Fast Differentiable Sorting and Ranking

Mathieu Blondel [1] **Olivier Teboul** [1] **Quentin Berthet** [1] **Josip Djolonga** [1]

Abstract

The sorting operation is one of the most basic and commonly used building blocks in computer programming. In machine learning, it is commonly used for robust statistics. However, seen as a function, it is piecewise linear and as a result includes many kinks at which it is non-differentiable. More problematic is the related ranking operator, commonly used for order statistics and ranking metrics. It is a piecewise constant function, meaning that its derivatives are null or undefined. While numerous works have proposed differentiable proxies to sorting and ranking, they do not achieve the $O(n \log n)$ time complexity one would expect from sorting and ranking operations. In this paper, we propose the first differentiable sorting and ranking operators with $O(n \log n)$ time and $O(n)$ space complexity. Our proposal in addition enjoys exact computation and differentiation. We achieve this feat by constructing differentiable sorting and ranking operators as projections onto the permutahedron, the convex hull of permutations, and using a reduction to isotonic optimization. Empirically, we confirm that our approach is an order of magnitude faster than existing approaches and showcase two novel applications: differentiable Spearman's rank correlation coefficient and soft least trimmed squares.

1. Introduction

Modern deep learning architectures are built by composing parameterized functional blocks (including loops and conditionals) and are trained end-to-end using gradient backpropagation. This has motivated the term **differentiable programming**, recently popularized, among others, by LeCun (2018). Despite great empirical successes, many operations commonly used in computer programming remain poorly differentiable or downright pathological, limiting the set of architectures for which a gradient can be computed.

We focus in this paper on two such operations: **sorting** and **ranking**. Sorting returns the given input vector with its values re-arranged in monotonic order. It plays a key role to handle outliers in robust statistics, as in least-quantile (Rousseeuw, 1984) or trimmed (Rousseeuw & Leroy, 2005) regression. As a piecewise linear function, however, the sorted vector contains many kinks where it is non-differentiable. In addition, when used in composition with other functions, sorting often induces non-convexity, thus rendering model parameter optimization difficult.

The ranking operation, on the other hand, outputs the positions, or ranks, of the input values in the sorted vector. A workhorse of order statistics (David & Nagaraja, 2004), ranks are used in several metrics, including Spearman's rank correlation coefficient (Spearman, 1904), top-k accuracy and normalized discounted cumulative gain (NDCG). As piecewise constant functions, ranks are unfortunately much more problematic than sorting: their derivatives are null or undefined, preventing gradient backpropagation. For this reason, a large body of work has studied differentiable proxies to ranking. While several works opt to approximate ranking metrics directly (Chapelle & Wu, 2010; Adams & Zemel, 2011; Lapin et al., 2016), others introduce "soft" ranks, which can then be plugged into any differentiable loss function. Taylor et al. (2008) use a random perturbation technique to compute expected ranks in $O(n^3)$ time, where n is the dimensionality of the vector to rank. Shortly after, Qin et al. (2010) propose a simple method based on comparing pairwise distances between values, thereby taking $O(n^2)$ time. This method is refined by Grover et al. (2019) using unimodal row-stochastic matrices. Lastly, Cuturi et al. (2019) adopt an optimal transport viewpoint of sorting and ranking. Their method is based on differentiating through the iterates of the Sinkhorn algorithm (Sinkhorn & Knopp, 1967), thereby costing $O(Tn^2)$ time, where T is the number of Sinkhorn iterations. Alas, none of these approaches achieve the $O(n \log n)$ time complexity one would expect from sorting and ranking operations.

In this paper, we propose the first differentiable sorting and ranking operators with $O(n \log n)$ time and $O(n)$ memory complexity. Our proposals enjoy **exact** computation and differentiation (i.e., they do not involve differentiating through

[1]Google Research, Brain team. Correspondence to: Mathieu Blondel <mblondel@google.com>, Olivier Teboul <oliviert@google.com>, Quentin Berthet <qberthet@google.com>, Josip Djolonga <josipd@google.com>.

the iterates of an approximate algorithm). We achieve this feat by casting differentiable sorting and ranking as projections onto the permutahedron, the convex hull of all permutations, and using a reduction to isotonic optimization. While the permutahedron had been used for learning before (Yasutake et al., 2011; Ailon et al., 2016; Blondel, 2019), it had not been used to define fast differentiable operators. The rest of the paper is organized as follows.

- We review the necessary background (§2) and show how to cast sorting and ranking as linear programs over the permutahedron, the convex hull of all permutations (§3).
- We introduce regularization in these linear programs, which turns them into projections onto the permutahedron and allows us to define differentiable sorting and ranking operators. We analyze the properties of these operators, such as their asymptotic behavior (§4).
- Using a reduction to isotonic optimization, we achieve $O(n \log n)$ computation and $O(n)$ differentiation of our operators, a key technical contribution of this paper (§5).
- We show that our approach is an order of magnitude faster than existing approaches and showcase two novel applications: differentiable Spearman's rank coefficient and soft least trimmed squares (§6).

2. Preliminaries

In this section, we define the notation that will be used throughout this paper. Let $\boldsymbol{\theta} \coloneqq (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$. We will think of $\boldsymbol{\theta}$ as a vector of scores or "logits" produced by a model, i.e., $\boldsymbol{\theta} \coloneqq g(\boldsymbol{x})$ for some $g \colon \mathcal{X} \to \mathbb{R}^n$ and some $\boldsymbol{x} \in \mathcal{X}$. For instance, in a label ranking setting, $\boldsymbol{\theta}$ may contain the score of each of n labels for the features $\boldsymbol{x}$.

We denote a **permutation** of $[n]$ by $\sigma = (\sigma_1, \dots, \sigma_n)$ and its inverse by σ^{-1}. For convenience, we will sometimes use $\pi \coloneqq \sigma^{-1}$. If a permutation σ is seen as a vector, we denote it with bold, $\boldsymbol{\sigma} \in [n]^n$. We denote the set of $n!$ permutations of $[n]$ by Σ. Given a permutation $\sigma \in \Sigma$, we denote the version of $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$ permuted according to σ by $\boldsymbol{\theta}_\sigma \coloneqq (\theta_{\sigma_1}, \dots, \theta_{\sigma_n}) \in \mathbb{R}^n$. We define the reversing permutation by $\rho \coloneqq (n, n-1, \dots, 1)$ or $\boldsymbol{\rho}$ in vector form. Given a set $\mathcal{S} \subseteq [n]$ and a vector $\boldsymbol{v} \in \mathbb{R}^n$, we denote the restriction of $\boldsymbol{v}$ to $\mathcal{S}$ by $\boldsymbol{v}_\mathcal{S} \coloneqq (v_i \colon i \in \mathcal{S}) \in \mathbb{R}^{|\mathcal{S}|}$.

We define the **argsort** of $\boldsymbol{\theta}$ as the indices sorting $\boldsymbol{\theta}$, i.e.,

$$\sigma(\boldsymbol{\theta}) \coloneqq (\sigma_1(\boldsymbol{\theta}), \dots, \sigma_n(\boldsymbol{\theta})),$$

where $\theta_{\sigma_1(\boldsymbol{\theta})} \geq \dots \geq \theta_{\sigma_n(\boldsymbol{\theta})}$. If some of the coordinates of $\boldsymbol{\theta}$ are equal, we break ties arbitrarily. We define the **sort** of $\boldsymbol{\theta}$ as the values of $\boldsymbol{\theta}$ in descending order, i.e.,

$$s(\boldsymbol{\theta}) \coloneqq \boldsymbol{\theta}_{\sigma(\boldsymbol{\theta})}.$$

We define the **rank** of $\boldsymbol{\theta}$ as the function evaluating at coordinate j to the position of θ_j in the descending sort (smaller rank $r_j(\boldsymbol{\theta})$ means that θ_j has higher value). It is formally equal to the argsort's inverse permutation, i.e.,

$$r(\boldsymbol{\theta}) \coloneqq \sigma^{-1}(\boldsymbol{\theta}).$$

For instance, if $\theta_3 \geq \theta_1 \geq \theta_2$, then $\sigma(\boldsymbol{\theta}) = (3, 1, 2)$, $s(\boldsymbol{\theta}) = (\theta_3, \theta_1, \theta_2)$ and $r(\boldsymbol{\theta}) = (2, 3, 1)$. All three operations can be computed in $O(n \log n)$ time. Note that throughout this paper, we use descending order for convenience. The ascending order counterparts are easily obtained by $\sigma(-\boldsymbol{\theta})$, $-s(-\boldsymbol{\theta})$ and $r(-\boldsymbol{\theta})$, respectively.

3. Sorting and ranking as linear programs

We show in this section how to cast sorting and ranking operations as linear programs over the permutahedron. To that end, we first formulate the argsort and ranking operations as optimization problems over the set of permutations Σ.

Lemma 1. *Discrete optimization formulations*

For all $\boldsymbol{\theta} \in \mathbb{R}^n$ and $\boldsymbol{\rho} \coloneqq (n, n-1, \dots, 1)$, we have

$$\sigma(\boldsymbol{\theta}) = \operatorname*{argmax}_{\sigma \in \Sigma} \langle \boldsymbol{\theta}_\sigma, \boldsymbol{\rho} \rangle, \text{ and} \tag{1}$$

$$r(\boldsymbol{\theta}) = \operatorname*{argmax}_{\pi \in \Sigma} \langle \boldsymbol{\theta}, \boldsymbol{\rho}_\pi \rangle. \tag{2}$$

A proof is provided in §B.1. To obtain continuous optimization problems, we introduce the permutahedron induced by a vector $\boldsymbol{w} \in \mathbb{R}^n$, the convex hull of permutations of $\boldsymbol{w}$:

$$\mathcal{P}(\boldsymbol{w}) \coloneqq \operatorname{conv}(\{\boldsymbol{w}_\sigma \colon \sigma \in \Sigma\}) \subset \mathbb{R}^n.$$

A well-known object in combinatorics (Bowman, 1972; Ziegler, 2012), the permutahedron of $\boldsymbol{w}$ is a convex polytope, whose vertices correspond to permutations of $\boldsymbol{w}$. It is illustrated in Figure 1. In particular, when $\boldsymbol{w} = \boldsymbol{\rho}$, $\mathcal{P}(\boldsymbol{w}) = \operatorname{conv}(\Sigma)$. With this defined, we can now derive linear programming formulations of sort and ranks.

Proposition 1. *Linear programming formulations*

For all $\boldsymbol{\theta} \in \mathbb{R}^n$ and $\boldsymbol{\rho} \coloneqq (n, n-1, \dots, 1)$, we have

$$s(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{P}(\boldsymbol{\theta})} \langle \boldsymbol{y}, \boldsymbol{\rho} \rangle, \text{ and} \tag{3}$$

$$r(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{P}(\boldsymbol{\rho})} \langle \boldsymbol{y}, -\boldsymbol{\theta} \rangle. \tag{4}$$

A proof is provided in §B.2. The key idea is to perform a change of variable to "absorb" the permutation in (1) and (2) into a permutahedron. From the fundamental theorem of linear programming (Dantzig et al., 1955, Theorem 6), an optimal solution of a linear program is always achieved at a vertex of the convex polytope. Thus, an optimal solution over a permutahedron is always a permutation. Interestingly, $\boldsymbol{\theta}$ appears in the constraints and $\boldsymbol{\rho}$ appears in the objective for sorting, while this is the opposite for ranking.

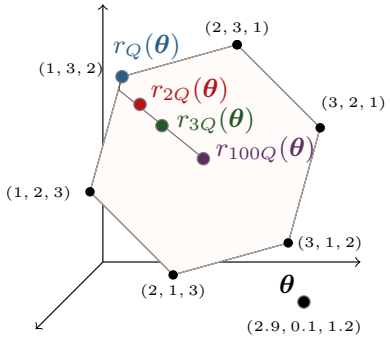

Figure 1. **Illustration of the permutahedron** $\mathcal{P}(\boldsymbol{\rho})$, whose vertices are permutations of $\boldsymbol{\rho} = (3, 2, 1)$. In this example, the ranks of $\boldsymbol{\theta} = (2.9, 0.1, 1.2)$ are $r(\boldsymbol{\theta}) = (1, 3, 2)$. In this case, our proposed soft rank $r_{\varepsilon Q}(\boldsymbol{\theta})$ with $\varepsilon = 1$ is exactly equal to $r(\boldsymbol{\theta})$. When $\varepsilon \to \infty$, $r_{\varepsilon Q}(\boldsymbol{\theta})$ converges towards the centroid of the permutahedron. The gray line indicates the regularization path of $r_{\varepsilon Q}(\boldsymbol{\theta})$ between these two regimes, when varying ε.

