最优化算法作业 5

陈文轩

更新：May 21, 2025

1. 对于 group lasso 问题，请使用交替线性极小化方法求解，写出迭代公式。Group lasso 问题如下：

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \sum_{g=1}^{m} \sqrt{n_g}\|\beta_g\|_2$$

where：

- $y \in \mathbb{R}^n$ 和 $X \in \mathbb{R}^{n\times p}$ 是给定的数据，
- 变量 $\beta \in \mathbb{R}^p, \lambda > 0$ 是给定的参数，
- 下标集 $\{1, 2, \ldots, p\}$ 被分为不相交的 m 组，g 是分组下标，β_g 是对应分组 g 的分量，n_g 是分组 g 的分量维度。

解

首先进行外推操作：$\hat{\beta}_g^{k-1} = \beta_g^{k-1} + w_g^{k-1}(\beta_g^{k-1} - \beta_g^{k-2}), w_g^{k-1} > 0$ 是外推参数。

然后计算梯度：$\hat{g}_g^k = \nabla_{\beta_g} f(\hat{\beta}^{k-1}) = X_g^\top (X\hat{\beta}_{g^*}^{k-1} - y)$，其中 X_g 是 X 的组 g 对应列子矩阵，$\hat{\beta}_{g^*}^{k-1} = (\beta_1^k, \ldots, \beta_{g-1}^k, \hat{\beta}_g^{k-1}, \beta_{g+1}^{k-1}, \ldots, \beta_m^{k-1})$。

记 $z_g^k = \hat{\beta}_g{}^{k-1} - \frac{1}{L_g^{k-1}}\hat{g}_g^k, (t)_+ = \max\{t, 0\}$，其中 $L_g^{k-1} > 0$ 为常数，需要求解子问题如下：

$$\begin{aligned}
\beta_g^k &= \arg\min_{\beta_g} \left\{ \langle \hat{g}_g^k, \beta_g - \hat{\beta}_g^{k-1}\rangle + \frac{L_g^{k-1}}{2}\|\beta_g - \hat{\beta}_g^{k-1}\|_2^2 + \lambda\sqrt{n_g}\|\beta_g\|_2 \right\} \\
&= \arg\min_{\beta_g} \left\{ \frac{L_g^{k-1}}{2} \left\| \beta_g - \left(\hat{\beta}_g^{k-1} - \frac{1}{L_g^{k-1}}\hat{g}_g^k \right) \right\|_2^2 + \lambda\sqrt{n_g}\|\beta_g\|_2 \right\} \\
&= \arg\min_{\beta_g} \left\{ \frac{\lambda\sqrt{n_g}}{L_g^{k-1}}\|\beta_g\|_2 + \frac{1}{2}\|\beta_g - z_g^k\|_2^2 \right\} = \operatorname{prox}_{\frac{\lambda\sqrt{n_g}}{L_g^{k-1}}\|\cdot\|_2}(z_g^k) \\
&= \left(1 - \frac{\lambda\sqrt{n_g}}{L_g^{k-1}}\right)_+ z_g^k
\end{aligned}$$

以上即为完整的交替近似线性极小化算法更新。

2. 给定原问题

$$\min_{w\in\mathbb{R}^d} P(w) = \frac{1}{n}\sum_{i=1}^{n} \phi_i(w^\top x_i) + \frac{\lambda}{2}\|w\|^2$$

其中 ϕ_i 为闭凸函数。

证明：对偶问题表述为：

$$\max_{\alpha\in\mathbb{R}^n} D(\alpha) = \frac{1}{n}\sum_{i=1}^{n} -\phi_i^*(-\alpha_i) - \frac{\lambda}{2}\left\|\frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i x_i\right\|^2$$

其中 $\phi_i^*(u) = \max_z(zu - \phi_i(z))$ 是 ϕ_i 的共轭函数。

解

把问题重写为 $\min_{w\in\mathbb{R}^d, u\in\mathbb{R}^n}\left(\frac{1}{n}\sum_{i=1}^{n}\phi_i(u_i) + \frac{\lambda}{2}\|w\|^2\right), \text{s.t. } u_i = x_i^\top w, i = 1,\dots,n,$

则 Lagrange 函数为 $L(w,u,\mu) = \frac{1}{n}\sum_{i=1}^{n}\phi_i(u_i) + \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{n}\mu_i(x_i^\top w - u_i)$。以下记 $n\mu_i = \alpha_i$，

此时 $L(w,u,\mu) = \mathcal{L}(w,u,\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left(\phi_i(u_i) + \alpha_i(x_i^\top w - u_i)\right) + \frac{\lambda}{2}\|w\|^2$，对 u, w 求极小：

考虑 u_i，即 $\min_{u_i} f_i(u_i) \coloneqq \min_{u_i}\frac{1}{n}\left(\phi_i(u_i) - \alpha_i u_i\right) = -\max_{u_i}\frac{1}{n}\left(u_i(-\alpha_i) - \phi_i(u_i)\right) = -\frac{1}{n}\phi_i^*(-\alpha_i)$。

