An Elementary Proof of the Near Optimality of LogSumExp Smoothing

Thabo Samakhoana* Benjamin Grimmer†

Abstract

We consider the design of smoothings of the (coordinate-wise) max function in $\mathbb{R}^d$ in the infinity norm. The LogSumExp function $f(x) = \ln(\sum_i^d \exp(x_i))$ provides a classical smoothing, differing from the max function in value by at most $\ln(d)$. We provide an elementary construction of a lower bound, establishing that every overestimating smoothing of the max function must differ by at least $\sim 0.8145 \ln(d)$. Hence, LogSumExp is optimal up to small constant factors. However, we provide strictly stronger smoothings showing the entropy-based LogSumExp approach is not exactly optimal. In small dimensions, we propose exactly optimal smoothings, attaining our lower bound.

1 Introduction

We consider the task of smoothing the coordinate-wise max function $\sigma_{\texttt{max}}(x) = \max_{i=1,\dots,d} x_i$ with respect to the infinity norm in d dimensions. Such smoothings play an important role in the acceleration of nonsmooth optimization algorithms [1] and recently in machine learning [2, 3]. In both of these settings, the function $f_{\texttt{LSE}}(x) = \ln(\sum_i^d \exp(x_i))$ has played a central role as the default "best" choice in the engineering of algorithms.

The canonical choice of the LogSumExp function $f_{\texttt{LSE}}$ (also known as the "softmax") is a convex overestimator of $\sigma_{\texttt{max}}$ that is 1-smooth with respect to the infinity norm:

$$\|\nabla f(x) - \nabla f(y)\|_1 \leq \|x - y\|_\infty \qquad \forall x, y \in \mathbb{R}^d. \tag{1.1}$$

Moreover, $f_{\texttt{LSE}}$ closely approximates the max function, having

$$\|f_{\texttt{LSE}} - \sigma_{\texttt{max}}\|_\infty = \ln(d)$$

where $\|f - \sigma\|_\infty := \sup_{x\in\mathbb{R}^d} |f(x) - \sigma(x)|$. Generally, we say that f is a δ-smoothing with respect to the infinity norm of the max function if f is convex and 1-smooth w.r.t. $\|\cdot\|_\infty$ with $\|f - \sigma_{\texttt{max}}\|_\infty \leq \delta$. Hence, $f_{\texttt{LSE}}$ is a $\ln(d)$-smoothing of $\sigma_{\texttt{max}}$.

Such δ-smoothings in the infinity norm play an important role for both of the applications mentioned above. Below we discuss these two core motivations:
(i) In nonsmooth optimization, a common form for a problem to take is minimizing a maximum of convex functions

$$\min_{y\in\mathbb{R}^n} \max_{i=1,\dots,d} g_i(y) = \sigma_{\texttt{max}} \circ g(y)$$

*Johns Hopkins University, Department of Applied Mathematics and Statistics, tsamakh1@jhu.edu
†Johns Hopkins University, Department of Applied Mathematics and Statistics, grimmer@jhu.edu

where $g(y) = [g_1(y), \ldots, g_d(y)]^T$. For general convex nonsmooth optimization, any first-order subgradient method seeking an ϵ-minimizer requires at least $\Omega(1/\epsilon^2)$ subgradient oracle queries. However, if each individual component objective g_i is M-Lipschitz and L-smooth with respect to the Euclidean norm in $\mathbb{R}^n$, then accelerated optimization algorithms from smooth convex optimization can be leveraged. In particular, [4] considered minimizing the smooth convex relaxation

$$\min_{y \in \mathbb{R}^n} \left(\frac{\epsilon}{2\delta} f\right) \circ \left(\frac{2\delta}{\epsilon} g\right)(y)$$

for a δ-smoothing f of $\sigma_{\texttt{max}}$. Since f is 1-smooth with respect to the infinity norm and each individual g_i is smooth and Lipschitz, the objective $y \mapsto \frac{\epsilon}{2\delta} f \circ \frac{2\delta}{\epsilon} g\,(y)$ above must be $(L + 2\delta M^2/\epsilon)$-smooth with respect to the Euclidean norm in $\mathbb{R}^n$ [4, Proposition 4.1]. Further, it differs from the original objective by at most $\epsilon/2$. As a result, by applying an accelerated algorithm for smooth convex optimization like Nesterov's method [5] to produce an $\epsilon/2$-minimizer, an ϵ-minimizer to the original problem can be produced with a first-order oracle complexity of at most

$$\sqrt{\frac{(L + 2\delta M^2/\epsilon)\|y_0 - y_\star\|^2}{\epsilon/2}} = \mathcal{O}\left(\frac{\sqrt{L\epsilon + \delta M^2}\|y_0 - y_\star\|}{\epsilon}\right). \tag{1.2}$$

Note the acceleration in the dependence on ϵ in the above $\mathcal{O}(1/\epsilon)$ rate over the general $\Omega(1/\epsilon^2)$ limit. Hence, δ-smoothings provide a direct reduction for accelerating structured nonsmooth optimization. Further, improvements in δ directly improve oracle complexities.

(ii) In neural network design, especially modern networks using the attention/transformer mechanism [6], it is often useful for the output of a layer in the network to be a vector of weights for a probability distribution. For example, this can correspond to the likelihood of outputting a given token. That is, one would like a mapping $g\colon \mathbb{R}^d \to \Delta_d$ where Δ_d is the probability simplex in dimension d. Gradients $\nabla f\colon \mathbb{R}^d \to \Delta_d$ of convex smoothings of the max function provide such a construction with desirable properties like monotonicity and Lipschitzness. Hence, there has been substantial interest in the machine learning literature in such smoothings [2, 3]. Since the outputs of the gradient map are probabilities (living in the simplex Δ_d), the typical norm to endow them with is the one-norm. Dually, the natural corresponding norm for the primal space is the infinity norm, as we study here. Note that by convexity, any δ-smoothing f has that

$$\langle \nabla f(x), x\rangle \geq \sigma_{\texttt{max}}(x) - 2\delta \qquad \forall x \in \mathbb{R}^d \tag{1.3}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner product, i.e., the expected value of sampling an element of x with probabilities $\nabla f(x)$ concentrates near the vector's maximum value. See [7] for a discussion of the potential importance of gains in δ.

For the task of smoothing in the infinity norm, in either of these disciplines, the LogSumExp function has emerged as the canonical standard choice. The fundamental nature of this choice is best seen by considering a dual perspective. Generally, convex conjugates $f^*(x) = \sup_{\lambda \in \mathbb{R}^d} \langle \lambda, x\rangle - f(\lambda)$ provide a duality between smooth convex functions and strongly convex functions. A Nesterov-style smoothing [1] of the max function is given by such a conjugate as

$$f(x) = \sup_{\lambda \in \Delta_d} \langle \lambda, x\rangle - h(\lambda) \tag{1.4}$$

for any 1-strongly convex "distance" function h on Δ_d in the one-norm. From this point of view, designing a smoothing corresponds to designing a strongly convex function on the simplex. If one takes $h(\lambda) = \sum_{i=1}^d \lambda_i \ln(\lambda_i)$ as the entropy function, a well-known strongly convex "distance" for the

probability simplex, then the resulting smoothing f is exactly $f_{\texttt{LSE}}$. Indeed, this development is at the origin of the LogSumExp function's usage in nonsmooth optimization [1]. Despite their ubiquity, the fundamental boundaries of what smoothings can attain remain underexplored.