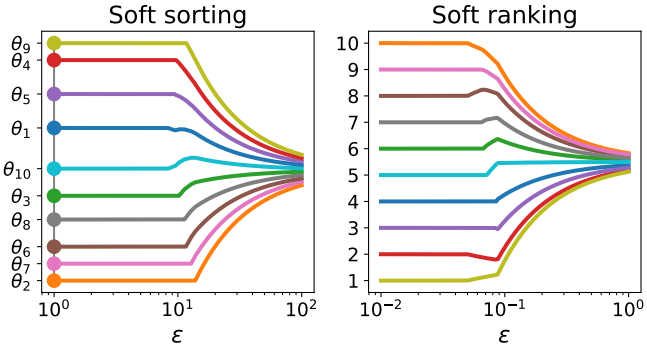

Figure 2. Illustration of the soft sorting and ranking operators, $s_{\varepsilon\Psi}(\boldsymbol{\theta})$ and $r_{\varepsilon\Psi}(\boldsymbol{\theta})$ for $\Psi = Q$; the results with $\Psi = E$ are similar. When $\varepsilon \to 0$, they converge to their "hard" counterpart. When $\varepsilon \to \infty$, they collapse into a constant, as proven in Prop.2.

Differentiability a.e. of sorting. For $s(\boldsymbol{\theta})$, the fact that $\boldsymbol{\theta}$ appears in the linear program constraints makes $s(\boldsymbol{\theta})$ piecewise linear and thus differentiable almost everywhere. When $\sigma(\boldsymbol{\theta})$ is unique at $\boldsymbol{\theta}$, $s(\boldsymbol{\theta}) = \boldsymbol{\theta}_{\sigma(\boldsymbol{\theta})}$ is differentiable at $\boldsymbol{\theta}$ and its Jacobian is the permutation matrix associated with $\sigma(\boldsymbol{\theta})$. When $\sigma(\boldsymbol{\theta})$ is not unique, we can choose any matrix in Clarkes generalized Jacobian, i.e., any convex combination of the permutation matrices associated with $\sigma(\boldsymbol{\theta})$.

Lack of differentiability of ranking. On the other hand, for $r(\boldsymbol{\theta})$, since $\boldsymbol{\theta}$ appears in the objective, a small perturbation to $\boldsymbol{\theta}$ may cause the solution of the linear program to jump to another permutation of $\boldsymbol{\rho}$. This makes $r(\boldsymbol{\theta})$ a discontinuous, piecewise constant function. This means that $r(\boldsymbol{\theta})$ has null or undefined partial derivatives, preventing its use within a neural network trained with backpropagation.

4. Differentiable sorting and ranking

As we have already motivated, our primary goal is the design of efficiently computable approximations to the sorting and ranking operators, that would smoothen the numerous kinks of the former, and provide useful derivatives for the latter. We achieve this by introducing strongly convex regularization in our linear programming formulations. This turns them into efficiently computable projection operators, which are differentiable and amenable to formal analysis.

Projection onto the permutahedron. Let $\boldsymbol{z}, \boldsymbol{w} \in \mathbb{R}^n$ and consider the linear program $\operatorname{argmax}_{\boldsymbol{\mu} \in \mathcal{P}(\boldsymbol{w})} \langle \boldsymbol{\mu}, \boldsymbol{z} \rangle$. Clearly, we can express $s(\boldsymbol{\theta})$ by setting $(\boldsymbol{z}, \boldsymbol{w}) = (\boldsymbol{\rho}, \boldsymbol{\theta})$ and $r(\boldsymbol{\theta})$ by setting $(\boldsymbol{z}, \boldsymbol{w}) = (-\boldsymbol{\theta}, \boldsymbol{\rho})$. Introducing quadratic regularization $Q(\boldsymbol{\mu}) \coloneqq \frac{1}{2}\|\boldsymbol{\mu}\|^2$ is considered by Martins & Astudillo (2016) over the unit simplex and by Niculae et al. (2018) over marginal polytopes. Similarly, adding Q to our linear program over the permutahedron gives

$$P_Q(\boldsymbol{z}, \boldsymbol{w}) \coloneqq \operatorname*{argmax}_{\boldsymbol{\mu} \in \mathcal{P}(\boldsymbol{w})} \langle \boldsymbol{z}, \boldsymbol{\mu} \rangle - Q(\boldsymbol{\mu}) = \operatorname*{argmin}_{\boldsymbol{\mu} \in \mathcal{P}(\boldsymbol{w})} \frac{1}{2}\|\boldsymbol{\mu} - \boldsymbol{z}\|^2,$$

i.e., the Euclidean projection of $\boldsymbol{z}$ onto $\mathcal{P}(\boldsymbol{w})$. We also consider entropic regularization $E(\boldsymbol{\mu}) \coloneqq \langle \boldsymbol{\mu}, \log \boldsymbol{\mu} - \mathbf{1} \rangle$, popularized in the optimal transport literature (Cuturi, 2013; Peyré & Cuturi, 2017). Subtly, we define

$$\begin{aligned} P_E(\boldsymbol{z}, \boldsymbol{w}) &\coloneqq \log \operatorname*{argmax}_{\boldsymbol{\mu} \in \mathcal{P}(e^{\boldsymbol{w}})} \langle \boldsymbol{z}, \boldsymbol{\mu} \rangle - E(\boldsymbol{\mu}) \\ &= \log \operatorname*{argmin}_{\boldsymbol{\mu} \in \mathcal{P}(e^{\boldsymbol{w}})} \operatorname{KL}(\boldsymbol{\mu}, e^{\boldsymbol{z}}), \end{aligned}$$

where $\operatorname{KL}(\boldsymbol{a}, \boldsymbol{b}) \coloneqq \sum_i a_i \log \frac{a_i}{b_i} - \sum_i a_i + \sum_i b_i$ is the Kullback-Leibler (KL) divergence between two positive measures $\boldsymbol{a} \in \mathbb{R}^n_+$ and $\boldsymbol{b} \in \mathbb{R}^n_+$. $P_E(\boldsymbol{z}, \boldsymbol{w})$ is therefore the *log* KL projection of $e^{\boldsymbol{z}}$ onto $\mathcal{P}(e^{\boldsymbol{w}})$. The purpose of $e^{\boldsymbol{w}}$ is to ensure that $\boldsymbol{\mu}$ always belongs to $\operatorname{dom}(E) = \mathbb{R}^n_+$ (since $\boldsymbol{\mu}$ is a convex combination of the permutations of $e^{\boldsymbol{w}}$) and that of the logarithm is to map $\boldsymbol{\mu}^\star$ back to $\mathbb{R}^n$.

More generally, we can use any strongly convex regularization Ψ under mild conditions. For concreteness, we focus our exposition in the main text on $\Psi \in \{Q, E\}$. We state all our propositions for these two cases and postpone a more general treatment to the appendix.

Soft operators. We now build upon these projections to define soft sorting and ranking operators. To control the regularization strength, we introduce a parameter $\varepsilon > 0$ which we multiply Ψ by (equivalently, divide $\boldsymbol{z}$ by).

For sorting, we choose $(\boldsymbol{z}, \boldsymbol{w}) = (\boldsymbol{\rho}, \boldsymbol{\theta})$ and therefore define the Ψ-regularized soft sort as

$$s_{\varepsilon\Psi}(\boldsymbol{\theta}) \coloneqq P_{\varepsilon\Psi}(\boldsymbol{\rho}, \boldsymbol{\theta}) = P_\Psi(\boldsymbol{\rho}/\varepsilon, \boldsymbol{\theta}). \tag{5}$$

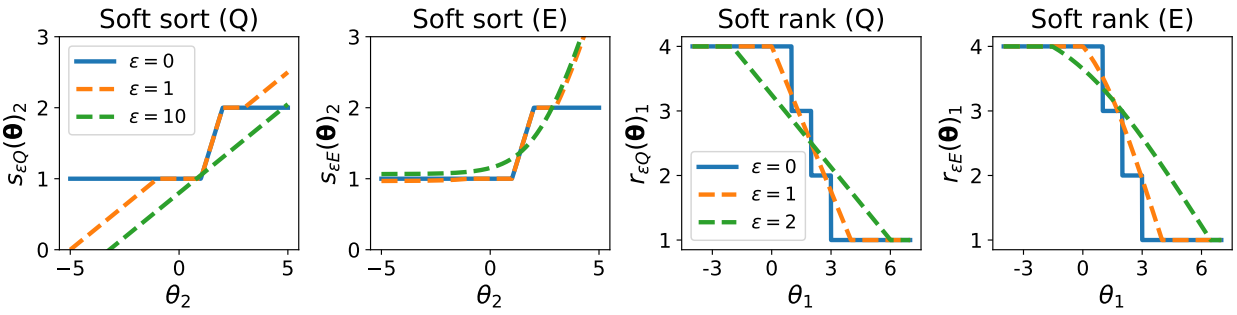

Figure 3. **Effect of the regularization parameter ε.** We take the vector $\boldsymbol{\theta} := (0, 3, 1, 2)$, vary one of its coordinates θ_i and look at how $[s_{\varepsilon\Psi}(\boldsymbol{\theta})]_i$ and $[r_{\varepsilon\Psi}(\boldsymbol{\theta})]_i$ change in response. For soft sorting with $\Psi = Q$, the function is still piecewise linear, like sorting. However, by increasing ε we reduce the number of kinks, and the function eventually converges to a mean (Proposition 2). With $\Psi = E$, the function tends to be even smoother. For soft ranking with $\Psi = Q$, the function is piecewise linear instead of piecewise constant for the "hard" ranks. With $\Psi = E$, the function again tends to be smoother though it may contain kinks.

For ranking, we choose $(\boldsymbol{z}, \boldsymbol{w}) = (-\boldsymbol{\theta}, \boldsymbol{\rho})$ and therefore define the Ψ-regularized soft rank as

$$r_{\varepsilon\Psi}(\boldsymbol{\theta}) := P_{\varepsilon\Psi}(-\boldsymbol{\theta}, \boldsymbol{\rho}) = P_{\Psi}(-\boldsymbol{\theta}/\varepsilon, \boldsymbol{\rho}). \tag{6}$$

We illustrate the behavior of both of these soft operations as we vary ε in Figures 2 and 3. As for the hard versions, the ascending-order soft sorting and ranking are obtained by negating the input as $-s_{\varepsilon\Psi}(-\boldsymbol{\theta})$ and $r_{\varepsilon\Psi}(-\boldsymbol{\theta})$, respectively.

Properties. We can further characterize these approximations. Namely, as we now formalize, they are differentiable a.e., and not only converge to the their "hard" counterparts, but also satisfy some of their properties for all ε.

Proposition 2. *Properties of $s_{\varepsilon\Psi}(\boldsymbol{\theta})$ and $r_{\varepsilon\Psi}(\boldsymbol{\theta})$*

1. **Differentiability.** *For all $\varepsilon > 0$, $s_{\varepsilon\Psi}(\boldsymbol{\theta})$ and $r_{\varepsilon\Psi}(\boldsymbol{\theta})$ are differentiable (a.e.) w.r.t. $\boldsymbol{\theta}$.*
2. **Order preservation.** *Let $\boldsymbol{s} := s_{\varepsilon\Psi}(\boldsymbol{\theta})$, $\boldsymbol{r} := r_{\varepsilon\Psi}(\boldsymbol{\theta})$ and $\sigma := \sigma(\boldsymbol{\theta})$. For all $\boldsymbol{\theta} \in \mathbb{R}^n$ and $0 < \varepsilon < \infty$, we have $s_1 \geq s_2 \geq \cdots \geq s_n$ and $r_{\sigma_1} \leq r_{\sigma_2} \leq \cdots \leq r_{\sigma_n}$.*
3. **Asymptotics.** *For all $\boldsymbol{\theta} \in \mathbb{R}^n$ without ties:*

$$s_{\varepsilon\Psi}(\boldsymbol{\theta}) \xrightarrow[\varepsilon\to 0]{} s(\boldsymbol{\theta}) \qquad r_{\varepsilon\Psi}(\boldsymbol{\theta}) \xrightarrow[\varepsilon\to 0]{} r(\boldsymbol{\theta})$$
$$\xrightarrow[\varepsilon\to\infty]{} f_\Psi(\boldsymbol{\theta})\mathbf{1} \qquad \xrightarrow[\varepsilon\to\infty]{} f_\Psi(\boldsymbol{\rho})\mathbf{1},$$

where $f_Q(\boldsymbol{u}) := mean(\boldsymbol{u})$, $f_E(\boldsymbol{u}) := \log f_Q(\boldsymbol{u})$.

