A Deferred proofs

A.1 Proof of Corollary 1

Consider two potential "bad" events: B_1, where $\|\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}\|_\infty > 1/2$, and B_2, where $\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_n\|_1 > \frac{2\|\hat{\boldsymbol{\mu}}_n\|_{1/2}^{1/2}}{\sqrt{n}} + 6\sqrt{\frac{\ln(4/\delta)}{2n}}$. Our assumption on the sample size n, together with the Dvoretzky-Kiefer-Wolfowitz inequality [Massart, 1990], implies that $\mathbb{P}(B_1) \le \delta/2$ and (1) implies that $\mathbb{P}(B_2) \le \delta/2$. Thus, with probability at least $1-\delta$, neither of B_1 or B_2 occurs, and we may invoke Theorem 1, from which the claim immediately follows.

□

A.2 Proof of Theorem 2

The following non-trivial fact [Lieb and Loss, 2001, Theorem 3.5 and Eq. (5) on p. 83] will be useful[6]:

$$\|\boldsymbol{\mu}^\downarrow - \boldsymbol{\nu}^\downarrow\|_p \quad \le \quad \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_p, \qquad p \in [1,\infty],\ \boldsymbol{\mu},\boldsymbol{\nu} \in \Delta_{\mathbb{N}}. \tag{11}$$

A result of Scheffé [1947] (more accurately credited to Riesz, 1928 [Kusolitsch, 2010]) implies that a sequence $\{\xi_{n\in\mathbb{N}}\} \subset \ell_1(\mathbb{N})$ converging pointwise to some $\xi \in \ell_1(\mathbb{N})$ also converges in ℓ_1 iff $\|\xi_n\|_1 \to \|\xi\|_1$. This immediately implies

Lemma 3. *If $\{\boldsymbol{\mu}_{n\in\mathbb{N}}\} \subset \Delta_{\mathbb{N}}$ converges pointwise to some $\boldsymbol{\mu} \in \Delta_{\mathbb{N}}$, then it also converges in ℓ_1.*

Berend et al. [2017, Lemma 1] showed that $\Delta_{\mathbb{N}}^{\downarrow(1)}[h]$ is compact under ℓ_1. We begin by extending this result to general α, p.

Lemma 4. *For all $\alpha \ge 1$, $p \in [1,\infty]$, and $h > 0$, the set $\Delta_{\mathbb{N}}^{\downarrow(\alpha)}[h]$ is compact under ℓ_p.*

Remark. This is quite false if either the non-increasing or the bounded-entropy condition is omitted. For a counterexample to the former, consider the sequence $\boldsymbol{\mu}_n \in \Delta_{\mathbb{N}}$ defined by $\boldsymbol{\mu}_n(i) = \mathbf{1}[i = n]$. For a counterexample to the latter, consider the sequence $\boldsymbol{\mu}_n \in \Delta_{\mathbb{N}}$, where $\boldsymbol{\mu}_n$ is uniform on $[n]$.

Proof. We closely follow the proof strategy of Berend et al. [2017, Lemma 1]. In a metric space, compactness and sequential compactness are equivalent. Let $\boldsymbol{\mu}_{n\in\mathbb{N}}$ be a sequence in $\Delta_{\mathbb{N}}^{\downarrow(\alpha)}[h]$. Since $[0,1]$ is compact, every $\{\boldsymbol{\mu}_n(i) : n \in \mathbb{N}\}$ has a convergent subsequence, and hence $\boldsymbol{\mu}_{n\in\mathbb{N}}$ has a pointwise convergent subsequence. There is thus no loss of generality in assuming that $\boldsymbol{\mu}_n \to \boldsymbol{\mu}$ pointwise. Obviously, $\boldsymbol{\mu}$ is non-negative and non-increasing. It remains to show that

(a) $\sum_{i\in\mathbb{N}} \boldsymbol{\mu}(i) = 1$,

(b) $\mathrm{H}^{(\alpha)}(\boldsymbol{\mu}) \le h$,

(c) $\|\boldsymbol{\mu}_n - \boldsymbol{\mu}\|_p \to 0$.

