Learning to Rank with Bregman Divergences and Monotone Retargeting

Sreangsu Acharyya, Oluwasanmi Koyejo and Joydeep Ghosh
University of Texas at Austin

Introduction

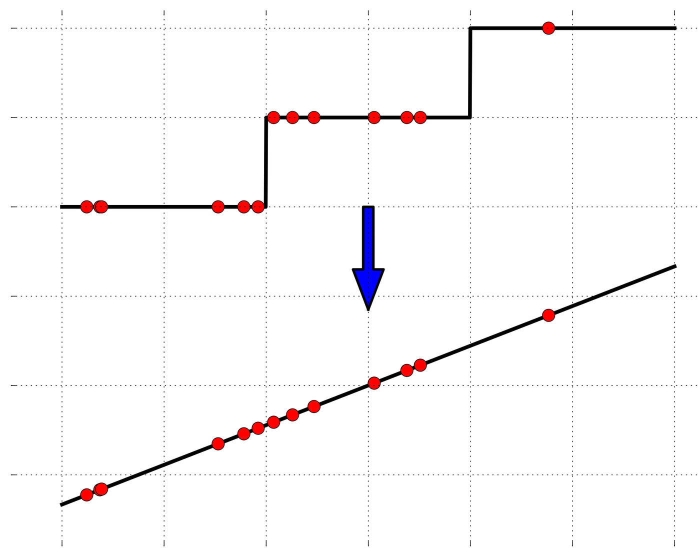
- Learns to rank-order items (from examples).
- Simple to implement, embarrassingly parallelizable, provably convergent, and attains global minimum under mild conditions.

Problem setup:

- For every query $q_i \in \mathcal{Q}$, there is a set of ordered items $\mathcal{V}_i \ni \{v_{i,j}\}$.
- The ordering is specified by a rank score vector $\tilde{\mathbf{r}}_i \in \mathbb{R}^{|\mathcal{V}_i|}$.
- Row j of feature matrix $\mathbf{A}_i$ is computed using the pair $\{q_i, v_{i,j}\}$.

Monotone Retargeting

- Prevalent approach: regress the scores $\tilde{\mathbf{r}}_i$.
- Our main Idea: no need to fit $\tilde{\mathbf{r}}_i$ exactly, sufficient to fit any score that preserves order.
- MR searches for an order preserving transformation of the target scores that may be easier for the regressor to fit.



Bregman Divergence

$$D_\phi\left(\mathbf{x}\middle|\middle|\mathbf{y}\right) \triangleq \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x}-\mathbf{y}, \nabla\phi(\mathbf{y})\rangle$$

- Squared L_2 metric, KL Divergence, GLM loglikelihood ...
- Unique class of cost functions statistically consistent with the normalized discounted gain (NDCG) [Ravikumar et al., 2011].
- We assume ϕ is separable.