The last property describes the behavior as $\varepsilon \to 0$ and $\varepsilon \to \infty$. Together with the proof of Proposition 2, we include in §B.3 a slightly stronger result. Namely, we derive an explicit value of ε below which our operators are exactly equal to their hard counterpart, and a value of ε above which our operators can be computed in closed form.

Convexification effect. Proposition 2 shows that $[s_{\varepsilon\Psi}(\boldsymbol{\theta})]_i$ and $[r_{\varepsilon\Psi}(\boldsymbol{\theta})]_i$ for all $i \in [n]$ converge to convex functions of $\boldsymbol{\theta}$ as $\varepsilon \to \infty$. This suggests that larger ε make the objective function increasingly easy to optimize (at the cost of departing from "hard" sorting or ranking). This behavior is also visible in Figure 3, where $[s_{\varepsilon Q}(\boldsymbol{\theta})]_2$ converges towards the mean f_Q, depicted by a straight line.

On tuning ε (or not). The parameter $\varepsilon > 0$ controls the trade-off between approximation of the original operator and "smoothness". When the model $g(\boldsymbol{x})$ producing the scores or "logits" $\boldsymbol{\theta}$ to be sorted/ranked is a homogeneous function, from (5) and (6), ε can be absorbed into the model. In our label ranking experiment, we find that indeed tuning ε is not necessary to achieve excellent accuracy. On the other hand, for top-k classification, we find that applying a logistic map to squash $\boldsymbol{\theta}$ to $[0, 1]^n$ and tuning ε is important, confirming the empirical finding of Cuturi et al. (2019).

Relation to linear assignment formulation. When using uniform weights on the inputs, the differentiable operators of Cuturi et al. (2019) are based on viewing sorting and ranking as linear assignment over the Birkhoff polytope $\mathcal{B} \subset \mathbb{R}^{n\times n}$, the convex hull of permutation matrices. To relate to their method, note that using the change of variable $\boldsymbol{y} = \boldsymbol{P}\boldsymbol{\rho}$ and $\mathcal{P}(\boldsymbol{\rho}) = \mathcal{B}\boldsymbol{\rho}$, we can rewrite (4) as $r(\boldsymbol{\theta}) = \boldsymbol{P}(\boldsymbol{\theta})\boldsymbol{\rho}$, where

$$\boldsymbol{P}(\boldsymbol{\theta}) := \operatorname*{argmax}_{\boldsymbol{P}\in\mathcal{B}} \langle \boldsymbol{P}\boldsymbol{\rho}, -\boldsymbol{\theta}\rangle.$$

Let $D(\boldsymbol{a}, \boldsymbol{b}) \in \mathbb{R}^{n\times n}$ be a distance matrix. Simple calculations show that if $[D(\boldsymbol{a}, \boldsymbol{b})]_{i,j} := \frac{1}{2}(a_i - b_j)^2$, then

$$\boldsymbol{P}(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{P}\in\mathcal{B}} \langle \boldsymbol{P}, D(-\boldsymbol{\theta}, \boldsymbol{\rho})\rangle.$$

Similarly, we can rewrite (3) as $s(\boldsymbol{\theta}) = \boldsymbol{P}(\boldsymbol{\theta})^\top\boldsymbol{\theta}$. To obtain a differentiable operator, Cuturi et al. (2019) (see also (Adams & Zemel, 2011)) propose to replace the permutation matrix $\boldsymbol{P}(\boldsymbol{\theta})$ by a doubly stochastic matrix $\boldsymbol{P}_{\varepsilon E}(\boldsymbol{\theta}) :=$

$\operatorname{argmin}_{\boldsymbol{P}\in\mathcal{B}} \langle \boldsymbol{P}, D(-\boldsymbol{\theta}, \boldsymbol{\rho})\rangle + \varepsilon E(\boldsymbol{P})$, which is computed approximately in $O(Tn^2)$ using Sinkhorn (1967). In comparison, our approach is based on regularizing $\boldsymbol{y} = \boldsymbol{P}\boldsymbol{\rho}$ with $\Psi \in \{Q, E\}$ directly, the key to achieve $O(n \log n)$ time and $O(n)$ space complexity, as we now show.

5. Fast computation and differentiation

As shown in the previous section, computing our soft sorting and ranking operators boils down to projecting onto a permutahedron. Our key contribution in this section is the derivation of an $O(n \log n)$ forward pass and an $O(n)$ backward pass (multiplication with the Jacobian) for these projections. Beyond soft sorting and ranking, this is an important sensitivity analysis question in its own right.

Reduction to isotonic optimization. We now show how to reduce the projections to isotonic optimization, i.e., with simple chain constraints, which is the key to fast computation and differentiation. We will w.l.o.g. assume that $\boldsymbol{w}$ is sorted in descending order (if not the case, we sort it first).

Proposition 3. *Reduction to isotonic optimization*

For all $\boldsymbol{z} \in \mathbb{R}^n$ and sorted $\boldsymbol{w} \in \mathbb{R}^n$ we have

$$P_\Psi(\boldsymbol{z}, \boldsymbol{w}) = \boldsymbol{z} - \boldsymbol{v}_\Psi(\boldsymbol{z}_{\sigma(\boldsymbol{z})}, \boldsymbol{w})_{\sigma^{-1}(\boldsymbol{z})}$$

where

$$\boldsymbol{v}_Q(\boldsymbol{s}, \boldsymbol{w}) \coloneqq \operatorname*{argmin}_{v_1 \geq \cdots \geq v_n} \frac{1}{2}\|\boldsymbol{v} - (\boldsymbol{s} - \boldsymbol{w})\|^2, \text{ and}$$
$$\boldsymbol{v}_E(\boldsymbol{s}, \boldsymbol{w}) \coloneqq \operatorname*{argmin}_{v_1 \geq \cdots \geq v_n} \langle e^{\boldsymbol{s}-\boldsymbol{v}}, \boldsymbol{1}\rangle + \langle e^{\boldsymbol{w}}, \boldsymbol{v}\rangle.$$

The function $\boldsymbol{v}_Q$ is classically known as isotonic regression. The fact that it can be used to solve the Euclidean projection onto $\mathcal{P}(\boldsymbol{w})$ has been noted several times (Negrinho & Martins, 2014; Zeng & Figueiredo, 2015). The reduction of Bregman projections, which we use here, to isotonic optimization was shown by Lim & Wright (2016). Unlike that study, we use the KL projection of $e^{\boldsymbol{z}}$ onto $\mathcal{P}(e^{\boldsymbol{w}})$, and not of $\boldsymbol{z}$ onto $\mathcal{P}(\boldsymbol{w})$, which simplifies many expressions. We include in §B.4 a simple unified proof of Proposition 3 based on Fenchel duality and tools from submodular optimization. We also discuss an interpretation of adding regularization to the primal linear program as relaxing the equality constraints of the dual linear program in §B.5.

Computation. As shown by Best et al. (2000), the classical pool adjacent violators (PAV) algorithm for isotonic regression can be extended to minimize any per-coordinate decomposable convex function $f(\boldsymbol{v}) = \sum_{i=1}^n f_i(v_i)$ subject to monotonicity constraints, which is exactly the form of the problems in Proposition 3. The algorithm repeatedly splits the coordinates into a set of contiguous blocks $\mathcal{B}_1, \ldots, \mathcal{B}_m$ that partition $[n]$ (their union is $[n]$ and $\max \mathcal{B}_j + 1 = \min \mathcal{B}_{j+1}$). It only requires access to an oracle that solves for each block $\mathcal{B}_j$ the sub-problem $\gamma(\mathcal{B}_j) = \operatorname{argmin}_{\gamma\in\mathbb{R}} \sum_{i\in\mathcal{B}_j} f_i(\gamma)$, and runs in **linear** time. Further, the solution has a clean block-wise constant structure, namely it is equal to $\gamma(\mathcal{B}_j)$ within block $\mathcal{B}_j$. Fortunately, in our case, as shown in §B.6, the function γ can be analytically computed, as

$$\gamma_Q(\mathcal{B}_j; \boldsymbol{s}, \boldsymbol{w}) \coloneqq \frac{1}{|\mathcal{B}_j|} \sum_{i\in\mathcal{B}_j} s_i - w_i, \text{ and} \tag{7}$$

$$\gamma_E(\mathcal{B}_j; \boldsymbol{s}, \boldsymbol{w}) \coloneqq \log \sum_{i\in\mathcal{B}_j} e^{s_i} - \log \sum_{i\in\mathcal{B}_j} e^{w_i}. \tag{8}$$

Hence, PAV returns an **exact** solution of both $\boldsymbol{v}_Q(\boldsymbol{s}, \boldsymbol{w})$ and $\boldsymbol{v}_E(\boldsymbol{s}, \boldsymbol{w})$ in $O(n)$ time (Best et al., 2000). This means that we do not need to choose a number of iterations or a level of precision, unlike with Sinkhorn. Since computing $P_Q(\boldsymbol{z}, \boldsymbol{w})$ and $P_E(\boldsymbol{z}, \boldsymbol{w})$ requires obtaining $\boldsymbol{s} = \boldsymbol{z}_{\sigma(\boldsymbol{\theta})}$ beforehand, the total computational complexity is $O(n \log n)$.

Differentiating isotonic optimization. The block-wise structure of the solution also makes its derivatives easy to analyze, despite the fact that we are differentiating the *solution* of an optimization problem. Since the coordinates of the solution in block $\mathcal{B}_j$ are all equal to $\gamma(\mathcal{B}_j)$, which in turn depends only on a subset of the parameters, the Jacobian has a simple block-wise form, which we now formalize.

Lemma 2. *Jacobian of isotonic optimization*

Let $\mathcal{B}_1, \ldots, \mathcal{B}_m$ be the ordered partition of $[n]$ induced by $\boldsymbol{v}_\Psi(\boldsymbol{s}, \boldsymbol{w})$ from Proposition 3. Then,

$$\frac{\partial \boldsymbol{v}_\Psi(\boldsymbol{s}, \boldsymbol{w})}{\partial \boldsymbol{s}} = \begin{bmatrix} \mathbf{B}_1^\Psi & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_m^\Psi \end{bmatrix} \in \mathbb{R}^{n\times n},$$

where $\mathbf{B}_j^\Psi \coloneqq \partial\gamma_\Psi(\mathcal{B}_j; \boldsymbol{s}, \boldsymbol{w})/\partial \boldsymbol{s} \in \mathbb{R}^{|\mathcal{B}_j|\times|\mathcal{B}_j|}$.

A proof is given in §B.7. The Jacobians w.r.t. $\boldsymbol{w}$ are entirely similar, thanks to the symmetry of (7) and (8).

In the quadratic regularization case, it was already derived by Djolonga & Krause (2017) that $\mathbf{B}_j^Q \coloneqq \mathbf{1}/|\mathcal{B}_j|$. The multiplication with the Jacobian, $\boldsymbol{\nu} \coloneqq \frac{\partial \boldsymbol{v}_Q(\boldsymbol{s},\boldsymbol{w})}{\partial \boldsymbol{s}}\boldsymbol{u}$ for some vector $\boldsymbol{u}$, can be computed as $\boldsymbol{\nu} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_m)$, where $\boldsymbol{\nu}_j = \operatorname{mean}(\boldsymbol{u}_{\mathcal{B}_j})\mathbf{1} \in \mathbb{R}^{|\mathcal{B}_j|}$. In the entropic regularization case, novel to our knowledge, we have $\mathbf{B}_j^E = \mathbf{1} \otimes \operatorname{softmax}(\boldsymbol{s}_{\mathcal{B}_j})$. Note that $\mathbf{B}_j^E$ is column-wise constant, so that the multiplication with the Jacobian $\boldsymbol{\nu} \coloneqq \frac{\partial \boldsymbol{v}_E(\boldsymbol{s},\boldsymbol{w})}{\partial \boldsymbol{s}}\boldsymbol{u}$, can be computed as $\boldsymbol{\nu}_j = \langle \operatorname{softmax}(\boldsymbol{s}_{\mathcal{B}_j}), \boldsymbol{u}_{\mathcal{B}_j}\rangle \mathbf{1} \in \mathbb{R}^{|\mathcal{B}_j|}$. In both cases, the multiplication with the Jacobian therefore takes $O(n)$ time.

