CS690OP Homework 2

Quoc Anh (Alan) Bui

February 2022

1 Proximal Operators [25 points + 5 points Extra Credit]

Compute the proximal operators for the following functions:

1. [6 points] $h(x) = \lambda \sum_i (x_i)_+ \qquad$ where $x_+ \overset{def}{=} \max(x, 0)$

   Answer:

   We have the proximal operator defined as:

$$\begin{aligned} \operatorname{prox}_{h,t}(x) &= \underset{z}{\operatorname{argmin}} \ \frac{1}{2t} \|z - x\|_2^2 \ + \ \lambda \sum_i (z_i)_+ \\ &= \underset{z}{\operatorname{argmin}} \ \sum_i \frac{1}{2t} (z_i - x_i)^2 \ + \ \lambda (z_i)_+ \end{aligned} \tag{1}$$

   Since this is componentwise, we can solve for each $\operatorname{prox}_{x_i,}$ instead:

$$\operatorname{prox}_{h,t,(}x_i) = \underset{z}{\operatorname{argmin}} \ \frac{1}{2t}(z_i - x_i)^2 \ + \ \lambda (z_i)_+$$

   z^* is obtained by solving $\dfrac{\partial}{\partial z_i} = 0$:

$$\begin{aligned} \frac{\partial}{\partial z_i} &= \frac{1}{t}(z_i - x_i) + \lambda \partial (z_i)_+ = 0 \\ &\therefore z_i - x_i + \lambda t \partial (z_i)_+ = 0 \end{aligned} \tag{2}$$

   We have the subgradient of (z_i) as following:

$$\partial (z_i)_+ = \begin{cases} 0 & \text{if } \mathrm{z}_i < 0 \\ 1 & \text{if } \mathrm{z}_i > 0 \\ [0,1] & \text{if } \mathrm{z}_i = 0 \end{cases} \tag{3}$$

   Plug this into the gradient, we observe that:

$$z_i^* = \begin{cases} z_i^* = x_i & \text{if } \mathrm{z}_i < 0 \iff x_i < 0 \\ z_i^* = x_i - \lambda t & \text{if } \mathrm{z}_i > 0 \iff x_i > \lambda t \\ z_i^* = 0 & \text{if } -x_i = \lambda t \partial (z_i) \iff x_i \in [0, \lambda t] \text{ since } \partial (z_i) \in [0,1] \end{cases} \tag{4}$$

Thus:

$$\boxed{z_i^* = \begin{cases} x_i & \text{for } \mathrm{x}_i < 0 \\ x_i - \lambda t & \text{for } \mathrm{x}_i > \lambda t \\ 0 & \text{for } x_i \in [0, \lambda t] \end{cases}} \tag{5}$$

2. [6 points] $h(x) = \lambda||x||_2^2$
   Answer:
   We have the proximal operator defined as following:

$$\text{prox}_{h,t}(x) = \underset{z}{\text{argmin}}\ \frac{1}{2t}\|z - x\|_2^2\ +\ \lambda\|z\|_2^2$$

   $h(x)$ is differentiable so z^* is achieved by solving $\frac{\partial}{\partial z} = 0$:

$$\begin{aligned} \frac{\partial}{\partial z} = \frac{1}{2t}(2z - 2x) + 2\lambda z = 0 \\ \frac{1}{t}(z - x) + 2\lambda z = 0 \\ z(1 + 2\lambda t) = x \\ \therefore\ \boxed{z^* = \frac{x}{1 + 2\lambda t}} \end{aligned} \tag{6}$$

3. [6 points] $h(x) = \lambda||x||_\infty$
   Answer:
   Using Moreau decomposition, we have:

$$\text{prox}_f(x) = x - \text{prox}_{f^*}(x)$$

   where f^* is the convex (Fenchel) conjugate of $f(x)$. Claim that for $f(x) = \|x\|$ then $f^(x) = \begin{cases} 0 & \text{if } \|x\|_* \leq 1 \\ \infty & \text{if } \|x\|_* > 1 \end{cases}$.
   where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

