Entropy-Smooth Convex Optimization Cannot Be Accelerated

Jacob M. Aguirre* Dmitrii M. Ostrovskii†

Abstract

We prove an $\Omega(L/T)$ lower bound for the convergence rate of minimization in the class of functions that are convex and L-smooth relative to negative entropy on the standard d-simplex, valid for every first-order method when $d = \Omega(T^2)$. In particular, this shows that mirror descent is optimal up to a logarithmic factor in this class. This may be surprising due to the fact that accelerated methods are readily available under the assumption of smoothness in ℓ_1-norm. While Dragomir et al. (Mathematical Programming, 2022) have already showed that acceleration might be impossible under relative smoothness, their prox-function is pathological and constructed together with the hard instance. In contrast, we show non-acceleration for a specific prox-function with particularly favorable structure. We also extend the result to the quantum setting, proving the same lower bound in the class of functions L-smooth relative to the negative von Neumann entropy on the spectrahedron of $d \times d$ Hermitian positive-semidefinite matrices with unit trace.

1 Introduction

In this paper, we study convex optimization problems on the standard simplex $\Delta_d \subset \mathbb{R}^d$, of the form

$$\min_{x \in \Delta_d} f(x). \tag{1.1}$$

More precisely, we are interested in the first-order oracle complexity[1] of the natural problem class in which the objective smoothness is measured with respect to negative Shannon entropy

$$h : \Delta_d \to \mathbb{R}, \qquad h(x) = \sum_{k=1}^{d} (x)_k \log (x)_k, \tag{1.2}$$

where $(x)_k$ denotes the kth entry of x and we use the standard convention $0 \log 0 = 0$. Here, "smoothness of one function with respect to another" refers to the notion of *relative smoothness*, as defined in [2] and [14], which we now recall in the form adapted to our setting. Any function ϕ strictly convex on a domain $X \subseteq \mathbb{R}^d$ and C^1 in its relative interior, defines the Bregman divergence

$$D_\phi : X \times \mathrm{ri}(X) \to \mathbb{R}_+, \qquad D_\phi(x, u) = \phi(x) - \phi(u) - \langle \nabla \phi(u), x - u \rangle,$$

a directed measure of discrepancy of an arbitrary point $x \in X$ from the "center" point $u \in \mathrm{ri}(X)$. (In the sequel, X is assumed to lie in some affine subspace of $\mathbb{R}^d$, and we write $\nabla f(x)$ for the gradient

*Georgia Institute of Technology, H. Milton Stewart School of Industrial and Systems Engineering (ISyE), Atlanta, USA. Email: aguirre@gatech.edu.

†Georgia Institute of Technology, School of Mathematics & H. Milton Stewart School of Industrial and Systems Engineering (ISyE), Atlanta, USA. Email: ostrov@gatech.edu.

[1] In the classical sense of Nemirovski and Yudin [15]; we shall briefly recap their complexity framework in Section 2.

of $f : X \to \mathbb{R}$ restricted to the affine hull of X; a more detailed discussion is deferred to Section 2.) For $L > 0$, a function $f : X \to \mathbb{R}$ in $C^1(\mathrm{ri}(X))$ is called *L-smooth relative to ϕ* if

$$f(x) - f(u) - \langle \nabla f(u), x - u \rangle \leqslant L D_\phi(x, u) \qquad \forall (x, u) \in X \times \mathrm{ri}(X). \tag{1.3}$$

The inequality in (1.3) can be recast as $D_f(x, u) \leqslant L D_\phi(x, u)$, so (1.3) is equivalent to the convexity of $L\phi - f$. When $X = \mathbb{R}^d$ and $\phi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, the associated Bregman divergence is $\frac{1}{2}\|x - u\|_2^2$, and L-relative smoothness reduces to the usual Euclidean smoothness, i.e., L-Lipschitzness of ∇f. When $(X, \phi) = (\Delta_d, h)$, the corresponding Bregman divergence gives the Kullback–Leibler divergence. For our purposes, it is convenient to treat the latter as an extended-value function over $\Delta_d \times \Delta_d$,

$$D_h(x, u) := \begin{cases} \displaystyle\sum_{k \in \operatorname{supp} x} (x)_k \log \frac{(x)_k}{(u)_k}, & \operatorname{supp} x \subseteq \operatorname{supp} u, \\ +\infty, & \text{otherwise}, \end{cases} \tag{1.4}$$

with the convention $0 \log 0 = 0$. Geometrically, $D_h(x, u) < +\infty$ occurs when u is "at least as interior" as x; on $\Delta_d \times \mathrm{ri}(\Delta_d)$ this definition coincides with the usual Bregman divergence generated by $h(\cdot)$.

We are interested in the (Δ_d, h) case, so let us formally define the corresponding class of functions.

Definition 1. *The class $\mathsf{Ent}_d(L)$ consists of all functions $f : \Delta_d \to \mathbb{R}$ that are convex, continuously differentiable on $\mathrm{ri}(\Delta_d)$, and satisfy the following inequalities for all $(x, u) \in \Delta_d \times \mathrm{ri}(\Delta_d)$:*

$$0 \leqslant f(x) - f(u) - \langle \nabla f(u), x - u \rangle \leqslant L D_h(x, u). \tag{1.5}$$

The first inequality in (1.5) is the usual first-order convexity certificate, whereas the second one is the condition of L-smoothness relative to h on Δ_d. Note that differentiability is only required on $\mathrm{ri}(\Delta_d)$; yet, functions in $\mathsf{Ent}_d(L)$ are finite and continuous on the closed simplex. In particular, $h \in \mathsf{Ent}_d(1)$.

The notion of relative smoothness, proposed simultaneously by Bauschke et al. [2] and Lu et al. [14], is motivated by the simple yet striking observation: inequality (1.3) can be interpreted as the proximal descent lemma formulated directly in terms of the Bregman divergence at hand. Using this observation, [2] and [14] generalized the classical analysis of mirror descent [15], showing that

$$f(x_T) - \min_{x \in X} f(x) \leqslant \frac{L D_\phi(x^\star, x_0)}{T} \qquad \forall x^\star \in \operatorname*{Argmin}_{x \in X} f(x) \tag{1.6}$$

after T iterations of the algorithm initialized from x_0, all without using *any norm* on the domain. Specializing this to the entropy on Δ_d with initialization at the barycenter $d^{-1}\mathbf{1}_d$, we get the bound

$$f(x_T) - \min_{x \in \Delta_d} f(x) \leqslant \frac{L \log d}{T} \qquad \forall f \in \mathsf{Ent}_d(L). \tag{1.7}$$

To put this into perspective, the classic analysis of mirror descent proceeds by choosing a norm $\|\cdot\|$ such that ϕ is 1-strongly convex w.r.t. $\|\cdot\|$ on X, which implies convexity and L-smoothness in $\|\cdot\|$,

$$0 \leqslant f(x) - f(u) - \langle \nabla f(u), x - u \rangle \leqslant \frac{L}{2}\|x - u\|^2 \qquad \forall x, u \in X,$$

and using 1-strong convexity of ϕ at some stage of the analysis. This route still results in (1.6), but only in the smaller class of functions: indeed, the above inequality implies (1.3) but not vice versa. In particular, in the relevant to us case (Δ_d, h), the suitable norm is $\|x\|_1 = \sum_{i=1}^d |(x)_i|$; adopting the name $\mathsf{Lip}_d(L)$ for the corresponding class of functions on Δ_d, i.e. those satisfying the inequalities

$$0 \leqslant f(x) - f(u) - \langle \nabla f(u), x - u \rangle \leqslant \frac{L}{2}\|x - u\|_1^2 \qquad \forall x, u \in \Delta_d, \tag{1.8}$$

the inclusion $\mathsf{Lip}_d(L) \subsetneq \mathsf{Ent}_d(L)$ follows from the 1-strong convexity of h w.r.t. ℓ_1-norm, that is Pinsker's inequality [19]. In these terms, the guarantee in (1.7) extends from $\mathsf{Lip}_d(L)$ to $\mathsf{Ent}_d(L)$.

One may ask whether other classical results for proximal algorithms admit similar generalizations. In particular, a very natural question, mentioned already in [2], [14] and more explicitly discussed by R.-A. Dragomir in his PhD thesis [5], is whether acceleration "à la Nesterov" [16] is possible under relative smoothness. Indeed, in the class of functions convex and L-smooth on a set X with respect to a given norm $\|\cdot\|$, one can obtain $O(LD_\phi(x^\star, x_0)T^{-2})$ error using any potential that is 1-strongly convex with respect to the norm $\|\cdot\|$; see [17]. [2] With this in mind, we are led to the question:

> *Given a convex domain $X \subseteq \mathbb{R}^d$ and a potential ϕ, is there a first-order algorithm with guaranteed $O(T^{-2})$ convergence in the corresponding class of relatively smooth functions?*

One may also specialize this question to *concrete* domain-potential pairs. In particular, for (Δ_d, h)

$$f(x_T) - \min_{x\in\Delta_d} f(x) = O\left(\frac{L\log d}{T^2}\right) \qquad \forall f \in \mathsf{Lip}_d(L), \tag{1.9}$$

and the question we are left with is the following one:

Can the guarantee in (1.9) be extended to the class $\mathsf{Ent}_d(L)$?

To the best of our knowledge, both these questions are open. Our paper resolves the second question.

Our results. Focusing on the case of (Δ_d, h), we answer the acceleration question in the negative. A rigorous statement of our result relies on the basic notions of black-box complexity theory [15], to be recapped in Section 2. The simplified formulation, presented next, suffices to make our point.

Theorem 1. *Let $L > 0$, $T \geqslant 1$, and $d = \Omega(T^2)$. For any deterministic method that makes T queries $x_1, \dots, x_T \in \mathrm{ri}(\Delta_d)$ of the oracle $x \mapsto (f(x), \nabla f(x))$ and returns $x_{T+1} \in \Delta_d$, there exists $f(\cdot) \in \mathsf{Ent}_d(L)$ such that*

$$f(x_{T+1}) - \min_{x\in\Delta_d} f(x) > \frac{L}{4(T+1)}. \tag{1.10}$$

This result shows that, in contrast to $\mathsf{Lip}_d(L)$, accelerated $O(T^{-2})$ convergence cannot be attained on $\mathsf{Ent}_d(L)$. Moreover, the convergence guarantee (1.7) of entropic mirror descent with stepsize $1/L$ is optimal—up to a logarithmic factor—if the dimension is large enough, namely when $d = \Omega(T^2)$.

Let us make several remarks regarding Theorem 1.

R1. As mentioned, Theorem 1 leaves a logarithmic gap between the best available upper and lower bounds for the class $\mathsf{Ent}_d(L)$; cf. (1.7). We expect that the lower bound in (1.10) is loose, and the missing $\log d$ factor can be recovered. A promising approach is outlined in Section 5.

R2. When $d = \exp(T)$, the right-hand side of (1.10) reads as $O(LT^{-2}\log d)$, matching the upper bound in (1.9) and diverging from (1.7) by T. Therefore, acceleration is formally not ruled out for such "extremely high-dimensional" problems; however, this regime is of limited practical interest anyway, since even a single iteration of a deterministic first-order method is prohibitive.

R3. The quadratic dependence of the dimension on T is crucial for our resisting oracle construction. This gives an $O(T^{-2})$ "safety margin" that ensures consistency of the transcript after T steps.

[2] We write $g = O(f)$ or $f = \Omega(g)$ if there exists a universal constant $c > 0$ such that $g \leqslant cf$ for all argument values.

