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 Rademacher's 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).