   Proof: By the definition we have: $f(x) = \|x\|$ and $f^*(x) = \sup_y \left(y^T x - \|y\|\right)$. And by the definition of dual norm we have:

$$\|x\|_* = \sup_{\|y\| \leq 1} y^T x$$

$$\therefore \forall x \text{ and } \forall y : x^T y = \|y\| \left(x^T \frac{y}{\|y\|}\right) \leq \|y\|\,\|x\|_*$$

   If $\|x\|_* \leq 1$, then:

$$x^T y \leq \|y\|\,\|x\|_* = \|y\|$$

   The equality holds when $y = 0$, thus:

$$\boxed{f^*(x) = \sup_y\ (y^T x - \|y\|) = 0 \quad \text{if } \|x\|_* \leq 1}$$

   If $\|x\|_* > 1$, then:

$$\|x\|_* = \sup_{\|y\| \leq 1} y^T x > 1$$

$$\therefore \exists y : \|y\| \leq 1 \quad \text{and} \quad y^T x > 1$$

$$\therefore f^*(x) = \sup_y \; y^T x - \|y\| > 0$$

Let $y = tz$ where $t \in \mathbb{R}$:

$$f^*(x) = y^T x - \|y\| = t(z^T x - \|z\|)$$

And as $t \to \infty \;:\; f^*(x) \to \infty$

$$\boxed{f^*(x) = \infty \quad \text{if} \;\; \|x\|_* > 1}$$

Plug this back to Moreau decomposition, we have:

$$\text{prox}_{f^*}(x) = \underset{z}{\text{argmin}} \, \frac{1}{2} \|z - x\|_2^2 \; + \; \mathbb{I}(\|z\|_* \leq 1) \tag{7}$$

And we know that $L1$ norm is the dual norm of L_∞ norm, we have:

$$\text{prox}_{f^*}(x) = \underset{z}{\text{argmin}} \; \frac{1}{2} \|z - x\|_2^2 \; + \; \mathbb{I}(\|z\|_1 \leq 1)$$

This is basically a projection onto $L1$ unit norm ball. Thus:

$$\text{prox}_{\|\cdot\|_\infty}(x) = x - \text{proj}_{\|\cdot\|_1 \leq 1}(x)$$

$$\boxed{\text{prox}_{\|\cdot\|_\infty, t} = x - \text{proj}_{\|\cdot\|_1 \leq t\lambda}(x)}$$

Note: Professor confirms that for this question we don't need to explicitly shows what projection onto $L1$ unit norm ball is, although it is trivial.

4. [7 points] $h(x) = \lambda ||x||_0$

Answer:

We have the proximal operator defined as following:

$$\begin{aligned} \text{prox}_{h,t}(x) &= \underset{z}{\text{argmin}} \;\; frac12t \|z - x\|_2^2 \; + \; \lambda \|z\|_0 \\ &= \underset{z}{\text{argmin}} \;\; \frac{1}{2t} \sum_i (z_i - x_i)^2 \; + \; \lambda \mathbb{I}(z_i \neq 0) \end{aligned} \tag{8}$$

Due to its componentwise, we can solve for $\text{prox}_{h,t}(x_i)$:

$$\begin{aligned} \text{prox}_{h,t}(x_i) &= \underset{z_i}{\text{argmin}} \;\; \frac{1}{2t}(z_i - x_i)^2 \; + \; \lambda \mathbb{I}(z_i \neq 0) \\ &= \underset{z_i}{\text{argmin}} \begin{cases} \frac{1}{2t} x_i^2 & \text{if} \;\; \text{z}_i = 0 \\ \frac{1}{2t}(z_i - x_i)^2 + \lambda & \text{if} \;\; \text{z}_i \neq 0 \end{cases} \end{aligned} \tag{9}$$

If $z_i = 0$ then $\text{argmin}_{z_i}(f) = x_i$ and $\min(f) = \lambda$. If $z_i = 0$ then $\min(f) = \dfrac{1}{2t} x_i^2$. Thus:

$$\boxed{z_i^* = \text{prox}_{h,t}(x_i) = \begin{cases} 0 & \text{if} \;\; |x_i| \leq \sqrt{2\lambda t} \\ |x_i| & \text{if} \;\; |x_i| > \sqrt{2\lambda t} \end{cases}}$$

This is Hard Thresholding operator.