R4. Restriction to interior queries is natural: indeed, f might be non-differentiable on the relative boundary; this is not a pathological case either, as shown by the scaled negative entropy Lh.

Theorem 1 admits a noncommutative generalization, pertaining to functions on the "spectraplex"

$$\mathbf{\Delta}_d := \{X \in \mathbf{H}_+^d : \operatorname{tr}(X) = 1\}$$

where $\mathbf{H}_+^d$ is the Hermitian positive-semidefinite cone. In this setting, defined rigorously in Section 4, we introduce the class $\mathsf{Ent}_d^{\mathsf{H}}(L)$ of functions on $\mathbf{\Delta}_d$ that are convex, $C^1(\operatorname{ri}(\mathbf{\Delta}_d))$, and L-smooth relative to the negative von Neumann entropy $\operatorname{tr}(X \log X)$, which is the spectral analog of $h(x)$. It turns out that this noncommutative setting can be reduced to the diagonal case, corresponding to the class $\mathsf{Ent}_d(L)$ and the setting of Theorem 5. This is done by embedding the hard instance of Theorem 1 diagonally, and using Lindblad's inequality [13] to argue that the composition $F(\cdot) = f(\operatorname{diag}(\cdot))$ of $f(\cdot) \in \mathsf{Ent}_d(L)$ and the diagonal extraction map belongs to $\mathsf{Ent}_d^{\mathsf{H}}(L)$. The final ingredient is the observation that the diagonal of the (matrix) transcript of an arbitrary first-order method run on F emulates the transcript of some first-order method run on f. We defer further details to Section 4.

Finally, as a minor contribution, in Section 5 we find an explicit form of the pointwise minimal interpolant in the class $\mathsf{Ent}_d(L)$ for given interpolation data $\mathcal{I} = (x_i, f_i, \xi_i)_{i=1}^N$, defined as the function $f \in \mathsf{Ent}_d(L)$ that interpolates $\mathcal{I}$ to first order, i.e. satisfies $f(x_i) = f_i$ and $\nabla f(x_i) = \xi_i$ for all $i \in [N]$, and is no larger than any other such function at every point $x \in \Delta_d$. This result follows easily from the results in R.-A. Dragomir's PhD thesis [5]; we record it due to its pivotal role in the promising approach of improving Theorem 1. Further details are deferred to Section 5.

Summary of the approach. Our construction is based on the *right Bregman–Moreau envelope* [4] applied to an incremental coordinatewise construction in the spirit of Guzmán and Nemirovski [9]. The right Bregman-Moreau envelope replaces the inf-convolution smoothing operator of [9] as the smoothing mechanism. This smoothing mechanism heavily relies upon the f-divergence properties of the KL divergence, namely its joint and separate convexity in both arguments; as demonstrated in [3], among Bregman divergences, these properties occur only in the Euclidean and KL cases. Crucially, the right envelope of a convex function is convex and smooth with respect to the potential (see [4]); meanwhile, the local smoothing of [9] cannot be employed, as it would give an instance in $\mathsf{Lip}_d(L)$, and the latter class *does* admit acceleration.

Related work. In addition to [9], closely relevant to ours is the work of Dragomir, Taylor, d'Aspremont, and Bolte [6], who showed the following: there *exists a pair* (f, ϕ) in which ϕ is strictly convex on the positive orthant $\mathbb{R}_+^d$, f is convex and smooth relative to ϕ, and any first-order method has worst-case convergence rate no better than $\Omega(1/T)$. Their result is based on so-called performance estimation techniques (e.g. [7], [21]), allowing one to find worst-case instances in infinite-dimensional functional classes by reducing the corresponding optimization problem to a (finite-dimensional) semidefinite program (SDP), dubbed "performance estimation program" (PEP). Usually, the PEP methodology is limited to Euclidean geometry, since the SDP representation arises from the Gram matrix that juxtaposes the candidate iterates x_k and gradients $g_k = \nabla f(x_k)$ in the transcript; as a result, PEPs are poorly suited for dealing with non-Euclidean geometries, where the terms of the form $\|x_k - x_l\|^2$ and $\|g_k - g_l\|_*^2$ cannot be expressed linearly via dot products. The authors of [6] elegantly sidestepped this limitation by allowing the potential itself to vary with f, so that the PEP constructs a worst-case *pair* (f, ϕ). However, the resulting potential is quite pathological—as one might expect with its adversarial origin—and it may very well be that some *specific* pairs (X, ϕ) do admit acceleration. While our result shows this is not the case

for (Δ_d, h), other highly structured settings remain open, most notably that of the logarithmic-barrier potential $\phi(x) = -\sum_{i=1}^d \log(x)_i$ on $\mathbb{R}^d_+$.

Chapter 5 of Dragomir's PhD thesis [5] provides exact first-order interpolation conditions for the class $\mathsf{Ent}_d^+(L)$ of functions on the nonnegative orthant $\mathbb{R}^d_+$ that are convex and smooth relative to the unnormalized entropy $h_{\mathrm{e}}(x) := \sum_{k=1}^d (x)_k \log(x)_k - (x)_k$; such conditions are the inequalities imposed on the interpolation data $\mathcal{I} = (x_i, f_i, \xi_i)_{i=1}^N$, whose validity is equivalent to the existence of a function in $\mathsf{Ent}_d^+(L)$ that interpolates $\mathcal{I}$. Conceptually, these conditions are counterparts of SDP-representable interpolation conditions in the Euclidean case, which are the crux of the PEP framework, and their existence crucially relies upon the joint convexity of KL divergence (see [5, Lem. 5.3.3 and Rem. 3]). In Section 5, we generalize these conditions for the class $\mathsf{Ent}_d(L)$ of entropy-smooth functions on Δ_d, and then use them to derive the pointwise minimal interpolant. It appears that these results might be leveraged to recover the logarithmic factor missing in (1.10). In Section 5 we further discuss this possibility in the light of the recent work of Florea and Nesterov [8].

Finally, we mention the triangular scaling exponent framework of [10], which seeks to relax the standard assumption of 1-strong convexity of ϕ w.r.t. a norm while retaining accelerated convergence. Our critique is that, while leading to locally adaptive algorithms that prove to be highly effective for optimization problems arising in some applications, the triangular scaling condition is hard to ensure in the worst case. In particular, Theorem 1 shows that this condition is not satisfied for $\mathsf{Ent}_d(L)$.

Roadmap. In Section 2 we collect the ingredients for the proof of Theorem 1. To that end, we recall the properties of the right Bregman-Moreau envelope and the associated smoothing operator, formalize the oracle complexity framework, and describe the family of hard instances subsequently used in the proof. In Section 3 we carry out the proof of Theorem 1, and in Section 4 we formulate and discuss its noncommutative generalization. In Section 5, we derive exact interpolation conditions and the form of pointwise minimal interpolant in $\mathsf{Ent}_d(L)$.

Notation. We denote $[d] := \{1, 2, \dots, d\}$. We let e_i be the ith canonical basis vector in $\mathbb{R}^d$ and use the concise notation $(x)_i := \langle x, e_i \rangle$ for the ith entry of $x \in \mathbb{R}^d$. We let 1_d be the all-ones vector in $\mathbb{R}^d$. As previously mentioned, we write ∇f for the tangent gradient of $f \in C^1(\mathrm{ri}(\Delta_d))$; in particular, $\nabla f(x) \in \mathrm{span}\{1_d\}^\perp$ for any $x \in \mathrm{ri}(\Delta_d)$. Additional notation is introduced as necessary.

2 Building blocks

Tangent gradients. In Sections 2–3, we work with functions defined on Δ_d and differentiable in $\mathrm{ri}(\Delta_d)$; this includes every $f \in \mathsf{Ent}_d(L)$ and, in particular, the hard instances presented in Section 2.3. While such functions may arise as restrictions of $C^1(\mathbb{R}^d_{++})$-functions (as, for example, h itself), this is not required in general. For such a function f, we might differentiate it at $x \in \mathrm{ri}(\Delta_d)$ intrinsically over the affine hull of Δ_d, by identifying the affine subspace $\mathrm{span}\{1_d\}^\perp + d^{-1}1_d$ with $\mathbb{R}^{d-1}$, via an affine homeomorphism (with orthogonal linear transformation), taking the full gradient of the resulting function in $\mathbb{R}^{d-1}$, and changing the basis to account for the transformation. We take this as the definition of the tangent gradient ∇f. Note that ∇f always belongs to the tangent space $\Theta_d := \mathrm{span}\{1_d\}^\perp$; moreover, if f is defined in a proper $\mathbb{R}^d$-neighborhood of x, then $\nabla f(x)$ coincides with the Euclidean projection onto Θ_d of the full gradient; denoting the latter with $\nabla_\star f$,

$$\nabla f(x) := \nabla_\star f(x) - d^{-1}\langle \nabla_\star f(x), 1_d \rangle 1_d, \qquad x \in \mathrm{ri}(\Delta_d). \tag{2.1}$$

In fact, even if f is defined only over Δ_d, this formula remains valid if we extend f to $\mathbb{R}^d_+$ with differentiability over $\mathbb{R}^d_{++}$. Such an extension is not unique, and $\nabla_* f$ in general depends on the

chosen extension; however, the dependence is only in the normal component $d^{-1}\langle \nabla_\star f(x), 1_d\rangle 1_d$, whereas $\nabla f(x)$ is an invariant depending only on the values of f in a Θ_d-neighborhood of $x \in \mathrm{ri}(\Delta_d)$.

2.1 Smoothing with right Bregman–Moreau envelope

Let $g : \Delta_d \to \mathbb{R}$ be convex and continuous. Adapting the standard definition to our setting, *the right Bregman–Moreau envelope* of g with parameter $\eta > 0$ is defined by its values on the simplex:

$$S_\eta[g](x) := \min_{u \in \Delta_d} \left\{ g(u) + \eta^{-1} D_h(x, u) \right\}, \qquad x \in \Delta_d. \tag{2.2}$$

For $x \in \mathrm{ri}(\Delta_d)$, the minimizer is unique and attained on $\mathrm{ri}(\Delta_d)$; this defines $\mathrm{prox}_{\eta g} : \mathrm{ri}(\Delta_d) \to \mathrm{ri}(\Delta_d)$,

$$\mathrm{prox}_{\eta g}(x) = \operatorname*{argmin}_{u \in \Delta_d} \left\{ \eta g(u) + D_h(x, u) \right\}, \qquad x \in \mathrm{ri}(\Delta_d), \tag{2.3}$$

called *the prox-mapping* of $\eta g(\cdot)$.

Next, we collect some properties of the right Bregman–Moreau envelope and proximal mapping.

