Appendix

A More examples of generalized entropies

In this section, we give two more examples of generalized entropies: squared norm entropies and Rényi entropies.

Squared norm entropies. Inspired by Niculae and Blondel (2017), as a simple extension of the Gini index (7), we consider the following generalized entropy based on squared q-norms:

$$\mathsf{H}_q^{\text{SQ}}(\boldsymbol{p}) \coloneqq \frac{1}{2}(1 - \|\boldsymbol{p}\|_q^2) = \frac{1}{2} - \frac{1}{2}\left(\sum_{j=1}^d p_j^q\right)^{\frac{2}{q}}.$$

The constant term $\frac{1}{2}$, omitted by Niculae and Blondel (2017), ensures satisfaction of A.1. For $q \in (1, 2]$, it is known that the squared q-norm is strongly convex w.r.t. $\|\cdot\|_q$ (Ball et al., 1994), implying that $(-\mathsf{H}_q^{\text{SQ}})^*$, and therefore $L_{-\mathsf{H}_q^{\text{SQ}}}$, is smooth. Although $\widehat{\boldsymbol{y}}_{-\mathsf{H}_q^{\text{SQ}}}(\boldsymbol{\theta})$ cannot be solved in closed form for $q \in (1, 2)$, it can be solved efficiently using projected gradient descent methods.

Rényi $\boldsymbol{\beta}$-entropies. Rényi entropies (Rényi, 1961) are defined for any $\beta \geq 0$ as:

$$\mathsf{H}_\beta^{\text{R}}(\boldsymbol{p}) \coloneqq \frac{1}{1-\beta}\log\sum_{j=1}^d p_j^\beta.$$

Unlike Shannon and Tsallis entropies, Rényi entropies are not separable, with the exception of $\beta \to 1$, which also recovers Shannon entropy as a limit case. The case $\beta \to +\infty$ gives $\mathsf{H}_\beta^{\text{R}}(\boldsymbol{p}) = -\log\|\boldsymbol{p}\|_\infty$. For $\beta \in [0, 1]$, Rényi entropies satisfy assumptions A.1–A.3; for $\beta > 1$, Rényi entropies fail to be concave. They are however pseudo-concave (Mangasarian, 1965), meaning that, for all $\boldsymbol{p}, \boldsymbol{q} \in \triangle^d$, $\langle \nabla \mathsf{H}_\beta^{\text{R}}(\boldsymbol{p}), \boldsymbol{q} - \boldsymbol{p}\rangle \leq 0$ implies $\mathsf{H}_\beta^{\text{R}}(\boldsymbol{q}) \leq \mathsf{H}_\beta^{\text{R}}(\boldsymbol{p})$. This implies, among other things, that points $\boldsymbol{p} \in \triangle^d$ with zero gradient are maximizers of $\langle \boldsymbol{p}, \boldsymbol{\theta}\rangle + \mathsf{H}_\beta^{\text{R}}(\boldsymbol{p})$, which allows us to compute the predictive distribution $\widehat{\boldsymbol{y}}_{-\mathsf{H}_\beta^{\text{R}}}$ with gradient-based methods.

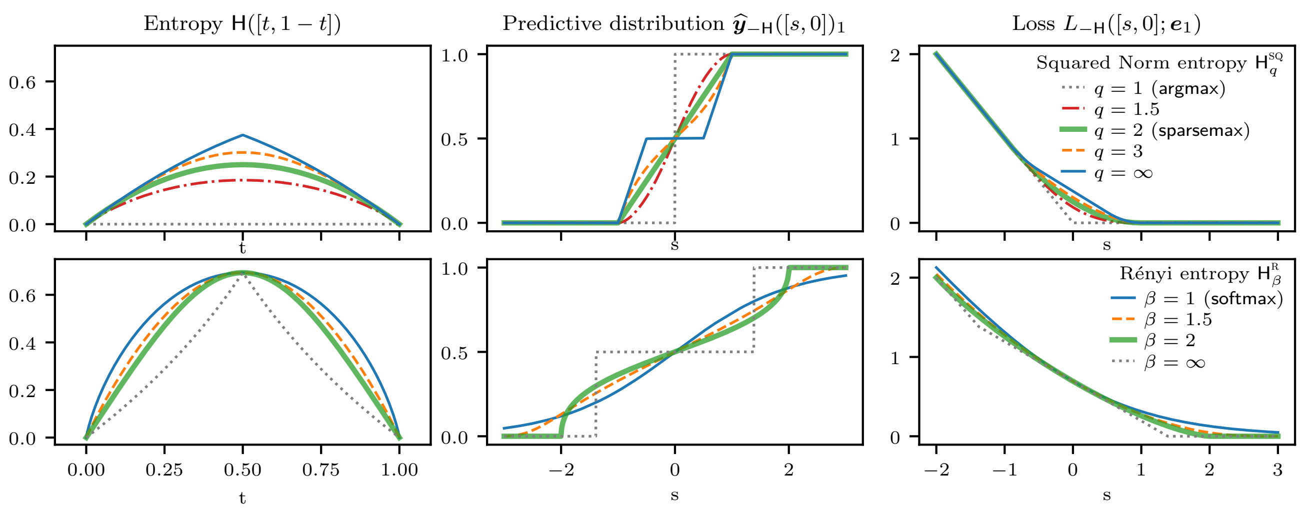

Figure 3: Squared norm and Rényi entropies, together with the distributions and losses they generate.