To show (a), assume, for a contradiction, that $\sum_{i\in\mathbb{N}} \boldsymbol{\mu}(i) > 1$. Then there must be an $i_0 \in \mathbb{N}$ such that $\sum_{i=1}^{i_0} \boldsymbol{\mu}(i) > 1$. But the latter must then hold for all $\boldsymbol{\mu}_n$ with n sufficiently large, which contradicts $\boldsymbol{\mu}_n \in \Delta_{\mathbb{N}}$. Now assume $\varepsilon := 1 - \sum_{i\in\mathbb{N}} \boldsymbol{\mu}(i) > 0$. For any $i_0 \in \mathbb{N}$, we have $\sum_{i=1}^{i_0} \boldsymbol{\mu}_n(i) < 1 - \varepsilon/2$ for all sufficiently large n. Now every $\boldsymbol{\nu} \in \Delta_{\mathbb{N}}^{\downarrow}$ satisfies $\boldsymbol{\nu}(i) \le \frac{1}{i}(\boldsymbol{\nu}(1) + \boldsymbol{\nu}(2) + \ldots + \boldsymbol{\nu}(i)) \le \frac{1}{i}$. Hence,

$$\sum_{i=i_0+1}^{\infty} \boldsymbol{\mu}_n(i)\,|\log \boldsymbol{\mu}_n(i)|^\alpha \ge \sum_{i=i_0+1}^{\infty} \boldsymbol{\mu}_n(i)(\log i_0)^\alpha > \frac{\varepsilon}{2}(\log i_0)^\alpha.$$

[6] The result is stated for functions in $f \in L_2(\mathbb{R}^n)$ and their symmetric-decreasing rearrangements f^*, but the specialization to discrete distributions is straightforward. We convert $\boldsymbol{\mu}$ to a function $f : \mathbb{R}_+ \to \mathbb{R}_+$ via $f(x) = \mu(\lceil x \rceil)$ and $\boldsymbol{\nu}$ to $g(x)$ analogously. A direct calculation then shows that $\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_p = \|f - g\|_p$ and $\|\boldsymbol{\mu}^\downarrow - \boldsymbol{\nu}^\downarrow\|_p = \|f^* - g^*\|_p$, to which the result from Lieb and Loss [2001] applies to yield (11).

Choosing i_0 sufficiently large makes the latter expression exceed h, violating the assumption $\boldsymbol{\mu}_n \in \Delta_{\mathbb{N}}^{\downarrow(\alpha)}[h]$. Thus (a) holds.

To show (b), assume for a contradiction that $\mathrm{H}^{(\alpha)}(\boldsymbol{\mu}) > h$ — and, in particular, $\sum_{i=1}^{i_0} \boldsymbol{\mu}(i) \left|\log \boldsymbol{\mu}(i)\right|^\alpha > h$ for some $i_0 \in \mathbb{N}$. But the latter must hold for all $\boldsymbol{\mu}_n$ with n sufficiently large, a contradiction.

Finally, to show (c), we invoke Lemma 3: if $\{\boldsymbol{\mu}_{n\in\mathbb{N}}\} \subset \Delta_{\mathbb{N}}$ converges pointwise to some $\boldsymbol{\mu} \in \Delta_{\mathbb{N}}$, then it also converges in ℓ_1. Since ℓ_1 dominates every ℓ_p, $p > 1$, this proves (c). □

Next, we examine the continuity of $\mathrm{H}(\cdot)$ on $\Delta_{\mathbb{N}}^{\downarrow(\alpha)}[h]$ under ℓ_p.

Lemma 5. *Fix $h > 0$, $\alpha > 1$, and $p \in [1, \infty]$. If $\{\boldsymbol{\mu}_{n\in\mathbb{N}}\} \subset \Delta_{\mathbb{N}}^{\downarrow(\alpha)}[h]$ converges in ℓ_p, then its limit is some $\boldsymbol{\mu} \in \Delta_{\mathbb{N}}^{\downarrow(\alpha)}[h]$ and furthermore, $\mathrm{H}(\boldsymbol{\mu}_n) \to \mathrm{H}(\boldsymbol{\mu})$. In other words, $\mathrm{H}(\cdot)$ is continuous on $\Delta_{\mathbb{N}}^{\downarrow(\alpha)}[h]$ under ℓ_p.*