Lemma 2. *For $\eta > 0$, the prox-mapping $\mathrm{prox}_{\eta g}(\cdot)$ is well-defined and continuous on $\mathrm{ri}(\Delta_d)$.*

Proof. As a continuous function on Δ_d, g is bounded from below. From this and the expression (1.4) for KL divergence, we see that for $x \in \mathrm{ri}(\Delta_d)$, minimization in (2.2) can be restricted to $\mathrm{ri}(\Delta_d)$. The same expression shows that the Hessian of $D_h(x, \cdot)$ is diagonal with positive entries over $\mathrm{ri}(\Delta_d)$, so the objective in (2.2) is strictly convex and the minimizer is unique.[3] Continuity of $\mathrm{prox}_{\eta g}(\cdot)$ follows from the optimal-set mapping theorem of Rockafellar and Wets (see [20, Example 5.22]), applied to

$$\Phi(u, x) := g(u) + \eta^{-1} D_h(x, u) + \delta_{\Delta_d}(u), \qquad x \in \mathrm{ri}(\Delta_d).$$

Since Δ_d is compact, the function Φ is proper, lower semicontinuous, and level-bounded in u locally uniformly in x. Since $\Phi(\mathrm{prox}_{\eta g}(\bar{x}), x)$ is continuous in x at $\bar{x}$, the cited theorem implies continuity of the value function at any $\bar{x} \in \mathrm{ri}(\Delta_d)$, and outer continuity of the corresponding minimizing set-mapping. Since this mapping outputs a singleton, this is the continuity of $x \mapsto \mathrm{prox}_{\eta g}(x)$. □

Lemma 3. *The function $S_\eta[g](\cdot)$ is convex on Δ_d and continuously differentiable on $\mathrm{ri}(\Delta_d)$, with*

$$\nabla S_\eta[g](x) = -\eta^{-1}(\log \mathrm{prox}_{\eta g}(x) - \log x), \tag{2.4}$$

where $\log(\cdot)$ is applied entrywise. Moreover, $S_\eta[g](\cdot)$ satisfies the inequality for $x \in \mathrm{ri}(\Delta_d)$ and $y \in \Delta_d$:

$$S_\eta[g](y) \leqslant S_\eta[g](x) + \langle \nabla S_\eta[g](x), y - x\rangle + \eta^{-1} D_h(y, x). \tag{2.5}$$

In particular, $S_\eta[g](\cdot) \in \mathsf{Ent}(\eta^{-1})$.

Proof. By the perspective rule applied coordinatewise to $x \mapsto x \log x$, extended-value KL divergence (1.4) is jointly convex, whence convexity of $S_\eta[g]$ follows by the partial minimization rule. For (2.4), apply Danskin's theorem and the first-order optimality condition to (2.2) with $x \in \mathrm{ri}(\Delta_d)$. Finally, invoking (2.2) at x and y, with $u := \mathrm{prox}_{\eta g}(x)$ as a feasible point in the latter case, gives

$$\eta S_\eta[g](y) - \eta S_\eta[g](x) \;\leqslant\; D_h(y, u) - D_h(x, u) \;=\; D_h(y, x) + \langle \log x - \log u, y - x\rangle. \tag{2.6}$$

We used the three-point identity in the final step. Combining this with (2.4) and (2.1) gives (2.5). □

[3] Note that $D_h(x, \cdot)$ is infinitely differentiable as a function over $\mathbb{R}^d_{++}$, hence we may consider its full Hessian here.

Next, we establish *locality* of the smoothing mechanism based upon the right Bregman–Moreau envelope, ensuring that the envelopes of locally coinciding functions (locally) agree to first order.

Lemma 4. *Let $g, \tilde{g} : \Delta_d \to \mathbb{R}$ be finite, continuous, convex, and such that $\tilde{g} \geqslant g$ on Δ_d. Fix arbitrary $\hat{x} \in \mathrm{ri}(\Delta_d)$ and let $\hat{u} := \mathrm{prox}_{\eta g}(\hat{x})$. If $\tilde{g} = g$ in a relative neighborhood of $\hat{u}$, then*

$$S_\eta[\tilde{g}](x) = S_\eta[g](x) \qquad \forall x \in \mathcal{X} \tag{2.7}$$

in some relative neighborhood $\mathcal{X}$ of $\hat{x}$. In particular, $\nabla S_\eta[\tilde{g}](\hat{x}) = \nabla S_\eta[g](\hat{x})$.

Proof. Let $\mathcal{U}$ be a relative neighborhood of $\hat{u}$ where $\tilde{g} = g$. By continuity of $\mathrm{prox}_{\eta g}(\cdot)$, see Lemma 2, in some relative neighborhood $\mathcal{X}$ of $\hat{x}$ one has

$$\mathrm{prox}_{\eta g}(x) \in \mathcal{U} \qquad \forall x \in \mathcal{X}.$$

Now, fix arbitrary $x \in \mathcal{X}$ and put $u := \mathrm{prox}_{\eta g}(x)$. Since $\tilde{g} \geqslant g$ everywhere on Δ_d, by (2.2) we get

$$S_\eta[\tilde{g}](x) \geqslant S_\eta[g](x).$$

On the other hand, by invoking (2.2) for $\tilde{g}$ with u as a feasible point (note that $u \in \mathcal{U}$), we get

$$\begin{aligned} S_\eta[\tilde{g}](x) &\leqslant \tilde{g}(u) + \eta^{-1} D_h(x, u) \\ &= g(u) + \eta^{-1} D_h(x, u) = S_\eta[g](x). \end{aligned}$$

Thus $S_\eta[\tilde{g}] = S_\eta[g]$ on $\mathcal{X}$, and $\nabla S_\eta[\tilde{g}](\hat{x}) = \nabla S_\eta[g](\hat{x})$ follows from differentiability on $\mathrm{ri}(\Delta_d)$. □

2.2 Oracle complexity model and formal statement of the result

We work in the oracle complexity model of [15], adapted to account for the lack of differentiability on the boundary for objectives in $\mathsf{Ent}_d(L)$. Let us give a brief yet rigorous summary of this model.

Any $f \in \mathsf{Ent}_d(L)$ specifies a *first-order oracle* $\mathscr{O}_f$ which, when queried at any $x \in \mathrm{ri}(\Delta_d)$, returns

$$\mathscr{O}_f(x) := (f(x), \nabla f(x)) \tag{2.8}$$

where ∇f is the tangent gradient of f. As we explained in the beginning of Section 2, the use of ∇f is without loss of generality, since f is only defined over Δ_d whose affine hull is parallel to Θ_d. However, one could still ask if anything could be gained by allowing the objective to be defined on the whole orthant $\mathbb{R}^d_+$, with access to the full gradient oracle, while only requiring the relative smoothness condition in (1.5) to hold on Δ_d. To address this, in Appendix A we show that (1.10) remains valid for minimization in the class of 1-homogeneous extensions [18] of functions in $\mathsf{Ent}_d(L)$.

A deterministic *T-step first-order method* (FOM) makes T sequential queries $x_1, \dots, x_T \in \mathrm{ri}(\Delta_d)$ of such an $\mathscr{O}_f$, and outputs $x_{T+1} \in \Delta_d$. Thus, a T-step FOM is specified by a collection of mappings

$$\begin{aligned} &\Phi_t : \Omega^{t-1} \to \mathrm{ri}(\Delta_d), \qquad t \in [T]; \\ &\bar{\Phi}_{T+1} : \Omega^T \to \Delta_d, \end{aligned} \tag{2.9}$$

where $\Omega = \mathbb{R} \times \Theta_d$, and Ω^0 is a singleton (so Φ_1 merely selects a specific point $x_1 \in \mathrm{ri}(\Delta_d)$). We let $\mathsf{FOM}_d(T)$ be the class of all T-step FOMs, as per (2.9). When a given method $\mathscr{M} \in \mathsf{FOM}_d(T)$ is instantiated with an oracle $\mathscr{O}_f$ associated to a specific instance $f \in \mathsf{Ent}(L)$, the resulting sequence

$$x_1^{\mathscr{M}} = \Phi_1, \;\; x_2^{\mathscr{M}}(f) = \Phi_2(\mathscr{O}_f(x_1^{\mathscr{M}})), \;\; \dots, \;\; x_{T+1}^{\mathscr{M}}(f) = \bar{\Phi}_{T+1}(\mathscr{O}_f(x_1^{\mathscr{M}}), \dots, \mathscr{O}_f(x_T^{\mathscr{M}}(f))), \tag{2.10}$$

along with the responses $\mathscr{O}_f(x_1^{\mathscr{M}}), \mathscr{O}_f(x_2^{\mathscr{M}}(f)), \ldots, \mathscr{O}_f(x_T^{\mathscr{M}}(f))$, is called *the transcript* of $\mathscr{M}$ on f. Note that $\mathscr{M}$ selects each x_t before receiving the oracle response $\mathscr{O}_f(x_t)$, and the crucial property of a transcript is its *consistency*: the oracle responses in (2.10) correspond to the same $f \in \mathsf{Ent}_d(L)$.

To quantify the hardness of first-order optimization over $\mathsf{Ent}_d(L)$, we define its *minimax T-risk*

$$\mathsf{Risk}_d(T, L) := \inf_{\mathscr{M} \in \mathsf{FOM}_d(T)} \sup_{f \in \mathsf{Ent}_d(L)} \left\{ f(x_{T+1}^{\mathscr{M}}(f)) - \min_{x \in \Delta_d} f(x) \right\}. \tag{2.11}$$

With this definition at hand, we are now in the position to give the rigorous statement of Theorem 1.

Theorem 5. *For all $L > 0$, $T \geqslant 1$ and $d \geqslant 8(T+1)^2 + T$, the minimax T-risk of $\mathsf{Ent}_d(L)$ satisfies*

$$\mathsf{Risk}_d(T, L) > \frac{L}{4(T+1)}. \tag{2.12}$$

To prove Theorem 5, we shall proceed via the "resisting oracle" approach: interacting with arbitrary $\mathscr{M} \in \mathsf{FOM}_d(T)$, we shall construct a sequence $\omega_1, \ldots, \omega_T \in \Omega$, with $\omega_t = \omega_t(x_1, \ldots, x_t)$, which, along with the associated sequence of queries $x_1 = \Phi_1,\ x_2 = \Phi_2(\omega_1), \ldots, x_{T+1} = \bar{\Phi}_{T+1}(\omega_1, \ldots, \omega_T)$, corresponds to a consistent transcript, i.e., matches some $f \in \mathsf{Ent}_d(L)$ in the sense that $\omega_t = \mathscr{O}_f(x_t^{\mathscr{M}})$. Our hard instance will be a maximum of affine functions, smoothed via the right Bregman-Moreau envelope to put it in $\mathsf{Ent}_d(L)$. In what follows, we first describe the general family of such nonsmooth functions and study their envelopes, then present the adaptive construction and complete the proof.

Before we proceed, a simple remark is in order. The minimax T-risk is 1-homogeneous in L, i.e.

$$\mathsf{Risk}_d(T, L) = L\mathsf{Risk}_d(T, 1), \tag{2.13}$$

as seen by noting that $f \in \mathsf{Ent}_d(1)$ implies $Lf \in \mathsf{Ent}_d(L)$, and that passing from f to Lf does not influence consistency of a transcript. Since the right-hand side of (2.12) has the same homogeneity, it suffices to treat the $L = 1$ case; in other words, exhibit $f \in \mathsf{Ent}(1)$ that certifies (2.12) with $L = 1$.

2.3 Hard instances and their properties

Our hard instances are based on nonsmooth functions that are maxima of shifted coordinate atoms

$$g(u) = \max_{k \in [K]} \{-(u)_{j_k} - b_k\}, \tag{2.14}$$

indexed by a finite set $[K]$, with distinct coordinates j_k and offsets $b_k \geqslant 0$. For brevity, we shall refer to such functions as *backbones*. In a backbone, each atom is 1-Lipschitz in ℓ_∞-norm, and so is the whole backbone; this will be used later on. By definition, the set of *g-active atoms* at $u \in \mathrm{ri}(\Delta_d)$ is

$$A_g(u) := \{k \in [K] :\ -(u)_{j_k} - b_k = g(u)\},$$

and $\{j_k : k \in A_g(u)\}$ are *g-active coordinates at u*. To obtain a legitimate $f \in \mathsf{Ent}_d(1)$, we take the right Bregman–Moreau envelope of a backbone g. Indeed, by Lemma 3 we have $\eta S_\eta[g](\cdot) \in \mathsf{Ent}_d(1)$ for any $\eta > 0$; later on, we shall fix $\eta = 1/2$.