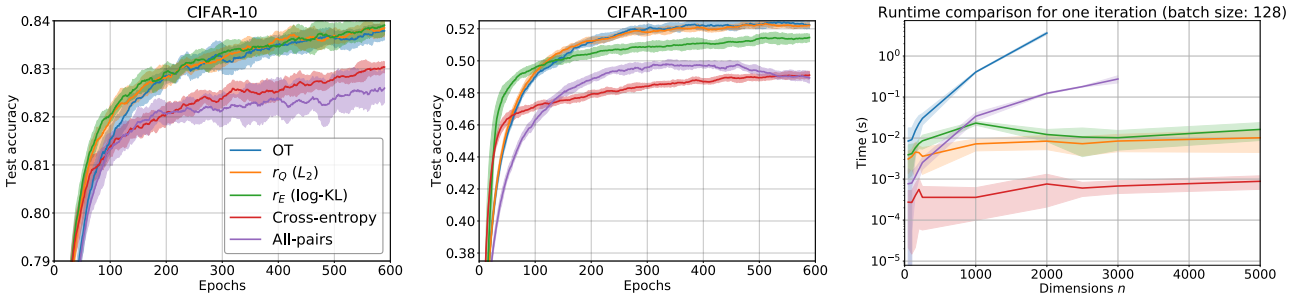

Figure 4. **Left, center:** Accuracy comparison on CIFAR-10, CIFAR-100 ($n = 10$, $n = 100$). **Right:** Runtime comparison for one batch computation with backpropagation disabled. OT and All-pairs go out-of-memory starting from $n = 2000$ and $n = 3000$, respectively. With backpropagation enabled, the runtimes are similar but OT and All-pairs go out-of-memory at $n = 1000$ and $n = 2500$, respectively.

There are interesting differences between the two forms of regularization. For quadratic regularization, the Jacobian only depends on the partition $\mathcal{B}_1, \dots, \mathcal{B}_m$ (not on $\boldsymbol{s}$) and the blocks have constant value. For entropic regularization, the Jacobian does depend on $\boldsymbol{s}$ and the blocks are constant column by column. Both formulations are averaging the incoming gradients, one uniformly and the other weighted.

Differentiating the projections. We now combine Proposition 3 with Lemma 2 to characterize the Jacobians of the projections onto the permutahedron and show how to multiply arbitrary vectors with them in linear time.

Proposition 4. *Jacobian of the projections*

Let $P_\Psi(\boldsymbol{z}, \boldsymbol{w})$ be defined in Proposition 3. Then,

$$\frac{\partial P_\Psi(\boldsymbol{z}, \boldsymbol{w})}{\partial \boldsymbol{z}} = \boldsymbol{J}_\Psi(\boldsymbol{z}_{\sigma(\boldsymbol{z})}, \boldsymbol{w})_{\sigma^{-1}(\boldsymbol{z})},$$

where $\boldsymbol{J}_\pi$ is the matrix obtained by permuting the rows and columns of $\boldsymbol{J}$ according to π, and where

$$\boldsymbol{J}_\Psi(\boldsymbol{s}, \boldsymbol{w}) := \boldsymbol{I} - \frac{\partial \boldsymbol{v}_\Psi(\boldsymbol{s}, \boldsymbol{w})}{\partial \boldsymbol{s}}.$$

Again, the Jacobian w.r.t. $\boldsymbol{w}$ is entirely symmetric. Unlike the Jacobian of isotonic optimization, the Jacobian of the projection is not block diagonal, as we need to permute its rows and columns. We can nonetheless multiply with it in linear time by using the simple identity $(\boldsymbol{J}_\pi)\boldsymbol{z} = (\boldsymbol{J}\boldsymbol{z}_{\pi^{-1}})_\pi$, which allows us to reuse the $O(n)$ multiplication with the Jacobian of isotonic optimization.

Differentiating $s_{\varepsilon\Psi}$ and $r_{\varepsilon\Psi}$. With the Jacobian of $P_\Psi(\boldsymbol{z}, \boldsymbol{w})$ w.r.t. $\boldsymbol{z}$ and $\boldsymbol{w}$ at hand, differentiating $s_{\varepsilon\Psi}$ and $r_{\varepsilon\Psi}$ boils down to a mere application of the chain rule to (5) and (6). To summarize, we can multiply with the Jacobians of our soft operators in $O(n)$ time and space.

6. Experiments

We present in this section our empirical findings. We will release in the near future JAX, PyTorch and Tensorflow implementations of our soft operators building upon a highly-optimized C++ implementation of the PAV algorithm.

6.1. Top-k classification loss function

Experimental setup. To demonstrate the effectiveness of our proposed soft rank operators as a drop-in replacement for exisiting ones, we reproduce the top-k classification experiment of Cuturi et al. (2019). The authors propose a loss for top-k classification between a ground truth class $y \in [n]$ and a vector of soft ranks $\boldsymbol{r} \in \mathbb{R}^n$, which is higher if the predicted soft ranks correctly place y in the top-k elements. We compare the following soft operators

- OT (Cuturi et al., 2019): The $O(Tn^2)$ optimal transport formulation discussed in §4.
- All-pairs (Qin et al., 2010): observing that $[r(\boldsymbol{\theta})]_i$ can be written as $\sum_{j=1}^n \mathbb{1}[\theta_i > \theta_j]$, one can obtain soft ranks in $O(n^2)$ by replacing the indicator function with a sigmoid.
- Proposed: our $O(n \log n)$ soft ranks r_Q and r_E. Although not used in this experiment, for top-k ranking, the complexity can be reduced to $O(n \log k)$ by computing P_Ψ using the algorithm of Lim & Wright (2016).

We use the CIFAR-10 and CIFAR-100 datasets, with $n = 10$ and $n = 100$ classes, respectively. Following Cuturi et al. (2019), we use a vanilla CNN (4 Conv2D with 2 max-pooling layers, ReLU activation, 2 fully connected layers with batch norm on each), the ADAM optimizer (Kingma & Ba, 2014) with a constant step size of 10^{-4}, and set $k = 1$. Similarly to Cuturi et al. (2019), we found that squashing the scores $\boldsymbol{\theta}$ to $[0, 1]^n$ with a logistic map was beneficial.

Results. Our empirical results, averaged over 12 runs, are shown in Figure 4 (left, center). On both CIFAR-10 and CIFAR-100, our soft rank formulations achieve comparable accuracy to the OT formulation, though significantly faster,

as we elaborate below. Confirming the results of Cuturi et al. (2019), we found that the soft top-k loss slightly outperforms the classical cross-entropy (logistic) loss for these two datasets. However, we did not find that the All-pairs formulation could outperform the cross-entropy loss.

The training times for 600 epochs on CIFAR-100 were 29 hours (OT), 21 hours (r_Q), 23 hours (r_E) and 16 hours (All-pairs). Training times on CIFAR-10 were similar. While our soft operators are several hours faster than OT, they are slower than All-pairs, despite its $O(n^2)$ complexity. This is due the fact that, with $n = 100$, All-pairs is very efficient on GPUs, while our PAV implementation runs on CPU.

6.2. Runtime comparison: effect of input dimension

To measure the impact of the dimensionality n on the runtime of each method, we designed the following experiment.

Experimental setup. We generate score vectors $\boldsymbol{\theta} \in \mathbb{R}^n$ randomly according to $\mathcal{N}(0, 1)$, for n ranging from 100 up to 5000. For fair comparison with GPU implementations (OT, All-pairs, Cross-entropy), we create a batch of 128 such vectors and we compare the time to compute soft ranking operators on this batch. We run this experiment on top of TensorFlow (Abadi et al., 2016) on a six core Intel Xeon W-2135 with 64 GBs of RAM and a GeForce GTX 1080 Ti.

Results. Run times for one batch computation with backpropagation disabled are shown in Figure 4 (Right). While their runtime is reasonable in small dimension, OT and All-pairs scale quadratically with respect to the dimensionality n (note the log scale on the y-axis). Although slower than a softmax, our formulations scale well, with the dimensionality n having negligible impact on the runtime. OT and All-pairs go out-of-memory starting from $n = 2000$ and $n = 3000$, respectively. With backpropagation enabled, they go out-of-memory at $n = 1000$ and $n = 2500$, due to the need for recording the computational graph. This shows that the lack of memory available on GPUs is problematic for these methods. In contrast, our approaches only require $O(n)$ memory and comes with the theoretical Jacobian (they do not rely on differentiating through iterates). They therefore suffer from no such issues.

6.3. Label ranking via soft Spearman's rank correlation coefficient

We now consider the label ranking setting where supervision is given as full rankings (e.g., $2 \succ 1 \succ 3 \succ 4$) rather than as label relevance scores. The goal is therefore to learn to predict permutations, i.e., a function $f_{\boldsymbol{w}} \colon \mathcal{X} \to \Sigma$. A classical metric between ranks is Spearman's rank correlation coefficient, defined as the Pearson correlation coefficient between the ranks. Maximizing this coefficient is equivalent to minimizing the squared loss between ranks. A naive idea would be therefore to use as loss $\frac{1}{2}\|\boldsymbol{r} - r(\boldsymbol{\theta})\|^2$, where $\boldsymbol{\theta} = g_{\boldsymbol{w}}(\boldsymbol{x})$. This is unfortunately a discontinuous function of $\boldsymbol{\theta}$. We therefore propose to rather use $\frac{1}{2}\|\boldsymbol{r} - r_\Psi(\boldsymbol{\theta})\|^2$, hence the name differentiable Spearman's rank correlation coefficient. At test time, we replace r_Ψ with r, which is justified by the order-preservation property (Proposition 2).

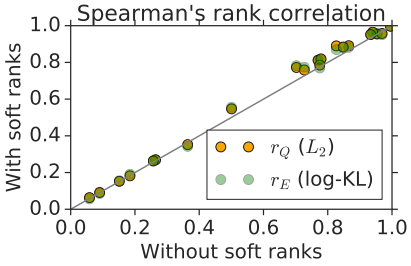

Figure 5. **Label ranking accuracy** with and without soft rank layer. Each point above the line represents a dataset where our soft rank layer improves Spearman's rank correlation coefficient.

Experimental setup. We consider the 21 datasets from (Hüllermeier et al., 2008; Cheng et al., 2009), which has both semi-synthetic data obtained from classification problems, and real biological measurements. Following (Korba et al., 2018), we average over two 10-fold validation runs, in each of which we train on 90% and evaluate on 10% of the data. Within each repetition, we run an internal 5-fold cross-validation to grid-search for the best parameters. We consider linear models of the form $g_{\boldsymbol{w}}(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + w_0$, and for ablation study we drop the soft ranking layer r_Ψ.

Results. Due to the large number of datasets, we choose to present a summary of the results in Figure 5. We postpone detailed results to the appendix (Table 1). Out of 21 datasets, introducing a soft rank layer with $\Psi = Q$ works better on 15 datasets, similarly on 4 and worse on 2 datasets. We can thus conclude that even for such simple model, introducing our layer is beneficial, and even achieving state of the art results on some of the datasets (full details in the appendix).

6.4. Robust regression via soft least trimmed squares

We explore in this section the application of our soft sorting operator $s_{\varepsilon\Psi}$ to robust regression. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y_1, \ldots, y_n \in \mathcal{Y} \subseteq \mathbb{R}$ be a training set of input-output pairs. Our goal is to learn a model $g_{\boldsymbol{w}} \colon \mathbb{R}^d \to \mathbb{R}$ that predicts outputs from inputs, where $\boldsymbol{w}$ are model parameters. We focus on $g_{\boldsymbol{w}}(\boldsymbol{x}) \coloneqq \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ for simplicity. We further assume that a certain proportion of examples are outliers including some label noise, which makes the task of robustly estimating $g_{\boldsymbol{w}}$ particularly challenging.