In this work, we derive limits on the process of designing smooth approximations of the coordinate-wise max function. That is, we seek to derive lower bounds on the infinite-dimensional optimization problem of finding the best δ-smoothing in d dimensions

$$\delta^\star(d) := \begin{cases} \min_f & \|f - \sigma_{\texttt{max}}\|_\infty \\ \text{s.t.} & f\colon \mathbb{R}^d \to \mathbb{R} \text{ is convex and 1-smooth in } \|\cdot\|_\infty. \end{cases}$$

Equivalently, one could consider optimizing only over overestimators of $\sigma_{\texttt{max}}$, as $f_{\texttt{LSE}}$ does, denoted

$$\delta^\star_{\texttt{over}}(d) := \begin{cases} \min_f & \|f - \sigma_{\texttt{max}}\|_\infty \\ \text{s.t.} & f\colon \mathbb{R}^d \to \mathbb{R} \text{ is convex and 1-smooth in } \|\cdot\|_\infty \\ & f(x) \geq \sigma_{\texttt{max}}(x)\ \forall x \in \mathbb{R}^d. \end{cases}$$

By adding simple shifts, it is immediate that $\delta^\star(d) = \delta^\star_{\texttt{over}}(d)/2$. The existing results for the LogSumExp function $f_{\texttt{LSE}}$ discussed above then establish an upper bound for this problem of $\delta^\star(d) \leq 0.5\ln(d)$ for all d. We consider the complementary question of lower bounds.

1.1 Our Contributions

A lower bound on $\delta^\star(d)$ of $\ln(d)/8$ can be extracted from two related lower bounds in the literature. Epasto et al. [7] provided a lower bound via limits on any Lipschitz mappings satisfying (1.3). Their approach relies on a game-theoretic argument establishing the existence of a Nash equilibrium from an application of Glicksberg's theorem and a careful compactification approach.

Proposition 1.1 (Theorem 4.4, Epasto et al. [7])**.** *For any dimension $d = 2^{2^k}$ for an integer $k \geq 0$, one has*

$$\delta^\star(d) \geq \frac{1}{8}\ln(d).$$

Alternatively, via a reduction to a statistical prediction game of [8] in online learning, one can derive an asymptotically matching lower bound for all d (proven in Appendix A for completeness).

Proposition 1.2. *For any dimension $d \geq 2$, one has*

$$\delta^\star(d) \geq \left(\frac{1}{8} + o(1)\right)\ln(d).$$

These bounds leave a factor of four gap with the LogSumExp upper bound of $0.5\ln(d)$ and rely on nontrivial equilibrium existence and statistical arguments, respectively. As our main result, we further improve on this, deriving stronger lower bounds directly via an elementary constructive approach relying only on classical inequalities for smooth convex functions and a combinatorial recurrence. At its core, our analysis works inductively by considering elementary inequalities between certain sparse points $x^{(j_0)}, x^{(j_1)}, \ldots, x^{(j_k)} \in \mathbb{R}^d$, each with nonzero components only in the first j_ℓ entries, for certain subsets $\{j_0, j_1, \ldots, j_{k-1}, j_k\} \subseteq \{1, \ldots, d\}$. For each such subset, we derive a lower bound. Optimizing over these bounds, we arrive at the following maximal partition sum quantity

$$\gamma(d) := \begin{cases} \max\limits_{\{j_0, j_1, \ldots, j_{k-1}, j_k\} \subseteq \{1,\ldots,d\}} & \sum_{\ell=1}^k \left(1 - \frac{j_{\ell-1}}{j_\ell}\right)^2 \\ \text{s.t.} & 1 = j_0 < j_1 < \cdots < j_{k-1} < j_k = d. \end{cases} \tag{1.5}$$

Section 1.3 shows this quantity can be understood as optimizing over a structured family of proofs to build the strongest lower bound.

Our main result shows that this provides a lower bound on the optimal smoothing gap $\delta^\star(d)$. Our proof of Theorem 1.1 in fact provides a stronger lower bound, applying to any p-norm smoothing.

Theorem 1.1. *For any dimension $d \geq 2$, one has*

$$\delta^\star(d) \geq \gamma(d).$$

Further, $\lim_{d\to\infty} \frac{\gamma(d)}{\ln(d)} = 2\beta(1-\beta) \geq 0.40726$ *where β is the unique root of $2\beta \ln \beta - \beta + 1$ in $(0,1)$.*

Given the $0.5\ln(d)$ upper bound due to the LogSumExp function, this theorem determines the exact optimal gap down asymptotically to a narrow numerical range of $[0.40726, 0.5]$. The closeness of these bounds provides a case for the strong performance of the LogSumExp smoothing: It leaves at most a factor of 20% "on the table" in terms of performance.

However, although nearly matching our bound, $f_{\texttt{LSE}}$ fails to attain our lower bound. As a final result, we examine the tightness of our elementary lower bound. For any dimension d, we provide a new smoothing, strictly outperforming $0.5\ln(d)$. This disproves the optimality of $f_{\texttt{LSE}}$ for every d. In the case of $d = 2, 3$, our lower bound construction via $\gamma(d)$ is exactly tight, proven by an exactly optimal smoothing attaining $\delta^\star(d)$.

Theorem 1.2. *For any dimension $d \geq 2$, one has $\delta^\star(d) < 0.5\ln(d)$. For $d = 2$ and $d = 3$, the optimal smoothing value is exactly $\delta^\star(d) = \gamma(d)$.*

The use of such further optimized smoothings offers constant gains in worst-case guarantees for accelerated nonsmooth optimization directly by the previously discussed reduction to smooth optimization techniques of [4]. Smoothings attaining our theoretical lower bound saturate what is possible by such reductions, meaning any further progress in accelerated nonsmooth optimization will require approaches distinct from existing smoothing frameworks. Resolving the exact optimal smoothing f for $d \geq 4$ is left as an interesting future direction.

1.2 Related Works

Constructions of Smoothings. Moreau envelopes [9] are perhaps the most classic form of smoothing in optimization. General recipes for constructing smoothings have been extensively explored. Nesterov [1] does so via dualizations of the form (1.4). Beck and Teboulle [4] provided an alternative asymptotic smoothing approach applicable for smoothing any sublinear function. Epasto et al. [7] proposed a general piecewise linear form for constructing gradients of a smoothing of the max function, yielding a piecewise quadratic family of f. In areas of machine learning, alternatives to the LogSumExp/softmax function have been of recent interest. The sparsemax [2], defined by projections onto the simplex and corresponding to the Moreau envelope of the max function, typically has sparse gradients, which can improve computation and interpretability. In reinforcement learning, the mellowmax smoothing [3], defined as a translation of the LogSumExp, has found widespread interest.

Prior Lower Bound Theory for Smoothings. As discussed above, lower bound proofs can be extracted from the literature [7, 8] via the game theoretic or statistical reductions in Propositions 1.1 and 1.2 respectively. Our work here circumvents the need for such heavy machinery by using entirely constructive, elementary proof techniques. Moreover, our tailored, direct approach yields tighter lower bounds, being exact for $d = 2, 3$ at least.

