

Entropy-Smooth Convex Optimization Cannot Be Accelerated

Jacob M. Aguirre

Joint work with Dmitrii M. Ostrovskii

School of ISyE, Georgia Tech

October 27, 2026

TACOS

Acknowledgments: Radu-Alexandru Dragomir, Adrien Taylor, Alexandre dAspremont, and Arkadi Nemirovski

[arXiv:2607.27476](https://arxiv.org/abs/2607.27476)

Goal

- **Question (open since 2017–18):** Can Nesterov's acceleration, $O(1/T) \rightarrow O(1/T^2)$, be achieved under *relative smoothness* with respect to entropy on Δ_d ?
- **Answer: No!** For $d \gtrsim T^2$, every deterministic first-order method suffers worst-case error $> L/(4(T+1))$, whereas Mirror Descent achieves $L \log d/T$.
- **Mechanism:** along FGM's coupling, with weight $\theta_t \asymp 1/t$, KL shrinks like θ_t , not θ_t^2 ; entropy smoothness allows curvature $\asymp 1/x_i$ on small coordinates, where the hard instances live.

Outline

1. **Relative smoothness:** definition, ℓ_1 vs. KL paradox.
2. **Main Result:** oracle model, $\Omega(L/T)$ lower bound.
3. **Proof:** nonsmooth coordinate game + right Bregman–Moreau envelope.
4. **Extensions and open problems:** quantum density matrices, the $\log d$ gap.

Simplex geometry and entropy

- **Problem:** on the probability simplex Δ_d ,

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

- **Conventions:**

- f is convex on Δ_d and C^1 on $\text{ri } \Delta_d := \{x \in \Delta_d : x_i > 0\}$;
- $\nabla f(x) \in \Theta_d := \mathbf{1}^\perp$ is the **tangent gradient**.

- **Reference function** (throughout the talk): negative entropy $h(x) := \sum_{i=1}^d x_i \log x_i$.
- **Bregman divergence** of h is the **KL divergence**:

$$D_h(x, u) = \sum_{i=1}^d x_i \log \frac{x_i}{u_i}.$$

- **Two facts:**

- **Bounded radius:** $\max_{x \in \Delta_d} D_h(x, \frac{1}{d} \mathbf{1}) = \log d$.
- **Three-point identity:**

$$D_h(y, u) - D_h(x, u) = D_h(y, x) + \langle \log x - \log u, y - x \rangle.$$

Relative smoothness

- **Idea:** measure the smoothness of f by a reference function ϕ directly, instead of comparing ϕ with a norm.
- f is **L -smooth relative to ϕ** (Bauschke et al. 2017; Lu et al. 2018) if $L\phi - f$ is convex; equivalently,

$$\nabla^2 f \preceq L \nabla^2 \phi, \quad D_f \leq L D_\phi.$$

- **Examples of ϕ :**
 - $\frac{1}{2} \|\cdot\|_2^2$: the classical smooth class;
 - $\frac{\|x\|_2^{r+2}}{r+2} + \frac{\|x\|_2^2}{2}$ if $\|\nabla^2 f(x)\| \leq L(1 + \|x\|_2^r)$ (Nesterov 2019);
 - $-\sum_i \log x_i$: Poisson inverse problems, D-optimal design;
 - h on Δ_d : this talk.

Definition (Entropy-smooth class Ent_d)

Convex $f : \Delta_d \rightarrow \mathbb{R}$, C^1 on $\text{ri } \Delta_d$, with $0 \leq D_f(x, u) \leq D_h(x, u)$ on $\Delta_d \times \text{ri } \Delta_d$. By homogeneity, $L = 1$ throughout: $\text{Ent}_d(L) := L \cdot \text{Ent}_d$.