The classical ridge regression can be cast as

$$\min_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} \ell_i(\boldsymbol{w}) + \frac{1}{2\varepsilon}\|\boldsymbol{w}\|^2, \tag{9}$$

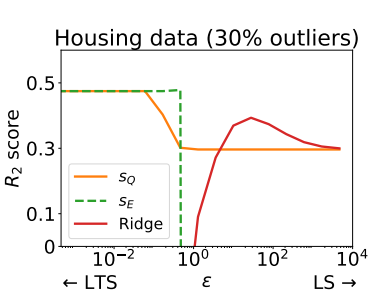

Figure 6. Empirical validation of interpolation between LTS and LS.

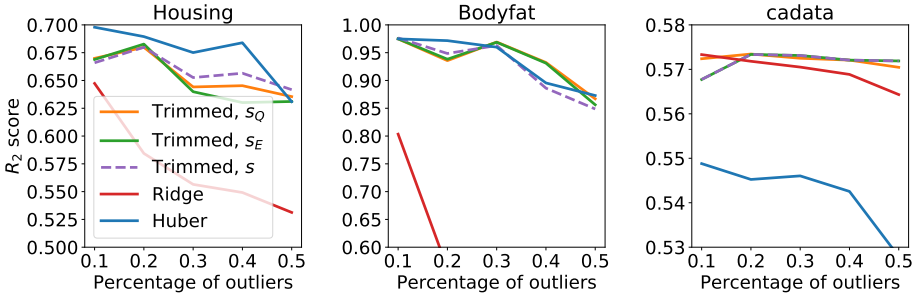

Figure 7. R_2 score (higher is better) averaged over 10 train-time splits for increasing percentage of outliers. Hyper-parameters are tuned by 5-fold cross-validation.

where $\ell_i(\boldsymbol{w}) := \frac{1}{2}(y_i - g_{\boldsymbol{w}}(\boldsymbol{x}_i))^2$. In order to be robust to label noise, we propose instead to sort the losses (from larger to smaller) and to ignore the first k ones. Introducing our soft sorting operator, this can be formulated as

$$\min_{\boldsymbol{w}} \frac{1}{n-k} \sum_{i=k+1}^{n} \ell_i^{\varepsilon}(\boldsymbol{w}) \tag{10}$$

where $\ell_i^{\varepsilon}(\boldsymbol{w}) := [s_{\varepsilon\Psi}(\boldsymbol{\ell}(\boldsymbol{w}))]_i$ is the i^{th} loss in the soft sort, and $\boldsymbol{\ell}(\boldsymbol{w}) \in \mathbb{R}^n$ is the loss vector that gathers $\ell_i(\boldsymbol{w})$ for each $i \in [n]$. Solving (10) with existing $O(n^2)$ soft sorting operators could be particularly computationally prohibitive, since here n is the number of training samples.

When $\varepsilon \to 0$, we have $s_{\varepsilon\Psi}(\boldsymbol{\ell}(\boldsymbol{w})) \to s(\boldsymbol{\ell}(\boldsymbol{w}))$ and (10) is known as least trimmed squares (LTS) (Rousseeuw, 1984; Rousseeuw & Leroy, 2005). When $\varepsilon \to \infty$, we have, from Proposition 2, $s_{\varepsilon\Psi}(\boldsymbol{\ell}(\boldsymbol{w})) \to \text{mean}(\boldsymbol{\ell}(\boldsymbol{w}))\mathbf{1}$ and therefore both (9) and (10) converge to the least squares (LS) objective, $\min_{\boldsymbol{w}} \frac{1}{n}\sum_{i=1}^{n} \ell_i(\boldsymbol{w})$. To summarize, our proposed objective (10), dubbed soft least trimmed squares, **interpolates** between least trimmed squares ($\epsilon \to 0$) and least squares ($\epsilon \to \infty$), as also confirmed empirically in Figure 6.

Experimental setup. To empirically validate our proposal, we compare cross-validated results for increasing percentage of outliers of the following methods:

- Least trimmed squares, with truncation parameter k,
- Soft least trimmed squares (10), with truncation parameter k and regularization parameter ε,
- Ridge regression (9), with regularization parameter ε,
- Huber loss (Huber, 1964) with regularization parameter ε and threshold parameter τ, as implemented in scikit-learn (Pedregosa et al., 2011).

We consider datasets from the LIBSVM archive (Fan & Lin, 2011). We hold out 20% of the data as test set and use the rest as training set. We artificially create outliers, by adding noise to a certain percentage of the training labels, using $y_i \leftarrow y_i + e$, where $e \sim \mathcal{N}(0, 5 \times \text{std}(\boldsymbol{y}))$. We do not add noise to the test set. For all methods, we use L-BFGS (Liu & Nocedal, 1989), with a maximum of 300 iterations. For hyper-parameter optimization, we use 5-fold cross-validation. We choose k from $\{\lceil 0.1n \rceil, \lceil 0.2n \rceil, \ldots, \lceil 0.5n \rceil\}$, ε from 10 log-spaced values between 10^{-3} and 10^4, and τ from 5 linearly spaced values between 1.3 and 2. We repeat this procedure 10 times with a different train-test split, and report the averaged R_2 scores (a.k.a. coefficient of determination).

Results. The averaged R_2 scores (higher is better) are shown in Figure 7. On all datasets, the accuracy of ridge regression deteriorated significantly with increasing number of outliers. Least trimmed squares (hard or soft) performed slightly worse than the Huber loss on housing, comparably on bodyfat and much better on cadata. We found that hard least trimmed squares (i.e., $\varepsilon = 0$) worked well on all datasets, showing that regularization is less important for sorting operators (which are piecewise linear) than for ranking operators (which are piecewise constant). Nevertheless, regularization appeared useful in some cases. For instance, on cadata, the cross-validation procedure picked $\varepsilon > 1000$ when the percentage of outliers is less than 20%, and $\varepsilon < 10^{-3}$ when the percentage of outliers is larger than 20%. This is confirmed visually on Figure 7, where the soft sort with $\Psi = Q$ works slightly better than the hard sort with few outliers, then performs comparably with more outliers. The interpolation effect enabled by ε therefore allows some adaptivity to the (unknown) percentage of outliers.

7. Conclusion

Building upon projections onto permutahedra, we constructed differentiable sorting and ranking operators. We derived **exact** $O(n \log n)$ computation and $O(n)$ differentiation of these operators, a key technical contribution of this paper. We demonstrated that our operators can be used as a drop-in replacement for existing $O(n^2)$ ones, with an order-of-magnitude speed-up. We also showcased two applications enabled by our soft operators: label ranking with differentiable Spearman's rank correlation coefficient and robust regression via soft least trimmed squares.

Acknowledgements

We are grateful to Marco Cuturi and Jean-Philippe Vert for useful discussions, and to Carlos Riquelme for comments on a draft of this paper.

References

Abadi, M. et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation*, pp. 265–283, 2016.

Adams, R. P. and Zemel, R. S. Ranking via sinkhorn propagation. *arXiv e-prints*, 2011.

Ailon, N., Hatano, K., and Takimoto, E. Bandit online optimization over the permutahedron. *Theoretical Computer Science*, 650:92–108, 2016.

Bach, F. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends® in Machine Learning*, 6(2-3):145–373, 2013.

Best, M. J., Chakravarti, N., and Ubhaya, V. A. Minimizing separable convex functions subject to simple chain constraints. *SIAM Journal on Optimization*, 10(3):658–672, 2000.

Blondel, M. Structured prediction with projection oracles. In *Proc. of NeurIPS*, 2019.

Blondel, M., Martins, A. F., and Niculae, V. Learning classifiers with Fenchel-Young losses: Generalized entropies, margins, and algorithms. In *Proc. of AISTATS*, 2019.

Blondel, M., Martins, A. F., and Niculae, V. Learning with Fenchel-Young losses. *arXiv preprint arXiv:1901.02324*, 2019.

Bowman, V. Permutation polyhedra. *SIAM Journal on Applied Mathematics*, 22(4):580–589, 1972.

Chapelle, O. and Wu, M. Gradient descent optimization of smoothed information retrieval metrics. *Information retrieval*, 13(3):216–235, 2010.

Cheng, W., Hühn, J., and Hüllermeier, E. Decision tree and instance-based learning for label ranking. In *International Conference on Machine Learning (ICML-09)*, 2009.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proc. of NeurIPS*, 2013.

Cuturi, M., Teboul, O., and Vert, J.-P. Differentiable ranking and sorting using optimal transport. In *Proc. of NeurIPS*, 2019.

Dantzig, G. B., Orden, A., and Wolfe, P. The generalized simplex method for minimizing a linear form under linear inequality restraints. *Pacific Journal of Mathematics*, 5 (2):183–195, 1955.

David, H. A. and Nagaraja, H. N. Order statistics. *Encyclopedia of Statistical Sciences*, 2004.

Djolonga, J. and Krause, A. Differentiable learning of submodular models. In *Proc. of NeurIPS*, pp. 1013–1023, 2017.

Edmonds, J. Submodular functions, matroids, and certain polyhedra. In *Combinatorial structures and their applications*, pp. 69–87, 1970.

Fan, R.-E. and Lin, C.-J. LIBSVM datasets, 2011.

Grover, A., Wang, E., Zweig, A., and Ermon, S. Stochastic optimization of sorting networks via continuous relaxations. *arXiv preprint arXiv:1903.08850*, 2019.

Huber, P. J. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Hüllermeier, E., Fürnkranz, J., Cheng, W., and Brinker, K. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172, 2008.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Korba, A., Garcia, A., and d'Alché Buc, F. A structured prediction approach for label ranking. In *Proc. of NeurIPS*, 2018.

Lapin, M., Hein, M., and Schiele, B. Loss functions for top-k error: Analysis and insights. In *Proc. of CVPR*, 2016.

LeCun, Y. Deep Learning est mort. Vive Differentiable Programming!, 2018.

Lim, C. H. and Wright, S. J. Efficient bregman projections onto the permutahedron and related polytopes. In *Proc. of AISTATS*, pp. 1205–1213, 2016.

Liu, D. C. and Nocedal, J. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

Martins, A. F. and Astudillo, R. F. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proc. of ICML*, 2016.

Negrinho, R. and Martins, A. Orbit regularization. In *Proc. of NeurIPS*, 2014.

Niculae, V., Martins, A. F., Blondel, M., and Cardie, C. SparseMAP: Differentiable sparse structured inference. In *Proc. of ICML*, 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Peyré, G. and Cuturi, M. *Computational Optimal Transport*. Foundations and Trends in Machine Learning, 2017.

Qin, T., Liu, T.-Y., and Li, H. A general approximation framework for direct optimization of information retrieval measures. *Information retrieval*, 13(4):375–397, 2010.

Rousseeuw, P. J. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

Rousseeuw, P. J. and Leroy, A. M. *Robust regression and outlier detection*, volume 589. John wiley & sons, 2005.

Sinkhorn, R. and Knopp, P. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Spearman, C. The proof and measurement of association between two things. *American Journal of Psychology*, 1904.

Suehiro, D., Hatano, K., Kijima, S., Takimoto, E., and Nagano, K. Online prediction under submodular constraints. In *International Conference on Algorithmic Learning Theory*, pp. 260–274, 2012.

Taylor, M., Guiver, J., Robertson, S., and Minka, T. Softrank: optimizing non-smooth rank metrics. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 77–86, 2008.

Yasutake, S., Hatano, K., Kijima, S., Takimoto, E., and Takeda, M. Online linear optimization over permutations. In *International Symposium on Algorithms and Computation*, pp. 534–543. Springer, 2011.

Zeng, X. and Figueiredo, M. A. The ordered weighted ℓ_1 norm: Atomic formulation and conditional gradient algorithm. In *Proc. of SPARS*, 2015.

Ziegler, G. M. *Lectures on polytopes*, volume 152. Springer Science & Business Media, 2012.

Appendix

A. Additional empirical results

We include in this section the detailed label ranking results on the same 21 datasets as considered by Hüllermeier et al. (2008) as well as Cheng et al. (2009).