If one instead considers optimal smoothings with respect to the two-norm, requiring instead that

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \|x - y\|_2 \qquad \forall x, y \in \mathbb{R}^d,$$

then a theory for exactly optimal smoothings of general sublinear functions was given by the authors in [10]. In this case, the optimal difference between a smooth convex f and $\sigma_{\texttt{max}}$ is bounded by a constant for all d (in particular being at most $1/4$, see [10, Section 4.2.3]). As a result, $f_{\texttt{LSE}}$ is fundamentally suboptimal not just by constants but in a big-O for two-norm smoothing.

Complexity Lower Bound Theory for Optimization. Often, the complexity theory for first-order methods is independent of the ambient problem dimension d. As exceptions to this, Guzman and Nemirovski [11] showed a log term for Hölder smooth convex optimization over p-norm balls outside the classical Euclidean setting. Carmon and Hinder [12] showed log factors in problem dimension are a fundamental price of adaptivity for stochastic optimization when certain problem parameters are unknown. For the particular setting of minimizing a finite maximum considered here, our Theorem 1.1 establishes a smoothing barrier: no smoothing-based reduction, i.e. (1.2), can have a dimension-independent complexity or any dependence better than $\mathcal{O}(\sqrt{\ln(d)})$.

In online convex optimization, algorithms often must have regret scale at least with $\Omega(\sqrt{\ln(d)})$. This is proven by a reduction to coin flipping games [8] and attained up to constant factors by approaches using entropic regularization, dual to the LogSumExp function. See the surveys [13, Section 2.5] and [14, Section 7]. Our proof of Proposition 1.2 reduces to this prior art.

Generalization to the Maximum Eigenvalue Function. Lastly, we note our theory immediately provides lower bounds on smoothings of the maximum eigenvalue function $\sigma_\lambda(X) = \lambda_{\texttt{max}}(X)$ defined for all symmetric matrices $X \in \mathbb{R}^{d\times d}$ with respect to the spectral norm. This follows by noting that for any $x \in \mathbb{R}^d$, $\sigma_\lambda(\mathrm{diag}(x)) = \sigma_{\texttt{max}}(x)$. As a result, any spectral-norm 1-smoothing of σ_λ must differ by at least $\gamma(d)$ somewhere. This is the optimal order as $f(X) = \ln(\sum_i^d \exp(\lambda_i(X)))$ provides a 1-smooth overestimator of σ_λ differing by at most $\ln(d)$. This generalization of the $f_{\texttt{LSE}}$ smoothings has played a role in solving semidefinite programs, for example, using the radial framework of Renegar [15] and Grimmer [16].

1.3 The Lower Bound $\gamma(d)$ as Optimization over Structured Proofs

Before formally proving Theorem 1.1, here we provide an informal progression of ideas leading to $\gamma(d)$ as a plausible candidate lower bound on $\delta^\star(d)$. To build lower bounding inequalities, we consider certain inequalities which must be satisfied for any function that is convex and 1-smooth in $\|\cdot\|_\infty$. For a differentiable function f, these two properties are equivalent to the following being nonnegative for every $x, y \in \mathbb{R}^d$,

$$Q_{x,y} := f(x) - f(y) - \langle \nabla f(y), x - y\rangle - \frac{1}{2}\|\nabla f(x) - \nabla f(y)\|_1^2 \geq 0.$$

See [17, Theorem 5.8]. We aim to construct our lower bound by considering well-chosen combinations of these inequalities at various points x and y.

In particular, we aim to show that some points $y^{(d)}$ and $y^{(1)}$ must have a large difference in f values (larger than $\sigma_{\texttt{max}}$). This suffices to provide a lower bound as

$$\left[f(y^{(d)}) - f(y^{(1)})\right] - \left[\sigma_{\texttt{max}}(y^{(d)}) - \sigma_{\texttt{max}}(y^{(1)})\right] \leq 2\|f - \sigma_{\texttt{max}}\|_\infty.$$

So showing $\left[f(y^{(d)}) - f(y^{(1)})\right] - \left[\sigma_{\texttt{max}}(y^{(d)}) - \sigma_{\texttt{max}}(y^{(1)})\right] - 2\gamma \geq 0$ would prove $\|f - \sigma_{\texttt{max}}\|_\infty \geq \gamma$.

We prove such a nonnegativity result by introducing intermediate points $y^{(1)}, \ldots, y^{(d)}$ between which we will consider the nonnegative $Q_{\cdot,\cdot}$ quantities. In particular, we restrict attention to scaled

points $y^{(j)} = \alpha x^{(j)}$ with $x^{(j)} = (1/j, \ldots, 1/j, 0, \ldots, 0)$, having j nonzero entries followed by zeros. In our formal proof below, we will take a limit as α tends to infinity, but for motivating our approach, considering a fixed large value will suffice. For ease, denote $Q_{y^{(i)},y^{(j)}}$ by $Q_{i,j}$.

Given these points, one may hope for a proof of the target nonnegativity that selects nonnegative multipliers $\lambda_{i,j}$ for each $i > j$ such that the following identity holds

$$\sum_{i>j} \lambda_{i,j} Q_{i,j} = \left[f(y^{(d)}) - f(y^{(1)})\right] - \left[\sigma_{\texttt{max}}(y^{(d)}) - \sigma_{\texttt{max}}(y^{(1)})\right] - 2\gamma \tag{1.6}$$

for any values of $f(\cdot)$ and $\nabla f(\cdot)$ at the points $y^{(j)}$. Since the left-hand side is a sum of nonnegative quantities, this identity would establish the right-hand side is nonnegative. Optimizing over such proofs, the strongest lower bound is given by the λ yielding the maximum value of γ.

As a further simplification, one may guess that the gradients at the points $y^{(j)}$, for α large enough, converge towards $(1/j, \ldots, 1/j, 0, \ldots, 0)$. Indeed, in (2.5), we will find this can be taken without loss via a symmetry argument. Under this asymptotic simplification, we have for $i > j$ that

$$Q_{i,j} = \left[f(y^{(i)}) - \sigma_{\texttt{max}}(y^{(i)})\right] - \left[f(y^{(j)}) - \sigma_{\texttt{max}}(y^{(j)})\right] - 2\left(1 - \frac{j}{i}\right)^2.$$

After fixing the values of $y^{(j)}$ and (via asymptotic reasoning) values of gradients $\nabla f(y^{(j)})$, both sides of the identity (1.6) are affine in $f_j = f(y^{(j)})$. So a collection of multipliers λ satisfies (1.6) if and only if the coefficients agree on each term: Namely, for the coefficients on f_j to match, we need

$$\sum_{i=1}^{j-1} \lambda_{j,i} - \sum_{i=j+1}^{d} \lambda_{i,j} = \begin{cases} 1 & \text{if } j = d, \\ -1 & \text{if } j = 1, \\ 0 & \text{otherwise,} \end{cases}$$

and then, for the constant terms to match,

$$\sum_{i>j} \lambda_{i,j} \left(1 - \frac{j}{i}\right)^2 = \gamma.$$

Consequently, calculation of the strongest lower bound, proven by this structured template, amounts to optimizing over a polyhedron: maximize the induced value of γ over the set of nonnegative λ with the needed identity. This polyhedron can be viewed combinatorially as routing one unit of flow from 1 to d with $\lambda_{i,j}$ flowing from j to i. Our definition of $\gamma(d)$ in (1.5) is a discrete formulation, where each partition $1 = j_0 < j_1 < \cdots < j_{k-1} < j_k = d$ corresponds to the extreme point with $\lambda_{j_{\ell+1},j_\ell} = 1$ for $\ell = 0, \ldots k-1$ and all other multipliers equal to zero.

