Proximal Methods

October 9, 2024

1 Moreau decomposition

In this section, we will explore some applications of duality in settings related to proximal gradient methods. First, recall the definition of a proximal operator:

$$\text{prox}_f(v) = \arg\min_x \left(\frac{1}{2} \|x - v\|_2^2 + f(x) \right).$$

A useful fact for manipulating and extending proximal operators is known as **Moreau decomposition**. It states that the following relationship always holds:

$$v = \text{prox}_f(v) + \text{prox}_{f^*}(v),$$

where

$$f^*(y) = \max_x \left(y^\top x - f(x) \right).$$

Moreau's decomposition is "the main relationship between proximal operators and duality" and follows from the properties of sub-gradients and conjugate functions.

Notice that this is a generalization of orthogonal decomposition. Let L be a subspace of a vector space U. For any $v \in U$, we have

$$v = \Pi_L(v) + \Pi_{L^\perp}(v).$$

To illustrate the usefulness of this decomposition, we review a simple example. If $f(x) = \|x\|$, then $f^*(y) = I_B(y)$, where $B = \{z\colon \|z\|_* \leq 1\}$ is a unit ball according to the dual norm. By Moreau decomposition,

$$\begin{aligned} v &= \text{prox}_f(v) + \text{prox}_{f^*}(v) \\ &= \text{prox}_{\|\cdot\|}(v) + \text{prox}_{I_B}(v), \end{aligned}$$

where

$$\begin{aligned} \text{prox}_{I_B}(v) &= \arg\min_x \left(\frac{1}{2}\|x - v\|_2^2 + I_B(x)\right) \\ &= \arg\min_x \frac{1}{2}\|x - v\|_2^2 \text{ s.t. } x \in B \\ &= \Pi_B(v). \end{aligned}$$

It follows that

$$\text{prox}_{\|\cdot\|}(v) = v - \text{prox}_{I_B}(v) = v - \Pi_B(v).$$

2 Extending the Moreau Decomposition

Starting from the identity

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

we want to derive a similar identity when we replace f by λf for some $\lambda > 0$. We want to show that

$$\text{prox}_{\lambda f}(v) = v - \text{prox}_{(\lambda f)^*}(v) = v - \lambda \text{prox}_{f^*/\lambda}\left(v/\lambda\right).$$

First, we find the convex conjugate of λf:

$$\begin{aligned}
(\lambda f)^*(v) &= \max_y \left(v^\top y - \lambda f(y)\right) \\
&= \max_y \lambda \left(\frac{v}{\lambda}^\top y - f(y)\right) \\
&= \lambda \max_y \left(\frac{v}{\lambda}^\top y - f(y)\right) \\
&= \lambda f^* \left(\frac{v}{\lambda}\right).
\end{aligned}$$

Then, we get

$$\begin{aligned}
\text{prox}_{(\lambda f)^*}(v) &= \arg\min_y \left[(\lambda f)^*(y) + \frac{1}{2}\left\|y - v\right\|_2^2\right] \\
&= \arg\min_y \left[\lambda f^* \left(\frac{y}{\lambda}\right) + \frac{1}{2}\left\|y - v\right\|_2^2\right] \\
&= \arg\min_y \left[f^* \left(\frac{y}{\lambda}\right) + \frac{1}{2\lambda}\left\|y - v\right\|_2^2\right].
\end{aligned}$$

Now, we write $y = \lambda z$ to get

$$\begin{aligned}
\text{prox}_{(\lambda f)^*}(v) &= \arg\min_{\lambda z} \left[f^*\left(z\right) + \frac{1}{2\lambda}\left\|\lambda z - v\right\|_2^2\right] \\
&= \lambda \arg\min_z \left[f^*\left(z\right) + \frac{\lambda}{2}\left\|z - \frac{v}{\lambda}\right\|_2^2\right] \\
&= \lambda \text{prox}_{f^*/\lambda}\left(\frac{v}{\lambda}\right).
\end{aligned}$$

Finally, we have the identity

$$\text{prox}_{\lambda f}(v) = v - \text{prox}_{(\lambda f)^*}(v) = v - \lambda \text{prox}_{f^*/\lambda}(v/\lambda) .$$

If $f = \|\cdot\|$ is a general norm on $\mathbb{R}^n$, then

$$f^*(v) = I_B(v) = \begin{cases} 0 & \text{if } \|v\|_* \leq 1, \\ \infty & \text{otherwise.} \end{cases}$$

where $B = \{x : \|x\|_* \leq 1\}$ is the unit-ball in $(\mathbb{R}^n, \|\cdot\|_*)$. Observe that

$$f^*/\lambda = I_B/\lambda = I_B.$$

Then by Moreau decomposition, we get:

$$\text{prox}_{\lambda\|\cdot\|}(v) = v - \lambda \Pi_B\left(\frac{v}{\lambda}\right).$$