For entropic regularization E, in addition to r_E, we also consider an alternative formulation. Since $\boldsymbol{\rho}$ is already strictly positive, instead of using the log-projection onto $\mathcal{P}(e^{\boldsymbol{\rho}})$, we can directly use the projection onto $\mathcal{P}(\boldsymbol{\rho})$. In our notation, this can be written as $\tilde{r}_{\varepsilon E}(\boldsymbol{\theta}) = \tilde{r}_E(\boldsymbol{\theta}/\varepsilon)$, where

$$\tilde{r}_E(\boldsymbol{\theta}) := \operatorname*{argmin}_{\boldsymbol{\mu} \in \mathcal{P}(\boldsymbol{\rho})} \mathrm{KL}(\boldsymbol{\mu}, e^{-\boldsymbol{\theta}}) = e^{P_E(-\boldsymbol{\theta}, \log \boldsymbol{\rho})}.$$

Spearman's rank correlation coefficient for each method, averaged over 5 runs, is shown in the table below.

Dataset	r_Q (L_2)	r_E (log-KL)	$\tilde{r}_E$ (KL)	No projection
fried	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
wine	0.96 ± 0.03 (-0.01)	0.95 ± 0.04 (-0.02)	0.96 ± 0.03 (-0.01)	0.97 ± 0.02
authorship	0.96 ± 0.01	0.95 ± 0.01	0.95 ± 0.01	0.95 ± 0.01
pendigits	0.96 ± 0.00 (+0.02)	0.96 ± 0.00 (+0.02)	0.96 ± 0.00 (+0.02)	0.94 ± 0.00
segment	0.95 ± 0.01 (+0.02)	0.95 ± 0.00 (+0.02)	0.95 ± 0.01 (+0.02)	0.93 ± 0.01
glass	0.89 ± 0.04 (+0.03)	0.88 ± 0.05 (+0.02)	0.89 ± 0.04 (+0.03)	0.87 ± 0.05
vehicle	0.88 ± 0.02 (+0.04)	0.88 ± 0.02 (+0.03)	0.89 ± 0.02 (+0.04)	0.85 ± 0.03
iris	0.89 ± 0.07 (+0.06)	0.87 ± 0.07 (+0.04)	0.87 ± 0.07 (+0.05)	0.83 ± 0.09
stock	0.82 ± 0.02 (+0.04)	0.81 ± 0.02 (+0.03)	0.83 ± 0.02 (+0.05)	0.78 ± 0.02
wisconsin	0.79 ± 0.03 (+0.01)	0.77 ± 0.03 (-0.01)	0.79 ± 0.03 (+0.01)	0.78 ± 0.03
elevators	0.81 ± 0.00 (+0.04)	0.81 ± 0.00 (+0.04)	0.81 ± 0.00 (+0.04)	0.77 ± 0.00
vowel	0.76 ± 0.03 (+0.03)	0.77 ± 0.01 (+0.05)	0.78 ± 0.02 (+0.05)	0.73 ± 0.02
housing	0.77 ± 0.03 (+0.07)	0.78 ± 0.02 (+0.08)	0.77 ± 0.03 (+0.07)	0.70 ± 0.03
cpu-small	0.55 ± 0.01 (+0.05)	0.56 ± 0.01 (+0.05)	0.54 ± 0.01 (+0.04)	0.50 ± 0.02
bodyfat	0.35 ± 0.07 (-0.01)	0.34 ± 0.07 (-0.02)	0.34 ± 0.08 (-0.02)	0.36 ± 0.07
calhousing	0.27 ± 0.01 (+0.01)	0.27 ± 0.01	0.27 ± 0.01 (+0.01)	0.26 ± 0.01
diau	0.26 ± 0.02	0.26 ± 0.02	0.26 ± 0.02	0.26 ± 0.02
spo	0.18 ± 0.02	0.19 ± 0.02 (+0.01)	0.18 ± 0.02	0.18 ± 0.02
dtt	0.15 ± 0.04	0.16 ± 0.04	0.14 ± 0.04 (-0.01)	0.15 ± 0.04
cold	0.09 ± 0.03	0.09 ± 0.03	0.10 ± 0.03 (+0.01)	0.09 ± 0.04
heat	0.06 ± 0.02	0.06 ± 0.02	0.06 ± 0.02	0.06 ± 0.02

Table 1. Detailed results of our label ranking experiment. Blue color indicates better Spearman rank correlation coefficient compared to using no projection. Red color indicates worse coeffcient.

B. Proofs

B.1. Proof of Lemma 1 (Discrete optimization formulation)

For the first claim, we have for all $\boldsymbol{w} \in \mathbb{R}^n$ such that $w_1 > w_2 > \cdots > w_n$

$$\sigma(\boldsymbol{\theta}) = \operatorname*{argmax}_{\sigma \in \Sigma} \langle \boldsymbol{\theta}_\sigma, \boldsymbol{w} \rangle \tag{11}$$

and in particular for $\boldsymbol{w} = \boldsymbol{\rho}$. The second claim follows from

$$\sigma(\boldsymbol{\theta}) = \operatorname*{argmax}_{\sigma \in \Sigma} \langle \boldsymbol{\theta}, \boldsymbol{w}_{\sigma^{-1}} \rangle = \operatorname*{argmax}_{\pi^{-1} \in \Sigma} \langle \boldsymbol{\theta}, \boldsymbol{w}_\pi \rangle = \left(\operatorname*{argmax}_{\pi \in \Sigma} \langle \boldsymbol{\theta}, \boldsymbol{w}_\pi \rangle \right)^{-1}.$$

B.2. Proof of Proposition 1 (Linear programming formulations)

Let us prove the first claim. The key idea is to absorb $\boldsymbol{\theta}_\sigma$ in the permutahedron. Using (11), we obtain for all $\boldsymbol{\theta} \in \mathbb{R}^n$ and for all $\boldsymbol{w} \in \mathbb{R}^n$ such that $w_1 > \cdots > w_n$

$$\boldsymbol{\theta}_{\sigma(\boldsymbol{\theta})} = \operatorname*{argmax}_{\boldsymbol{\theta}_\sigma \,:\, \sigma \in \Sigma} \langle \boldsymbol{\theta}_\sigma, \boldsymbol{w} \rangle = \operatorname*{argmax}_{\boldsymbol{y} \in \Sigma(\boldsymbol{\theta})} \langle \boldsymbol{y}, \boldsymbol{w} \rangle = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{P}(\boldsymbol{\theta})} \langle \boldsymbol{y}, \boldsymbol{w} \rangle,$$

where in the second equality we used $\mathcal{P}(\boldsymbol{\theta}) = \operatorname{conv}(\Sigma(\boldsymbol{\theta}))$ and the fundamental theorem of linear programming (Dantzig et al., 1955, Theorem 6). For the second claim, we have similarly

$$\boldsymbol{w}_{r(\boldsymbol{\theta})} = \operatorname*{argmax}_{\boldsymbol{w}_\pi \,:\, \pi \in \Sigma} \langle \boldsymbol{\theta}, \boldsymbol{w}_\pi \rangle = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{P}(\boldsymbol{w})} \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle.$$

Setting $\boldsymbol{w} = \boldsymbol{\rho}$ and using $\boldsymbol{\rho}_{r(\boldsymbol{\theta})} = \boldsymbol{\rho}_{\sigma^{-1}(\boldsymbol{\theta})} = \sigma^{-1}(-\boldsymbol{\theta}) = r(-\boldsymbol{\theta})$ proves the claim.

B.3. Proof of Proposition 2 (Properties of soft sorting and ranking operators)

Differentiability. Let $\mathcal{C}$ be a closed convex set and let $\boldsymbol{\mu}^\star(\boldsymbol{z}) \coloneqq \operatorname{argmax}_{\boldsymbol{\mu} \in \mathcal{C}} \langle \boldsymbol{\mu}, \boldsymbol{z} \rangle - \Psi(\boldsymbol{z})$. If Ψ is strongly convex over $\mathcal{C}$, then $\boldsymbol{\mu}^\star(\boldsymbol{z})$ is Lipschitz continuous. By Rademachers theorem, $\boldsymbol{\mu}^\star(\boldsymbol{z})$ is differentiable almost everywhere. Furthermore, since $P_\Psi(\boldsymbol{z}, \boldsymbol{w}) = \nabla\Psi(\boldsymbol{\mu}^\star(\boldsymbol{z}))$ with $\mathcal{C} = \mathcal{P}(\nabla\Psi^{-1}(\boldsymbol{w}))$, $P_\Psi(\boldsymbol{z}, \boldsymbol{w})$ is differentiable a.e. as long as Ψ is twice differentiable, which is the case when $\Psi \in \{Q, E\}$.

Order preservation. Proposition 1 of Blondel et al. (2019) shows that $\boldsymbol{\mu}^\star(\boldsymbol{z})$ and $\boldsymbol{z}$ are sorted the same way. Furthermore, since $P_\Psi(\boldsymbol{z}, \boldsymbol{w}) = \nabla\Psi(\boldsymbol{\mu}^\star(\boldsymbol{z}))$ with $\mathcal{C} = \mathcal{P}(\nabla\Psi^{-1}(\boldsymbol{w}))$ and since $\nabla\Psi$ is monotone, $P_\Psi(\boldsymbol{z}, \boldsymbol{w})$ is sorted the same way as $\boldsymbol{z}$, as well. Let $\boldsymbol{s} = s_{\varepsilon\Psi}(\boldsymbol{\theta})$ and $\boldsymbol{r} = r_{\varepsilon\Psi}(\boldsymbol{\theta})$. From the respective definitions, this means that $\boldsymbol{s}$ is sorted the same way as $\boldsymbol{\rho}$ (i.e., it is sorted in descending order) and $\boldsymbol{r}$ is sorted the same way as $-\boldsymbol{\theta}$, which concludes the proof.

Asymptotic behavior. We will now characterize the behavior for sufficiently small and large regularization strength ε. Note that rather than multiplying the regularizer Ψ by $\varepsilon > 0$, we instead divide $\boldsymbol{s}$ by ε, which is equivalent.

Lemma 3. *Analytical solutions of isotonic optimization in the limit regimes*

If $\varepsilon \leq \varepsilon_{min}(\boldsymbol{s}, \boldsymbol{w}) \coloneqq \min_{i \in [n-1]} \frac{s_i - s_{i+1}}{w_i - w_{i+1}}$, *then*

$$\boldsymbol{v}_Q(\boldsymbol{s}/\varepsilon, \boldsymbol{w}) = \boldsymbol{v}_E(\boldsymbol{s}/\varepsilon, \boldsymbol{w}) = \boldsymbol{s}/\varepsilon - \boldsymbol{w}.$$

If $\varepsilon > \varepsilon_{max}(\boldsymbol{s}, \boldsymbol{w}) \coloneqq \max_{i<j} \frac{s_i - s_j}{w_i - w_j}$, *then*

$$v_Q(\boldsymbol{s}/\varepsilon, \boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^n (s_i/\varepsilon - w_i)\mathbf{1} \quad \textit{and} \quad v_E(\boldsymbol{s}/\varepsilon, \boldsymbol{w}) = (\mathit{LSE}(\boldsymbol{s}/\varepsilon) - \mathit{LSE}(\boldsymbol{w}))\mathbf{1},$$

where $\mathit{LSE}(\boldsymbol{x}) \coloneqq \log \sum_i e^{x_i}$.

Proof. We start with the $\varepsilon \leq \varepsilon_{\min}(\boldsymbol{s}, \boldsymbol{w})$ case. Recall that $\boldsymbol{s}$ is sorted in descending order. Therefore, since we chose ε sufficiently small, the vector $\boldsymbol{v} = \boldsymbol{s}/\varepsilon - \boldsymbol{w}$ is sorted in descending order as well. This means that $\boldsymbol{v}$ is feasible, i.e., it belongs to the constraint sets in Proposition 3. Further, note that $v_i = \gamma_Q(\{i\}; \boldsymbol{s}/\varepsilon, \boldsymbol{w}) = \gamma_E(\{i\}; \boldsymbol{s}/\varepsilon, \boldsymbol{w}) = s_i/\varepsilon - w_i$ so that $\boldsymbol{v}$ is the optimal solution if we drop the constraints, which completes the argument.