From Theorems 1.1 and 1.2, this lower bounding approach is quite strong, being within a small factor of the known upper bound and exact for $d = 2$ and $d = 3$. Preliminary numerics with $d = 4$ using sample points beyond $y^{(1)}, \ldots, y^{(4)}$ may provide stronger lower bounds. However, for such selections, we cannot use symmetry to determine exact gradients, making the resulting optimization much less tractable. We expect tighter bounds to follow from identifying new sample points with tractable proof-optimizations.

2 Proof of Theorem 1.1 – An Elementary Lower Bound

We prove the first result of $\delta^\star(d) \geq \gamma(d)$ more generally than the case of ∞-norms of interest to this paper, deriving a bound for any p-norm smoothing, where $p \geq 1$. Letting q denote the dual norm exponent (i.e., $1/p + 1/q = 1$), a function is 1-smooth with respect to $\|\cdot\|_p$ if $\|\nabla f(x) - \nabla f(y)\|_q \leq \|x - y\|_p$. Correspondingly, for any $1 \leq p \leq \infty$, define

$$\delta_p^\star(d) := \begin{cases} \min_f & \|f - \sigma_{\texttt{max}}\|_\infty \\ \text{s.t.} & f\colon \mathbb{R}^d \to \mathbb{R} \text{ is convex and 1-smooth in } \|\cdot\|_p \end{cases}$$

and for $p > 1$, define

$$\gamma_p(d) := \begin{cases} \max & \sum_{\ell=1}^k \frac{1}{4}\left(j_{\ell-1}\left(\frac{1}{j_{\ell-1}} - \frac{1}{j_\ell}\right)^q + (j_\ell - j_{\ell-1})\left(\frac{1}{j_\ell}\right)^q\right)^{2/q} \\ \text{s.t.} & 1 = j_0 < j_1 < \cdots < j_{k-1} < j_k = d \end{cases}$$

with $\gamma_1(d)$ defined by

$$\gamma_1(d) := \begin{cases} \max & \sum_{\ell=1}^k \frac{1}{4}\left(\max\left\{\frac{1}{j_{\ell-1}} - \frac{1}{j_\ell}, \frac{1}{j_\ell}\right\}\right)^2 \\ \text{s.t.} & 1 = j_0 < j_1 < \cdots < j_{k-1} < j_k = d \end{cases}$$

Notice that $\gamma_p(d)$ converges monotonically to $\gamma_1(d)$ and to $\gamma_\infty(d) = \gamma(d)$ as p tends to one and infinity, respectively, (and so conversely, q tends to infinity and one). As a first observation, we note that without loss of generality, it suffices to restrict our attention to functions f that are invariant under permutations. That is, functions with $f(x) = f(Px)$ for all $x \in \mathbb{R}^d$ and $P \in \Pi_d$ where Π_d is the set of all $d \times d$ permutation matrices.

Lemma 2.1. *For any dimension $d \geq 2$, one has*

$$\delta_p^\star(d) = \hat{\delta}_p^\star(d) := \begin{cases} \min_f & \|f - \sigma_{\texttt{max}}\|_\infty \\ \text{s.t.} & f\colon \mathbb{R}^d \to \mathbb{R} \textit{ is convex and 1-smooth in } \|\cdot\|_p \\ & f(x) = f(Px)\ \forall x \in \mathbb{R}^d, P \in \Pi_d. \end{cases}$$

Proof of Lemma 2.1. Clearly $\delta_p^\star(d) \leq \hat{\delta}_p^\star(d)$ as it has further constrained the search for an optimal smoothing. We establish the reverse direction by showing that for any feasible f for $\delta_p^\star(d)$, a permutation invariant, convex, 1-smooth in the p-norm $\hat{f}$ can be constructed that is no farther from $\sigma_{\texttt{max}}$. To this end, consider any convex, 1-smooth function f with respect to $\|\cdot\|_p$. Define

$$\hat{f}(x) = \frac{1}{d!}\sum_{P \in \Pi_d} f(Px).$$

It is clear that for each $P \in \Pi_d$, the map $x \mapsto f(Px)$ is convex and 1-smooth in the p-norm.[1] As a result, $\hat{f}$ is convex and 1-smooth in the p-norm. Finally, we establish $\hat{f}$ is no farther from $\sigma_{\texttt{max}}$ than f is as

$$\begin{aligned} \|f - \sigma_{\texttt{max}}\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x) - \sigma_{\texttt{max}}(x)| &= \sup_{x \in \mathbb{R}^d} \sup_{P \in \Pi_d} |f(Px) - \sigma_{\texttt{max}}(Px)| \\ &\geq \sup_{x \in \mathbb{R}^d} \left|\frac{1}{d!}\sum_{P \in \Pi_d} f(Px) - \sigma_{\texttt{max}}(Px)\right| \\ &= \|\hat{f} - \sigma_{\texttt{max}}\|_\infty \end{aligned}$$

where in the very last equality we have used the permutation invariance of $\sigma_{\texttt{max}}$. □

[1] Applying the chain rule and the simple observation $\|P\|_{p \to p} = 1$ gives the smoothness result.

Next, we show that the subdifferential $\partial f(x) := \{\zeta \mid f(y) \geq f(x) + \langle \zeta, y - x\rangle \ \forall y\}$ of any convex approximator must lie in the simplex $\Delta_d \subset \mathbb{R}^d$.

Lemma 2.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. If $\|f - \sigma_{\texttt{max}}\|_\infty < \infty$, then for any $x \in \mathbb{R}^d$ and $\zeta \in \partial f(x)$, it holds that $\sigma_{\texttt{max}}(y) \geq \langle \zeta, y\rangle$ for all $y \in \mathbb{R}^d$. In particular, $\partial f(x) \subseteq \Delta_d$ for all $x \in \mathbb{R}^d$.*

Proof of Lemma 2.2. Let $\delta = \|f - \sigma_{\texttt{max}}\|_\infty$ and fix $\zeta \in \partial f(x)$ for some arbitrary $x \in \mathbb{R}^d$. For any $y \in \mathbb{R}^d$ and any $\varepsilon > 0$, we have

$$\begin{aligned}\sigma_{\texttt{max}}(y) = \varepsilon\sigma_{\texttt{max}}(y/\varepsilon) &\geq \varepsilon[f(y/\varepsilon) - \delta] \geq \varepsilon[f(x) + \langle \zeta, y/\varepsilon - x\rangle - \delta] \\ &= \langle \zeta, y\rangle + \varepsilon[f(x) - \langle \zeta, x\rangle - \delta]\end{aligned}$$

where the second inequality is the subgradient inequality. Taking the limit as $\varepsilon \to 0$ gives $\sigma_{\texttt{max}}(y) \geq \langle \zeta, y\rangle$. Since x, y, and $\zeta \in \partial f(x)$ were all arbitrary, this proves the first statement. The second statement follows directly from the first. □

Now let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex, permutation invariant function that is 1-smooth in $\|\cdot\|_p$ and satisfies $\|f - \sigma_{\texttt{max}}\|_\infty < \infty$. For $j = 1, \ldots, d$, let $x^{(j)} \in \mathbb{R}^d$ denote the point $(1/j, \ldots, 1/j, 0, \ldots, 0)$ with first j entries set equal to $1/j$. For any $\alpha > 0$ and $i, j = 1, \ldots, d$, convexity and p-norm smoothness of f [17, Theorem 5.8] implies

$$\begin{aligned}Q_{i,j}(\alpha) := f(\alpha x^{(i)}) - f(\alpha x^{(j)}) - \left\langle \nabla f(\alpha x^{(j)}), \alpha x^{(i)} - \alpha x^{(j)} \right\rangle \\ - \frac{1}{2}\|\nabla f(\alpha x^{(i)}) - \nabla f(\alpha x^{(j)})\|_q^2 \geq 0.\end{aligned} \tag{2.1}$$