Reminder: ℓ_1 -smoothness and acceleration

- **Classical ℓ_1 -smooth class Lip_d :** $0 \leq D_f(x, u) \leq \frac{1}{2} \|x - u\|_1^2$. The **Fast Gradient Method** (FGM, Nesterov 2005) with reference function h , 1-strongly convex in $\|\cdot\|_1$ (**Pinsker**), achieves $O(\log d / T^2)$ on Lip_d .
- **FGM's proof uses homogeneity of the norm** to tie a far-away sequence to two close ones: for $x_\theta := (1 - \theta)a + \theta b$, $u_\theta := (1 - \theta)a + \theta c$ and $\theta = \theta_t \asymp 1/t$,
$$\frac{1}{2} \|x_\theta - u_\theta\|_1^2 = \theta^2 \frac{1}{2} \|b - c\|_1^2.$$
- **No norm, no homogeneity:** $D_h(x_\theta, u_\theta) \stackrel{(*)}{=} \theta D_h(b, c)$ for $a = e_3$ and $\text{supp } b, \text{supp } c \subseteq \{1, 2\}$.
- **Triangle scaling** (Hanzely, Richtárik & Xiao 2021) imposes the effect of homogeneity: $D_h(x_\theta, u_\theta) \leq G \theta^\gamma D_h(b, c)$ yields $O(1/T^\gamma)$, and $\gamma = 2$ is full acceleration. By $(*)$, h admits no $\gamma > 1$: this blocks the *proof* of acceleration, not acceleration itself.

Why Ent_d is much wider: boundary curvature

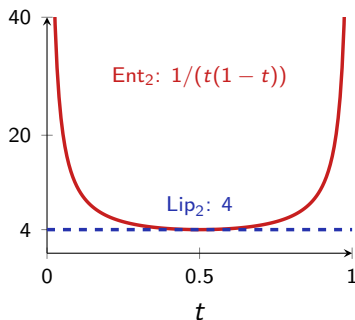
“Curvature explodes at the boundary.”

- By Pinsker: $\text{Lip}_d \subsetneq \text{Ent}_d$.
- Second-order test ($v \in \Theta_d$):

$$0 \leq v^\top \nabla^2 f(x) v \leq \sum_{i=1}^d \frac{v_i^2}{x_i}.$$

- Locally near $\frac{1}{d}\mathbf{1}$, both classes look Euclidean.
- Near $\partial\Delta_d$, Ent_d allows **curvature**
 $\asymp 1/x_i \rightarrow \infty$.

Allowed curvature on
 $x = (t, 1 - t) \in \Delta_2$



$h(x)$ itself lies on the red curve.

Benchmark: Mirror Descent needs no norm

“The standard proof of Mirror Descent uses the norm in one step only.”

- **Entropic Mirror Descent** from $x_1 := \frac{1}{d}\mathbf{1}$:

$$x_{t+1} = \operatorname{argmin}_{x \in \Delta_d} \{ \langle \nabla f(x_t), x \rangle + D_h(x, x_t) \} \iff (x_{t+1})_i \propto (x_t)_i e^{-(\nabla f(x_t))_i}.$$

Theorem (Bauschke et al. 2017; Lu et al. 2018)

For any $f \in \text{Ent}_d$, after T gradient calls, $f(x_{T+1}) - f^* \leq \frac{D_h(x^*, x_1)}{T} \leq \frac{\log d}{T}$.

- **The one step:** the three-point identity implies

$$\langle \nabla f(x_t), x_{t+1} - y \rangle \leq D_h(y, x_t) - D_h(y, x_{t+1}) - D_h(x_{t+1}, x_t),$$

and $D_f(x_{t+1}, x_t) \leq D_h(x_{t+1}, x_t)$ cancels the last term exactly. On Lip_d , this bound came from $\frac{1}{2}\|x_{t+1} - x_t\|_1^2$ and Pinsker.

Main question

Function class	Mirror Descent	Accelerated rate?
Lip_d	$O(\log d / T)$	$O(\log d / T^2)$ (FGM)
Ent_d	$\log d / T$	???