Remark. We note that $\mathrm{H}(\cdot)$ is not continuous on $\Delta_{\mathbb{N}}^{\downarrow(1)}[h]$ under ℓ_p, $p \in [1, \infty]$, as evidenced by the sequence $\boldsymbol{\mu}_n = (1 - \varepsilon_n, \varepsilon_n/n, \ldots, \varepsilon/n, 0, 0, \ldots)$, with support size $n + 1$. We can choose ε_n so that $\mathrm{H}(\boldsymbol{\mu}_n) = h$, but of course the limiting $\boldsymbol{\mu}$ has $\mathrm{H}(\boldsymbol{\mu}) = 0$ (see Example 1 in Berend et al. [2017]).

Proof. It follows from Lemma 4 that the limiting $\boldsymbol{\mu}$ belongs to $\Delta_{\mathbb{N}}^{\downarrow(\alpha)}[h]$. Further, Lemma 3 implies that $\boldsymbol{\mu}_n \to \boldsymbol{\mu}$ in ℓ_1. Invoking the continuity result in Theorem 1 proves the claim. □

Proof of Theorem 2. It follows from Lemma 5 that $\mathrm{H}(\cdot)$ is continuous on $\Delta_{\mathbb{N}}^{\downarrow(\alpha)}[h]$ under ℓ_p. Since, by Lemma 4, $\Delta_{\mathbb{N}}^{\downarrow(\alpha)}[h]$ is compact under ℓ_p, it follows that $\mathrm{H}(\cdot)$ is uniformly continuous on $\Delta_{\mathbb{N}}^{\downarrow(\alpha)}[h]$: there is a function F such that

$$\left|\mathrm{H}(\boldsymbol{\mu}) - \mathrm{H}(\boldsymbol{\nu})\right| \;\leq\; F(\left\|\boldsymbol{\mu} - \boldsymbol{\nu}\right\|_p, h, \alpha, p), \qquad \boldsymbol{\mu}, \boldsymbol{\nu} \in \Delta_{\mathbb{N}}^{\downarrow(\alpha)}[h]$$

and $\varepsilon_n := \left\|\boldsymbol{\mu}_n - \boldsymbol{\nu}_n\right\|_p \to 0 \implies F(\varepsilon_n, h, \alpha, p) \to 0$. Now, for all $\boldsymbol{\mu}, \boldsymbol{\nu} \in \Delta_{\mathbb{N}}^{(\alpha)}[h]$ we have

$$\begin{aligned} \left|\mathrm{H}(\boldsymbol{\mu}) - \mathrm{H}(\boldsymbol{\nu})\right| &= \left|\mathrm{H}(\boldsymbol{\mu}^\downarrow) - \mathrm{H}(\boldsymbol{\nu}^\downarrow)\right| \\ &\leq F(\|\boldsymbol{\mu}^\downarrow - \boldsymbol{\nu}^\downarrow\|_p, h, \alpha, p). \end{aligned}$$

It follows from (11) that $\left\|\boldsymbol{\mu}_n - \boldsymbol{\nu}_n\right\|_p \to 0 \implies \|\boldsymbol{\mu}^\downarrow - \boldsymbol{\nu}^\downarrow\|_p \to 0$, which concludes the proof. □

B Auxiliary results

Proposition 2. *For $K \geq 2$ and $\alpha \geq 1$,*

$$\max\left\{\log K, (\alpha/\mathrm{e})\right\}^\alpha \;\leq\; \max_{\boldsymbol{\mu}\in\Delta_K} \mathrm{H}^{(\alpha)}(\boldsymbol{\mu}) \;\leq\; \max\left\{\alpha, \log K\right\}^\alpha + (\alpha/\mathrm{e})^\alpha.$$

We will need the following useful (and likely known) result.

Lemma 6 (folklore)**.** *Suppose that $0 < a < 1$ and $f : [0, 1] \to \mathbb{R}$ is strictly concave on $[0, a]$ and strictly convex on $[a, 1]$. Define the function $F : \Delta_K \to \mathbb{R}$ by*

$$F(\boldsymbol{\mu}) = \sum_{i=1}^{K} f(\boldsymbol{\mu}(i)).$$

Then any maximizer $\boldsymbol{\mu}^\star$ of F is either the uniform distribution or else has exactly 1 "heavy" mass $v \in [a, 1]$ and $K - 1$ identical "light" masses $(1 - v)/(K - 1)$.