3 From Proximal to Projection

Euclidean norm penalty. Here, $f = f^* = \|\cdot\|_2$. We project v onto the Euclidean unit ball B as follows:

$$\Pi_B(v) = \begin{cases} v/\|v\|_2 & \text{if } \|v\|_2 > 1 \\ 0 & \text{if } \|v\|_2 \leq 1. \end{cases}$$

We get:

$$\begin{aligned}
\text{prox}_{\lambda\|\cdot\|_2}(v) &= v - \lambda \Pi_B\left(\frac{v}{\lambda}\right) \\
&= \begin{cases} (1 - \lambda/\|v\|_2)\, v & \text{if } \|v\|_2 \geq \lambda \\ 0 & \text{if } \|v\|_2 < \lambda \end{cases} \\
&= (1 - \lambda/\|v\|_2)_+\, v,
\end{aligned}$$

where

$$(z)_+ = \begin{cases} z & \text{if } z > 0 \\ 0 & \text{if } z \leq 0\,. \end{cases}$$

Group lasso penalty. This is how you compute proximal for each group in **group lasso**. For $x \in \mathbb{R}^p$,

$$f(x) = \sum_{g=1}^{G} w_g \|x_g\|_2$$

where $\{1, ..., p\}$ is partitioned into G groups. We get

$$\begin{aligned}
\text{prox}_{\lambda f}(v) &= \arg\min_x \frac{1}{2t}\|v - x\|_2^2 + \lambda f(x) \\
&= \arg\min_x \frac{1}{2t}\|v - x\|_2^2 + \lambda \sum_{g=1}^{G} w_g \|x_g\|_2.
\end{aligned}$$

So, for $g \in \{1, ..., G\}$,

$$\begin{aligned}[\text{prox}_{\lambda f}(v)]_g &= \arg\min_{x_g} \frac{1}{2t}\|v_g - x_g\|_2^2 + \lambda w_g \|x_g\|_2 \\ &= \text{prox}_{\lambda w_g \|x_g\|_2}(v_g) \\ &= \left(1 - \frac{t\lambda w_g}{\|v_g\|_2}\right)_+ v_g.\end{aligned}$$

Sparse group lasso penalty. For $x \in \mathbb{R}^p$, the sparse group lasso penalty is

$$f(x) = \sum_{g=1}^{G} w_g \left[(1-\alpha)\|x_g\|_2 + \alpha\|x_g\|_1\right]$$

where $\{1, ..., p\}$ is partitioned into G groups. This is how you compute proximal for each group in sparse **group lasso**. We get

$$\text{prox}_{\lambda f}(v) = \arg\min_{x} \frac{1}{2t}\|v - x\|_2^2 + \lambda f(x)$$

So, for $g \in \{1, ..., G\}$, define $\tau = t\lambda w_g$

$$\begin{aligned}[\text{prox}_{\lambda f}(v)]_g &= \arg\min_{x_g} \frac{1}{2t}\|v_g - x_g\|_2^2 + \lambda w_g \left[(1-\alpha)\|x_g\|_2 + \alpha\|x_g\|_1\right] \\ &= \text{prox}_{\lambda f}(v_g) \\ &= \left(1 - \frac{(1-\alpha)\tau}{\|S_{\alpha\tau}(v_g)\|_2}\right)_+ S_{\alpha\tau}(v_g),\end{aligned}$$

where

$$S_{\alpha\tau}(v_g) = \text{sgn}(v_g)(|v_g| - \alpha\tau)_+.$$

l^1 and l^∞ norms penalty. When $f = \|\cdot\|_1$, then $f^* = I_B$, $B = \{x : \|x\|_\infty \leq 1\}$. We project onto the ∞-norm unit ball B as follows:

$$(\Pi_B(v))_i = \begin{cases} 1 & : \; v_i > 1 \\ v_1 & : \; |v_i| \leq 1 \\ -1 & : \; v_i < -1. \end{cases}$$

We get an alternative way of getting the proximal operator of lasso

$$\text{prox}_{\lambda f}(v) = \text{prox}_{\lambda\|\cdot\|_1}(v) = v - \lambda\,\Pi_B\left(\frac{v}{\lambda}\right).$$

So

$$\left[\text{prox}_{\lambda f}(v)\right]_i = \begin{cases} v_i - \lambda & : \; v_i > \lambda \\ 0 & : \; |v_i| \leq \lambda \\ v_i + \lambda & : \; v_i < \lambda. \end{cases}$$

When $f = \|\cdot\|_\infty$, then $f^* = I_B$, $B = \{x : \|x\|_1 \leq 1\}$. See paper for how to project on B.