Nesterov (2019): the relative-smoothness rate is “not as good as for the Fast Gradient Methods,” while its field of applications is much wider.

Question (Bauschke et al. 2017; Lu et al. 2018; Dragomir 2021):

Can the $O(\log d / T^2)$ rate be extended from Lip_d to Ent_d ?

- **Dragomir, Taylor, d’Aspremont & Bolte (2022):** $\Omega(1/T)$ for a ϕ built together with f against the method. Neither result implies the other.
- **Entropy is special:** KL is **jointly convex** (Bauschke & Borwein 2001), which could have served an upper bound.
- **This talk:** for the entropy, the answer is **NO**.

Black-box oracle model (Nemirovski & Yudin 1983)

- **Local first-order oracle:** $\mathcal{O}_f(x) := (f(x), \nabla f(x)) \in \Omega := \mathbb{R} \times \Theta_d$ at $x \in \text{ri } \Delta_d$;
local: $f = g$ near $x \Rightarrow \mathcal{O}_f(x) = \mathcal{O}_g(x)$ (Guzmán & Nemirovski 2015).
- **Deterministic T -step algorithm** $\mathcal{M} \in \text{FOM}_d(T)$: arbitrary *search rules*
 $\Phi_t : \Omega^{t-1} \rightarrow \Delta_d$ (**no span assumption**):

$$x_t = \Phi_t(\mathcal{O}_f(x_1), \dots, \mathcal{O}_f(x_{t-1})), \quad t = 1, \dots, T+1;$$

search points $x_1, \dots, x_T$, approximate solution x_{T+1} .

- **Minimax risk:**

$$\text{Risk}_d(T) := \inf_{\mathcal{M} \in \text{FOM}_d(T)} \sup_{f \in \text{Ent}_d} [f(x_{T+1}) - f^*] \leq \frac{\log d}{T} \quad (\text{Mirror Descent}).$$

Main theorem: no acceleration on Ent_d

Theorem (Aguirre & Ostrovskii 2026, Thm. 5)

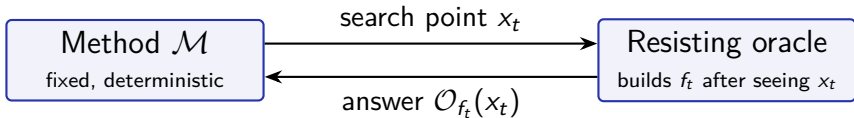
For all $T \geq 1$ and $d \geq 8(T+1)^2 + T$, the minimax T -risk of Ent_d satisfies

$$\text{Risk}_d(T) > \frac{1}{4(T+1)}.$$

- R1.** A $\log d$ gap remains between the bounds; we expect the lower bound to be loose and the missing factor to be recoverable...
- R2.** When $d = \exp(T)$, we have $O(T^{-2} \log d)$, matching FGM on Lip_d and diverging from MD by T .
- R3.** The quadratic dependence of d on T is crucial for our resisting oracle: it gives an $O(T^{-2})$ “safety margin” that ensures consistency of the transcript after T steps.

Resisting oracle

- Fix $\mathcal{M} \in \text{FOM}_d(T)$. Build the hard instance *along the trajectory* of $\mathcal{M}$: x_t is the t -th search point of $\mathcal{M}$ as applied to f_{t-1} ; f_t coincides with f_{t-1} near $x_1, \dots, x_{t-1}$.



- Two requirements on the final function $f \in \text{Ent}_d$:**
 - (C) Consistency:** f and f_t are indistinguishable near x_t (local oracle) $\implies \mathcal{M}$ as applied to f has trajectory $x_1, \dots, x_T$ and approximate solution x_{T+1} .
 - (G) Gap:** $f(x_{T+1}) - f^* > \frac{1}{4(T+1)}$.
- Two layers:**
 - Layer 1:** nonsmooth coordinate game on piecewise-linear functions g_t .
 - Layer 2:** smoothing $g_t \mapsto f_t \in \text{Ent}_d$ that preserves locality.