Proof. A standard "smoothing" argument [Loh, 2013] shows that if two masses $u \leq v$ occur in the interval $(a, 1)$, there is an $\varepsilon > 0$ such that $f(u - \varepsilon) + f(v + \varepsilon) > f(u) + f(v)$. In other words, such masses can be pushed apart (keeping their sum fixed) to increase the value of F, until one of them reaches the boundary of $[a, 1]$. Furthermore, since $0 < a < u < v$ and $u + v \leq 1$, repeated

iteration of the "pushing apart" operation will hit the left endpoint (i.e., a) rather than the right one (i.e., 1). Having exhausted the "pushing apart" process, we are left with one "heavy" mass $v \in [a, 1]$ and $K-1$ "lighter" ones in $[0, a]$. But concavity implies that F will be maximized by pulling the lighter masses in (as opposed to pushing them apart), which amounts to replacing each of them by the average of the $K-1$ values. □

Proof of Proposition 2. Choosing $\boldsymbol{\mu}$ to be the uniform distribution yields $\mathrm{H}^{(\alpha)}(\boldsymbol{\mu}) = \log^{\alpha} K$, and choosing $\boldsymbol{\mu}$ such that $v := \boldsymbol{\mu}(1) = \mathrm{e}^{-\alpha}$ yields $\mathrm{H}^{(\alpha)}(\boldsymbol{\mu}) \geq v \log(1/v)^{\alpha} = (\alpha/\mathrm{e})^{\alpha}$. Thus, the lower bound is proven and it only remains to prove the upper bound.

Let $\boldsymbol{\mu}^{\star}$ be a maximizer for given α, K. Recall the function $\mathrm{h}^{(\alpha)}(z) = z \log^{\alpha}(1/z)$ and note that it is strictly concave on $[0, \mathrm{e}^{-(\alpha-1)}]$ and strictly convex on $[\mathrm{e}^{-(\alpha-1)}, 1]$. Then Lemma 6 shows that $\boldsymbol{\mu}^{\star}$ will either be uniform or else attains at most one value $v \in [\mathrm{e}^{-(\alpha-1)}, 1]$ in the convex interval, with the remaining values equal to $\frac{1-v}{K-1} \in [0, \mathrm{e}^{-(\alpha-1)}]$ in the concave interval. Only the latter case is non-trivial:

$$\mathrm{H}^{(\alpha)}(\boldsymbol{\mu}^{\star}) \quad = \quad v \left(\log \frac{1}{v} \right)^{\alpha} + (1-v) \left(\log \frac{K-1}{1-v} \right)^{\alpha}$$

for some v satisfying

$$0 < \frac{1-v}{K-1} \leq \mathrm{e}^{-(\alpha-1)} \leq v < 1. \tag{12}$$

Now $v \left(\log \frac{1}{v} \right)^{\alpha}$ is maximized over $[0, 1]$ by $v = \mathrm{e}^{-\alpha}$, which yields the value $(\alpha/\mathrm{e})^{\alpha}$.

To bound the second term, $g(v) := (1-v) \left(\log \frac{K-1}{1-v} \right)^{\alpha}$, we consider two cases: (i) $K - 1 < \mathrm{e}^{\alpha}$ and (ii) $K - 1 \geq \mathrm{e}^{\alpha}$. In case (i), g is maximized by $v^{\star} = 1 - (K-1)/\mathrm{e}^{\alpha}$ and

$$g(v^{\star}) = (1-v^{\star}) \left(\log \frac{K-1}{1-v^{\star}} \right)^{\alpha} \leq \left(\log \frac{K-1}{1-v^{\star}} \right)^{\alpha} = \alpha^{\alpha}.$$

In case (ii), g is monotonically decreasing in v. The constraint $\frac{1-v}{K-1} \leq v$ from (12) implies $v \geq 1/K$, so in this case,

$$g(v) \leq \left(\log \frac{K-1}{1-1/K} \right)^{\alpha} = \log^{\alpha} K.$$

This proves the upper bound.

□