考虑 w，即 $\min_w g(w) \coloneqq \min_w\left(\frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n}\alpha_i x_i^\top w\right), \nabla g = \lambda w + \frac{1}{n}\sum_{i=1}^{n}\alpha_i x_i = 0$

$$\Rightarrow w = -\frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i x_i, g(w) = \frac{\lambda}{2}\left\|-\frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i x_i\right\|^2 + \frac{1}{\lambda n^2}\left\|\sum_{i=1}^{n}\alpha_i x_i\right\|^2 = -\frac{\lambda}{2}\left\|\frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i x_i\right\|^2。$$

相加以上函数，对偶问题为 $\max_\alpha D(\alpha) = -\frac{1}{n}\sum_{i=1}^{n}\phi_i^*(-\alpha_i) - \frac{\lambda}{2}\left\|\frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i x_i\right\|^2$。

3. 给定矩阵 $A\in\mathbb{R}^{m\times n}$，使用随机梯度法求解如下问题，写出迭代具体公式

$$\min_{x\in\mathbb{R}^{n\times r}} \operatorname{tr}(x^\top A^\top A x).$$

解

记 $f(x) = \operatorname{tr}(x^\top A^\top A x) = \|Ax\|_F^2 = \sum_{i=1}^{m}\|a_i^\top x\|_2^2$，其中 $a_i^\top$ 是 A 的第 i 行，则 $\nabla f(x) = 2A^\top A x$。

均匀抽取一行 $a_{i_k}^\top$ 来近似梯度 $g_k = 2m a_{i_k} a_{i_k}^\top x_k$，由 $\mathbb{E}(g_k) = 2m\left(\frac{1}{m}\sum_{i=1}^{m} a_i a_i^\top\right)x = 2A^\top A x$，

有 g_k 是 $\nabla f(x)$ 的无偏估计。假设 η_k 是步长序列，则迭代为 $x_{k+1} = x_k - \eta_k g_k$。

4. 在最小二乘问题中出现的平方误差损失函数非常适合使用随机梯度方法进行最小化。我们的问题是：

$$\min_{\theta\in\mathbb{R}^n}\frac{1}{2}\|X\theta - y\|_2^2 = \frac{1}{2}\sum_{i=1}^{m}(x_i^\top\theta - y_i)^2,$$

其中 $x_i^\top$ 是 $X\in\mathbb{R}^{m\times n}$ 的第 i 行，$y\in\mathbb{R}^m$。我们可以将这个目标函数写为：

$$f(\theta) = \sum_{i=1}^{m} f(\theta),$$

其中

$$f_i(\theta) = \frac{1}{2}(x_i^\top \theta - y_i)^2，对于 i = 1, \ldots, m.$$

然后随机梯度方法给出更新规则：

$$\theta_{k+1} = \theta_k - \eta_k \nabla f_{s[k]}(\theta_k)$$

其中 η_k 是步长（也称为学习率），$s[k] \in \{1, \ldots, m\}$，通常是通过从集合 $\{1, \ldots, m\}$ 中随机抽取一个数字。

(a). 假设 $\{x_i\}_{i=1}^m$ 是一组相互正交的向量。找到一个固定的步长 η，使得随机梯度方法收敛到最小二乘问题的解。

(b). 如果没有条件上述条件，即 $\{x_i\}_{i=1}^m$ 并不相互正交，那么随机梯度法的收敛需要什么条件？

解

(a). 记最小二乘解为 $\theta^* = (X^\top X)^{-1} X^\top y$，误差 $e_k = \theta_k = \theta^*$，梯度 $\nabla f_i(\theta) = (x_i^\top \theta - y_i) x_i$.
设 $s[k] = i$，则 $e_{k+1} = e_k - \eta(x_i^\top \theta - y_i)x_i = e_k - \eta(x_i^\top e_k)x_i$。由于 $\{x_i\}$ 正交，
有分解 $e_k = \sum_{j=1}^{m} \frac{x_j^\top e_k}{\|x_i\|_2^2} x_j + e^\perp$，其中 $e^\perp$ 与所有 x_i 正交，更新时不变。更新后结果为
$e_{k+1} = \sum_{j \neq i} \frac{x_j^\top e_k}{\|x_i\|_2^2} x_j + \frac{x_i^\top e_k}{\|x_i\|_2^2}(1 - \|x_i\|_2^2 \eta) x_j + e^\perp$，若误差减少，则 $(1 - \|x_i\|_2^2 \eta) < 1, \forall i$。
此时需要 $\eta \in \left(0, \frac{2}{\max\limits_i \|x_i\|_2^2}\right)$，此时取固定步长 $\eta = \frac{1}{\max\limits_i \|x_i\|_2^2}$ 即可。

(b). 此时需要 $X^\top X$ 非奇异，即 $X^\top X$ 正定，此时目标函数强凸，解唯一。
随机梯度均匀采样时其无偏且二阶矩有界。记 $L = \lambda_{\max}(X^\top X)$，则以下学习率收敛：

- 满足 $\eta \in \left(0, \frac{2}{L}\right)$ 的常数学习率；
- 满足 $\sum_{k=0}^{\infty} \eta_k = \infty, \sum_{k=0}^{\infty} \eta_k^2 < \infty$ 的递减学习率。