For $i > j \geq 1$, a direct calculation gives

$$\begin{aligned}Q_{i,j}(\alpha) &= f(\alpha x^{(i)}) - f(\alpha x^{(j)}) - \alpha \left\langle \nabla f(\alpha x^{(j)}), \frac{j}{i} x^{(j)} - x^{(j)} \right\rangle \\ &\quad - \frac{1}{2}\|\nabla f(\alpha x^{(i)}) - \nabla f(\alpha x^{(j)})\|_q^2 - \alpha \left\langle \nabla f(\alpha x^{(j)}), x^{(i)} - \frac{j}{i} x^{(j)} \right\rangle \\ &\leq f(\alpha x^{(i)}) - f(\alpha x^{(j)}) - \left(\frac{j}{i} - 1\right) \left\langle \nabla f(\alpha x^{(j)}), \alpha x^{(j)} \right\rangle \\ &\quad - \frac{1}{2}\|\nabla f(\alpha x^{(i)}) - \nabla f(\alpha x^{(j)})\|_q^2 \\ &= f(\alpha x^{(i)}) - \sigma_{\texttt{max}}(\alpha x^{(i)}) - \left[f(\alpha x^{(j)}) - \sigma_{\texttt{max}}(\alpha x^{(j)})\right] \\ &\quad - \frac{1}{2}\|\nabla f(\alpha x^{(i)}) - \nabla f(\alpha x^{(j)})\|_q^2 \\ &\quad - \left(1 - \frac{j}{i}\right) \left[\sigma_{\texttt{max}}(\alpha x^{(j)}) - \left\langle \nabla f(\alpha x^{(j)}), \alpha x^{(j)} \right\rangle\right] \\ &\leq f(\alpha x^{(i)}) - \sigma_{\texttt{max}}(\alpha x^{(i)}) - \left[f(\alpha x^{(j)}) - \sigma_{\texttt{max}}(\alpha x^{(j)})\right] \\ &\quad - \frac{1}{2}\|\nabla f(\alpha x^{(i)}) - \nabla f(\alpha x^{(j)})\|_q^2\end{aligned} \tag{2.2}$$

where the first inequality holds because $\left\langle \nabla f(\alpha x^{(j)}), x^{(i)} - \frac{j}{i} x^{(j)} \right\rangle \geq 0$ since the components of $x^{(i)} - \frac{j}{i} x^{(j)}$ are nonnegative and Lemma 2.2 ensures $\nabla f(\alpha x^{(j)}) \in \Delta_d$, the second equality holds

because $\frac{j}{i}\sigma_{\texttt{max}}(\alpha x^{(j)}) = \alpha/i = \sigma_{\texttt{max}}(\alpha x^{(i)})$, and the last inequality follows from Lemma 2.2 and the assumption that $i > j$. Now fix an increasing sequence $j_0 = 1 < j_1 < \cdots < j_k = d$. Note that

$$\begin{aligned}
0 &\le \sum_{\ell=1}^{k} Q_{j_\ell, j_{\ell-1}}(\alpha) \\
&\le f(\alpha x^{(d)}) - \sigma_{\texttt{max}}(\alpha x^{(d)}) - [f(\alpha x^{(1)}) - \sigma_{\texttt{max}}(\alpha x^{(1)})] \\
&\qquad - \sum_{\ell=1}^{k} \frac{1}{2}\|\nabla f(\alpha x^{(j_\ell)}) - \nabla f(\alpha x^{(j_{\ell-1})})\|_q^2 \\
&\le 2\|f - \sigma_{\texttt{max}}\|_\infty - \sum_{\ell=1}^{k} \frac{1}{2}\|\nabla f(\alpha x^{(j_\ell)}) - \nabla f(\alpha x^{(j_{\ell-1})})\|_q^2
\end{aligned}$$

where the first inequality is by (2.1), the second inequality holds because of inequality (2.2) and the fact that the terms $f(\alpha x^{(j_\ell)}) - \sigma_{\texttt{max}}(\alpha x^{(j_\ell)})$ and $\left[f(\alpha x^{(j_{\ell-1})}) - \sigma_{\texttt{max}}(\alpha x^{(j_{\ell-1})})\right]$ telescope, while the last inequality follows directly from the definition of $\|f - \sigma_{\texttt{max}}\|_\infty$. Rearranging and taking the limit as α tends to infinity gives

$$\|f - \sigma_{\texttt{max}}\|_\infty \ge \lim_{\alpha\to\infty} \sum_{\ell=1}^{k} \frac{1}{4}\|\nabla f(\alpha x^{(j_\ell)}) - \nabla f(\alpha x^{(j_{\ell-1})})\|_q^2. \tag{2.3}$$

Note that $\nabla f(Px) = P\nabla f(x)$ for all $P \in \Pi_d$ and $x \in \mathbb{R}^d$ because of the permutation invariance assumption. In particular, $\nabla f(\alpha x^{(j)}) = P\nabla f(\alpha x^{(j)})$ for all $P \in \Pi_d$ such that $P\alpha x^{(j)} = \alpha x^{(j)}$. Combining this with the fact that $\nabla f(\alpha x^{(j)}) \in \Delta_d$, we get that $\nabla f(\alpha x^{(d)}) = x^{(d)}$ and

$$\nabla f(\alpha x^{(j)}) = \left(\lambda_j(\alpha), \ldots, \lambda_j(\alpha), \frac{1 - j\lambda_j(\alpha)}{d-j}, \ldots, \frac{1 - j\lambda_j(\alpha)}{d-j}\right) \tag{2.4}$$

for all $\alpha > 0$ and $j = 1, \ldots, d-1$, where $\lambda_j(\alpha) \in [0, 1/j]$. Therefore $\lim_{\alpha\to\infty} \nabla f(\alpha x^{(d)}) = x^{(d)}$. Similarly, $\lim_{\alpha\to\infty} \nabla f(\alpha x^{(j)}) = x^{(j)}$ for all $j = 1, \ldots, d-1$, as the following calculation shows:

$$\begin{aligned}
0 \le \lim_{\alpha\to\infty} \frac{1}{j} - \lambda_j(\alpha) &= \lim_{\alpha\to\infty} \frac{1}{\alpha}\left(\frac{\alpha}{j} - \alpha\lambda_j(\alpha)\right) \\
&= \lim_{\alpha\to\infty} \frac{1}{\alpha}\left(\sigma_{\texttt{max}}(\alpha x^{(j)}) - \left\langle \nabla f(\alpha x^{(j)}), \alpha x^{(j)} \right\rangle\right) \\
&\le \lim_{\alpha\to\infty} \frac{1}{\alpha}\left(\|f - \sigma_{\texttt{max}}\|_\infty + f(\alpha x^{(j)}) - \left\langle \nabla f(\alpha x^{(j)}), \alpha x^{(j)} \right\rangle\right) \\
&\le \lim_{\alpha\to\infty} \frac{1}{\alpha}\left(\|f - \sigma_{\texttt{max}}\|_\infty + f(0)\right) \\
&= 0
\end{aligned} \tag{2.5}$$