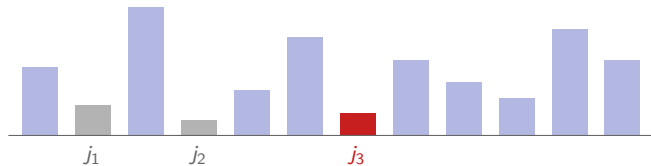
Layer 1: the nonsmooth coordinate game

- **Goal:** piecewise-linear g_t whose later changes are *invisible* near earlier search points, with a gap of order $1/T$ at x_{T+1} .

- Build piecewise-linear **functions** from coordinate **atoms** $u \mapsto -u_j - b$:

$$j_t \in \operatorname{argmin}_{i \notin \{j_1, \dots, j_{t-1}\}} (x_t)_i, \quad a_t(u) := -u_{j_t} - (t-1)\delta, \quad g_t := \max\{a_1, \dots, a_t\}.$$

- **Rule:** at search point x_t , penalize the “least-loaded” unused coordinate j_t (smallest mass $(x_t)_i$), shifted down by $(t-1)\delta$.



search point $x_3 \in \Delta_{12}$; bars are the masses $(x_3)_i$, which sum to 1

grey: used coordinates j_1, j_2

red: least-loaded unused coordinate, $(x_3)_{j_3} \leq \frac{1}{d-2}$ (pigeonhole)

Layer 1: why the rule works

“Later atoms are invisible at old search points, yet the last one costs the approximate solution $1/T$.”

1. **Small mass (pigeonhole):** $\geq d - T$ coordinates are unused $\implies (x_t)_{j_t} \leq \varepsilon := \frac{1}{d-T}$.
2. **(C) Consistency:** at an old search point x_t , $g_t(x_t) \geq -\varepsilon - (t-1)\delta$, while $a_{t+1}, \dots, a_{T+1} \leq -t\delta$; $\delta > \varepsilon$ keeps them strictly below g_t near x_t .
3. **(G) Gap:** j_{T+1} , chosen at x_{T+1} , keeps $g_{T+1}(x_{T+1}) \geq -\varepsilon - T\delta$, while the uniform $\bar{x}$ on $\{j_1, \dots, j_{T+1}\}$ has $g_{T+1}(\bar{x}) = -\frac{1}{T+1}$.

Nonsmooth gap with $\varepsilon = 1/(d - T)$ and $\delta := 2\varepsilon$, which satisfies (C):

$$\begin{aligned} g_{T+1}(x_{T+1}) - \min_{\Delta_d} g_{T+1} &\geq g_{T+1}(x_{T+1}) - g_{T+1}(\bar{x}) \stackrel{(G)}{\geq} \frac{1}{T+1} - \varepsilon - T\delta \\ &= \frac{1}{T+1} - \frac{2T+1}{d-T} \geq \frac{1}{2(T+1)} \quad \text{for } d \geq 4(T+1)^2 + T. \end{aligned}$$

Layer 2: Why standard norm-based smoothing fails

“A norm cannot certify our class.”

- **Guzmán & Nemirovski (2015)**: inf-convolution with a norm-smooth kernel of radius χ .

Property	Guzmán–Nemirovski	What we need on Δ_d
Smoothness	$O(1/\chi)$ -smooth in $\ \cdot\ $	$\text{Ent}_d(O(1))$ (relative to h)
Locality	g within $O(\chi)$ of x	g near a controllable point
Smoothing loss	$O(\chi)$	$\lesssim \varepsilon + T\delta$ at x_{T+1}

- **Obstacle**: if the hard f lies in $\text{Lip}_d(C)$, FGM attains $\frac{4C \log d}{T(T+3)}$, which forces $C \gtrsim T / \log d$.