B Experiment details and additional empirical results

Benchmark datasets. The datasets we used in §6 are summarized below.

Table 3: Dataset statistics

Dataset	Type	Train	Dev	Test	Features	Classes	Avg. labels
Birds	Audio	134	45	172	260	19	2
Cal500	Music	376	126	101	68	174	25
Emotions	Music	293	98	202	72	6	2
Mediamill	Video	22,353	7,451	12,373	120	101	5
Scene	Images	908	303	1,196	294	6	1
SIAM TMC	Text	16,139	5,380	7,077	30,438	22	2
Yeast	Micro-array	1,125	375	917	103	14	4

The datasets can be downloaded from `http://mulan.sourceforge.net/datasets-mlc.html` and `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

Sparse label proportion estimation on synthetic data. We follow Martins and Astudillo (2016) and generate a document $\boldsymbol{x} \in \mathbb{R}^p$ from a mixture of multinomials and label proportions $\boldsymbol{y} \in \triangle^d$ from a multinomial. The number of words in $\boldsymbol{x}$ and labels in $\boldsymbol{y}$ is sampled from a Poisson distribution — see Martins and Astudillo (2016) for a precise description of the generative process. We use 1200 samples as training set, 200 samples as validation set and 1000 samples as test set. We tune $\lambda \in \{10^{-6}, 10^{-5}, \dots, 10^{0}\}$ and $\alpha \in \{1.0, 1.1, \dots, 2.0\}$ against the validation set. We report the Jensen-Shannon divergence in Figure 4. Results using the mean squared error (MSE) were entirely similar. When the number of classes is 10, we see that Tsallis and sparsemax losses perform almost exactly the same, both outperforming softmax. When the number of classes is 50, Tsallis losses outperform both sparsemax and softmax.

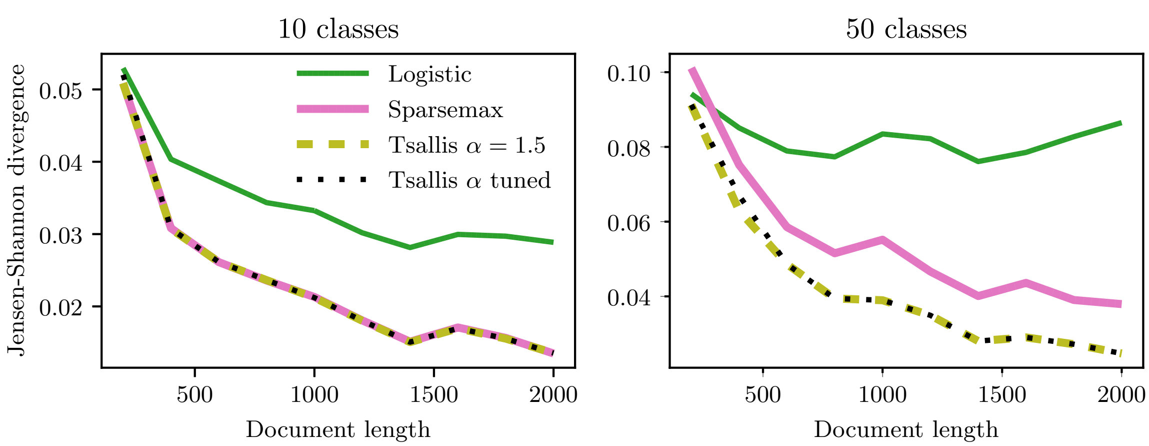

Figure 4: Jensen-Shannon divergence between predicted and true label proportions, when varying document length, of various losses generated by a Tsallis entropy.

C Proofs

In this section, we give proofs omitted from the main text.

C.1 Proof of Proposition 1

Effect of a permutation. Let Ω be symmetric. We first prove that Ω^* is symmetric as well. Indeed, we have

$$\Omega^*(\boldsymbol{P}\boldsymbol{\theta}) = \sup_{\boldsymbol{p} \in \mathsf{dom}(\Omega)} (\boldsymbol{P}\boldsymbol{\theta})^\top \boldsymbol{p} - \Omega(\boldsymbol{p}) = \sup_{\boldsymbol{p} \in \mathsf{dom}(\Omega)} \boldsymbol{\theta}^\top \boldsymbol{P}^\top \boldsymbol{p} - \Omega(\boldsymbol{P}^\top \boldsymbol{p}) = \Omega^*(\boldsymbol{\theta}).$$

The last equality was obtained by a change of variable $\boldsymbol{p}' = \boldsymbol{P}^\top \boldsymbol{p}$, from which $\boldsymbol{p}$ is recovered as $\boldsymbol{p} = \boldsymbol{P}\boldsymbol{p}'$, which proves $\nabla\Omega^*(\boldsymbol{P}\boldsymbol{p}) = \boldsymbol{P}\nabla\Omega^*(\boldsymbol{p})$.