Formulation: Block Coordinate Descent in r and w

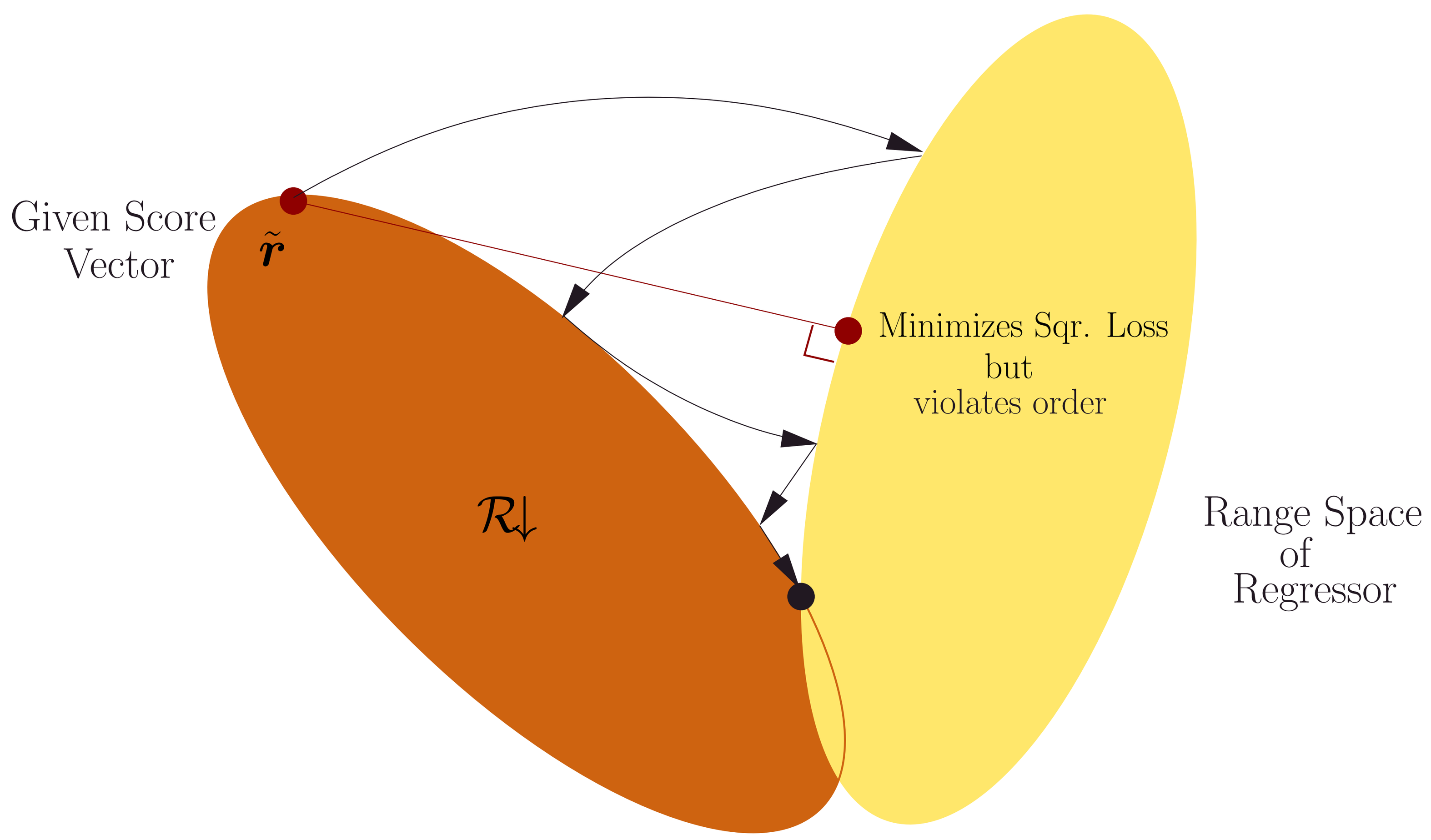
$$\min_{\mathbf{w}, \mathbf{r}_i \in \mathcal{R}\downarrow_i \cap \Delta} \sum_{i=1}^{|\mathcal{Q}|} \frac{1}{|\mathcal{V}_i|} D_\phi\left(\mathbf{r}_i \middle|\middle| (\nabla\phi)^{-1}(\mathbf{A}_i\mathbf{w})\right) + \frac{C}{2}||\mathbf{w}||^2,$$

$$\text{s.t. } \mathcal{R}\downarrow_i = \{\mathbf{r} \,|\, \exists \mathbf{M} \in \mathcal{M} \text{ s.t. } \mathbf{M}(\tilde{\mathbf{r}}_i) = \mathbf{r}\},$$
$$\mathcal{M} = \text{the set of all monotonic transformations.}$$

- When $\mathbf{0} \in \text{dom}\,\phi(\cdot)$, $\mathbf{r}_i$ should be bounded away from $\mathbf{0}$.
- For such cost functions, we constrain $\mathbf{r}_i \in \Delta_o = \mathcal{R}\downarrow \cap \Delta$.

Lemma: The set Δ_o of all discrete probability distributions of dimension d that are in descending order is the image $T\mathbf{x}$ s.t. $\mathbf{x} \in \Delta$ where T is an upper triangular matrix generated from the vector $\mathbf{v}_\Delta = \{1, \frac{1}{2} \cdots \frac{1}{d}\}$ such that $T(i,:) = \{0\}^{i-1} \times \mathbf{v}_\Delta(i:)$.