where the first inequality uses $\lambda_j(\alpha) \in [0, 1/j]$, the second equality holds by equation (2.4) and the definition of $x^{(j)}$, the second inequality uses $\sigma_{\texttt{max}}(\alpha x^{(j)}) - f(\alpha x^{(j)}) \le \|f - \sigma_{\texttt{max}}\|_\infty$, and the last inequality uses the subgradient inequality $f(0) \ge f(\alpha x^{(j)}) + \left\langle \nabla f(\alpha x^{(j)}), 0 - \alpha x^{(j)} \right\rangle$. As a result, for

all $1 \le q \le \infty$, we have

$$\begin{aligned}
&\|f - \sigma_{\mathtt{max}}\|_\infty \\
&\ge \sum_{\ell=1}^{k} \frac{1}{4} \| \lim_{\alpha\to\infty} \nabla f(\alpha x^{(j_\ell)}) - \lim_{\alpha\to\infty} \nabla f(\alpha x^{(j_{\ell-1})}) \|_q^2 \\
&= \sum_{\ell=1}^{k} \frac{1}{4} \|x^{(j_\ell)} - x^{(j_{\ell-1})}\|_q^2 \\
&= \begin{cases} \sum_{\ell=1}^{k} \frac{1}{4} \left(j_{\ell-1} \left(\frac{1}{j_{\ell-1}} - \frac{1}{j_\ell} \right)^q + (j_\ell - j_{\ell-1}) \left(\frac{1}{j_\ell} \right)^q \right)^{2/q} & \text{if } 1 \le q < \infty \\ \sum_{\ell=1}^{k} \frac{1}{4} \left(\max \left\{ \frac{1}{j_{\ell-1}} - \frac{1}{j_\ell}, \frac{1}{j_\ell} \right\} \right)^2 & \text{if } q = \infty \end{cases}
\end{aligned}$$

where the inequality follows from (2.3). Since the sequence $j_0 = 1 < j_1 < \cdots < j_k = d$ was arbitrary, maximizing over the sequences gives $\|f - \sigma_{\mathtt{max}}\|_\infty \ge \gamma_p(d)$ for all $1 \le p \le \infty$ and for all d. Hence $\delta_p^\star(d) \ge \gamma_p(d)$ for all d. When $p = \infty$ and $q = 1$, the objective reduces to $\sum_{\ell=1}^{k} \left(1 - \frac{j_{\ell-1}}{j_\ell}\right)^2$, giving the theorem's first claim.

Finally, we prove our asymptotic claim for $p = \infty$ by showing that

$$2\beta(1-\beta)\ln(d) - \frac{2(d-1)}{d} \le \gamma(d) \le 2\beta(1-\beta)\ln(d) \tag{2.6}$$

where β is the unique solution to $2\beta \ln \beta - \beta + 1 = 0$ in $(0, 1)$. (Table 1 provides numerical values of $\gamma(d)$ and these bounds, where we see $\gamma(d)$ closely follows this upper bound.) Numerically $\beta \approx 0.28467$, implying that $\gamma(d) \sim 2\beta(1-\beta)\ln(d) \approx 0.40726 \ln(d)$. The key observation in showing this is to note that $\gamma(d)$ is characterized by the recursive relation

$$\gamma(d) = \max_{1 \le i < d} \left\{ \gamma(i) + \left(1 - \frac{i}{d}\right)^2 \right\}, \qquad \gamma(1) = 0.$$

We prove both the upper and lower bounds inductively. At $d = 1$ and $d = 2$, one can verify directly. The upper bound then follows inductively for $d > 2$ by considering a continuous relaxation as

$$\begin{aligned}
\gamma(d) &\le \max_{1 \le i < d} \left\{ 2\beta(1-\beta)\ln(i) + \left(1 - \frac{i}{d}\right)^2 \right\} \\
&\le \max_{\phi \in (0,1)} \left\{ 2\beta(1-\beta)\ln(\phi) + (1-\phi)^2 \right\} + 2\beta(1-\beta)\ln(d) \\
&= \left(2\beta(1-\beta)\ln(\beta) + (1-\beta)^2 \right) + 2\beta(1-\beta)\ln(d) \\
&= 2\beta(1-\beta)\ln(d)
\end{aligned}$$

where the first line applies our inductive hypothesis, the second substitutes $i = \phi d$, the third notes that $\phi = \beta$ optimizes the expression, and the final line uses the definition of β. The lower bound

Table 1: Numerical values of $\gamma(d)$ and the bounds in (2.6), rounded to five digits.

d	$2\beta(1-\beta)\ln(d) - \frac{2(d-1)}{d}$	$\gamma(d)$	$2\beta(1-\beta)\ln(d)$
2	−0.71771	0.25000	0.28229
3	−0.88591	0.44444	0.44743
4	−0.93541	0.56250	0.56459
5	−0.94453	0.64000	0.65547
10	−0.86224	0.93444	0.93776
100	−0.10448	1.86950	1.87552
1000	0.81528	2.80460	2.81328
10000	1.75124	3.74724	3.75104

follows inductively for $d > 2$ by considering the simple selection $i = \lceil \beta d \rceil$ as

$$\begin{aligned}
\gamma(d) &\geq \gamma(\lceil \beta d \rceil) + \left(1 - \frac{\lceil \beta d \rceil}{d}\right)^2 \\
&\geq 2\beta(1-\beta)\ln(\lceil \beta d \rceil) + \left(1 - \frac{\lceil \beta d \rceil}{d}\right)^2 - \frac{2(\lceil \beta d \rceil - 1)}{\lceil \beta d \rceil} \\
&\geq 2\beta(1-\beta)\ln(d) + 2\beta(1-\beta)\ln(\beta) + \left(1 - \frac{\beta d + 1}{d}\right)^2 - \frac{2\beta d}{\beta d + 1} \\
&= 2\beta(1-\beta)\ln(d) - \frac{2(1-\beta)}{d} + \frac{1}{d^2} - \frac{2\beta d}{\beta d + 1} \\
&\geq 2\beta(1-\beta)\ln(d) - \frac{2(d-1)}{d}
\end{aligned}$$

where the second line applies our inductive hypothesis, the third applies bounds of $\beta d \leq \lceil \beta d \rceil \leq \beta d + 1$, the fourth cancels terms using the definition of β, and the final collects and simplifies terms. □

3 Proof of Theorem 1.2 – Improved Smoothings

We prove the strict inequality $\delta^\star(d) < 0.5\ln(d)$ for $d \geq 3$ by demonstrating that a strictly better smoothing than LogSumExp exists with

$$\|f - \sigma_{\texttt{max}}\|_\infty \leq \frac{1}{2}\left[\ln(d-1) + \tfrac{1}{d(d-2)}\ln(d-1)\right] < \frac{1}{2}\ln(d).$$

The strict inequality above follows by noting that

$$\begin{aligned}
\ln(d) &= \ln(d-1) + \int_{d-1}^{d} \frac{1}{t}\,\mathrm{dt} \\
&> \ln(d-1) + \frac{1}{d} \\
&\geq \ln(d-1) + \frac{1}{d(d-2)}\ln(d-1)
\end{aligned} \tag{3.1}$$

where the strict inequality uniformly lower bounds the integrand and the second inequality uses that $\ln(1+t) \leq t$ at $t = d-2$.