Order preservation. Since Ω^* is convex, the gradient operator $\nabla\Omega^*$ is monotone, i.e.,

$$(\boldsymbol{\theta}' - \boldsymbol{\theta})^\top (\boldsymbol{p}' - \boldsymbol{p}) \geq 0$$

for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$, $\boldsymbol{p} = \nabla\Omega^*(\boldsymbol{\theta})$ and $\boldsymbol{p}' = \nabla\Omega^*(\boldsymbol{\theta}')$. Let $\boldsymbol{\theta}'$ be obtained from $\boldsymbol{\theta}$ by swapping two coordinates, i.e., $\theta'_j = \theta_i$, $\theta'_i = \theta_j$, and $\theta'_k = \theta_k$ for any $k \notin \{i, j\}$. Then, since Ω is symmetric, we obtain:

$$2(\theta_j - \theta_i)(p_j - p_i) \geq 0,$$

which implies $\theta_i > \theta_j \Rightarrow p_i \geq p_j$ and $p_i > p_j \Rightarrow \theta_i \geq \theta_j$. To fully prove the claim, we need to show that the last inequality is strict: to do this, we simply invoke $\nabla\Omega^*(\boldsymbol{P}\boldsymbol{p}) = \boldsymbol{P}\nabla\Omega^*(\boldsymbol{p})$ with a matrix $\boldsymbol{P}$ that permutes i and j, from which we must have $\theta_i = \theta_j \Rightarrow p_i = p_j$.

Gradient mapping. This follows directly from Danskin's theorem (Danskin, 1966). See also Bertsekas (1999, Proposition B.25).

Temperature scaling. This immediately follows from properties of the argmax operator.

C.2 Proof of Proposition 3

We set $\Omega \coloneqq \Psi + I_{\mathcal{C}}$.

Bregman projections. If Ψ is Legendre type, then $\nabla\Psi(\nabla\Psi^*(\boldsymbol{\theta})) = \boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \mathsf{int}(\mathsf{dom}(\Psi^*))$, where $\mathsf{int}(\mathcal{D})$ denotes the interior of $\mathcal{D}$. Using this and our assumption that $\mathsf{dom}(\Psi^*) = \mathbb{R}^d$, we get for all $\boldsymbol{\theta} \in \mathbb{R}^d$:

$$B_\Psi(\boldsymbol{p} || \nabla\Psi^*(\boldsymbol{\theta})) = \Psi(\boldsymbol{p}) - \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle + \langle \boldsymbol{\theta}, \nabla\Psi^*(\boldsymbol{\theta}) \rangle - \Psi(\nabla\Psi^*(\boldsymbol{\theta})). \tag{13}$$

The last two terms are independent of $\boldsymbol{p}$ and therefore

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \mathop{\mathsf{argmax}}_{\boldsymbol{p} \in \mathcal{C}} \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle - \Psi(\boldsymbol{p}) = \mathop{\mathsf{argmin}}_{\boldsymbol{p} \in \mathcal{C}} B_\Psi(\boldsymbol{p} || \nabla\Psi^*(\boldsymbol{\theta})),$$

where $\mathcal{C} \subseteq \mathsf{dom}(\Psi)$. The r.h.s. is the Bregman projection of $\nabla\Psi^*(\boldsymbol{\theta}) = \widehat{\boldsymbol{y}}_\Psi(\boldsymbol{\theta})$ onto $\mathcal{C}$.

Difference of Bregman divergences. Let $\boldsymbol{p} = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$. Using (13), we obtain

$$\begin{aligned} B_\Psi(\boldsymbol{y} || \nabla\Psi^*(\boldsymbol{\theta})) - B_\Psi(\boldsymbol{p} || \nabla\Psi^*(\boldsymbol{\theta})) &= \Psi(\boldsymbol{y}) - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle + \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle - \Psi(\boldsymbol{p}) \\ &= \Omega(\boldsymbol{y}) - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle + \Omega^*(\boldsymbol{\theta}) \\ &= L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}), \end{aligned} \tag{14}$$

where we assumed $\boldsymbol{y} \in \mathcal{C}$ and $\mathcal{C} \subseteq \mathsf{dom}(\Psi)$, implying $\Psi(\boldsymbol{y}) = \Omega(\boldsymbol{y})$.

If $\mathcal{C} = \mathsf{dom}(\Psi)$ (i.e., $\Omega = \Psi$), then $\boldsymbol{p} = \nabla\Psi^*(\boldsymbol{\theta})$ and $B_\Psi(\boldsymbol{p} || \nabla\Psi^*(\boldsymbol{\theta})) = 0$. We thus get the **composite form** of Fenchel-Young losses

$$B_\Omega(\boldsymbol{y} || \nabla\Omega^*(\boldsymbol{\theta})) = B_\Omega(\boldsymbol{y} || \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})) = L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}).$$

Bound. Let $\boldsymbol{p} = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$. Since $\boldsymbol{p}$ is the Bregman projection of $\nabla\Psi^*(\boldsymbol{\theta})$ onto $\mathcal{C}$, we can use the well-known Pythagorean theorem for Bregman divergences (see, e.g., Banerjee et al. (2005, Appendix A)) to obtain for all $\boldsymbol{y} \in \mathcal{C} \subseteq \mathsf{dom}(\Psi)$:

$$B_\Psi(\boldsymbol{y}||\boldsymbol{p}) + B_\Psi(\boldsymbol{p}||\nabla\Psi^*(\boldsymbol{\theta})) \leq B_\Psi(\boldsymbol{y}||\nabla\Psi^*(\boldsymbol{\theta})).$$