Next, we prove some regularity properties of the right Bregman-Moreau envelope of a backbone. We begin with a lemma that controls the entrywise growth of the prox-mapping associated with ηg.

Lemma 6. *Let g be as in (2.14), and $\eta < 1$. For all $x \in \mathrm{ri}(\Delta_d)$ and $j \in [d]$,*

$$0 < \eta(\mathrm{prox}_{\eta g}(x))_j \leqslant \frac{\eta}{1-\eta}(x)_j. \tag{2.15}$$

In particular, for $\eta \leqslant \frac{1}{2}$ one has $0 < \eta\,\mathrm{prox}_{\eta g}(x) \leqslant x$ entrywise.

Proof. Let $\hat{u} := \mathrm{prox}_{\eta g}(x)$. The left inequality in (2.15), i.e. that $\mathrm{prox}_{\eta g}(x) \in \mathrm{ri}(\Delta_d)$, is immediate: if $\hat{u}$ has a zero entry, then $D_h(x, \hat{u}) = +\infty$ by (1.4), which contradicts the optimality of $\hat{u}$. For the right inequality, since the minimum in (2.2) is attained on $\mathrm{ri}(\Delta_d)$, the optimality condition writes as

$$\exists \mu \in \mathbb{R} : \quad \partial g(\hat{u}) \ni \eta^{-1} \frac{x}{\hat{u}} - \mu \mathbf{1}_d, \tag{2.16}$$

where the division is entrywise. Now, let $(j_k)_{k \in A_g(\hat{u})}$ be the active part of the finite sequence $(j_k)_{k \in [K]}$ in (2.14), and consider arbitrary subgradient $\xi(\hat{u}) = -\sum_{k \in A_g(\hat{u})} w_k e_{j_k}$ of $g(\cdot)$ at $\hat{u}$, where $w_k \geqslant 0$ and $\sum_{k \in A_g(\hat{u})} w_k = 1$. Letting λ_i, for $i \in [d]$, be the sum of all active multipliers for each coordinate i,

$$\lambda_i := \sum_{k \,\in\, A_g(\hat{u})} w_k \, \delta_{i j_k},$$

we can write $\xi(\hat{u}) = -\sum_{i \in [d]} \lambda_i e_i$ where $(\lambda_1, \dots, \lambda_d) \in \Delta_d$ (in particular, $\lambda_i \leqslant 1$). Plugging in (2.16) the relevant subgradient $\xi(\hat{u})$—i.e., one realizing the inclusion in (2.16)—and rearranging, we get

$$\frac{(x)_i}{(\hat{u})_i} = \eta(\mu - \lambda_i), \quad i \in [d].$$

Multiplying by $(\hat{u})_i$, summing over i, and using that $x, \hat{u} \in \Delta_d$, we see that $\eta(\mu - \hat{m}_\lambda) = 1$, where $\hat{m}_\lambda := \sum_{i \in [d]} \lambda_i (\hat{u})_i$ is the λ-weighted average of the entries of $\hat{u}$. Combining with the above,

$$\frac{(x)_i}{(\hat{u})_i} = 1 + \eta(\hat{m}_\lambda - \lambda_i), \quad i \in [d]. \tag{2.17}$$

Noting that $\hat{m}_\lambda \geqslant 0$ and $\lambda_i \leqslant 1$, we get $(\hat{u})_i \leqslant (1-\eta)^{-1}(x)_i$, that is the right inequality in (2.15). □

The next lemma controls the value decrease of a backbone caused by smoothing.

Lemma 7. *Let g be as in (2.14), $x \in \mathrm{ri}(\Delta_d)$, and $\hat{u} = \mathrm{prox}_{\eta g}(x)$. Furthermore, suppose that $(\hat{u})_{j_k} \leqslant M$ for all g-active coordinates at $\hat{u}$, i.e. for all $k \in A_g(\hat{u})$. Then*

$$S_\eta[g](x) \geqslant g(x) - \eta M. \tag{2.18}$$

Proof. Rearranging (2.17) gives $(x)_i - (\hat{u})_i = \eta(\hat{m}_\lambda - \lambda_i)(\hat{u})_i$ in terms of $\hat{m}_\lambda = \sum_{i \in [d]} \lambda_i (\hat{u})_i$, the quantity introduced in the proof of Lemma 6. Whence

$$|(x)_i - (\hat{u})_i| = \eta |\hat{m}_\lambda - \lambda_i| (\hat{u})_i \,. \tag{2.19}$$

Since $\hat{m}_\lambda$ is a convex combination of the active-coordinate values $\{(\hat{u})_{j_k}, k \in A_g(\hat{u})\}$, we have $\hat{m}_\lambda \leqslant M$ by the premise of the lemma. Let us show that

$$\|x - \hat{u}\|_\infty \leqslant \eta M. \tag{2.20}$$

To that end, since $\hat{m}_\lambda, \lambda_i \in [0, 1]$, we get $|(x)_i - (\hat{u})_i| \leqslant \eta M$ for active coordinates $i \in \{j_k, k \in A_g(\hat{u})\}$. Meanwhile, for inactive coordinates $\lambda_i = 0$, so (2.19) and $(\hat{u})_i \leqslant 1$ result in $|(x)_i - (\hat{u})_i| \leqslant \eta \hat{m}_\lambda \leqslant \eta M$. This verifies (2.20). Now, by the 1-Lipschitzness of a backbone, it follows that $g(\hat{u}) \geqslant g(x) - \eta M$. In turn, this implies $S_\eta[g](x) = g(\hat{u}) + \eta^{-1} D_h(x, \hat{u}) \geqslant g(\hat{u}) \geqslant g(x) - \eta M$, as claimed in (2.18). □

3 Proof of Theorem 1

In this section, we prove Theorem 5, and Theorem 1 along with it. To that end, we first implement the resisting oracle announced in Section 2.2: interacting with an arbitrary method $\mathscr{M} \in \mathsf{FOM}_d(T)$, cf. (2.9)–(2.10), after each query we add a new atom at the currently smallest-mass coordinate, with offset increased in constant increments. Smoothing via the right Bregman–Moreau envelope with $\eta = \frac{1}{2}$, while producing an instance in $\mathsf{Ent}_d(1)$ by Lemma 3, ensures that a new atom is strictly dominated near all the previous queries; this effect is attained through Lemma 7 and offsets. This guarantees transcript consistency: the final instance matches the earlier oracle answers.

The argument proceeds in three stages. In Section 3.1, we detail the construction. In Section 3.2, we prove a locality lemma ensuring that new atoms are dominated by the previous ones, and deduce transcript consistency. In Section 3.3, we bound the suboptimality gap; this is done by augmenting the exposed face with an extra dimension and using the center of the augmented face.

3.1 Constructing the hard instance

Recall the assumption $d \geqslant 8(T+1)^2 + T$ in the premise. Throughout, fix $\mathscr{M} \in \mathsf{FOM}_d(T)$ and set

$$\varepsilon_T := \frac{1}{8(T+1)^2}\,, \qquad \rho := \frac{2}{3}\,, \qquad \delta_T := (2+\rho)\varepsilon_T = \frac{1}{3(T+1)^2}\,. \tag{3.1}$$

We shall select distinct coordinates $j_1, \dots, j_T, j_{T+1}$ incrementally, in response to $\mathscr{M}$'s queries.

Rounds $t \in [T]$. Once $\mathscr{M}$ has issued a query $x_t \in \mathrm{ri}(\Delta_d)$ determined via (2.10) by the previous answers, we construct the answer to x_t by selecting a coordinate where x_t has the smallest mass:

$$j_t \in \operatorname*{Argmin}_{i \in [d]\setminus E_t} \langle x_t, e_i\rangle, \qquad E_t := \{j_1, \dots, j_{t-1}\}, \tag{3.2}$$

where $E_1 = \emptyset$ by convention. We then let

$$a_t(x) := -\langle x, e_{j_t}\rangle - (t-1)\delta_T, \qquad g_t(x) := \max_{s\in[t]} a_s(x), \qquad f_t := \tfrac{1}{2} S_{1/2}[g_t] \tag{3.3}$$

and return

$$\mathscr{O}_{f_t}(x_t) = (f_t(x_t), \nabla f_t(x_t)), \tag{3.4}$$

cf. (2.8), as the answer to query x_t. Note that g_t, cf. (3.3), is a backbone in the sense of (2.14). This procedure specifies the resisting oracle at rounds $t \in [T]$, corresponding to $\mathscr{M}$'s queries $x_1, \dots, x_T$.

Final round. To finalize the construction and exhibit the hard instance, we use the candidate minimizer $x_{T+1} \in \Delta_d$ returned by $\mathscr{M}$ after the final oracle call. Namely, we invoke (3.2) with $t = T+1$, producing

$$j_{T+1} \in \operatorname*{Argmin}_{i \in [d]\setminus E_T} \langle x_{T+1}, e_i\rangle;$$

we then let $g_{T+1}(x) := \max\{g_T(x), a_{T+1}(x)\}$ with $a_{T+1}(x) := -\langle x, e_{j_{T+1}}\rangle - T\delta_T$, cf. (3.3), and take

$$f_{T+1} := \tfrac{1}{2} S_{1/2}[g_{T+1}] \tag{3.5}$$

as the hard instance. Notice that g_{T+1} is still a backbone, therefore $f_{T+1} \in \mathsf{Ent}_d(1)$ by Lemma 3.

Before we begin the proof of transcript consistency, let us record one implication of Lemma 6.

Proposition 8. *For every $t \in [T]$, selection rule (3.2) guarantees the following for ε_T as in (3.1):*

$$\langle \mathrm{prox}_{\frac{1}{2}g_t}(x_t), e_{j_t}\rangle \leqslant 2\langle x_t, e_{j_t}\rangle \leqslant 2\varepsilon_T. \tag{3.6}$$

Moreover, the right-hand inequality in (3.6) extends to $t = T+1$, i.e. it holds that $\langle x_{T+1}, e_{j_{T+1}}\rangle \leqslant \varepsilon_T$.

Proof. For any $t \in [T+1]$, the cardinality of exposed support $E_t = \{j_1, \dots, j_{t-1}\}$ is at most T, so that of $H_t := [d] \setminus E_t$ is at least $d - T \geqslant {\varepsilon_T}^{-1}$, cf. (3.1). As such, $\min_{i \in H_t} \langle x_t, e_i \rangle > \varepsilon_T$ would imply

$$\langle x_t, 1_d \rangle \geqslant \sum_{i \in H_t} \langle x_t, e_i \rangle \geqslant |H_t| \min_{i \in H_t} \langle x_t, e_i \rangle > 1,$$

contradicting $x_t \in \Delta_d$. This proves the right-hand inequality in (3.6) for all $t \in [T+1]$. For $t \in [T]$, we additionally have $x_t \in \mathrm{ri}(\Delta_d)$, and the left-hand inequality in (3.6) follows by Lemma 6. □

3.2 Ensuring transcript consistency

In this section, we establish transcript consistency for the construction presented in Section 3.1. The crux of the argument is that the "fresh" atom introduced at round t remains strictly dominated ("locally invisible") in a vicinity $\mathcal{U}_s$ of $\hat{u}_s = \mathrm{prox}_{\frac{1}{2}g_s}(x_s)$; thus, $g_t = g_s$ over $\mathcal{U}_s$. By Lemma 4, this gives local coincidence of $f_s = S_{1/2}[g_s]$ and $f_{T+1} = S_{1/2}[g_{T+1}]$ in some relative neighborhood of x_s.