To construct this improvement, consider Nesterov-style smoothings

$$f_{A,B}(x) := \sup_{\lambda \in \Delta_d} \langle \lambda, x \rangle - h_{A,B}(\lambda). \tag{3.2}$$

Here the dual function $h_{A,B}$ is defined as the closed convex proper function[2]

$$h_{A,B}(\lambda) := \begin{cases} \frac{1}{A^2} \sum_{i=1}^d (A\lambda_i + B) \ln(A\lambda_i + B) - C & \text{if } \lambda \in \Delta_d \\ +\infty & \text{otherwise} \end{cases} \tag{3.3}$$

for any $A > 0, B \geq 0$ with $A + dB = 1$ and additive constant C defined as

$$C = \frac{1}{2}(U_{A,B} + L_{A,B})$$

where $U_{A,B} = \frac{(A+B)\ln(A+B) + (d-1)B\ln(B)}{A^2}$ and $L_{A,B} = \frac{A+dB}{A^2} \ln(\frac{A}{d} + B)$.

Note that $f_{1,0} = f_{\texttt{LSE}} - \frac{1}{2}\ln(d)$, so this family includes LogSumExp (centered) as a special case. The following lemma allows us to select A and B more judiciously, yielding the claimed strictly better smoothing.

Lemma 3.1. *For any $d \geq 2$ and $A > 0, B \geq 0$ with $A + dB = 1$, the function $f_{A,B}$ in (3.2) is 1-smooth in $\|\cdot\|_\infty$ and*

$$\|f_{A,B} - \sigma_{\texttt{max}}\|_\infty \leq \frac{1}{2}(U_{A,B} - L_{A,B}).$$

Proof of Lemma 3.1. First, we verify smoothness by showing $h_{A,B}$ is 1-strongly convex with respect to $\|\cdot\|_1$. We do this by reducing to the 1-strong convexity of the entropy $h(z) = \sum_{i=1}^d z_i \ln(z_i)$ for $z \in \Delta_d$ (and $+\infty$ otherwise). That is, we know that for all $z, z' \in \Delta_d$ and $t \in [0,1]$,

$$h(tz + (1-t)z') \leq th(z) + (1-t)h(z') - \frac{1}{2}t(1-t)\|z - z'\|_1^2.$$

For any $\lambda, \lambda' \in \Delta_d$, consider the translated values with $z = A\lambda + Be$ and $z' = A\lambda' + Be$ where e denotes the all ones vector. Since $A + Bd = 1$, we have $z, z' \in \Delta_d$. Then dividing the above inequality by A^2 and subtracting C from both sides, we have

$$h_{A,B}(t\lambda + (1-t)\lambda') \leq th_{A,B}(\lambda) + (1-t)h_{A,B}(\lambda') - \frac{1}{2}t(1-t)\|\lambda - \lambda'\|_1^2.$$

Hence, $h_{A,B}$ is 1-strongly convex with respect to $\|\cdot\|_1$.

Next, we bound $\|f_{A,B} - \sigma_{\texttt{max}}\|_\infty$. Observe that for any $x \in \mathbb{R}^d$, the difference $f_{A,B}(x) - \sigma_{\texttt{max}}(x)$ is bounded above and below by

$$\begin{aligned} -\sup_{\hat{\lambda} \in \Delta_d} h_{A,B}(\hat{\lambda}) &= \sup_{\lambda \in \Delta_d} \langle \lambda, x \rangle - \sup_{\hat{\lambda} \in \Delta_d} h_{A,B}(\hat{\lambda}) - \sigma_{\texttt{max}}(x) \\ &\leq f_{A,B}(x) - \sigma_{\texttt{max}}(x) \\ &\leq \sup_{\lambda \in \Delta_d} \langle \lambda, x \rangle - \inf_{\hat{\lambda} \in \Delta_d} h_{A,B}(\hat{\lambda}) - \sigma_{\texttt{max}}(x) \\ &= -\inf_{\hat{\lambda} \in \Delta_d} h_{A,B}(\hat{\lambda}) \end{aligned}$$

[2]Here we use the convention that $0\ln(0) = 0$.

where the equalities use the fact that $\sigma_{\max}(x) = \sup_{\lambda\in\Delta_d} \langle\lambda, x\rangle$, while the inequalities follow directly from the definition of $f_{A,B}$. This implies that

$$\|f_{A,B} - \sigma_{\max}\|_\infty \le \max\left\{\sup_{\lambda\in\Delta_d} h_{A,B}(\lambda), -\inf_{\lambda\in\Delta_d} h_{A,B}(\lambda)\right\}.$$

Now, since $h_{A,B}$ is convex and permutation invariant, the maximum over Δ_d is attained at the vertices of Δ_d, while the minimum is attained at $(\frac{1}{d},\dots,\frac{1}{d})$. Plugging in these $h_{A,B}$ values, the maximum value is $U_{A,B}-C$ and the minimum value is $L_{A,B}-C$. Simplifying terms, both components of the above maximum are equal to the claimed bound of $\frac{1}{2}(U_{A,B} - L_{A,B})$. □

Consider the selection $A = \frac{d-2}{d-1} > 0$ and $B = \frac{1}{d(d-1)} > 0$. By the above lemma, the resulting $f_{A,B}$ provides a feasible smoothing, proving

$$\begin{aligned}
\delta^\star(d) &\le \|f_{A,B} - \sigma_{\max}\|_\infty \\
&\le \frac{1}{2}\left[\frac{(A+B)\ln(A+B) + (d-1)B\ln(B)}{A^2} - \frac{A+dB}{A^2}\ln\left(\frac{A}{d}+B\right)\right] \\
&= \frac{1}{2A^2}\left[\frac{d-1}{d}\ln\left(\frac{d-1}{d}\right) + \frac{1}{d}\ln\left(\frac{1}{d(d-1)}\right) - \ln\left(\frac{1}{d}\right)\right] \\
&= \frac{1}{2A^2}\frac{d-2}{d}\ln(d-1) \\
&= \frac{1}{2}\left[\ln(d-1) + \frac{1}{d(d-2)}\ln(d-1)\right] < \frac{1}{2}\ln(d)
\end{aligned}$$

where the final strict inequality is by (3.1).

Now we restrict to $d = 2, 3$ and provide a stronger, exactly optimal smoothing. Note that $(1-1/d)^2$ is a lower bound on $\gamma(d)$, and hence $\delta^\star(d)$, for any d, established by considering Theorem 1.1 with the simple sequence of length two $j_\ell : 1, d$. For $d = 2, 3$, we find that this lower bound is attained, i.e., $\delta^\star(d) = \gamma(d) = (1 - 1/d)^2$. In both cases, it suffices to consider a Nesterov-style smoothing using a quadratic distance function[3]

$$f_d(x) = \sup_{\lambda\in\Delta_d} \langle\lambda, x\rangle - \frac{c_d}{2}(\|\lambda\|_2^2 - 1) - \gamma(d)$$

where Δ_d is the simplex in $\mathbb{R}^d$ and $c_d = \max_{w\in\mathbb{R}^d:\ \sum w_i=0} \|w\|_1^2/\|w\|_2^2$. Here c_d is chosen exactly to be the needed constant to ensure that $\frac{c_d}{2}(\|\lambda\|_2^2 - 1)$ is 1-strongly convex in the one-norm on the simplex. Below, we briefly calculate c_d.