Using (14), we obtain for all $\boldsymbol{y} \in \mathcal{C} \subseteq \mathsf{dom}(\Psi)$:

$$0 \leq B_\Psi(\boldsymbol{y}||\boldsymbol{p}) = B_\Omega(\boldsymbol{y}||\boldsymbol{p}) \leq L_\Omega(\boldsymbol{\theta};\boldsymbol{y}).$$

Since Ω is a l.s.c. proper convex function, from Proposition 2, we immediately get

$$\boldsymbol{p} = \boldsymbol{y} \Leftrightarrow L_\Omega(\boldsymbol{\theta};\boldsymbol{y}) = 0 \Leftrightarrow B_\Omega(\boldsymbol{y}||\boldsymbol{p}) = 0.$$

C.3 Proof of Proposition 4

The two facts stated in Proposition 4 (H is always non-negative and maximized by the uniform distribution) follow directly from Jensen's inequality. Indeed, for all $\boldsymbol{p} \in \triangle^d$:

- $\mathsf{H}(\boldsymbol{p}) \geq \sum_{j=1}^d p_j \mathsf{H}(\boldsymbol{e}_j) = 0$;
- $\mathsf{H}(\mathbf{1}/d) = \mathsf{H}\left(\sum_{\boldsymbol{P}\in\mathcal{P}} \frac{1}{d!}\boldsymbol{P}\boldsymbol{p}\right) \geq \sum_{\boldsymbol{P}\in\mathcal{P}} \frac{1}{d!}\mathsf{H}(\boldsymbol{P}\boldsymbol{p}) = \mathsf{H}(\boldsymbol{p})$,

where $\mathcal{P}$ is the set of $d \times d$ permutation matrices. Strict concavity ensures that $\boldsymbol{p} = \mathbf{1}/d$ is the unique maximizer.

C.4 Proof of Proposition 5

Let $\Omega(\boldsymbol{p}) = \sum_{j=1}^d g(p_j) + I_{\triangle^d}(\boldsymbol{p})$, where $g\colon [0,1] \to \mathbb{R}_+$ is a non-negative, strictly convex, differentiable function. Therefore, g' is strictly monotonic on $[0,1]$, thus invertible. We show how computing $\nabla(\Omega)^*$ reduces to finding the root of a monotonic scalar function, for which efficient algorithms are available.

From strict convexity and the definition of the convex conjugate,

$$\nabla\Omega^*(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{p}\in\triangle^d} \langle \boldsymbol{p}, \boldsymbol{\theta}\rangle - \sum_j g(p_j).$$

The constrained optimization problem above has Lagrangian

$$\mathcal{L}(\boldsymbol{p}, \boldsymbol{\nu}, \tau) \coloneqq \sum_{j=1}^d g(p_j) - \langle \boldsymbol{\theta} + \boldsymbol{\nu}, \boldsymbol{p}\rangle + \tau(\mathbf{1}^\top \boldsymbol{p} - 1).$$

A solution $(\boldsymbol{p}^\star, \boldsymbol{\nu}^\star, \tau^\star)$ must satisfy the KKT conditions

$$\begin{cases} g'(p_j) - \theta_j - \nu_j + \tau = 0 \qquad \forall j \in [d] \\ \langle \boldsymbol{p}, \boldsymbol{\nu}\rangle = 0 \\ \boldsymbol{p} \in \triangle^d,\ \boldsymbol{\nu} \geq 0. \end{cases} \tag{15}$$

Let us define

$$\tau_{\min} \coloneqq \max(\boldsymbol{\theta}) - g'(1) \quad \text{and} \quad \tau_{\max} \coloneqq \max(\boldsymbol{\theta}) - g'\left(\frac{1}{d}\right).$$

Since g is strictly convex, g' is increasing and so $\tau_{\min} < \tau_{\max}$. For any $\tau \in [\tau_{\min}, \tau_{\max}]$, we construct $\boldsymbol{\nu}$ as

$$\nu_j \coloneqq \begin{cases} 0, & \theta_j - \tau \geq g'(0) \\ g'(0) - \theta_j + \tau, & \theta_j - \tau < g'(0) \end{cases}$$

By construction, $\nu_j \geq 0$, satisfying dual feasability. Injecting ν into (15) and combining the two cases, we obtain

$$g'(p_j) = \max\{\theta_j - \tau,\ g'(0)\}. \tag{16}$$

We show that i) the stationarity conditions have a unique solution given τ, and ii) $[\tau_{\min}, \tau_{\max}]$ forms a sign-changing bracketing interval, and thus contains $\tau^\star$, which can then be found by one-dimensional search. The solution verifies all KKT conditions, thus is globally optimal.