Lemma 9. *For all $1 \leqslant s < t \leqslant T+1$, the following holds.*

1. *The point $\hat{u}_s = \mathrm{prox}_{\frac{1}{2}g_s}(x_s)$ satisfies*

$$a_t(\hat{u}_s) \leqslant g_s(\hat{u}_s) - \rho\varepsilon_T. \tag{3.7}$$

2. *As a result, there is a relative neighborhood $\mathcal{X}_s \subseteq \mathrm{ri}(\Delta_d)$ of x_s where one has $f_t(x) = f_s(x)$.*

Proof. **1°**. By (3.3), a backbone dominates its last atom, so $g_s(\hat{u}_s) \geqslant a_s(\hat{u}_s) = -\langle \hat{u}_s, e_{j_s} \rangle - (s-1)\delta_T$. When combined with the bound $\langle \hat{u}_s, e_{j_s} \rangle \leqslant 2\varepsilon_T$ furnished by Proposition 8, cf. (3.6), this gives

$$g_s(\hat{u}_s) \geqslant -2\varepsilon_T - (s-1)\delta_T.$$

On the other hand, since $\hat{u}_s$ has nonnegative entries, $a_t(\hat{u}_s) = -\langle \hat{u}_s, e_{j_t} \rangle - (t-1)\delta_T \leqslant -(t-1)\delta_T$. Subtracting the two estimates and using the facts that $t > s$ and $\delta_T = (2+\rho)\varepsilon_T$, cf. (3.1), we get

$$g_s(\hat{u}_s) - a_t(\hat{u}_s) \;\geqslant\; (t-s)\delta_T - 2\varepsilon_T \;\geqslant\; \delta_T - 2\varepsilon_T \;=\; \rho\varepsilon_T.$$

2°. Our plan is to invoke Lemma 4. Fix $s < t$ and consider the neighborhood $\mathcal{U}_s$ of $\hat{u}_s$ as follows:

$$\mathcal{U}_s := \left\{ u \in \mathrm{ri}(\Delta_d) : \|u - \hat{u}_s\|_\infty \leqslant \frac{\rho\varepsilon_T}{4} \right\}.$$

Combining (3.7) with 1-Lipschitzness of $a_t(\cdot)$ and $g_s(\cdot)$ in ℓ_∞-norm, one has for all $u \in \mathcal{U}_s$ and $t > s$:

$$g_s(u) - a_t(u) \geqslant g_s(\hat{u}_s) - a_t(\hat{u}_s) - \frac{\rho\varepsilon_T}{2} \overset{(3.7)}{\geqslant} \rho\varepsilon_T - \frac{\rho\varepsilon_T}{2} = \frac{\rho\varepsilon_T}{2} > 0.$$

Thus, all future atoms a_t, $t > s$, are strictly dominated by g_s in the relative neighborhood $\mathcal{U}_s$ of $\hat{u}_s$ (note that we used that $\hat{u}_s$ is in the relative interior of Δ_d). Applying this to $t = s+1, \dots, T+1$,

$$g_{T+1}(u) = \max\left\{ g_s(u), \max_{s < t \leqslant T+1} a_t(u) \right\} = g_s(u) \qquad \forall u \in \mathcal{U}_s.$$

Meanwhile, we have $g_{T+1} \geqslant g_s$ pointwise on Δ_d, directly from the definitions of g_{T+1} and g_s, cf. (3.3). As such, we may invoke Lemma 4 with $g = g_s$, $\tilde{g} = g_{T+1}$, $\hat{x} = x_s$ and $\hat{u} = \hat{u}_s$; this gives a relative neighborhood $\mathcal{X}_s \subseteq \mathrm{ri}(\Delta_d)$ in which $S_{1/2}[g_{T+1}] = S_{1/2}[g_s]$, that is $f_{T+1} = f_s$ (cf. (3.2)–(3.5)). □

Proposition 10 (Transcript consistency)**.** *The queries $x_1, \dots, x_T$, the oracle answers returned in (3.4), and the reported point x_{T+1} coincide with the transcript of $\mathscr{M}$ run on the objective f_{T+1}.*

Proof. Fix any $s \in [T]$. By Lemma 9 with $t = T+1$, we get $f_{T+1} = f_s$ in a relative neighborhood $\mathcal{X}_s$ of x_s; in particular $f_{T+1}(x_s) = f_s(x_s)$. Moreover, since the tangent gradient of $f \in \mathsf{Ent}_d(1)$ at x_s is determined by the values of f in a Θ_d-neighborhood of x_s (cf. (2.1)), we have $\nabla f_{T+1}(x_s) = \nabla f_s(x_s)$ and

$$\mathscr{O}_{f_{T+1}}(x_s) = \mathscr{O}_{f_s}(x_s) \qquad \forall s \in [T]. \tag{3.8}$$

We argue by induction that, when run on f_{T+1}, $\mathscr{M}$ queries exactly $x_1, \dots, x_T$ and reports x_{T+1}. For the base case, the first query $x_1 = \Phi_1$ is fixed by $\mathscr{M}$ independently of the objective, cf. (2.10), so it coincides with the constructed one. Assume, for some $2 \leqslant t \leqslant T$, that the first $t-1$ queries $x_1, \dots, x_{t-1}$ coincide with those in the transcript of $\mathscr{M}$ run on f_{T+1}, and the corresponding constructed answers $\mathscr{O}_{f_1}(x_1), \dots, \mathscr{O}_{f_{t-1}}(x_{t-1})$ coincide with those in the transcript, $\mathscr{O}_{f_{T+1}}(x_1), \dots, \mathscr{O}_{f_{T+1}}(x_{t-1})$. By (2.10), the next query of $\mathscr{M}$ given the constructed data is $x_t = \Phi_t(\mathscr{O}_{f_1}(x_1), \dots, \mathscr{O}_{f_{t-1}}(x_{t-1}))$, and the next query in the transcript is $\Phi_t(\mathscr{O}_{f_{T+1}}(x_1), \dots, \mathscr{O}_{f_{T+1}}(x_{t-1}))$, identical by the induction premise. Now due to (3.8), the next constructed answer $\mathscr{O}_{f_t}(x_t)$ is identical to the corresponding answer $\mathscr{O}_{f_{T+1}}(x_t)$ in the transcript. This advances the induction. Finally, the same argument applied for $t = T+1$, with the report map $\bar{\Phi}_{T+1}$ in place of Φ_t, gives consistency of the reported point. □

3.3 Bounding the suboptimality gap

In this section, we complete the proof of Theorem 5 by estimating the suboptimality gap of f_{T+1} at the reported point x_{T+1}. To this end, we first bound $f_{T+1}(x_{T+1})$ from below, by using Lemmas 6–7 and Proposition 8; then we exhibit a comparator whose support includes an unexplored direction.

1°: *Lower bound at the reported point.* Recall that $\langle x_{T+1}, e_{j_{T+1}} \rangle \leqslant \varepsilon_T$ by Proposition 8, whence

$$\begin{aligned} g_{T+1}(x_{T+1}) \geqslant a_{T+1}(x_{T+1}) &= -\langle x_{T+1}, e_{j_{T+1}} \rangle - T\delta_T \\ &\geqslant -\varepsilon_T - T\delta_T. \end{aligned} \tag{3.9}$$

Let us bound each g_{T+1}-active coordinate (of $\hat{u}_{T+1}$) at the proximal point $\hat{u}_{T+1}$ of x_{T+1}, i.e. $\langle \hat{u}_{T+1}, e_{j_t} \rangle$ for $t \in [T+1]$ such that $a_t(\hat{u}_{T+1}) = g_{T+1}(\hat{u}_{T+1})$. For such t, the final atom is dominated at $\hat{u}_{T+1}$: one has $a_t(\hat{u}_{T+1}) \geqslant a_{T+1}(\hat{u}_{T+1})$, that is, $-\langle \hat{u}_{T+1}, e_{j_t} \rangle - (t-1)\delta_T \geqslant -\langle \hat{u}_{T+1}, e_{j_{T+1}} \rangle - T\delta_T$. Rearranging,

$$\langle \hat{u}_{T+1}, e_{j_t} \rangle \leqslant \langle \hat{u}_{T+1}, e_{j_{T+1}} \rangle + (T+1-t)\delta_T \leqslant \langle \hat{u}_{T+1}, e_{j_{T+1}} \rangle + T\delta_T.$$

Lemma 6, applied to $g = g_{T+1}$ and $x = x_{T+1}$, gives $\langle \hat{u}_{T+1}, e_{j_{T+1}} \rangle \leqslant 2\langle x_{T+1}, e_{j_{T+1}} \rangle \leqslant 2\varepsilon_T$, whence

$$\langle \hat{u}_{T+1}, e_{j_t} \rangle \leqslant 2\varepsilon_T + T\delta_T$$

for every g_{T+1}-active coordinate j_t at $\hat{u}_{T+1}$. As such, the premise of Lemma 7 holds for $g = g_{T+1}$ and $x = x_{T+1}$, with $M = 2\varepsilon_T + T\delta_T$. Combining that lemma (cf. (2.18)) with (3.9) results in

$$S_{1/2}[g_{T+1}](x_{T+1}) \overset{(2.18)}{\geqslant} g_{T+1}(x_{T+1}) - \frac{M}{2} \overset{(3.9)}{\geqslant} -2\varepsilon_T - \frac{3T}{2}\delta_T. \tag{3.10}$$

2°: *Upper bound at the comparator.* Consider $\bar{x} = \frac{1}{T+1}\sum_{t\in[T+1]} e_{j_t}$, i.e. uniform on the selected coordinates; note that this includes the coordinate j_{T+1} revealed *after* $\mathscr{M}$'s reported point x_{T+1}. Clearly, $\bar{x} \in \Delta_d$. At $\bar{x}$, each atom evaluates as $a_t(\bar{x}) = -\langle \bar{x}, e_{j_t} \rangle - (t-1)\delta_T = -\frac{1}{T+1} - (t-1)\delta_T$. Since $(t-1)\delta_T \geqslant 0$, we get

$$g_{T+1}(\bar{x}) = \max_{t\in[T+1]} a_t(\bar{x}) \leqslant -\frac{1}{T+1}.$$

Since $\bar{x}$ is feasible in (2.2), this bound extends to the envelope: $S_{1/2}[g_{T+1}](\bar{x}) \leqslant g_{T+1}(\bar{x}) = -\frac{1}{T+1}$. Combining this with (3.10), we get

$$S_{1/2}[g_{T+1}](x_{T+1}) - \min_{x\in\Delta_d} S_{1/2}[g_{T+1}](x) \geqslant \frac{1}{T+1} - 2\varepsilon_T - \frac{3T}{2}\delta_T.$$

Plugging in the parameter values $\varepsilon_T := \frac{1}{8(T+1)^2}$ and $\delta_T = \frac{1}{3(T+1)^2}$ from (3.1), and using (3.5), we get

$$\begin{aligned} f(x_{T+1}) - \min_{x\in\Delta_d} f(x) \geqslant \frac{1}{2}\left(\frac{1}{T+1} - 2\varepsilon_T - \frac{3T}{2}\delta_T\right) &= \frac{1}{2(T+1)}\left(1 - \frac{1}{4(T+1)} - \frac{T}{2(T+1)}\right) \\ &= \frac{2T+3}{8(T+1)^2} > \frac{1}{4(T+1)}. \end{aligned}$$

This completes the proof in the case of $L = 1$. Recall that the general case follows by homogeneity. □

4 Quantum extension

Let $\mathbf{H}^d$ be the vector space of Hermitian $d \times d$ matrices with inner product $\langle U, V\rangle := \mathrm{tr}(UV)$, and let $\mathbf{H}^d_+$ be the positive-semidefinite cone in $\mathbf{H}^d$. Define the spectrahedron of density matrices

$$\boldsymbol{\Delta}_d := \{X \in \mathbf{H}^d_+ : \mathrm{tr}(X) = 1\}. \tag{4.1}$$