Next, we tackle the $\varepsilon > \varepsilon_{\max}(\boldsymbol{s}, \boldsymbol{w})$ case. Note that the claimed solutions are exactly $\gamma_Q([n]; \boldsymbol{s}, \boldsymbol{w})$ and $\gamma_E([n]; \boldsymbol{s}, \boldsymbol{w})$, so the claim will immediately follow if we show that $[n]$ is an optimal partition. The PAV algorithm (cf. §B.6) merges at each iteration any two neighboring blocks B_1, B_2 that violate $\gamma_\Psi(B_1; \boldsymbol{s}/\varepsilon, \boldsymbol{w}) \geq \gamma_\Psi(B_2; \boldsymbol{s}/\varepsilon, \boldsymbol{w})$, starting from the partitions consisting of singleton sets. Let $k \in \{1, \dots, n-1\}$ be the iteration number. We claim that the two blocks, $B_1 = \{1, 2, \dots, k\}$ and $B_2 = \{k+1\}$, will always be violating the constraint, so that they can be merged. Note that in the quadratic case, they can be merged only if

$$\sum_{i=1}^{k} (s_i/\varepsilon - w_i)/k < s_{k+1}/\varepsilon - w_{k+1},$$

which is equivalent to

$$\sum_{i=1}^{k} \frac{s_i - s_{k+1}}{k\varepsilon} < \sum_{i=1}^{k} (w_i - w_{k+1}),$$

which is indeed satisfied when $\varepsilon > \varepsilon_{\max}(\boldsymbol{s}, \boldsymbol{w})$. In the KL case, they can be merged only if

$$\begin{aligned}
\log \sum_{i=1}^{k} e^{s_i/\varepsilon} - \log \sum_{i=1}^{k} e^{w_i} < s_{k+1}/\varepsilon - w_{k+1} &\iff \log \sum_{i=1}^{k} e^{s_i/\varepsilon} - s_{k+1}/\varepsilon < \log \sum_{i=1}^{k} e^{w_i} - w_{k+1} \\
&\iff \log \sum_{i=1}^{k} e^{s_i/\varepsilon} - \log e^{s_{k+1}/\varepsilon} < \log \sum_{i=1}^{k} e^{w_i} - \log e^{w_{k+1}} \\
&\iff \log \sum_{i=1}^{k} e^{(s_i - s_{k+1})/\varepsilon} < \log \sum_{i=1}^{k} e^{w_i - w_{k+1}} \\
&\iff \sum_{i=1}^{k} e^{(s_i - s_{k+1})/\varepsilon} < \sum_{i=1}^{k} e^{w_i - w_{k+1}}.
\end{aligned}$$

This will be true if the i^{th} term on the left-hand side is smaller than the i^{th} term on the right-hand side, i.e., when $(s_i - s_{k+1})/\varepsilon < w_i - w_{k+1}$, which again is implied by the assumption. □

We can now directly characterize the behavior of the projection operator P_Ψ in the two regimes $\varepsilon \leq \varepsilon_{\min}(s(\boldsymbol{z}), \boldsymbol{w})$ and $\varepsilon > \varepsilon_{\max}(s(\boldsymbol{z}), \boldsymbol{w})$. This in turn implies the results for both the soft ranking and sorting operations using (5) and (6).

Proposition 5. *Analytical solutions of the projections in the limit regimes*

If $\varepsilon \leq \varepsilon_{min}(s(\boldsymbol{z}), \boldsymbol{w})$*, then*

$$P_\Psi(\boldsymbol{z}/\varepsilon, \boldsymbol{w}) = \boldsymbol{w}_{\sigma^{-1}(\boldsymbol{z})}.$$

If $\varepsilon > \varepsilon_{max}(s(\boldsymbol{z}), \boldsymbol{w})$*, then*

$$\begin{aligned}
P_Q(\boldsymbol{z}/\varepsilon, \boldsymbol{w}) &= \boldsymbol{z}/\varepsilon - \mathit{mean}(\boldsymbol{z}/\varepsilon - \boldsymbol{w})\mathbf{1}, \text{ and} \\
P_E(\boldsymbol{z}/\varepsilon, \boldsymbol{w}) &= \boldsymbol{z}/\varepsilon - \mathit{LSE}(\boldsymbol{z}/\varepsilon)\mathbf{1} + \mathit{LSE}(\boldsymbol{w})\mathbf{1}.
\end{aligned}$$

Therefore, in these two regimes, we do not even need PAV to compute the optimal projection.

B.4. Proof of Proposition 3 (Reduction to isotonic optimization)

Before proving Proposition 3, we need the following three lemmas.

Lemma 4. *Technical lemma*

Let $f\colon \mathbb{R} \to \mathbb{R}$ be convex, $v_1 \geq v_2$ and $s_2 \geq s_1$. Then, $f(s_1 - v_1) + f(s_2 - v_2) \geq f(s_2 - v_1) + f(s_1 - v_2)$.

Proof. Note that $s_2 - v_2 \geq s_2 - v_1 \geq s_1 - v_1$ and $s_2 - v_2 \geq s_1 - v_2 \geq s_1 - v_1$. This means that we can express $s_2 - v_1$ and $s_1 - v_2$ as a convex combination of the endpoints of the line segment $[s_1 - v_1, s_2 - v_2]$, namely

$$s_2 - v_1 = \alpha(s_2 - v_2) + (1 - \alpha)(s_1 - v_1) \quad \text{and} \quad s_1 - v_2 = \beta(s_2 - v_2) + (1 - \beta)(s_1 - v_1).$$

Solving for α and β gives $\alpha = 1 - \beta$. From the convexity of f, we therefore have

$$f(s_2 - v_1) \leq \alpha f(s_2 - v_2) + (1 - \alpha) f(s_1 - v_1) \quad \text{and} \quad f(s_1 - v_2) \leq (1 - \alpha) f(s_2 - v_2) + \alpha f(s_1 - v_1).$$

Summing the two proves the claim. □

Lemma 5. *Dual formulation of a regularized linear program*

Let $\boldsymbol{\mu}^\star = \operatorname{argmax}_{\boldsymbol{\mu} \in \mathcal{C}} \langle \boldsymbol{\mu}, \boldsymbol{z} \rangle - \Psi(\boldsymbol{\mu})$, where $\mathcal{C} \subseteq \mathbb{R}^n$ is a closed convex set and Ψ is strongly convex. Then, the corresponding dual solution is $\boldsymbol{u}^\star = \operatorname{argmin}_{\boldsymbol{u} \in \mathbb{R}^n} \Psi^*(\boldsymbol{z} - \boldsymbol{u}) + s_{\mathcal{C}}(\boldsymbol{u})$, where $s_{\mathcal{C}}(\boldsymbol{u}) \coloneqq \sup_{\boldsymbol{y} \in \mathcal{C}} \langle \boldsymbol{y}, \boldsymbol{u} \rangle$ is the support function of $\mathcal{C}$. Moreover, $\boldsymbol{\mu}^\star = \nabla \Psi^*(\boldsymbol{z} - \boldsymbol{u}^\star)$.

Proof. The result is well-known and we include the proof for completeness. Let us define the Fenchel conjugate of a function $\Omega\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ by

$$\Omega^*(\boldsymbol{z}) \coloneqq \sup_{\boldsymbol{\mu} \in \mathbb{R}^n} \langle \boldsymbol{\mu}, \boldsymbol{z} \rangle - \Omega(\boldsymbol{\mu}).$$

Let $\Omega \coloneqq \Psi + \Phi$, where Ψ is strongly convex and Φ is convex. We have

$$\Omega^*(\boldsymbol{z}) = (\Psi + \Phi)^*(\boldsymbol{z}) = \inf_{\boldsymbol{u} \in \mathbb{R}^n} \Phi^*(\boldsymbol{u}) + \Psi^*(\boldsymbol{z} - \boldsymbol{u}),$$

which is the infimal convolution of Φ^* with Ψ^*. Moreover, $\nabla \Omega^*(\boldsymbol{z}) = \nabla \Psi^*(\boldsymbol{z} - \boldsymbol{u}^\star)$. The results follows from choosing $\Phi(\boldsymbol{\mu}) = I_{\mathcal{C}}(\boldsymbol{\mu})$ and noting that $I_{\mathcal{C}}^* = s_{\mathcal{C}}$. □

For instance, with $\Psi = Q$, we have $\Psi^* = Q$, and with $\Psi = E$, we have $\Psi^* = \exp$.

The next lemma shows how to go further by choosing $\mathcal{C}$ as the base polytope $\mathcal{B}(F)$ associated with a cardinality-based submodular function F, of which the permutahedron is a special case. The polytope is defined as (see, e.g., Bach (2013))

$$\mathcal{B}(F) \coloneqq \left\{ \boldsymbol{\mu} \in \mathbb{R}^n \colon \sum_{i \in \mathcal{S}} \mu_i \leq F(\mathcal{S}) \; \forall \mathcal{S} \subseteq [n], \sum_{i=1}^n \mu_i = F([n]) \right\}.$$

Lemma 6. *Reducing dual formulation to isotonic regression*

Let $F(\mathcal{S}) = g(|\mathcal{S}|)$ for some concave g. Let $\mathcal{B}(F)$ be its corresponding base polytope. Let σ be a permutation of $[n]$ such that $\boldsymbol{z} \in \mathbb{R}^n$ is sorted in descending order, i.e., $z_{\sigma_1} \geq z_{\sigma_2} \geq \cdots \geq z_{\sigma_n}$. Assume $\Psi(\boldsymbol{\mu}) = \sum_{i=1}^n \psi(\mu_i)$, where ψ is convex. Then, the dual solution $\boldsymbol{u}^\star$ from Lemma 5 is equal to $\boldsymbol{v}^\star_{\sigma^{-1}}$, where

$$\begin{aligned} \boldsymbol{v}^\star &= \operatorname*{argmin}_{v_1 \geq \cdots \geq v_n} \Psi^*(\boldsymbol{z}_\sigma - \boldsymbol{v}) + \langle \boldsymbol{f}_\sigma, \boldsymbol{v} \rangle \\ &= - \operatorname*{argmin}_{v'_1 \leq \cdots \leq v'_n} \Psi^*(\boldsymbol{v}'_\sigma + \boldsymbol{z}) - \langle \boldsymbol{f}_\sigma, \boldsymbol{v}' \rangle. \end{aligned}$$

Proof. The support function $s_{\mathcal{B}(F)}(\boldsymbol{u})$ is known as the Lovász extension of F. For conciseness, we use the standard notation $f(\boldsymbol{u}) \coloneqq s_{\mathcal{B}(F)}(\boldsymbol{u})$. Applying Lemma 5, we obtain

$$\boldsymbol{u}^\star = \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^n} \Psi^*(\boldsymbol{z}-\boldsymbol{u}) + f(\boldsymbol{u}).$$

Using the "greedy algorithm" of Edmonds (1970), we can compute $f(\boldsymbol{u})$ as follows. First, choose a permutation σ that sorts $\boldsymbol{u}$ in descending order, i.e., $u_{\sigma_1} \geq u_{\sigma_2} \geq \cdots \geq u_{\sigma_n}$. Then a maximizer $\boldsymbol{f} \in \mathcal{B}(F) \subseteq \mathbb{R}^n$ is obtained by forming $\boldsymbol{f}_\sigma = (f_{\sigma_1}, \ldots, f_{\sigma_n})$, where

$$f_{\sigma_i} \coloneqq F(\{\sigma_1, \ldots, \sigma_i\}) - F(\{\sigma_1, \ldots, \sigma_{i-1}\}).$$

Moreover, $\langle \boldsymbol{f}, \boldsymbol{u} \rangle = f(\boldsymbol{u})$.