Solving the stationarity conditions. Since g is strictly convex, its derivative g' is continuous and strictly increasing, and is thus a one-to-one mapping between $[0,1]$ and $[g'(0), g'(1)]$. Denote by $(g')^{-1} \colon [g'(0), g'(1)] \to [0,1]$ its inverse. If $\theta_j - \tau \geq g'(0)$, we have

$$\begin{aligned} g'(0) \leq g'(p_j) = \theta_j - \tau &\leq \max(\boldsymbol{\theta}) - \tau_{\min} \\ &= \max(\boldsymbol{\theta}) - \max(\boldsymbol{\theta}) + g'(1) \\ &= g'(1). \end{aligned}$$

Otherwise, $g'(p_j) = g'(0)$. This verifies that the r.h.s. of (16) is always within the domain of $(g')^{-1}$. We can thus apply the inverse to both sides to solve for p_j, obtaining

$$p_j(\tau) = (g')^{-1}(\max\{\theta_j - \tau,\ g'(0)\}). \tag{17}$$

Strict convexity implies the optimal $\boldsymbol{p}^\star$ is unique; it can be seen that $\tau^\star$ is also unique. Indeed, assume optimal $\tau_1^\star, \tau_2^\star$. Then, $\boldsymbol{p}(\tau_1^\star) = \boldsymbol{p}(\tau_2^\star)$, so $\max(\boldsymbol{\theta} - \tau_1^\star, g'(0)) = \max(\boldsymbol{\theta} - \tau_2^\star, g'(0))$. This implies either $\tau_1^\star = \tau_2^\star$, or $\boldsymbol{\theta} - \tau^\star_{\{1,2\}} \leq g'(0)$, in which case $\boldsymbol{p} = \mathbf{0} \notin \triangle^d$, which is a contradiction.

Validating the bracketing interval. Consider the primal infeasability function $\phi(\tau) \coloneqq \langle \boldsymbol{p}(\tau), \mathbf{1} \rangle - 1$; $\boldsymbol{p}(\tau)$ is primal feasible iff $\phi(\tau) = 0$. We show that ϕ is decreasing on $[\tau_{\min}, \tau_{\max}]$, and that it has opposite signs at the two extremities. From the intermediate value theorem, the unique root $\tau^\star$ must satisfy $\tau^\star \in [\tau_{\min}, \tau_{\max}]$.

Since g' is increasing, so is $(g')^{-1}$. Therefore, for all j, $p_j(\tau)$ is decreasing, and so is the sum $\phi(\tau) = \sum_j p_j(\tau) - 1$. It remains to check the signs at the boundaries.

$$\begin{aligned} \sum_i p_i(\tau_{\max}) &= \sum_i (g')^{-1}(\max\{\theta_i - \max(\boldsymbol{\theta}) + g'\left(1/d\right),\ g'(0)\}) \\ &\leq d\ (g')^{-1}(\max\{g'\left(1/d\right),\ g'(0)\}) \\ &= d\ (g')^{-1}\left(g'\left(1/d\right)\right) = 1, \end{aligned}$$

where we upper-bounded each term of the sum by the largest one. At the other end,

$$\begin{aligned} \sum_i p_i(\tau_{\min}) &= \sum_i (g')^{-1}(\max\{\theta_i - \max(\boldsymbol{\theta}) + g'(1),\ g'(0)\}) \\ &\geq (g')^{-1}(\max\{g'(1),\ g'(0)\}) \\ &= (g')^{-1}(g'(1)) = 1, \end{aligned}$$

using that a sum of non-negative terms is no less than its largest term. Therefore, $\phi(\tau_{\min}) \geq 0$ and $\phi(\tau_{\max}) \leq 0$. This implies that there must exist $\tau^\star$ in $[\tau_{\min}, \tau_{\max}]$ satisfying $\phi(\tau^\star) = 0$. The corresponding triplet $(\boldsymbol{p}(\tau^\star), \boldsymbol{\nu}(\tau^\star), \tau^\star)$ thus satisfies all of the KKT conditions, confirming that it is the global solution.

Algorithm 1 is an example of a bisection algorithm for finding an approximate solution; more advanced root finding methods can also be used. We note that the resulting algorithm resembles the method provided in Krichene et al. (2015), with a non-trivial difference being the order of the thresholding and $(-g)^{-1}$ in Eq. (17).

Algorithm 1: Bisection for $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \nabla\Omega^*(\boldsymbol{\theta})$

Input: $\boldsymbol{\theta} \in \mathbb{R}^d$, $\Omega(\boldsymbol{p}) = I_{\triangle^d} + \sum_i g(p_i)$
$\boldsymbol{p}(\tau) \coloneqq (g')^{-1}(\max\{\boldsymbol{\theta} - \tau, g'(0)\})$
$\phi(\tau) \coloneqq \langle \boldsymbol{p}(\tau), \mathbf{1} \rangle - 1$
$\tau_{\min} \leftarrow \max(\boldsymbol{\theta}) - g'(1)$;
$\tau_{\max} \leftarrow \max(\boldsymbol{\theta}) - g'\left(1/d\right)$
$\tau \leftarrow (\tau_{\min} + \tau_{\max})/2$
while $|\phi(\tau)| > \epsilon$
  if $\phi(\tau) < 0$ $\tau_{\max} \leftarrow \tau$
  else $\tau_{\min} \leftarrow \tau$
  $\tau \leftarrow (\tau_{\min} + \tau_{\max})/2$
Output: $\nabla\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \approx \boldsymbol{p}(\tau)$

C.5 Proof of Proposition 6

We start by proving the following lemma.