Alternating Projections



Universality of Minimizers

Theorem 1: For $\mathcal{R}\downarrow \subset \mathbb{R}^d$ the set of vectors with descending ordered components, the minimizer $\mathbf{y}^* = \underset{\mathbf{y}\in\mathcal{R}\downarrow}{\text{Argmin}} D_\phi\left(\mathbf{x}\middle|\middle|\mathbf{y}\right)$ is independent of $\phi(\cdot)$.

Corollary 2: If dom $\psi(\cdot) = \mathbb{R}^d$ where $\psi(\cdot)$ is the conjugate of $\phi(\cdot)$, then:

$$\text{Argmin}_{\mathbf{r}\in\mathcal{R}\downarrow\cap\text{dom}\,\phi} D_\phi\left(\mathbf{r}\middle|\middle|(\nabla\phi)^{-1}(\mathbf{x})\right) = (\nabla\phi)^{-1}(\mathbf{z}^*),$$

where $\mathbf{z}^* = \text{Argmin}_{\mathbf{z}\in\mathcal{R}\downarrow} ||\mathbf{x}-\mathbf{z}||^2$ (Reduction to squared loss minimization).

Joint Convexity and Global Minimum

The cost function is related to the gap in the Fenchel-Young inequality given by:

$$D_\phi\left(\mathbf{r}\middle|\middle|(\nabla\phi)^{-1}(\mathbf{y})\right) = (\phi)^*(\mathbf{y}) + \phi(\mathbf{r}) - \langle \mathbf{r}, \mathbf{y}\rangle$$

Theorem 3: For any twice differentiable strictly convex $\phi(\cdot)$ with a differentiable conjugate $(\phi)^*(\cdot)$, the gap is jointly convex if and only if $\phi(\mathbf{r}) = c||\mathbf{r}||^2 \;\; \forall\, c > 0$.

Sufficiency of Sorting

- Here, we assume that the items are totally ordered, though the finer ordering between similar items is not visible to the ranking algorithm.

Theorem 4: If $r_1 \geq r_2$ and $y_1 \geq y_2$, then $D_\phi\left(\begin{bmatrix} r_1 \\ r_2\end{bmatrix}\middle|\middle|\begin{bmatrix} y_1 \\ y_2\end{bmatrix}\right) \leq D_\phi\left(\begin{bmatrix} r_1 \\ r_2\end{bmatrix}\middle|\middle|\begin{bmatrix} y_2 \\ y_1\end{bmatrix}\right)$ and $D_\phi\left(\begin{bmatrix} y_1 \\ y_2\end{bmatrix}\middle|\middle|\begin{bmatrix} r_1 \\ r_2\end{bmatrix}\right) \leq D_\phi\left(\begin{bmatrix} y_2 \\ y_1\end{bmatrix}\middle|\middle|\begin{bmatrix} r_1 \\ r_2\end{bmatrix}\right)$. (Extend to $\mathbf{r} \in \mathbb{R}^d$ using induction over d.)

- Thus, no need to solve linear assignment problem in an inner loop.

Algorithm for Partially Hidden Order

$$\mathbb{P}_i^{t+1} = \underset{\mathbb{P}}{\text{Argmin}}\, D_\phi\left(T\mathbf{x}_i^t\middle|\middle|(\nabla\phi)^{-1}\left(\mathbb{P}\mathbf{A}_i\mathbf{w}^t + \beta_i^t\right)\right) \;\; \forall i \text{ in parallel}$$

$$\mathbf{x}_i^{t+1} = \underset{\mathbf{x}\in\Delta}{\text{Argmin}}\, D_\phi\left(T\mathbf{x}\middle|\middle|(\nabla\phi)^{-1}\left(\mathbb{P}_i^{t+1}\mathbf{A}_i\mathbf{w}^t + \beta_i^t\right)\right) \;\; \forall i \text{ in parallel}$$

$$\mathbf{w}^{t+1} = \underset{\mathbf{w}}{\text{Argmin}} \sum_{i=1}^{|\mathcal{Q}|} D_\phi\left(T\mathbf{x}_i^{t+1}\middle|\middle|(\nabla\phi)^{-1}\left(\mathbb{P}_i^{t+1}\mathbf{A}_i\mathbf{w} + \beta_i^t\right)\right) + \frac{C}{2}||\mathbf{w}||^2$$