The set $\mathrm{ri}(\boldsymbol{\Delta}_d)$ consists of positive-definite unit-trace matrices in $\mathbf{H}^d$. To formulate the result, it is convenient to use the notion of spectral functions. Recall that any symmetric function $\phi : \mathbb{R}^d \to \mathbb{R}$ defines the spectral function $\phi \circ \lambda$ on $\mathbf{H}^d$, where $\lambda(X) = (\lambda_1(X), \ldots, \lambda_d(X))$ is the ordered spectrum of X. It is well-known (e.g. [12, Thm. 1.1]) that the gradient of $\phi \circ \lambda$ at $X = Q\mathrm{Diag}(\lambda(X))Q^\dagger$ reads

$$\nabla[\phi \circ \lambda](X) = Q\mathrm{Diag}(\nabla\phi(\lambda(X)))Q^\dagger$$

where $A^\dagger$ is the conjugate transpose of a matrix A and $\mathrm{Diag}(\cdot)$ maps $x \in \mathbb{R}^d$ to the diagonal $d \times d$ matrix with x on the diagonal; in other words, $\nabla[\phi \circ \lambda](X)$ is a Hermitian matrix with the same eigenbasis as X, whose spectrum is given by the gradient of ϕ evaluated at the spectrum of X. In particular, negative entropy (1.2) gives the negative von Neumann entropy $h(\lambda(X)) = \mathrm{tr}(X \log X)$. It is also well-known (e.g. [1]) that strict convexity of $\phi \circ \lambda$ is equivalent to that of ϕ; this allows to extend Bregman divergences to $\mathbf{H}^d$ via $D_{\phi\circ\lambda}(X, U) := \phi(\lambda(X)) - \phi(\lambda(U)) - \langle\nabla[\phi \circ \lambda](U), X - U\rangle$. In particular, *(Umegaki's) quantum relative entropy* between $X \in \boldsymbol{\Delta}_d$ and $U \in \mathrm{ri}(\boldsymbol{\Delta}_d)$, as given by

$$D_{h\circ\lambda}(X, U) := \mathrm{tr}(X(\log X - \log U)), \tag{4.2}$$

extends KL divergence in the above sense [11] and can be further extended to $\boldsymbol{\Delta}_d \times \boldsymbol{\Delta}_d$ as in (1.4). As such, the appropriate generalization of $\mathsf{Ent}_d(L)$ is the class $\mathsf{Ent}^{\mathsf{H}}_d(L)$ of convex functions $F : \boldsymbol{\Delta}_d \to \mathbb{R}$ continuously differentiable on $\mathrm{ri}(\boldsymbol{\Delta}_d)$ and satisfying the following inequalities:

$$0 \leqslant F(X) - F(U) - \langle\nabla F(U), X - U\rangle \leqslant LD_{h\circ\lambda}(X, U) \qquad \forall (X, U) \in \boldsymbol{\Delta}_d \times \mathrm{ri}(\boldsymbol{\Delta}_d). \tag{4.3}$$

Here, as in (1.5), we let ∇F be the tangent gradient of F on $\mathrm{ri}(\boldsymbol{\Delta}_d)$, so that $\nabla F(U) \in \mathrm{span}\{I_d\}^\perp$.

We can now state a noncommutative generalization of Theorem 1.

Theorem 11. *Let d, L, T be as in the premise of Theorem 5. For every deterministic method that makes T queries $X_1, \ldots, X_T \in \mathrm{ri}(\boldsymbol{\Delta}_d)$ of the oracle $X \mapsto (F(X), \nabla F(X))$ and returns $X_{T+1} \in \boldsymbol{\Delta}_d$, there exists $F(\cdot) \in \mathsf{Ent}^{\mathsf{H}}_d(L)$ such that*

$$F(X_{T+1}) - \min_{X\in\boldsymbol{\Delta}_d} F(X) > \frac{L}{4(T+1)}. \tag{4.4}$$

Proof sketch. **1°**. First of all, we observe that the function $F : \mathbf{\Delta}_d \to \mathbb{R}$ given by $F(X) = f(\mathrm{diag}(X))$, where f is the hard instance $f \in \mathsf{Ent}(L)$ from Theorem 5 and $\mathrm{diag}(X) = (X_{11}, \dots, X_{dd})$, belongs to the class $\mathsf{Ent}_d^{\mathsf{H}}(L)$. Indeed, convexity of F follows from that of f, since $\mathrm{diag}(\cdot)$ is a linear mapping. On the other hand, defining the diagonal truncation operator $\Pi(H) := \mathrm{Diag}(\mathrm{diag}(H))$, we have

$$D_h(\mathrm{diag}(X), \mathrm{diag}(U)) = D_{h\circ\lambda}(\Pi(X), \Pi(U)) \leqslant D_{h\circ\lambda}(X, U) \qquad \forall (X, U) \in \mathbf{\Delta}_d \times \mathrm{ri}(\mathbf{\Delta}_d). \tag{4.5}$$

Here, the identity holds by (4.2) and (1.4); the final estimate is by Lindblad's data processing inequality [13, 22]. Meanwhile, by the chain rule $\nabla F(U)$ is a diagonal matrix with $\nabla f(\mathrm{diag}(U))$ on the main diagonal; as a result, we have the following in terms of $x = \mathrm{diag}(X)$ and $u = \mathrm{diag}(U)$,

$$F(X) - F(U) - \langle \nabla F(U), X - U\rangle = f(x) - f(u) - \langle \nabla f(u), x - u\rangle \leqslant L D_h(x, u) \leqslant L D_{h\circ\lambda}(X, U),$$

as required in (4.3). As such, $F(\cdot) = f(\mathrm{diag}(\cdot))$ indeed belongs to $\mathsf{Ent}_d^{\mathsf{H}}(L)$ whenever $f \in \mathsf{Ent}_d(L)$.

2°. To complete the proof, we note that a method $\mathscr{M}$ with queries $X_t \mapsto (F(X_t), \nabla F(X_t))$, when run on $F(\cdot) = f(\mathrm{diag}(\cdot))$, receives answers $(f(x_t), \mathrm{Diag}(\nabla f(x_t)))$ where $x_t = \mathrm{diag}(X_t)$. Since the matrix $\mathrm{Diag}(\nabla f(x_t))$ contains the same information as $\nabla f(x_t)$, the sequence $(x_t, f(x_t), \nabla f(x_t))_{t\in[T]}$ is the transcript of *some* $\mathscr{M} \in \mathsf{FOM}_d(T)$ run on $f \in \mathsf{Ent}_d(L)$ and returning x_{T+1}. It remains to pick f as in Theorem 5 and combine (1.10) with the fact that $\min_{X\in\mathbf{\Delta}_d} f(\mathrm{diag}(X)) = \min_{x\in\Delta_d} f(x)$. □

In Appendix B we give an explicit minimax formulation of Theorem 11 analogous to Theorem 5, together with the full presentation of the emulation argument in step **2°** of the above proof sketch.

5 Pointwise minimal interpolant

A promising approach towards improving Theorem 1 would adapt the primal-dual estimate function (PDEF) framework of Florea and Nesterov [8] to the class $\mathsf{Ent}_d(L)$. This framework is built around the *optimal interpolating lower model*, defined as the pointwise-minimal function in the class of convex functions with L-Lipschitz gradient (w.r.t. $\|\cdot\|_2$), that interpolates the given data $\{(x_s, f_s, \xi_s)\}_{s\in[t]}$. The idea is that such a model can be updated incrementally: the next primal point x_{t+1} is selected by minimizing the current model $\hat{f}_t$; a resisting oracle then selects an answer (f_{t+1}, ξ_{t+1}) compatible with $\hat{f}_t$, and the model is updated to incorporate the new point. Here, we make the first step towards implementing this program, by deriving an explicit form of the optimal interpolating lower model for the class $\mathsf{Ent}_d(L)$, as a consequence of the results of Dragomir [5, Chap. 5]. We note that the remaining step would be to construct a resisting oracle that gives the tightest final lower bound.

Given the data $\mathcal{I} = \{(x_i, f_i, \xi_i)\}_{i\in[N]}$ with $(x_i, f_i, \xi_i) \in \mathrm{ri}(\Delta_d) \times \mathbb{R} \times \Theta_d$, we define the quantities

$$r_i := f_i - \langle \xi_i, x_i\rangle, \qquad w_i := \exp(L^{-1}(r_i \mathbf{1}_d + \xi_i)), \qquad W_{\mathcal{I}} := \mathrm{conv}\{w_1, \dots, w_N\}. \tag{5.1}$$

(In the sequel, $\exp(\cdot)$, $\log(\cdot)$ and division are entrywise.) The next lemma, derived from [5, Thm. 5.11] and proved in Appendix C, provides explicit interpolation conditions for the class $\mathsf{Ent}_d(L)$.

Lemma 12. *The existence of $f \in \mathsf{Ent}_d(L)$ that satisfies $f(x_i) = f_i$ and $\nabla f(x_i) = \xi_i$ for all $i \in [N]$ is equivalent to the following inequalities in terms of (5.1):*

$$\langle x_i, w_j / w_i\rangle \leqslant 1 \qquad \forall i, j \in [N]. \tag{5.2}$$

Using these interpolation conditions, we now derive the pointwise minimal interpolant in $\mathsf{Ent}_d(L)$.

Proposition 13. *Assume $\mathcal{I} = \{(x_i, f_i, \xi_i)\}_{i\in[N]}$ satisfies (5.2). The function $f_{\mathcal{I}} : \Delta_d \to \mathbb{R}$ defined by*

$$f_{\mathcal{I}}(x) := L \max_{w\in W_{\mathcal{I}}} \langle x, \log w\rangle$$

belongs to $\mathsf{Ent}_d(L)$, interpolates $\mathcal{I}$, and is pointwise minimal among all functions with these properties.

Proof. **1°**. Let us verify that $f_{\mathcal{I}} \in \mathsf{Ent}_d(L)$. The function $f_{\mathcal{I}}$ is convex, and one has

$$Lh(x) - f_{\mathcal{I}}(x) = L \min_{w\in W_{\mathcal{I}}} \sum_{k\in[d]} (x)_k \log \frac{(x)_k}{(w)_k}.$$

The summand is jointly convex in $((x)_k, (w)_k)$ as the perspective of a convex function $-\log(\cdot)$, so $Lh - f_{\mathcal{I}}$ is convex. For $x \in \mathrm{ri}(\Delta_d)$, strict concavity of $w \mapsto \langle x, \log w\rangle$ on $W_{\mathcal{I}}$ gives a unique maximizer $w(x)$ that continuously depends on x. Whence by Danskin's theorem, $f_{\mathcal{I}} \in C^1(\mathrm{ri}(\Delta_d))$ with $\nabla f_{\mathcal{I}}(x) = L\Pi_{\Theta_d}(\log w(x))$, and therefore $f_{\mathcal{I}} \in \mathsf{Ent}_d(L)$.

2°. We show that $f_{\mathcal{I}}$ interpolates $\mathcal{I}$. For all $w \in W_{\mathcal{I}}$, we have $\langle x_i, w/w_i\rangle \leqslant 1$ by (5.2), whence by concavity of $\log(\cdot)$,

$$\langle x_i, \log(w/w_i)\rangle \leqslant \log\langle x_i, w/w_i\rangle \leqslant 0.$$

Thus $w(x_i) = w_i$, therefore $f_{\mathcal{I}}(x_i) = L\langle x_i, \log w_i\rangle = f_i$ and $\nabla f_{\mathcal{I}}(x_i) = L\Pi_{\Theta_d}(\log w_i) = \xi_i$, cf. (5.1).