Lemma 1 *Let H satisfy assumptions A.1–A.3. Then:*

1. *We have $\boldsymbol{\theta} \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$ iff $\theta_k = (-\mathsf{H})^*(\boldsymbol{\theta})$. That is:*

$$\partial(-\mathsf{H})(\boldsymbol{e}_k) = \{\boldsymbol{\theta} \in \mathbb{R}^d \colon \theta_k \geq \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle + \mathsf{H}(\boldsymbol{p}),\ \forall \boldsymbol{p} \in \triangle^d\}.$$

2. *If $\boldsymbol{\theta} \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$, then, we also have $\boldsymbol{\theta}' \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$ for any $\boldsymbol{\theta}'$ such that $\theta'_k = \theta_k$ and $\theta'_i \leq \theta_i$, for all $i \neq k$.*

Proof of the lemma: Let $\Omega = -\mathsf{H}$. From Proposition 1 (order preservation), we can consider $\partial\Omega(\boldsymbol{e}_1)$ without loss of generality, in which case any $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{e}_1)$ satisfies $\theta_1 = \max_j \theta_j$. We have $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{e}_1)$ iff $\Omega(\boldsymbol{e}_1) = \langle \boldsymbol{\theta}, \boldsymbol{e}_1 \rangle - \Omega^*(\boldsymbol{\theta}) = \theta_1 - \Omega^*(\boldsymbol{\theta})$. Since $\Omega(\boldsymbol{e}_1) = 0$, we must have $\theta_1 = \Omega^*(\boldsymbol{\theta}) \geq \sup_{\boldsymbol{p} \in \triangle^d} \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle - \Omega(\boldsymbol{p})$, which proves part 1. To see 2, note that we have $\theta'_k = \theta_k \geq \langle \boldsymbol{\theta}, \boldsymbol{p} \rangle - \Omega(\boldsymbol{p}) \geq \langle \boldsymbol{\theta}', \boldsymbol{p} \rangle - \Omega(\boldsymbol{p})$, for all $\boldsymbol{p} \in \triangle^d$, from which the result follows. ■

We now proceed to the proof of Proposition 6. Let $\Omega = -\mathsf{H}$, and suppose that L_Ω has the separation margin property. Then, $\boldsymbol{\theta} = m\boldsymbol{e}_1$ satisfies the margin condition $\theta_1 \geq m + \max_{j \neq 1} \theta_j$, hence $L_\Omega(m\boldsymbol{e}_1, \boldsymbol{e}_1) = 0$. From the first part of Proposition 2, this implies $m\boldsymbol{e}_1 \in \partial\Omega(\boldsymbol{e}_1)$.

Conversely, let us assume that $m\boldsymbol{e}_1 \in \partial\Omega(\boldsymbol{e}_1)$. From the second part of Lemma 1, this implies that $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{e}_1)$ for any $\boldsymbol{\theta}$ such that $\theta_1 = m$ and $\theta_i \leq 0$ for all $i \geq 2$; and more generally we have $\boldsymbol{\theta} + c\mathbf{1} \in \partial\Omega(\boldsymbol{e}_1)$. That is, any $\boldsymbol{\theta}$ with $\theta_1 \geq m + \max_{i \neq 1} \theta_i$ satisfies $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{e}_1)$. From Proposition 2, this is equivalent to $L_\Omega(\boldsymbol{\theta}; \boldsymbol{e}_1) = 0$.

Let us now determine the margin of L_Ω, i.e., the smallest m such that $m\boldsymbol{e}_1 \in \partial\Omega(\boldsymbol{e}_1)$. From Lemma 1, this is equivalent to $m \geq mp_1 - \Omega(\boldsymbol{p})$ for any $\boldsymbol{p} \in \triangle^d$, i.e., $-\Omega(\boldsymbol{p})(1 - p_1) \leq m$. Note that by Proposition 1 the "most competitive" $\boldsymbol{p}$'s are sorted as $\boldsymbol{e}_1$, so we may write $p_1 = \|\boldsymbol{p}\|_\infty$ without loss of generality. The margin of L_Ω is the smallest possible such margin, given by (9).

C.6 Proof of Proposition 7

Let us start by showing that conditions 1 and 2 are equivalent. To show that 2 ⇒ 1, take an arbitrary $\boldsymbol{p} \in \triangle^d$. From Fenchel-Young duality and the Danskin's theorem, we have that $\nabla(-\mathsf{H})^*(\boldsymbol{\theta}) = \boldsymbol{p} \ \Rightarrow\ \boldsymbol{\theta} \in \partial(-\mathsf{H})(\boldsymbol{p})$, which implies the subdifferential set is non-empty everywhere in the simplex. Let us now prove that 1 ⇒ 2. Let $\Omega = -\mathsf{H}$, and assume that Ω has non-empty subdifferential everywhere in $\triangle^d$. We need to show that for any $\boldsymbol{p} \in \triangle^d$, there is some $\boldsymbol{\theta} \in \mathbb{R}^d$ such that $\boldsymbol{p} \in \mathsf{argmin}_{\boldsymbol{p}' \in \triangle^d} \Omega(\boldsymbol{p}') - \langle \boldsymbol{\theta}, \boldsymbol{p}' \rangle$. The Lagrangian associated with this minimization problem is:

$$\mathcal{L}(\boldsymbol{p}, \boldsymbol{\mu}, \lambda) = \Omega(\boldsymbol{p}) - \langle \boldsymbol{\theta} + \boldsymbol{\mu}, \boldsymbol{p} \rangle + \lambda(\mathbf{1}^\top \boldsymbol{p} - 1).$$