5. [Extra Credit: 5 points] $h(x) = \lambda||x||_2$

Answer:

Similar to (1.3), we know the dual norm of $L2$ is $L2$. Then by Moreau Decomposition we have:

$$\text{prox}_{\lambda t}(x) = x - \text{proj}_{\|\cdot\|_2 \le \lambda t}(x)$$

We know the projection on $L2$ unit norm ball is given by:

$$\text{proj}_{\|\cdot\|_2 \le 1}(x) = \begin{cases} \dfrac{x}{\|x\|_2} & \text{if } \|x\|_2 > 1 \\ x & \text{if } \|x\|_2 \le 1 \end{cases}$$

And:

$$\text{proj}_{\|\cdot\|_2 \le \lambda t}(x) = \begin{cases} \left(\dfrac{x}{\|x\|_2}\right)t\lambda & \text{if } \|x\|_2 > t\lambda \\ x & \text{if } \|x\|_2 \le t\lambda \end{cases}$$

Thus:

$$\begin{aligned} \text{prox}_{\|\cdot\|_2, t}(x) = x - \text{proj}_{\|\cdot\|_2 \le \lambda t}(x) &= \begin{cases} x - \dfrac{\lambda t}{\|x\|_2}x & \text{if } \|x\|_2 > \lambda t \\ 0 & \text{if } \|x\|_2 \le \lambda t \end{cases} \\ &\boxed{= \max\left(0, x - \dfrac{\lambda t}{\|x\|_2}x\right)} \end{aligned} \tag{10}$$

2 Proximal Operator Properties [25 points + 5 points Extra Credit]

1. [10 points] Prove that, for any convex function f,

$$\text{prox}_{,\lambda f}(x) = (I + \lambda \partial f)^{-1}(x). \tag{11}$$

Answer: We have the proximal operator defined as following:

$$\text{prox}_{\lambda f,} = \underset{z}{\text{argmin}} \ \frac{1}{2}\|z-x\|_2^2 \ + \ (z)$$

z^* is obtained by solving $\dfrac{\partial}{\partial z} = 0$:

$$\begin{aligned} \frac{\partial}{\partial z}(z-x) + \partial\lambda f(z) &= 0 \\ z + \lambda \partial f(z) &= x \\ (I + \lambda\partial f)(z) &= x \\ \boxed{z^* = (I+\lambda\partial f)^{-1}x} \end{aligned} \tag{12}$$

2. [15 points] Suppose that $f : E_1 \times E_2 \times \dots E_m \to (-\infty, \infty]$ is defined as

$$f(x_1, x_2, \dots x_m) = \sum_{i=1}^{m} f_i(x_i)$$

for any $x_i \in E_i$, $\forall i = 1 \dots m$.

Prove that, for any $x_1 \in E_1, x_2 \in E_2 \dots x_m \in E_m$

$$\text{prox}_{,f}(x_1, x_2 \dots x_m) = \text{prox}_{,f_1}(x_1) \times \text{prox}_{,f_2}(x_2) \times \dots \text{prox}_{,f_m}(x_m) \tag{13}$$

where $\times$ represents the cartesian product between sets.

Answer: We have the proximal operator defined as following:

$$\text{prox}_{f,}(x_1, \cdots, x_m) \ = \ \underset{z_1,\cdots,z_m}{\text{argmin}} \ \sum_{i=1}^{m} \frac{1}{2t}\|z_i - x_i\|_2^2 \ + \ f_i(z_i)$$

Due to its componentwise, we can solve for $\text{prox}_{f_i,}(x_i)$:

$$\begin{aligned} \text{prox}_{f_i,}(x_i) &= \underset{z_i}{\text{argmin}} \ \frac{1}{2t}\|z_i - x_i\| \ + \ f_i(x_i) \\ &= z_i^* \end{aligned} \tag{14}$$

Thus:

$$\begin{aligned} \text{prox}_{f,}(x_1, \cdots, x_m) \ &= \ (z_1^*, \cdots, z_m^*) \\ &\boxed{= \text{prox}_{f_1,}(x_1) \times \cdots \times \text{prox}_{f_m,}(x_m)} \end{aligned} \tag{15}$$

3. [Extra Credit: 5 points] Find the proximal operator of $g : R^n \to (-\infty, \infty]$ where

$$g(x) = \begin{cases} -\lambda \sum_{i=1}^{n} \log x_i & \text{if } x > 0 \\ \infty & \text{otherwise} \end{cases} \tag{16}$$