3°. Finally, consider arbitrary $F \in \mathsf{Ent}_d(L)$ that interpolates $\mathcal{I}$, i.e. $F(x_i) = f_i$ and $\nabla F(x_i) = \xi_i$ for all $i \in [N]$. Fix $x \in \mathrm{ri}(\Delta_d)$, and set $\xi := \nabla F(x)$ and $r := F(x) - \langle \xi, x\rangle$. Applying Lemma 12 to the interpolation data $\mathcal{I} \cup \{(x, F(x), \xi)\}$ we get

$$\exp(L^{-1}r_i)\langle x, \exp(L^{-1}(\xi_i - \xi))\rangle \leqslant \exp(L^{-1}r) \qquad \forall i \in [N] \tag{5.3}$$

where $r = F - \langle \xi, x\rangle$. As a result, for $w = \sum_{i\in[N]} \lambda_i w_i \in W_{\mathcal{I}}$ we get, by using the concavity of $\log(\cdot)$,

$$L\langle x, \log w\rangle \leqslant \langle \xi, x\rangle + L \log\left(\sum_{i\in[N]} \lambda_i \exp(L^{-1}r_i)\langle x, \exp(L^{-1}(\xi_i - \xi))\rangle\right) \overset{(5.3)}{\leqslant} \langle \xi, x\rangle + r = F(x).$$

Maximizing over w we get $f_{\mathcal{I}}(x) \leqslant F(x)$ on $\mathrm{ri}(\Delta_d)$. For $\bar{x}$ on the relative boundary of Δ_d, consider the segment $x(t) = (1-t)\bar{x} + ty$ with $y \in \mathrm{ri}(\Delta_d)$. Then by continuity of $f_{\mathcal{I}}$ and convexity of F we get $f_{\mathcal{I}}(\bar{x}) = \lim_{t\downarrow 0} f_{\mathcal{I}}(x(t)) \leqslant \limsup_{t\downarrow 0} F(x(t)) \leqslant \limsup_{t\downarrow 0} (1-t)F(\bar{x}) + tF(y) = F(\bar{x})$. □

Acknowledgments

J. M. Aguirre is supported by the NSF Graduate Research Fellowship under Grant No. DGE-2039655. D. M. Ostrovskii thanks Radu-Alexandru Dragomir, Adrien Taylor, and Alexandre d'Aspremont for interesting discussions pertaining to this problem.

A Objective extension to $\mathbb{R}^d_+$ with full-gradient oracle access

Tangent gradients and 1-homogeneous extension over $\mathbb{R}^d_+$. Recall that any function f defined on Δ_d and differentiable in $\mathrm{ri}(\Delta_d)$—including the hard instances presented in Section 2.3—admits the 1-homogeneous extension on $\mathbb{R}^d_+$ (see, e.g., [18]), defined as follows: setting $s(x) := \langle \mathbf{1}_d, x\rangle$,

$$f^+(x) = s(x)f(s(x)^{-1}x). \tag{A.1}$$

This function is differentiable in $\mathbb{R}^d_{++}$. An explicit calculation gives its full gradient for $x \in \mathrm{ri}(\Delta_d)$:

$$\nabla_\star f^+(x) = \nabla f(x) + (f(x) - \langle \nabla f(x), x\rangle)1_d. \tag{A.2}$$

Thus, the oracle $\mathscr{O}_f(x) = (f(x), \nabla f(x))$ is at least as strong as the full gradient oracle $x \mapsto \nabla_\star f^+(x)$ in the extended class $\{f^+ : f \in \mathsf{Ent}_d(L)\}$, in the sense that one can emulate the latter by using the former's answer at the same query point. Intuitively, this should imply that access to the full gradient of f^+ *cannot* lead to a faster convergence rate than access to the oracle $\mathscr{O}_f$, as $\mathscr{O}_f$ gives at least as much information as $\nabla_\star f^+$. The result we shall present next formalizes this comparison and leverages Theorem 5 to obtain a lower bound for the class of functions on the positive orthant, convex and smooth relative to the unnormalized negative entropy $h_{\mathrm{e}}(x) := \sum_{k\in[d]}(x)_k \log(x)_k - (x)_k$ on $\mathbb{R}^d_+$.

Define $\mathsf{Ent}^+_d(L)$ as the class of functions $\tilde{f} : \mathbb{R}^d_+ \to \mathbb{R}$ that are convex, $C^1(\mathbb{R}^d_{++})$, and such that

$$\tilde{f}(x) - \tilde{f}(u) - \langle \nabla_\star \tilde{f}(u), x - u\rangle \leqslant L D_{h_{\mathrm{e}}}(x, u) \quad \forall (x, u) \in \mathbb{R}^d_+ \times \mathbb{R}^d_{++}. \tag{A.3}$$

Since $\langle \nabla_\star \tilde{f}(x), v\rangle = \langle \nabla \tilde{f}(x), v\rangle$ for all tangent directions $v \in \Theta_d$, and $h_e(x) = h(x) - 1$ for all $x \in \Delta_d$, the restriction of $\tilde{f} \in \mathsf{Ent}^+_d(L)$ to Δ_d belongs to $\mathsf{Ent}_d(L)$. The next lemma gives a partial converse.

Lemma 14. *The 1-homogeneous extension $f^+(x)$ of any $f \in \mathsf{Ent}_d(L)$, cf. (A.1), belongs to $\mathsf{Ent}^+_d(L)$.*

Proof. f^+ is convex as the composition of the perspective transformation of f and a linear mapping. Clearly, $f^+ \in C^1(\mathbb{R}^d_{++})$. Finally, by writing

$$Lh_{\mathrm{e}}(x) - f^+(x) = s(x)(Lh(y) - f(y))\big|_{y=s(x)^{-1}x} + Ls(x)\log s(x) - Ls(x)$$

we see that $Lh_{\mathrm{e}} - f^+$ is convex, which is equivalent to (A.3). Indeed, the first term is convex as the composition of the perspective transformation of $Lh - f$ (which is convex) and a linear mapping. □

Similarly, we can extend the class of first-order methods, by considering collections of mappings

$$\begin{aligned} &\Psi_t : \mathbb{R}^{d(t-1)} \to \mathrm{ri}(\Delta_d), \qquad t \in [T]; \\ &\bar{\Psi}_{T+1} : \mathbb{R}^{dT} \to \Delta_d, \end{aligned} \tag{A.4}$$

where $\mathbb{R}^d$ replaces the oracle answer space $\Omega = \mathbb{R} \times \Theta_d$ in (2.9); this corresponds to minimizing $\tilde{f}$ over Δ_d. Let $\mathsf{FOM}^+_d(T)$ be the class of all methods defined by (A.4) and define the minimax risk

$$\mathsf{Risk}^+_d(T, L) := \inf_{\tilde{\mathscr{M}} \in \mathsf{FOM}^+_d(T)} \sup_{\tilde{f} \in \mathsf{Ent}^+_d(L)} \left\{ \tilde{f}(x^{\tilde{\mathscr{M}}}_{T+1}(\tilde{f})) - \min_{x\in\Delta_d} \tilde{f}(x) \right\}, \tag{A.5}$$

where the iterate sequence is generated by running $\tilde{\mathscr{M}}$ with the full gradient oracle $x \mapsto \nabla \tilde{f}(x)$, i.e.

$$x^{\tilde{\mathscr{M}}}_1 = \Psi_1, \;\; x^{\tilde{\mathscr{M}}}_2(\tilde{f}) = \Psi_2(\nabla \tilde{f}(x^{\tilde{\mathscr{M}}}_1)), \;\; \ldots, \;\; x^{\tilde{\mathscr{M}}}_{T+1}(\tilde{f}) = \bar{\Psi}_{T+1}(\nabla \tilde{f}(x^{\tilde{\mathscr{M}}}_1), \ldots, \nabla \tilde{f}(x^{\tilde{\mathscr{M}}}_T(\tilde{f}))). \tag{A.6}$$

Proposition 15. *The minimax risks defined in (A.5) and (2.11) satisfy $\mathsf{Risk}^+_d(T, L) \geqslant \mathsf{Risk}_d(T, L)$.*

Proof. Replacing $\mathsf{Ent}^+_d(L)$ with the smaller (due to Lemma 14) class $\{f^+ : f \in \mathsf{Ent}_d(L)\}$, we get

$$\begin{aligned} \mathsf{Risk}^+_d(T, L) &\geqslant \inf_{\tilde{\mathscr{M}} \in \mathsf{FOM}^+_d(T)} \sup_{f \in \mathsf{Ent}_d(L)} \left\{ f^+(x^{\tilde{\mathscr{M}}}_{T+1}(f^+)) - \min_{x\in\Delta_d} f^+(x) \right\} \\ &= \inf_{\tilde{\mathscr{M}} \in \mathsf{FOM}^+_d(T)} \sup_{f \in \mathsf{Ent}_d(L)} \left\{ f(x^{\tilde{\mathscr{M}}}_{T+1}(f^+)) - \min_{x\in\Delta_d} f(x) \right\} \geqslant \mathsf{Risk}_d(T, L). \end{aligned}$$

Here the identity holds since the mappings in (A.4) output points in Δ_d (where $f^+ = f$). For the last inequality, we use (A.2) and observe that the execution on f^+ of $\tilde{\mathscr{M}} \in \mathsf{FOM}_d^+(T)$, as per (A.4), produces the same query sequence as the execution on f of the method $\mathscr{M} \in \mathsf{FOM}_d(T)$, defined by

$$\begin{aligned}
\Phi_1 &:= \Psi_1, \\
\Phi_2(v_1, \xi_1) &:= \Psi_2\left(\xi_1 + (v_1 - \langle \xi_1, \Phi_1 \rangle)1_d\right), \\
\Phi_3(v_1, \xi_1, v_2, \xi_2) &:= \Psi_3\left(\xi_1 + (v_1 - \langle \xi_1, \Phi_1 \rangle)1_d,\ \ \xi_2 + (v_2 - \langle \xi_2, \Phi_2(v_1, \xi_1) \rangle)1_d\right), \\
&\vdots \\
\bar{\Phi}_{T+1}(v_1, \xi_1, \ldots, v_T, \xi_T) & \\
:= \bar{\Psi}_{T+1}(\xi_1 + (v_1 - &\langle \xi_1, \Phi_1 \rangle)1_d,\ \ldots,\ \xi_T + (v_T - \langle \xi_T, \Phi_T(v_1, \xi_1, \ldots, v_{T-1}, \xi_{T-1}) \rangle)1_d).
\end{aligned}$$

This can be verified by induction over t, in the same way as in the proof of Proposition 10. This fact implies the desired inequality, since in $\mathsf{Risk}_d(T, L)$ the infimum is over *all* methods in $\mathsf{FOM}_d(T)$. □