The KKT conditions are:

$$\begin{cases} 0 \in \partial_p \mathcal{L}(\boldsymbol{p}, \boldsymbol{\mu}, \lambda) = \partial\Omega(\boldsymbol{p}) - \boldsymbol{\theta} - \boldsymbol{\mu} + \lambda\mathbf{1} \\ \langle \boldsymbol{p}, \boldsymbol{\mu} \rangle = 0 \\ \boldsymbol{p} \in \triangle^d,\ \boldsymbol{\mu} \geq 0. \end{cases}$$

For a given $\boldsymbol{p} \in \triangle^d$, we seek $\boldsymbol{\theta}$ such that $(\boldsymbol{p}, \boldsymbol{\mu}, \lambda)$ are a solution to the KKT conditions for some $\boldsymbol{\mu} \geq 0$ and $\lambda \in \mathbb{R}$.

We will show that such $\boldsymbol{\theta}$ exists by simply choosing $\boldsymbol{\mu} = \mathbf{0}$ and $\lambda = 0$. Those choices are dual feasible and guarantee that the slackness complementary condition is satisfied. In this case, we have from the first condition that $\boldsymbol{\theta} \in \partial\Omega(\boldsymbol{p})$. From the assumption that Ω has non-empty subdifferential in all the simplex, we have that for any $\boldsymbol{p} \in \triangle^d$ we can find a $\boldsymbol{\theta} \in \mathbb{R}^d$ such that $(\boldsymbol{p}, \boldsymbol{\theta})$ are a dual pair, i.e., $\boldsymbol{p} = \nabla\Omega^*(\boldsymbol{\theta})$, which proves that $\nabla\Omega^*(\mathbb{R}^d) = \triangle^d$.

Next, we show that condition 1 ⇒ 3. Since $\partial(-\mathsf{H})(\boldsymbol{p}) \neq \varnothing$ everywhere in the simplex, we can take an arbitrary $\boldsymbol{\theta} \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$. From Lemma 1, item 2, we have that $\boldsymbol{\theta}' \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$ for $\theta'_k = \theta_k$ and $\theta'_j = \min_\ell \theta_\ell$; since

$(-\mathsf{H})^*$ is shift invariant, we can without loss of generality have $\theta' = m\boldsymbol{e}_k$ for some $m > 0$, which implies from Proposition 6 that L_Ω has a margin.

Let us show that, if $-\mathsf{H}$ is separable, then $3 \Rightarrow 1$, which establishes equivalence between all conditions 1, 2, and 3. From Proposition 6, the existing of a separation margin implies that there is some m such that $m\boldsymbol{e}_k \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$. Let $\mathsf{H}(\boldsymbol{p}) = \sum_{i=1}^d h(p_i)$, with $h : [0,1] \to \mathbb{R}_+$ concave. Due to assumption A.1, h must satisfy $h(0) = h(1) = 0$. Without loss of generality, suppose $\boldsymbol{p} = [\tilde{\boldsymbol{p}}; \boldsymbol{0}_k]$, where $\tilde{\boldsymbol{p}} \in \mathsf{relint}(\triangle^{d-k})$ and $\boldsymbol{0}_k$ is a vector with k zeros. We will see that there is a vector $\boldsymbol{g} \in \mathbb{R}^d$ such that $\boldsymbol{g} \in \partial(-\mathsf{H})(\boldsymbol{p})$, i.e., satisfying

$$-\mathsf{H}(\boldsymbol{p}') \geq -\mathsf{H}(\boldsymbol{p}) + \langle \boldsymbol{g}, \boldsymbol{p}' - \boldsymbol{p} \rangle, \quad \forall \boldsymbol{p}' \in \triangle^d. \tag{19}$$

Since $\tilde{\boldsymbol{p}} \in \mathsf{relint}(\triangle^{d-k})$, we have $\tilde{p}_i \in]0,1[$ for $i \in \{1, \dots, d-k\}$, hence $\partial(-h)(\tilde{p}_i)$ must be nonempty, since $-h$ is convex and $]0,1[$ is an open set. We show that the following $\boldsymbol{g} = (g_1, \dots, g_d) \in \mathbb{R}^d$ is a subgradient of $-\mathsf{H}$ at $\boldsymbol{p}$:

$$g_i = \begin{cases} \partial(-h)(\tilde{p}_i), & i = 1, \dots, d-k \\ m, & i = d-k+1, \dots, d. \end{cases}$$

By definition of subgradient, we have

$$-\psi(p_i') \geq -\psi(\tilde{p}_i) + \partial(-h)(\tilde{p}_i)(p_i' - \tilde{p}_i), \quad \text{for } i = 1, \dots, d-k. \tag{20}$$

Furthermore, since m upper bounds the separation margin of H, we have from Proposition 6 that $m \geq \frac{\mathsf{H}([1-p_i', p_i', 0, \dots, 0])}{1-\max\{1-p_i', p_i'\}} = \frac{h(1-p_i')+h(p_i')}{\min\{p_i', 1-p_i'\}} \geq \frac{h(p_i')}{p_i'}$ for any $p_i' \in]0,1]$. Hence, we have

$$-\psi(p_i') \geq -\psi(0) - m(p_i' - 0), \quad \text{for } i = d-k+1, \dots, d. \tag{21}$$

Summing all inequalities in Eqs. (20)–(21), we obtain the expression in Eq. (19), which finishes the proof.