Answer: When $x \leq 0$ then $g(x) = \infty$ then we cannot minimize the proximal operator function. Thus, we only care when $x > 0$:

$$g(x) = -\lambda \sum_{i}^{n} \log(x_i)$$

Then the proximal operator is defined as following:

$$\begin{aligned} \text{prox}_{g,t}(x) &= \underset{z}{\text{argmin}} \ \frac{1}{2t} \|z - x\|_2^2 \ + \ \lambda \sum_{i}^{n} - \log(x_i) \\ &= \underset{z}{\text{argmin}} \ \frac{1}{2t} \sum_{i}^{n} (z_i - x_i)^2 \ + \ + \lambda - \log(z_i) \end{aligned} \tag{17}$$

This is componentwise, we can solve for single element:

$$\text{prox}_{g_i,t}(x_i) = \underset{z_i}{\text{argmin}} \ \frac{1}{2t} (z_i - x_i)^2 \ + \ \lambda - \log(z_i)$$

$g_i(x)$ is differential, then z_i^* can be achieved by solving for $\frac{\partial}{\partial z_i} = 0$:

$$\begin{aligned} \frac{\partial}{\partial z_i} = \frac{1}{t}(z_i - x_i) - \frac{\lambda}{z_i} &= 0 \\ z_i - x_i - \frac{\lambda t}{z_i} &= 0 \\ z_i^2 - x_i z_i - \lambda t &= 0 \\ \boxed{z_i^* = \frac{x_i \pm \sqrt{x_i^2 + 4\lambda t}}{2}} \end{aligned} \tag{18}$$

3 Group Lasso Logistic Regression [50 points + 20 points EC]

Problem credit: [Tibshirani '19].

Suppose we have features $X \in \mathbb{R}^{n\times(p+1)}$ that we divide into J groups:

$$X = \left[\mathbf{1}\ X_{(1)}\ X_{(2)}\ \cdots\ X_{(J)}\right],$$

where $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^n$ and each $X_{(j)} \in \mathbb{R}^{n\times p_j}$.

To achieve sparsity over groups of features, rather than individual features, we can use a *group lasso* penalty. We write $\beta = (\beta_0, \beta_{(1)}, \ldots, \beta_{(J)}) \in \mathbb{R}^{p+1}$, where β_0 is an intercept term and each $\beta_{(j)} \in \mathbb{R}^{p_j}$.

Consider the problem

$$\min_{\beta}\ g(\beta) + \lambda \sum_{j=1}^{J} w_j \|\beta_{(j)}\|_2, \tag{19}$$

where g is a loss function and $\lambda \geq 0$ is a hyperparameter.

The penalty $h(\beta) = \lambda \sum_{j=1}^{J} w_j \|\beta_{(j)}\|_2$ is called the group lasso penalty.

A common choice for w_j is $\sqrt{p_j}$ to adjust for the group size.

1. [10 points] Derive the proximal operator $\text{prox}_{,h,t}(\beta)$ for the group lasso penalty defined above.

   Answer:

   We have the proximal operator defined as following:

$$\begin{aligned} \text{prox}_{h,t}(\beta) &= \underset{z}{\text{argmin}}\ \frac{1}{2t}\|z-\beta\|_2^2\ +\ \lambda\sum_{j=1}^{J} w_j\,\|\beta_j\|_2 \\ &= \underset{z}{\text{argmin}}\ \frac{1}{2t}\sum_{j=1}^{J}\|z_j-\beta_j\|_2^2\ +\ \lambda w_j\,\|\beta_j\|_2 \end{aligned} \tag{20}$$

   Due to its componentwise, for each group j:

$$\begin{aligned} \text{prox}_{h_j,t}(\beta_j) &= \underset{z_i}{\text{argmin}}\ \frac{1}{2t}\|z_i-x_i\|_2^2\ +\ \lambda w_j\,\|\beta_j\|_2 \\ &\boxed{= \max\left(0, 1-\frac{\lambda t w_j}{\|\beta_j\|_2}\right)\beta_j} \end{aligned} \tag{21}$$

   Derived earlier in (1.5) (noted that by defintion $\beta_j \in \mathbb{R}^{p_j}$ and thus we can use the derivation in 1.5)

2. [10 points] Let $y \in \{0,1\}^n$ be a binary label, and let g be the logistic loss

$$g(\beta) = -\sum_{i=1}^{n} y_i (X\beta)_i + \sum_{i=1}^{n} \log(1 + \exp\{(X\beta)_i\}),$$

Write out the steps for proximal gradient descent applied to the logistic group lasso problem (1) in explicit detail.