B Explicit formulation and emulation argument for Theorem 11

We first adapt the definition of first-order methods to make them in line with the class $\mathsf{Ent}_d^{\mathsf{H}}(L)$, cf. (4.3). Namely, we define the methods in $\mathsf{FOM}_d^{\mathsf{H}}(T)$ as collections of mappings (cf. (2.9)):

$$\begin{aligned}
\boldsymbol{\Phi_t} &: \boldsymbol{\Omega}^{t-1} \to \mathrm{ri}(\boldsymbol{\Delta}_d), \qquad t \in [T]; \\
\bar{\boldsymbol{\Phi}}_{\boldsymbol{T+1}} &: \boldsymbol{\Omega}^T \to \boldsymbol{\Delta}_d,
\end{aligned} \tag{B.1}$$

where $\boldsymbol{\Omega} = \mathbb{R} \times \mathrm{span}\{I_d\}^\perp$ and $\boldsymbol{\Omega}^0$ is a singleton. A given method in $\mathsf{FOM}_d^{\mathsf{H}}(T)$ sequentially queries

$$\mathscr{O}_F^{\mathsf{H}}(X) := (F(X), \nabla F(X)) \tag{B.2}$$

and uses the responses to form the sequence (cf. (2.10))

$$X_1 = \boldsymbol{\Phi}_1,\ \ X_2(F) = \boldsymbol{\Phi}_2(\mathscr{O}_F^{\mathsf{H}}(X_1)),\ \ \ldots,\ \ X_{T+1}(F) = \bar{\boldsymbol{\Phi}}_{T+1}(\mathscr{O}_F^{\mathsf{H}}(X_1),\ \ldots,\ \mathscr{O}_F^{\mathsf{H}}(X_T(F))). \tag{B.3}$$

In these terms, Theorem 11 claims that $\mathsf{Risk}_d(T, L)$, cf. (2.11), is a lower bound for the quantum risk

$$\mathsf{Risk}_d^{\mathsf{H}}(T, L) := \inf_{\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_T, \bar{\boldsymbol{\Phi}}_{T+1}} \sup_{F \in \mathsf{Ent}_d^{\mathsf{H}}(L)} \left\{ F(X_{T+1}(F)) - \min_{X \in \boldsymbol{\Delta}_d} F(X) \right\}, \tag{B.4}$$

where the infimum is over $\mathsf{FOM}_d^{\mathsf{H}}(T)$, cf. (B.1)–(B.3). By the argument in step ($\mathbf{1}^\circ$) of the proof sketch of Theorem 11, we have $(f \circ \mathrm{diag}) \in \mathsf{Ent}_d^{\mathsf{H}}(L)$, whence

$$\begin{aligned}
\mathsf{Risk}_d^{\mathsf{H}}(T, L) &\geqslant \inf_{\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_T, \bar{\boldsymbol{\Phi}}_{T+1}} \sup_{f \in \mathsf{Ent}_d(L)} \left\{ f(\mathrm{diag}(X_{T+1}(f \circ \mathrm{diag}))) - \min_{X \in \boldsymbol{\Delta}_d} f(\mathrm{diag}(X)) \right\} \\
&= \inf_{\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_T, \bar{\boldsymbol{\Phi}}_{T+1}} \sup_{f \in \mathsf{Ent}_d(L)} \left\{ f(\mathrm{diag}(X_{T+1}(f \circ \mathrm{diag}))) - \min_{x \in \Delta_d} f(x) \right\}.
\end{aligned} \tag{B.5}$$

Combining $\nabla[f \circ \mathrm{diag}](X) = \mathrm{Diag}(\nabla f(\mathrm{diag}(X)))$ with (B.2)–(B.3), for $X_t = X_t(f \circ \mathrm{diag})$ we get

$$X_t = \boldsymbol{\Phi}_t(f(\mathrm{diag}(X_1)), \mathrm{Diag}(\nabla f(\mathrm{diag}(X_1))), \ldots, f(\mathrm{diag}(X_{t-1})), \mathrm{Diag}(\nabla f(\mathrm{diag}(X_{t-1}))))$$

when $2 \leqslant t \leqslant T$, and

$$X_{T+1} = \bar{\boldsymbol{\Phi}}_{T+1}(f(\mathrm{diag}(X_1)), \mathrm{Diag}(\nabla f(\mathrm{diag}(X_1))), \ldots, f(\mathrm{diag}(X_T)), \mathrm{Diag}(\nabla f(\mathrm{diag}(X_T)))).$$

Since in both cases X_t enters the right-hand side only through the diagonal $x_t := \mathrm{diag}(X_t)$, we get

$$\begin{aligned} x_t(f) &= \Phi_t(f(x_1), \nabla f(x_1), \dots, f(x_{t-1}(f)), \nabla f(x_{t-1}(f))), \quad 2 \leqslant t \leqslant T; \\ x_{T+1}(f) &= \bar{\Phi}_{T+1}(f(x_1), \nabla f(x_1), \dots, f(x_T(f)), \nabla f(x_T(f))), \end{aligned}$$

with $\Phi_t : \Omega^{t-1} \to \mathrm{ri}(\Delta_d)$ defined by $\Phi_t(v_1, g_1, \dots, v_{t-1}, g_{t-1}) = \boldsymbol{\Phi}_t(v_1, \mathrm{Diag}(g_1), \dots, v_{t-1}, \mathrm{Diag}(g_{t-1}))$; ditto for $\bar{\Phi}_{T+1}$. These maps define a method in $\mathsf{FOM}_d(T)$ with transcript $(x_t, f(x_t), \nabla f(x_t))_{t\in[T]}$ consistent with $f \in \mathsf{Ent}_d(L)$; plugging the reported point $\mathrm{diag}(X_{T+1}(f \circ \mathrm{diag})) = x_{T+1}(f)$ in (B.5),

$$\mathsf{Risk}_d^{\mathsf{H}}(T, L) \geqslant \inf_{\Phi_1, \dots, \Phi_T, \bar{\Phi}_{T+1}} \sup_{f \in \mathsf{Ent}_d(L)} \left\{ f(x_{T+1}(f)) - \min_{x \in \Delta_d} f(x) \right\} = \mathsf{Risk}_d(T, L). \qquad \square$$

C Proof of Lemma 12

The gradient and convex conjugate of h_{e} are $\nabla_\star h_{\mathrm{e}}(x) = \log x$ and $h_{\mathrm{e}}^*(\xi) = \langle \mathbf{1}_d, \exp(\xi) \rangle$, respectively.

Lift $\mathcal{I} = \{(x_i, f_i, \xi_i)\}_{i\in[N]}$ to $\tilde{\mathcal{I}} := \{(x_i, f_i, \tilde{\xi}_i)\}_{i\in[N]}$, where $\tilde{\xi}_i := \xi_i + r_i \mathbf{1}_d$; cf. (5.1). By (A.2), we have $(f(x_i), \nabla f(x_i)) = (f_i, \xi_i)$ if and only if $(f^+(x_i), \nabla_\star f^+(x_i)) = (f_i, \tilde{\xi}_i)$. Meanwhile, [5, Thm. 5.11] gives equivalent conditions of the existence of $\tilde{f} \in \mathsf{Ent}_d^+(L)$ interpolating the data $\tilde{\mathcal{I}}$: for all $i, j \in [N]$,

$$A_{ij} := f_i - f_j - \langle \tilde{\xi}_j, x_i - x_j \rangle \geqslant L D_{h_{\mathrm{e}}^*}(\log x_i - L^{-1}(\tilde{\xi}_i - \tilde{\xi}_j), \log x_i). \tag{C.1}$$

Since $x_i, x_j \in \Delta_d$, we have $A_{ij} = f_i - f_j - \langle \xi_j, x_i - x_j \rangle$. Direct expansion of the right-hand side gives

$$\begin{aligned} D_{h_{\mathrm{e}}^*}(\log x_i + L^{-1}(\tilde{\xi}_j - \tilde{\xi}_i), \log x_i) &= \exp(L^{-1}(r_j - r_i)) \langle x_i, \exp(L^{-1}(\xi_j - \xi_i)) \rangle - \langle x_i, L^{-1}(\tilde{\xi}_j - \tilde{\xi}_i) \rangle - 1 \\ &= L^{-1} A_{ij} + \exp(L^{-1}(r_j - r_i)) \langle x_i, \exp(L^{-1}(\xi_j - \xi_i) \rangle - 1. \end{aligned}$$

Thus (C.1) is equivalent to $\exp(L^{-1}(r_j - r_i)) \langle x_i, \exp(L^{-1}(\xi_j - \xi_i)) \rangle \leqslant 1$, that is to (5.2); cf. (5.1).

Let us draw conclusions. If $f \in \mathsf{Ent}_d(L)$ interpolates $\mathcal{I}$, then there exists a function in $\mathsf{Ent}_d^+(L)$, namely the 1-homogeneous extension f^+ of f, that interpolates $\tilde{\mathcal{I}}$, so (5.2) holds. Conversely, (5.2) implies the existence of $\tilde{f} \in \mathsf{Ent}_d^+(L)$ that interpolates $\tilde{\mathcal{I}}$, but then the restriction f of $\tilde{f}$ to Δ_d interpolates $\mathcal{I}$: indeed, $\tilde{f}(x_i) = f(x_i)$ since $x_i \in \Delta_d$, and $\nabla f(x_i) = \Pi_{\Theta_d}(\nabla_\star \tilde{f}(x_i)) = \Pi_{\Theta_d}(\tilde{\xi}_i) = \xi_i$. $\square$

References

[1] M. Baes. Convexity and differentiability properties of spectral functions and spectral mappings on Euclidean Jordan algebras. *Linear Algebra and Its Applications*, 422(2-3):664–700, 2007.

[2] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[3] H. H. Bauschke and J. M. Borwein. Joint and separate convexity of the Bregman distance. In *Studies in Computational Mathematics*, volume 8, pages 23–36. Elsevier, 2001.

[4] H. H. Bauschke, M. N. Dao, and S. B. Lindstrom. Regularizing with Bregman–Moreau envelopes. *SIAM Journal on Optimization*, 28(4):3208–3228, 2018.

[5] R.-A. Dragomir. *Bregman Gradient Methods for Relatively Smooth Optimization*. Ph.D. thesis, Université Toulouse 1 Capitole, Toulouse, France, 2021.

[6] R.-A. Dragomir, A. B. Taylor, A. d'Aspremont, and J. Bolte. Optimal complexity and certification of Bregman first-order methods. *Mathematical Programming*, 194(1–2):41–83, 2022.

[7] Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: A novel approach. *Mathematical Programming*, 145(1–2):451–482, 2014.

[8] M. I. Florea and Yu. Nesterov. An optimal lower bound for smooth convex functions. *Foundations of Computational Mathematics*, pages 1–35, 2025.

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

[10] F. Hanzely, P. Richtárik, and L. Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *Computational Optimization and Applications*, 79:405–440, 2021.

[11] F. Hiai and D. Petz. The proper formula for relative entropy and its asymptotics in quantum probability. *Communications in Mathematical Physics*, 143(1):99–114, 1991.

[12] A. S. Lewis. Derivatives of spectral functions. *Mathematics of Operations Research*, 21(3):576–588, 1996.

[13] G. Lindblad. Completely positive maps and entropy inequalities. *Communications in Mathematical Physics*, 40(2):147–151, 1975.

[14] H. Lu, R. M. Freund, and Yu. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[15] A. S. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, 1983.

[16] Yu. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

[17] Yu. Nesterov and A. S. Nemirovski. On first-order algorithms for ℓ_1/nuclear norm minimization. *Acta Numerica*, 22:509–575, 2013.

[18] E. Y. Ovcharov. Proper scoring rules and Bregman divergence. *Bernoulli*, pages 53–79, 2018.

[19] M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Holden-Day, 1964.

[20] R. T. Rockafellar and R. J. B. Wets. *Variational analysis*. Springer, 1998.

[21] A. B. Taylor, J. M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1–2):307–345, 2017.

[22] A. Uhlmann. Relative entropy and the Wigner–Yanase–Dyson–Lieb concavity in an interpolation theory. *Communications in Mathematical Physics*, 54(1):21–32, 1977.