Exercise 6

Bowen Hua

October 27, 2017

1 Proximal Operators

1.1 (A)

We use these definitions of Moreau envelope and proximal operator:

$$E_\gamma f(x) = \min_z \left\{ f(z) + \frac{1}{2\gamma} \|z - x\|_2^2 \right\} \leq f(x)$$

$$\operatorname{prox} f(x) = \arg\min_z \left\{ f(z) + \frac{1}{2\gamma} \|z - x\|_2^2 \right\}$$

The proximal operator of a linear approximation at x_0 is derived as follows:

$$\begin{aligned} \operatorname{prox} \hat{f}(x; x_0) &= \arg\min_x \left\{ f(x_0) + (x - x_0)^T \nabla f(x_0) + \frac{1}{2\gamma} \|x - x_0\|_2^2 \right\} && (1) \\ &= \arg\min_x \left\{ x^T \nabla f(x_0) + \frac{1}{2\gamma} \|x - x_0\|_2^2 \right\} && (2) \end{aligned}$$

We then use the first order optimality condition of the minimization problem:

$$\nabla f(x_0) - \frac{1}{\gamma}(x - x_0) = 0$$

We get $x = x_0 + \gamma \nabla f(x_0)$, which is the gradient step for $f(x)$ with step size γ.

1.2 (B)

The proximal operator of $l(x)$ is

$$\operatorname*{prox}_{1/\gamma} l(x) = \arg\min_z \left\{ \frac{1}{2} z^T P z - q^T z + r + \frac{\gamma}{2} \|z - x\|_2^2 \right\}$$

If P is positive semidefinite, the minimization problem is convex, and we can use the first order optimality condition:

$$Pz - q + \gamma(z - x) = 0$$

which implies

$$z = (P + \gamma I)^{-1}(\gamma x + q)$$

where I is the identity matrix.

Suppose $(y|x) \sim N(Ax, \Omega^{-1})$. The negative log likelihood function of x can be written as

$$\begin{aligned} l(x) &\propto (y - Ax)^T \Omega (y - Ax) && (3) \\ &= (y^T \Omega y - y^T \Omega A x - x^T A^T \Omega y + x^T A^T \Omega A x) && (4) \\ &= (y^T \Omega y - 2y^T \Omega A x + x^T A^T \Omega A x) && (5) \end{aligned}$$

which can be written in the quadratic form.

1.3 (C)

Now we have $\phi(x) = \tau \|x\|_1$.

$$\operatorname*{prox}_{\gamma} \phi(x) = \arg\min_z \left\{ \tau \|z\|_1 + \frac{1}{2\gamma} \|z - x\|_2^2 \right\}$$

Since the problem is separable across each entry of z, we focus on the element-wise solution:

$$\arg\min_{z_i} \left\{ \frac{1}{2}(x_i - z_i)^2 + \gamma\tau|z_i| \right\}$$

From exercise 5, we showed that this is equal to the soft-thresholding function

$$\operatorname{sign}(x_i)(|x_i| - \gamma\tau)_+.$$

2 The proximal gradient method

2.1 (A)

The definition of $\hat{x}$ is: We look to minimize this approximation of the objective function.

$$\hat{x} = \arg\min_x \left\{ l(x_0) + (x - x_0)^T \nabla l(x_0) + \frac{1}{2\gamma} \|x - x_0\|_2^2 + \phi(x) \right\}$$

What we want to show is

$$\hat{x} = \arg\min_x \left\{ \phi(x) + \frac{1}{2\gamma} \|x - (x_0 - \gamma \nabla l(x_0))\|_2^2 \right\}$$

If we expand the squared norms in each of these two equations, we can see that the functions being minimized are only different by terms that are not a function of x. Therefore, the two minimizers are the same.

2.2 (B)

In the context of lasso regression, we have

$$l(\beta) = \frac{1}{N} \|y - X\beta\|_2^2$$

where N is the number of samples.

$$\nabla l(\beta) = \frac{2}{N}(X^T X\beta - X^T y)$$
$$\phi(\beta) = \lambda \|\beta\|_1 .$$

Now we put these variables into the equations we have for proximal gradient method, we get

$$\begin{aligned} u^{(t)} &= \beta^{(t)} - \gamma^{(t)} \nabla l(\beta^{(t)}) && (6) \\ &= \beta^{(t)} - \frac{2\gamma^{(t)}}{N}(X^T X\beta^{(t)} - X^T y) && (7) \end{aligned}$$

Also, we have

$$\beta^{(t+1)} = \operatorname*{prox}_{\gamma^{(t)}} \lambda \left\| u^{(t)} \right\|_1 \tag{8}$$

Component-wise, this is:

$$\beta_i^{(t+1)} = \operatorname{sign}(u_i^{(t)})(|u_i^{(t)}| - \gamma^{(t)}\lambda)_+$$

The most expensive computation in each iteration is the matrix-vector multiplication when computing the gradient that is needed for computing $u^{(t)}$.