2}\|Cx\|^2\right]=\frac{1}{t}(x-y)+C^TCx=\left(C^TC+\frac{1}{t}I\right)x-\frac{y}{t}$$

Therefore, setting the gradient to zero,

$$\text{prox}_{tf}(y)=\arg\min_x\left[\frac{1}{2t}\|x-y\|^2+\frac{1}{2}\|Cx\|^2\right]=(t\,C^TC+I)^{-1}y$$

(d) Observe that,

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

By setting the gradient to zero, since $t>0$, we observe that x is a positive scalar multiple of y. That is, $x=cy$ for some $c>0$. Then,

$$\frac{y}{t}=\left(\frac{1}{t}+\frac{1}{\|x\|}\right)x=\left(\frac{1}{t}+\frac{1}{c\,\|y\|}\right)cy$$

Therefore,

$$c=1-\frac{t}{\|y\|}$$

Finally,

$$\text{prox}_{tf}(y)=\arg\min_x\left[\frac{1}{2t}\|x-y\|^2+\|x\|\right]=\left(1-\frac{t}{\|y\|}\right)y$$

Problem 4

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

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

Solution

Problem 5

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 $m < n$ and this is an under determine system. Fortunately, we have the prior knowledge of x being sparse, by adding the ℓ_1 regularizer, we could recover the original signal.

Remark: we choose $\lambda = \|A^\top b\|_\infty/10$, the reason of it will be more clear when come to duality.

(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, can you recover the signal?

(c) Apply the accelerated proximal gradient algorithm, is it faster compare to (b)?

Solution

(b) Yes, we are able to recover the signal as shown in Figure 1.

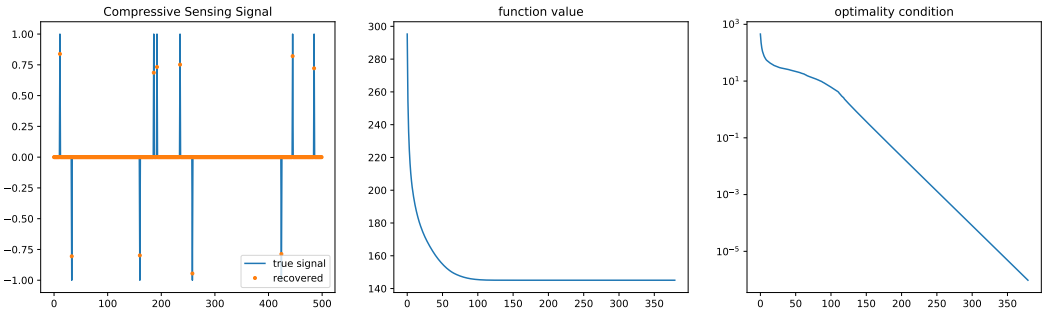

Figure 1: Proximal Gradient Descent

(c) Accelerated proximal gradient decent requires fewer iterations to terminate with the default condition.

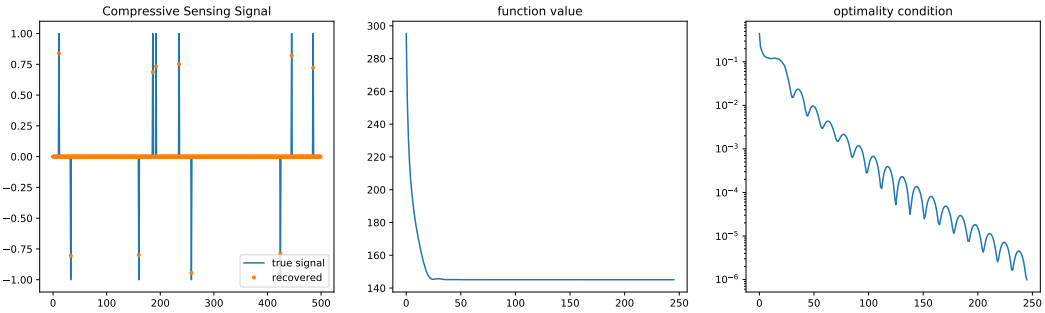

Figure 2: Accelerated Proximal Gradient Descent

Problem 6

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 try to use logistic regression to classify the "0" and "1" images from MNIST.

In this specific example, a_i is our image (vectorized), and b_i is the corresponding label. By solving the above optimization problem, we want to obtain an classifier, so that for a new image a_{new}, we could say

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

(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.

Solution

(b) Figure 3 shows the convergence of Gradient Descent, Accelerated Gradient Descent, and Newton's method on a logistic regression problem. We observe Newton's Method requires the fewest iterations.

Both Gradient Decent methods reached the maximum number of iterations without terminating. However, the standard Gradient Descent was worse both in terms of the objective function value and the optimality condition.

However, iterations may not correspond to "speed" since Newton's method requires a linear solve at each iteration.

(c) We have 100% accuracy.

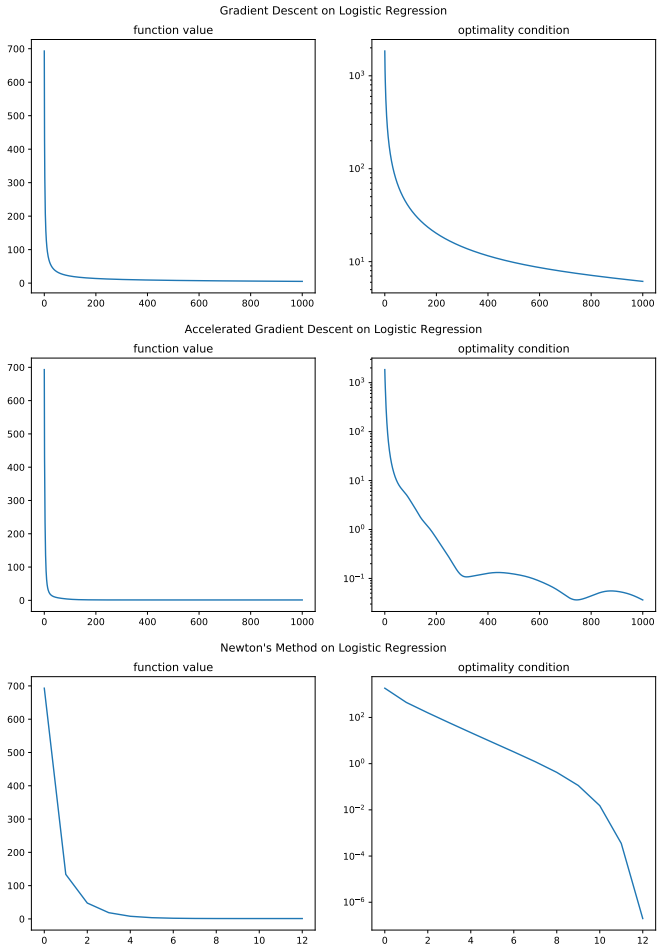

Figure 3: Convergence of Logistic Regression