By symmetry, we can assume that w attaining the max has k positive components taking value $a > 0$ and $d-k$ negative components taking value $-b < 0$ for some integer $1 \le k < d$. The constraint $\sum w_i = 0$ implies $ka - (d-k)b = 0$. We may scale w without altering the ratio, so we choose $a = d - k$ and $b = k$. This yields a candidate vector $w^{(k)}$ with k entries of $d-k$ and $d-k$ entries of $-k$. Note $\|w^{(k)}\|_1 = 2k(d-k)$ and $\|w^{(k)}\|_2^2 = dk(d-k)$. Substituting these into the objective function gives

$$\frac{\|w^{(k)}\|_1^2}{\|w^{(k)}\|_2^2} = \frac{4k^2(d-k)^2}{dk(d-k)} = \frac{4}{d}k(d-k).$$

When d is even, this is maximized by taking $k = d/2$, giving $c_d = d$. When d is odd, this is maximized at both $k = (d \pm 1)/2$, giving $c_d = d - \frac{1}{d}$.

[3]For $d \ge 4$, this construction is no longer optimal. It is dominated by the LogSumExp construction in terms of smoothing gap $\|f - \sigma_{\max}\|_\infty$ and both fail to match our lower bound $\gamma(d)$.

Dually, the 1-smoothness of this f_d in the infinity-norm follows from the 1-strong convexity of $\frac{c_d}{2}(\|\lambda\|_2^2 - 1)$ in the one-norm on the simplex. Similarly, the largest deviation between f_d and $\sigma_{\texttt{max}}$ is

$$\|f_d - \sigma_{\texttt{max}}\|_\infty = \frac{1}{2}\left(\max_{\lambda \in \Delta_d} \frac{c_d}{2}\|\lambda\|_2^2 - \min_{\lambda \in \Delta_d} \frac{c_d}{2}\|\lambda\|_2^2\right).$$

The maximum occurs at $\lambda = (1, 0, \ldots, 0)$ and minimum occurs at $\lambda = (1/d, \ldots, 1/d)$, giving $\|f_d - \sigma_{\texttt{max}}\|_\infty = \frac{c_d}{4}(1 - 1/d)$. At $d = 2$ and $d = 3$, this equals $1/4$ and $4/9$, respectively, exactly agreeing with our lower bound of $(1 - 1/d)^2$. □

Acknowledgments. This work was supported in part by the Air Force Office of Scientific Research under award number FA9550-23-1-0531. Benjamin Grimmer was additionally supported as a fellow of the Alfred P. Sloan Foundation.

References

[1] Yu Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, May 2005.

[2] André F. T. Martins and Ramón F. Astudillo. From softmax to sparsemax: a sparse model of attention and multi-label classification. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1614–1623. JMLR.org, 2016.

[3] Kavosh Asadi and Michael L. Littman. An alternative softmax operator for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 243–252, 2017.

[4] Amir Beck and Marc Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22:557–580, 2012.

[5] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017.

[7] Alessandro Epasto, Mohammad Mahdian, Vahab Mirrokni, and Manolis Zampetakis. Optimal approximation - smoothness tradeoffs for soft-max functions. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, 2020.

[8] Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, May 1997.

[9] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

[10] Thabo Samakhoana and Benjamin Grimmer. The optimal smoothings of sublinear functions and convex cones. *arXiv preprint arXiv:2508.06681*, 2025.

[11] Cristóbal Guzmán and Arkadi Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1–14, 2015.

[12] Yair Carmon and Oliver Hinder. The price of adaptivity in stochastic convex optimization. *Mathematical Programming*, Sep 2025.

[13] Shai Shalev-Shwartz. Online learning and online convex optimization. *Found. Trends Mach. Learn.*, 4(2):107–194, February 2012.

[14] Francesco Orabona. A modern introduction to online learning. *ArXiv*, abs/1912.13213, 2019.

[15] James Renegar. Accelerated first-order methods for hyperbolic programming. *Math. Program.*, 173(1-2):1–35, 2019.

[16] Benjamin Grimmer. Radial duality part ii: applications and algorithms. *Math. Program.*, 205(1–2):69–105, May 2023.

[17] Amir Beck. *First-Order Methods in Optimization*. SIAM-Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2017.

A Proof of Proposition 1.2 – A Statistical Lower Bound Reduction

Here, we take the dual perspective of designing a strongly convex function h on the simplex. Let $\text{Range}(h) = \max_{\lambda\in\Delta_d} h(\lambda) - \min_{\lambda\in\Delta_d} h(\lambda)$ denotes the range of h on the simplex. We claim that

$$\delta^\star(d) = \begin{cases} \min & \frac{1}{2}\text{Range}(h) \\ \text{s.t.} & h\colon \Delta_d \to \mathbb{R} \text{ is 1-strongly convex in the one-norm} \end{cases} \tag{A.1}$$

The claimed equivalence follows by considering the bijection $h = f^*$:

This bijection maps between the two problems' domains as (i) a convex function $f\colon \mathbb{R}^d \to \mathbb{R}$ has $\|f - \sigma_{\texttt{max}}\|_\infty < \infty$ if and only if f^* has domain Δ_d and (ii) a function f is 1-smooth with respect to the infinity norm if and only if f^* is 1-strongly convex with respect to the one-norm by [17, Theorem 5.26]. The equivalence of objective values under this bijection follows by considering the max function's representation $\sigma_{\texttt{max}}(x) = \sup_{\lambda\in\Delta_d}\langle\lambda, x\rangle$ as $f(x) = \sup_{\lambda\in\Delta_d}\langle\lambda, x\rangle - h(\lambda)$ must have

$$\sigma_{\texttt{max}}(x) - \max_{\lambda\in\Delta_d} h(\lambda) \le f(x) \le \sigma_{\texttt{max}}(x) - \min_{\lambda\in\Delta_d} h(\lambda).$$

Consequently, the deviation $\|f - \sigma_{\texttt{max}}\|_\infty$ is minimized when the range of h is centered around zero, yielding an error of $\frac{1}{2}(\max_{\lambda\in\Delta_d} h(\lambda) - \min_{\lambda\in\Delta_d} h(\lambda))$.

From this, the claimed lower bound follows directly from the standard lower bounds on "regret" in online learning. Cesa-Bianchi et al. [8, Theorem 3.2.3] provide a statistical argument[4] establishing that no algorithm can achieve a regret bound lower than $(1 + o(1))\sqrt{\ln(d)T/2}$. As an upper bound, Shalev-Shwartz [13, Theorem 2.11] establishes that for any 1-strongly convex regularizer h on the simplex, the "Follow-the-Regularized-Leader" algorithm has regret at most $\sqrt{2\text{Range}(h)T}$. Combining these upper and lower bounds on regret, it follows that

$$(1 + o(1))\sqrt{\ln(d)T/2} \le \sqrt{2\text{Range}(h)T} \iff \text{Range}(h) \ge (1/4 + o(1))\ln(d).$$

Then from the dual equivalence (A.1), $\delta^\star(d) \ge (1/8 + o(1))\ln(d)$. □

[4] Cesa-Bianchi et al. [8]'s lower bound is proven by considering the task of predicting a series of T coin flips with expert advice from d experts. Online learning seeks to make predictions with minimal deviation from the performance of the best expert, i.e., "regret". If the series of coin flips is entirely fair, then no algorithm can outperform on average, while the best of d random experts will statistically outperform the average.