C.7 Proof of Proposition 8

Define $\Omega = -\mathsf{H}$. Let us start by writing the margin expression (9) as a unidimensional optimization problem. This is done by noticing that the max-generalized entropy problem constrained to $\max(\boldsymbol{p}) = 1 - t$ gives $\boldsymbol{p} = \left[1-t, \frac{t}{d-1}, \dots, \frac{t}{d-1}\right]$, for $t \in \left[0, 1 - \frac{1}{d}\right]$ by a similar argument as the one used in Proposition 4. We obtain:

$$\operatorname{margin}(L_\Omega) = \sup_{t \in \left[0, 1-\frac{1}{d}\right]} \frac{-\Omega\left(\left[1-t, \frac{t}{d-1}, \dots, \frac{t}{d-1}\right]\right)}{t}.$$

We write the argument above as $A(t) = \frac{-\Omega(\boldsymbol{e}_1 + t\boldsymbol{v})}{t}$, where $\boldsymbol{v} := [-1, \frac{1}{d-1}, \dots, \frac{1}{d-1}]$. We will first prove that A is decreasing in $[0, 1 - \frac{1}{d}]$, which implies that the supremum (and the margin) equals $A(0)$. Note that we have the following expression for the derivative of any function $f(\boldsymbol{e}_1 + t\boldsymbol{v})$:

$$(f(\boldsymbol{e}_1 + t\boldsymbol{v}))' = \boldsymbol{v}^\top \nabla f(\boldsymbol{e}_1 + t\boldsymbol{v}).$$

Using this fact, we can write the derivative $A'(t)$ as:

$$A'(t) = \frac{-t\boldsymbol{v}^\top \nabla\Omega(\boldsymbol{e}_1 + t\boldsymbol{v}) + \Omega(\boldsymbol{e}_1 + t\boldsymbol{v})}{t^2} := \frac{B(t)}{t^2}.$$

In turn, the derivative $B'(t)$ is:

$$\begin{aligned} B'(t) &= -\boldsymbol{v}^\top \nabla\Omega(\boldsymbol{e}_1 + t\boldsymbol{v}) - t(\boldsymbol{v}^\top \nabla\Omega(\boldsymbol{e}_1 + t\boldsymbol{v}))' + \boldsymbol{v}^\top \nabla\Omega(\boldsymbol{e}_1 + t\boldsymbol{v}) \\ &= -t(\boldsymbol{v}^\top \nabla\Omega(\boldsymbol{e}_1 + t\boldsymbol{v}))' \\ &= -t\boldsymbol{v}^\top \nabla\nabla\Omega(\boldsymbol{e}_1 + t\boldsymbol{v})\boldsymbol{v} \\ &\leq 0, \end{aligned}$$

where we denote by $\nabla\nabla\Omega$ the Hessian of Ω, and used the fact that it is positive semi-definite, due to the convexity of Ω. This implies that B is decreasing, hence for any $t \in [0,1]$, $B(t) \leq B(0) = \Omega(\boldsymbol{e}_1) = 0$, where we used the fact $\|\nabla\Omega(\boldsymbol{e}_1)\| < \infty$, assumed as a condition of Proposition 7. Therefore, we must also have $A'(t) = \frac{B(t)}{t^2} \leq 0$ for any $t \in [0,1]$, hence A is decreasing, and $\sup_{t\in[0,1-1/d]} A(t) = \lim_{t\to 0+} A(t)$. By L'Hôpital's rule:

$$\begin{aligned}
\lim_{t\to 0+} A(t) &= \lim_{t\to 0+} (-\Omega(\boldsymbol{e}_1 + t\boldsymbol{v}))' \\
&= -\boldsymbol{v}^\top \nabla\Omega(\boldsymbol{e}_1) \\
&= \nabla_1\Omega(\boldsymbol{e}_1) - \frac{1}{d-1}\sum_{j\geq 2} \nabla_j\Omega(\boldsymbol{e}_1) \\
&= \nabla_1\Omega(\boldsymbol{e}_1) - \nabla_2\Omega(\boldsymbol{e}_1),
\end{aligned}$$

which proves the first part.

If Ω is separable, then $\nabla_j\Omega(\boldsymbol{p}) = -h'(p_j)$, in particular $\nabla_1\Omega(\boldsymbol{e}_1) = -h'(1)$ and $\nabla_2\Omega(\boldsymbol{e}_1) = -h'(0)$, yielding $\text{margin}(L_\Omega) = h'(0) - h'(1)$. Since h is twice differentiable, this equals $-\int_0^1 h''(t)dt$, completing the proof.