Experiments

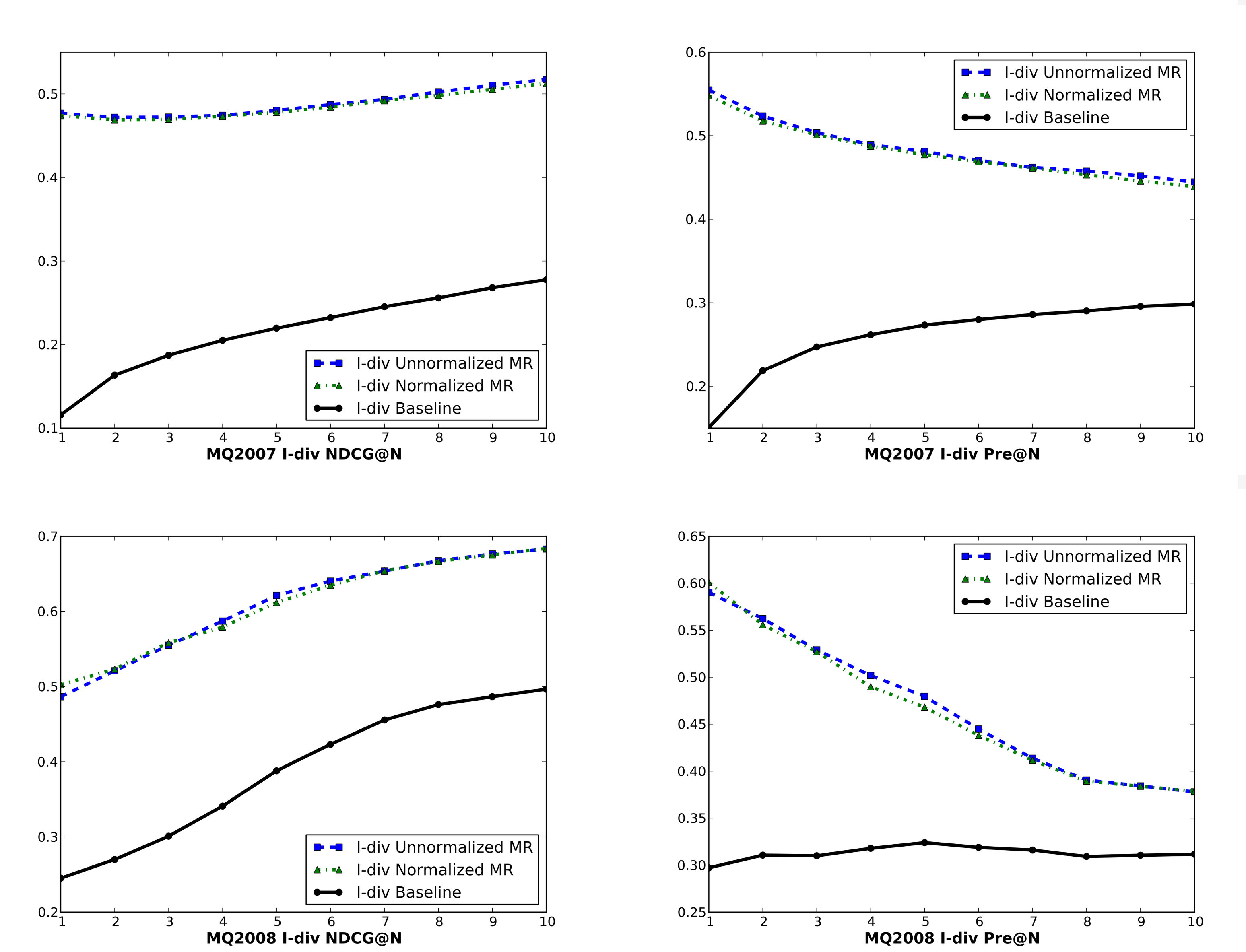


Figure: MR vs. NDCG consistent baseline

- MR improves NDCG performance over baseline algorithms specifically designed for optimizing NDCG.

Conclusion

- This work introduces a new family of cost functions for ranking.
- Listwise ranking model that can be easily optimized
- MR can globally optimize jointly over
  - regression parameters, and
  - all monotonic transformations
- MR has favorable statistical and optimization theoretic properties, and excellent empirical performance.