Theory for Rank2Plan

August 29, 2024

1 First Order Methods

While the linear programs we seek to solve for ranking are similar to that Dedieu et al. (2022), there are a few important differences. This in particular means we have to adapt the first order methods used by Dedieu et al. (2022) to find an initial low-accuracy solution. The following discussion is based on Section 4 of Dedieu et al. (2022).

The ranking problem can be described as given a set of feature vectors $X \in \mathbb{R}^{n\times p}$, and a set of pairs $P = \{(i,j)\}$, find a ranking function $r : \mathbb{R}^p \to \mathbb{R}$ such that $r(X_j) - r(X_i) \geq g_{i,j} \in \mathbb{R}_{\geq}$ with importance $s_{i,j} \in \mathbb{R}_{\geq}$.[1] We restrict r to be a linear function with zero bias, in which case the ranking problem can be formulated as the following linear program

$$\begin{aligned}\mathcal{P}_\lambda := \min_{\boldsymbol{\xi}\in\mathbb{R}^m,\boldsymbol{\beta}^+\in\mathbb{R}^p,\boldsymbol{\beta}^-\in\mathbb{R}^p} & \sum_{(i,j)\in P} s_{i,j}\xi_{i,j} + \lambda\sum_{i=1}^{p}\beta_i^+ + \lambda\sum_{i=1}^{p}\beta_i^- \\ \text{s.t.} & \boldsymbol{\beta}^+X_j^T - \boldsymbol{\beta}^-X_j^T - \boldsymbol{\beta}^+X_i^T + \boldsymbol{\beta}^-X_i^T \geq g_{i,j} - \xi_{i,j} \quad (i,j)\in P \\ & \boldsymbol{\xi} \geq 0, \boldsymbol{\beta}^+ \geq 0, \boldsymbol{\beta}^- \geq 0\end{aligned} \tag{1}$$

where $m = |P|$ and r is given by $r(\boldsymbol{x}) = \boldsymbol{\beta}\boldsymbol{x}^T$ where $\boldsymbol{\beta} = \boldsymbol{\beta}^+ - \boldsymbol{\beta}^-$.

1.1 Solving the Composite Form with Nesterov's Smoothing

For scaler u, we have that $\max(0,u) = \frac{1}{2}(u+|u|) = \max_{|w|\leq 1}\frac{1}{2}(u+wu)$ with the maximum achieved when $w = \text{sign}(u)$. Using this, the hinge loss in $\mathcal{P}_\lambda$ can thus be expressed as

$$\sum_{(i,j)\in P}(z_{i,j})_+ = \max_{\|\boldsymbol{w}\|_\infty\leq 1}\sum_{(i,j)\in P}\frac{1}{2}[z_{i,j} + w_{i,j}z_{i,j}], \tag{2}$$

where $z_{i,j} = s_{i,j}(g_{i,j} - (X_j^T\boldsymbol{\beta} - X_i^T\boldsymbol{\beta}))$. At this point, it is tempting to just divide by $s_{i,j}g_{i,j}$ to retrieve the same form as used in Dedieu et al. (2022), which would make everything really easy! Unfortunately, to preserve the original solution to the LP, we cannot divide different terms by a different value and so we must do the work ourselves.

One can obtain a smooth approximate of (2) as[2]

$$H^\tau(\boldsymbol{z}) := \max_{\|\boldsymbol{w}\|_\infty\leq 1}\sum_{(i,j)\in P}\frac{1}{2}[z_{i,j} + w_{i,j}z_{i,j}] - \frac{\tau}{2}\|w\|_2^2, \tag{3}$$

where τ controls the smoothness of H^τ and how well it approximates the original hinge loss (where $\tau = 0$).

[1] Note that the order of i and j means we prefer lower values

[2] We abuse the notation slightly and treat $\boldsymbol{w}$ and $\boldsymbol{z}$ as vectors even though they are indexed by the pair (i,j).

The following lemma, which is lemma 7 from Dedieu et al. (2022) and originally from Nesterov (2005), characterises H^τ.