Hierarchical grouped norms. Assume the variables $X_1, ..., X_p$ have a hierarchical structure. The variables are selected according to the following rule, for $i \in \{1, ..., p\}$:

$$\text{if } \beta_i \neq 0, \text{ then } \beta_j \neq 0 \text{ for all } \beta_j \in \text{ancestors}(\beta_i).$$

We define the following penalty:

$$\Omega(\beta) = \sum_{g \in G} w_g \,\|\, (\beta_g, \text{ descendents}(\beta_g)) \,\|_2,$$

where G is the set of all nodes. The proximal operator for this penalty is:

$$\text{prox}_{\lambda\Omega}(v) = \arg\min_{u\in\mathbb{R}^p} \frac{1}{2}\|v-u\|_2^2 + \lambda\Omega(u)$$

Dual of the proximal problem. Let $v \in \mathbb{R}^p$. Consider

$$\max_{\xi\in\mathbb{R}^{p\times|G|}} -\frac{1}{2}\left(\|(v - \sum_{g\in G}\xi^g)\|_2^2 - \|v\|_2^2\right)$$

such that for all $g \in G$, $\|\xi^g\|_* \leq \lambda w_g$ and $\xi_j^g = 0$ if $j \notin g$.

4 Applications

4.1 Multitask sparse learning

Data: K data sources $\{\mathbf{y}^{(k)}, \mathbf{X}^{(k)}\}_{k=1}^K$: k-th data has n_k observations

- **Response:** $\mathbf{y}^{(k)} = (y_1^{(k)}, \ldots, y_{n_k}^{(k)})^\intercal$
- **Predictors:** $\mathbf{X}^{(k)} = (\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_{n_k}^{(k)})^\intercal$
  - $\mathbf{x}_i^{(k)} = (x_{i1}^{(k)}, \ldots, x_{ip}^{(k)})^\intercal \in \mathbb{R}^p$

Model: For continuous data, assume

$$E(y_i^{(k)}|\mathbf{x}_i^{(k)}) = \mathbf{x}_i^{(k)\top}\boldsymbol{\beta}^{(k)},$$

and for binary outcome use logistic regression setting

$$\text{logit}\left[P\left(y_i^{(k)}=1|x_i^{(k)}\right)\right]=\mathbf{x}_i^{(k)\top}\boldsymbol{\beta}^{(k)},$$

where $y_i^{(k)}=\{-1,1\}$, $i=1,\dots,n_k$. Here

$$\boldsymbol{\beta}^{(k)}=(\beta_1^{(k)},\cdots,\beta_p^{(k)})^\top$$

is the coefficient vector for task k. Here $\beta_j^{(k)}$ is the j-th element of $\boldsymbol{\beta}^{(k)}$, for $j=1,\dots,p$. And the vector

$$\boldsymbol{\beta}_j=(\beta_j^{(1)},\cdots,\beta_j^{(K)})^\top,$$

contains the j-th elements of task 1 to task K. The whole coefficient can be written as a $p\times K$ matrix

$$\boldsymbol{\beta}=(\boldsymbol{\beta}_1^\top,\cdots,\boldsymbol{\beta}_p^\top)^\top\in\mathbb{R}^{p\times K}$$

To estimate $\boldsymbol{\beta}$, for continuous outcome we minimize an aggregated least squares loss function

$$\ell(\boldsymbol{\beta})=\frac{1}{n}\sum_{k=1}^{K}\left[\mathbf{y}^{(k)}-\mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)}\right]^\top\left[\mathbf{y}^{(k)}-\mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)}\right], \tag{1}$$

for binary outcome, we use

$$\ell(\boldsymbol{\beta})=\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left[1+\exp\left(-y_i^{(k)}\mathbf{x}_i^{(k)\top}\boldsymbol{\beta}^{(k)}\right)\right] \tag{2}$$

To consider structural sparsity

- Common relevant covariates across data sources
- Source-specific relevant covariates

$$
\begin{array}{c}
\begin{bmatrix} \boldsymbol{\beta}^{(1)} & \boldsymbol{\beta}^{(2)} & \cdots & \boldsymbol{\beta}^{(K)} \end{bmatrix} \\
\xrightarrow{K \text{ sources}}
\end{array}
$$

$$
\begin{bmatrix} \boldsymbol{\beta}_1^\intercal \\ \boldsymbol{\beta}_2^\intercal \\ \vdots \\ \boldsymbol{\beta}_p^\intercal \end{bmatrix} = \begin{bmatrix} \beta_1^{(1)} & \beta_1^{(2)} & \cdots & \beta_1^{(K)} \\ \beta_2^{(1)} & \beta_2^{(2)} & \cdots & \beta^{(K)} \\ \vdots & \vdots & & \vdots \\ \beta_p^{(1)} & \beta_p^{(2)} & \cdots & \beta_p^{(K)} \end{bmatrix}
$$

(a) Sparse (b) Group sparse (c) Group sparse *plus* sparse

We use composite L_1/L_2 penalty

$$
P_{\alpha,\lambda}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} v_j \left[(1-\alpha)||\boldsymbol{\beta}_j||_2 + \alpha||\boldsymbol{\beta}_j||_1 \right]
$$

Lasso: when $\alpha = 1$

- Group lasso: when $\alpha = 0$
- Sparse group lasso: when $0 < \alpha < 1$