Answer:

For each group j, we have:

$$\hat{y}_j = X_j \beta_j$$

We then take a look at $\nabla g(\beta_j)$. Since we are taking gradient wrt to β_j which is group j, the only terms in g that have effects on the gradient is those belong to group j:

$$g(\beta) = -\sum_{i \,\in\, \text{group } j} y_i (X\beta)_i \; + \sum_{i \,\in\, \text{group } j} \log(1 + exp\{(X\beta)_i\}$$

And this reduces down into simple Lasso problem, thus we have:

$$\boxed{\nabla g(\beta_j) = X_j^T \left(\frac{e^{(X\beta)_j}}{1 + e^{(X\beta)_j}} - y_j \right)}$$

Proximal Update:

At each iteration k, we have:

$$\beta^{(k)} \;=\; \text{prox}_{h,t_k}\left(\beta^{(k-1)} - t_k \nabla g(\beta^{(k-1)})\right) \tag{22}$$

Due to the fact that $\text{prox}_{h,t}(\beta)$ is componentwise, we can consider for single group j:

$$\text{prox}_{h_j,t_k}\left(\beta_j^{(k-1)} - t_k \nabla g(\beta_j^{(k-1)})\right) = \max\left(0,\; 1 - \frac{\lambda t_k w_j}{\left\|\beta_j^{(k-1)} - t_k \nabla g(\beta_j^{(k-1)})\right\|_2}\right) \cdot \left(\beta_j^{(k-1)} - t_k \nabla g(\beta_j^{(k-1)})\right) \tag{23}$$

3. [20 points] Now we'll use the logistic group lasso to classify a person's age group from their movie ratings. The movie ratings can be categorized into groups according to a movie's genre (e.g., all ratings for action movies can be grouped together). Load the training data in `trainRatings.txt`, `trainLabels.txt`. The features have already been arranged into groups and you can find information about this in `groupTitles.txt`, `groupLabelsPerRating.txt`.

   Solve the logistic group lasso problem in Eqn. (19) with regularization parameter $\lambda = 5$ by running proximal gradient descent for 1000 iterations with fixed step size $t = 10^{-4}$.

   Plot $f^{(k)} - f^\star$ versus k, where $f^{(k)}$ denotes the objective value at iteration k, and use as an optimal objective value $f^\star = 336.207$. Make sure the plot is on a semi-log scale (where the y-axis is in log scale).

   Answer: Given in Jupyter Notebook attached

4. [Extra Credit: 10 points] Implement Nesterov acceleration for the same problem. You should again run accelerated proximal gradient descent for 1000 iterations with fixed step size $t = 10^{-4}$. As before, produce a plot $f^{(k)} - f^\star$ versus k. Describe any differences you see in the criterion convergence curve.

5. [Extra Credit: 10 points] Implement backtracking line search (rather than a fixed step size), and rerun proximal gradient for 400 iterations, without acceleration. (Note this means 400 outer iterations; the backtracking loop itself can take several inner iterations.) You should set $\beta = 0.1$ and $\alpha = 0.5$.

   Produce a plot of $f^{(k)} - f^\star$ versus $i(k)$, where $i(k)$ counts the *total* number of iterations performed at outer iteration k (total, meaning the sum of the iterations in both the inner and outer loops).

   Note: since it makes for an easier comparison, you may show the convergence curves from the last 3 sub-problems on the same plot using a different color/marker for each curve.
   Answer: Given in Jupyter Notebook attached

6. [10 points] Finally, use the solution from one of the proximal gradient descent methods introduced in parts 3,4 and 5 to make predictions on the test set, available in `testRatings.txt`, `testLabels.txt`. Indicate which method you have used.

   What is the classification error?

   What movie genre are important for classifying whether a viewer is under 40 years old?
   Answer: Given in Jupyter Notebook attached