Lemma 1 *The function $\boldsymbol{z} \mapsto H^\tau(\boldsymbol{z})$ is an $O(\tau)$ approximation for the hinge loss $H^0(\boldsymbol{z})$, i.e. $H^0(\boldsymbol{z}) \in [H^\tau(\boldsymbol{z}), H^\tau(\boldsymbol{z}) + n\tau/2]$ for all $\boldsymbol{z}$. Furthermore, $H^\tau(\boldsymbol{z})$ has Lipschitz continuous gradient with parameter $1/(4\tau)$, i.e. $\|\nabla H^\tau(\boldsymbol{z}) - \nabla H^\tau(\boldsymbol{z'})\| \leq 1/(4\tau)\|\boldsymbol{z} - \boldsymbol{z'}\|_2$.*

Let us define,

$$F^\tau(\boldsymbol{\beta}) = \max_{\|\boldsymbol{w}\|_\infty \leq 1} \left\{ \sum_{(i,j)\in P} \frac{1}{2}[s_{i,j}(g_{i,j} - (X_j^T\boldsymbol{\beta} - X_i^T\boldsymbol{\beta})) + w_{i,j}s_{i,j}(g_{i,j} - (X_j^T\boldsymbol{\beta} - X_i^T\boldsymbol{\beta}))] - \frac{\tau}{2}\|\boldsymbol{w}\|_2^2 \right\}. \quad (4)$$

By Lemma 1, F^τ is a uniform $O(\tau)$-approximation of the hinge loss function. Its gradient is given by,

$$\nabla F^\tau(\boldsymbol{\beta}) = -\frac{1}{2} \sum_{(i,j)\in P} (1 + w_{i,j}^\tau)s_{i,j}(X_j - X_i) \in \mathbb{R}^p \quad (5)$$

where $\boldsymbol{w}^\tau$ is the optimal solution to (4) at $\boldsymbol{\beta}$. Further, $\boldsymbol{\beta} \mapsto \nabla F^\tau(\boldsymbol{\beta})$ is Lipschitz-continuous with parameter $C^\tau = \sigma_{\max}(\tilde{X}^T\tilde{X})/(4\tau)$ where $\tilde{X}$ is the matrix with rows given by $s_{i,j}(X_j - X_i)^T$ for $(i,j) \in P$ and $\sigma_{\max}$ denotes the maximum eigenvalue as

$$\begin{aligned}
\|\boldsymbol{z} - \boldsymbol{z'}\|_2^2 &= \sum_{(i,j)\in P} [s_{i,j}(g_{i,j} - (X_j^T\boldsymbol{\beta} - X_i^T\boldsymbol{\beta})) - (s_{i,j}(g_{i,j} - (X_j^T\boldsymbol{\beta'} - X_i^T\boldsymbol{\beta'})))]^2 \\
&= \sum_{(i,j)\in P} [s_{i,j}(X_j - X_i)^T(\boldsymbol{\beta'} - \boldsymbol{\beta})]^2 \\
&= \sum_{(i,j)\in P} [\tilde{X}_{i,j}^T(\boldsymbol{\beta'} - \boldsymbol{\beta})]^2 \\
&= \|\tilde{X}(\boldsymbol{\beta} - \boldsymbol{\beta'})\|_2^2 \\
&\leq \|\tilde{X}\|_2^2\|\boldsymbol{\beta} - \boldsymbol{\beta'}\|_2^2 \\
&= \sigma_{\max}(\tilde{X}^T\tilde{X})\|\boldsymbol{\beta} - \boldsymbol{\beta'}\|_2^2
\end{aligned}$$

The rest then follows directly in the exact same logic as Section 4 of Dedieu et al. (2022).

References

A. Dedieu, R. Mazumder, and H. Wang. Solving l1-regularized svms and related linear programs: Revisiting the effectiveness of column and constraint generation. *J. Mach. Learn. Res.*, 23:164:1–164:41, 2022. URL `http://jmlr.org/papers/v23/19-104.html`.

Y. E. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005. doi: 10.1007/S10107-004-0552-5.