Let us fix σ to the permutation that sorts $\boldsymbol{u}^\star$. Following the same idea as from (Djolonga & Krause, 2017), since the Lovász extension is linear on the set of all vectors that are sorted by σ, we can write

$$\operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^n} \Psi^*(\boldsymbol{z}-\boldsymbol{u}) + f(\boldsymbol{u}) = \operatorname*{argmin}_{u_{\sigma_1}\geq\cdots\geq u_{\sigma_n}} \Psi^*(\boldsymbol{z}-\boldsymbol{u}) + \langle \boldsymbol{f}, \boldsymbol{u} \rangle.$$

This is an instance of isotonic optimization, as we can rewrite the problem as

$$\operatorname*{argmin}_{v_1\geq\cdots\geq v_n} \Psi^*(\boldsymbol{z}-\boldsymbol{v}_{\sigma^{-1}}) + \langle \boldsymbol{f}, \boldsymbol{v}_{\sigma^{-1}} \rangle = \operatorname*{argmin}_{v_1\geq\cdots\geq v_n} \Psi^*(\boldsymbol{z}_\sigma-\boldsymbol{v}) + \langle \boldsymbol{f}_\sigma, \boldsymbol{v} \rangle, \tag{12}$$

with $\boldsymbol{u}^\star_\sigma = \boldsymbol{v}^\star \Leftrightarrow \boldsymbol{u}^\star = \boldsymbol{v}^\star_{\sigma^{-1}}$.

Let $\boldsymbol{s} \coloneqq \boldsymbol{z}_\sigma$. It remains to show that $s_1 \geq \cdots \geq s_n$, i.e., that $\boldsymbol{s}$ and the optimal dual variables $\boldsymbol{v}^\star$ are both in descending order. Suppose $s_j > s_i$ for some $i < j$. Let $\boldsymbol{s}'$ be a copy of $\boldsymbol{s}$ with s_i and s_j swapped. Since ψ^* is convex, by Lemma 4,

$$\Psi^*(\boldsymbol{s}-\boldsymbol{v}^\star) - \Psi^*(\boldsymbol{s}'-\boldsymbol{v}^\star) = \psi^*(s_i - v^\star_i) + \psi^*(s_j - v^\star_j) - \psi^*(s_j - v^\star_i) - \psi^*(s_i - v^\star_j) \geq 0,$$

which contradicts the assumption that $\boldsymbol{v}^\star$ and the corresponding σ are optimal. A similar result is proven by Suehiro et al. (2012, Lemma 1) but for the optimal primal variable $\boldsymbol{\mu}^\star$. □

We now prove Proposition 3. The permutahedron $\mathcal{P}(\boldsymbol{w})$ is a special case of $\mathcal{B}(F)$ with $F(\mathcal{S}) = \sum_{i=1}^{|\mathcal{S}|} w_i$ and $w_1 \geq w_2 \geq \cdots \geq w_n$. In that case, $\boldsymbol{f}_\sigma = (f_{\sigma_1}, \ldots, f_{\sigma_n}) = (w_1, \ldots, w_n) = \boldsymbol{w}$.

For $\mathcal{P}(\nabla\Psi^*(\boldsymbol{w}))$, we thus have $\boldsymbol{f}_\sigma = \nabla\Psi^*(\boldsymbol{w})$. Finally, note that if Ψ is Legendre-type, which is the case of both Q and E, then $\nabla\Psi^* = (\nabla\Psi)^{-1}$. Therefore, $\nabla\Psi(\boldsymbol{\mu}^\star) = \boldsymbol{z} - \boldsymbol{u}^\star$, which concludes the proof.

B.5. Relaxed dual linear program interpretation

We show in this section that the dual problem in Lemma 6 can be interpreted as the original dual linear program (LP) with relaxed equality constraints. Consider the primal LP

$$\max_{\boldsymbol{y}\in\mathcal{B}(F)} \langle \boldsymbol{y}, \boldsymbol{z} \rangle. \tag{13}$$

As shown by Bach (2013, Proposition 3.2), the dual LP is

$$\min_{\boldsymbol{\lambda}\in\mathcal{C}} \sum_{\mathcal{S}\subseteq\mathcal{V}} \lambda_\mathcal{S} F(\mathcal{S}) \tag{14}$$

where

$$\mathcal{C} \coloneqq \left\{ \boldsymbol{\lambda} \in \mathbb{R}^{2^\mathcal{V}} : \lambda_\mathcal{S} \geq 0 \; \forall \mathcal{S} \subset \mathcal{V}, \lambda_\mathcal{V} \in \mathbb{R}, z_i = \sum_{\mathcal{S}:\, i\in\mathcal{S}} \lambda_\mathcal{S} \; \forall i \in [n] \right\}.$$

Moreover, let σ be a permutation sorting $\boldsymbol{z}$ in descending order. Then, an optimal $\boldsymbol{\lambda}$ is given by (Bach, 2013, Proposition 3.2)

$$\lambda_\mathcal{S} = \begin{cases} z_{\sigma_i} - z_{\sigma_{i+1}} & \text{if } \mathcal{S} = \{\sigma_1, \ldots, \sigma_i\} \\ z_{\sigma_n} & \text{if } \mathcal{S} = \{\sigma_1, \ldots, \sigma_n\} \\ 0 & \text{otherwise.} \end{cases}$$

Now let us restrict to the support of $\boldsymbol{\lambda}$ and do the change of variable

$$\lambda_{\mathcal{S}} = \begin{cases} v_i - v_{i+1} & \text{if } \mathcal{S} = \{\sigma_1, \dots, \sigma_i\} \\ v_n & \text{if } \mathcal{S} = \{\sigma_1, \dots, \sigma_n\}. \end{cases}$$

The non-negativity constraints in $\mathcal{C}$ become $v_1 \geq v_2 \geq \cdots \geq v_n$ and the equality constraints in $\mathcal{C}$ become $\boldsymbol{z}_\sigma = \boldsymbol{v}$. Adding quadratic regularization $\frac{1}{2}\|\boldsymbol{y}\|^2$ in the primal problem (13) is equivalent to relaxing the dual equality constraints in (14) by smooth constraints $\frac{1}{2}\|\boldsymbol{z}_\sigma - \boldsymbol{v}\|^2$ (this can be seen by adding quadratic regularization to the primal variables of Bach (2013, Eq. (3.6))). For the dual objective (14), we have

$$\begin{aligned} \sum_{\mathcal{S} \subseteq \mathcal{V}} \lambda_{\mathcal{S}} F(\mathcal{S}) &= \sum_{i=1}^{n-1} (v_i - v_{i+1}) F(\{\sigma_1, \dots, \sigma_i\}) + v_n F(\{\sigma_1, \dots, \sigma_n\}) \\ &= \sum_{i=1}^{n} (F(\{\sigma_1, \dots, \sigma_i\}) - F(\{\sigma_1, \dots, \sigma_{i-1}\})) v_i \\ &= \langle \boldsymbol{f}_\sigma, \boldsymbol{v} \rangle, \end{aligned}$$

where in the second line we used (Bach, 2013, Eq. (3.2)). Altogether, we obtain $\min_{v_1 \geq \cdots \geq v_n} \frac{1}{2}\|\boldsymbol{z}_\sigma - \boldsymbol{v}\|^2 + \langle \boldsymbol{f}_\sigma, \boldsymbol{v} \rangle$, which is exactly the expression we derived in Lemma 6. The entropic case is similar.

B.6. Pool adjacent violators (PAV) algorithm

Let $g_1, \dots, g_n$ be convex functions. As shown in (Best et al., 2000; Lim & Wright, 2016),

$$\operatorname*{argmin}_{v_1 \geq \cdots \geq v_n} \sum_{i=1}^{n} g_i(v_i)$$

can be solved using a generalization of the PAV algorithm (note that unlike these works, we use decreasing constraints for convenience). All we need is a routine for solving, given some set $\mathcal{B}$ of indices, the "pooling" sub-problem

$$\operatorname*{argmin}_{\gamma \in \mathbb{R}} \sum_{i \in \mathcal{B}} g_i(\gamma).$$

Thus, we can use PAV to solve (12), as long as Ψ^* is separable. We now give the closed-form solution for two special cases. To simplify, we denote $\boldsymbol{s} \coloneqq \boldsymbol{z}_\sigma$ and $\boldsymbol{w} \coloneqq \boldsymbol{f}_\sigma$.

Quadratic regularization. We have $g_i(v_i) = \frac{1}{2}(s_i - v_i)^2 + v_i w_i$. We therefore minimize

$$\sum_{i \in \mathcal{B}} g_i(\gamma) = \sum_{i \in \mathcal{B}} \frac{1}{2}(s_i - \gamma)^2 + \gamma \sum_{i \in \mathcal{B}} w_i.$$

The closed-form solution is

$$\gamma_Q^\star(\boldsymbol{s}, \boldsymbol{w}; \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} (s_i - w_i).$$

Entropic regularization. We have $g_i(v_i) = e^{s_i - v_i} + v_i e^{w_i}$. We therefore minimize

$$\sum_{i \in \mathcal{B}} g_i(\gamma) = \sum_{i \in \mathcal{B}} e^{s_i - \gamma} + \gamma \sum_{i \in \mathcal{B}} e^{w_i}.$$

The closed-form solution is

$$\gamma_E^\star(\boldsymbol{s}, \boldsymbol{w}; \mathcal{B}) = -\log \frac{\sum_{i \in \mathcal{B}} w_i}{\sum_{i \in \mathcal{B}} e^{s_i}} = \mathrm{LSE}(\boldsymbol{s}_\mathcal{B}) - \mathrm{LSE}(\boldsymbol{w}_\mathcal{B}),$$

where $\mathrm{LSE}(\boldsymbol{x}) \coloneqq \log \sum_i e^{x_i}$.

Although not explored in this work, other regularizations are potentially possible, see, e.g., (Blondel et al., 2019).

B.7. Proof of Proposition 4 (Jacobian of isotonic optimization)

Let $\mathcal{B}_1, \dots, \mathcal{B}_m$ be the partition of $[n]$ induced by $\boldsymbol{v} \coloneqq \boldsymbol{v}_\Psi(\boldsymbol{s}, \boldsymbol{w})$. From the PAV algorithm, for all $i \in [n]$, there is a unique block $\mathcal{B}_l \in \{\mathcal{B}_1, \dots, \mathcal{B}_m\}$ such that $i \in \mathcal{B}_l$ and $v_i = \gamma_\Psi(\mathcal{B}_l; \boldsymbol{s}, \boldsymbol{w})$. Therefore, for all $i \in [n]$, we obtain

$$\frac{\partial v_i}{\partial s_j} = \begin{cases} \frac{\partial \gamma_\Psi(\mathcal{B}_l; \boldsymbol{s}, \boldsymbol{w})}{\partial s_j} & \text{if } i, j \in \mathcal{B}_l \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the Jacobian matrix is block diagonal, i.e.,

$$\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{s}} = \begin{bmatrix} \mathbf{B}_1^\Psi & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_m^\Psi \end{bmatrix}.$$

For the block $\mathcal{B}_l$, the non-zero partial derivatives form a matrix $\boldsymbol{B}_l^\Psi \in \mathbb{R}^{|\mathcal{B}_l| \times |\mathcal{B}_l|}$ such that each column is associated with one s_j and contains the value $\frac{\partial \gamma_\Psi(\mathcal{B}_l; \boldsymbol{s}, \boldsymbol{w})}{\partial s_j}$ (all values in a column are the same). For quadratic regularization, we have

$$\frac{\partial v_i}{\partial s_j} = \begin{cases} \frac{1}{|\mathcal{B}_l|} & \text{if } i, j \in \mathcal{B}_l \\ 0 & \text{otherwise.} \end{cases}$$

For entropic regularization, we have

$$\frac{\partial v_i}{\partial s_j} = \begin{cases} \frac{e^{s_j}}{\sum_{j' \in \mathcal{B}} e^{s_{j'}}} = \operatorname{softmax}(\boldsymbol{s}_{\mathcal{B}_l})_j & \text{if } i, j \in \mathcal{B}_l \\ 0 & \text{otherwise.} \end{cases}$$

The multiplication with the Jacobian uses the fact that each block is constant column-wise.

Remark. The expression above is for points $\boldsymbol{s}$ where $\boldsymbol{v}$ is differentiable. For points where $\boldsymbol{v}$ is not differentiable, we can take an arbitrary matrix in the set of Clarke's generalized Jacobians, the convex hull of Jacobians of the form $\lim_{\boldsymbol{s}_t \to \boldsymbol{s}} \partial \boldsymbol{v} / \partial \boldsymbol{s}_t$. The points of non-differentiability occur when a block of the optimal solution can be split up into two blocks with equal values. In that case, the two directional derivatives do not agree, but are derived for quadratic regularization by Djolonga & Krause (2017).