$\implies$ Certify $\text{Ent}_d(O(1))$ directly: the **right Bregman–Moreau envelope**.

The right Bregman–Moreau envelope

“Smooth g by minimizing over the second argument of KL.”

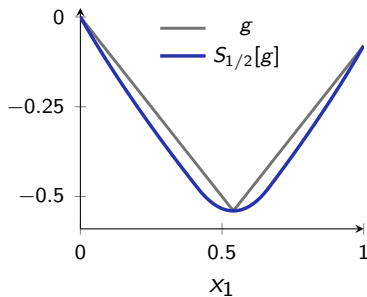
- For $\eta > 0$ (Bauschke, Dao & Lindstrom 2018):

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

- Its unique minimizer is the **prox point**:

$$\text{prox}_{\eta g}(x) := \operatorname{argmin}_{u \in \Delta_d} \{\eta g(u) + D_h(x, u)\}.$$

- $S_\eta[g] \leq g$ (take $u = x$).
- $S_\eta[g]$ is **convex**, since KL is **jointly convex**.



$$d = 2, \eta = \frac{1}{2},$$

$$g(u) = \max\{-u_1, -u_2 - 0.08\}.$$

Smoothness of the right envelope

“Every convex g , however nonsmooth, becomes entropy-smooth.”

Lemma (Envelope regularity; cf. Bauschke, Dao & Lindstrom 2018)

For convex continuous g and $\eta > 0$:

- (a) $\hat{u} := \text{prox}_{\eta g}(x) \in \text{ri } \Delta_d$ is unique and continuous in $x \in \text{ri } \Delta_d$.
- (b) $S_\eta[g]$ is convex, C^1 on $\text{ri } \Delta_d$, and $\nabla S_\eta[g](x) = \eta^{-1} \Pi_{\Theta_d}(\log x - \log \hat{u})$.
- (c) **Relative smoothness:** $\eta S_\eta[g] \in \text{Ent}_d$.

- **One-line proof of (c):** plug the suboptimal $\hat{u} = \text{prox}_{\eta g}(x)$ into $S_\eta[g](y)$; the three-point identity implies

$$\begin{aligned} \eta S_\eta[g](y) - \eta S_\eta[g](x) &\leq D_h(y, \hat{u}) - D_h(x, \hat{u}) \\ &= D_h(y, x) + \underbrace{\langle \log x - \log \hat{u}, y - x \rangle}_{= \eta \langle \nabla S_\eta[g](x), y - x \rangle} \end{aligned}$$

Locality is governed by the prox point

Lemma (Locality transfer)

Let $g \leq \tilde{g}$, $\hat{x} \in \text{ri } \Delta_d$ and $\hat{u} := \text{prox}_{\eta g}(\hat{x})$. If $\tilde{g} = g$ near $\hat{u}$, then $S_\eta[\tilde{g}] = S_\eta[g]$ near $\hat{x}$; in particular, their oracle answers at $\hat{x}$ coincide.

- **Proof:** for x near $\hat{x}$, $u(x) := \text{prox}_{\eta g}(x)$ stays near $\hat{u}$ by continuity; hence $\tilde{g}(u(x)) = g(u(x))$ and
$$S_\eta[g](x) \leq S_\eta[\tilde{g}](x) \leq \tilde{g}(u(x)) + \eta^{-1}D_h(x, u(x)) = S_\eta[g](x).$$
- **In the game:** to keep the answer at a past search point x_t , later atoms must stay below g_t near the **prox point** $\hat{u}_t := \text{prox}_{\eta g_t}(x_t)$.

Why KL smoothing works: multiplicative stability

- A single atom $g(u) = -u_1$, at x with $x_1 \ll 1$:

Euclidean prox: additive shift

$$u = \operatorname{argmin}_{u \in \Delta_d} \left\{ \eta g(u) + \frac{1}{2} \|u - x\|_2^2 \right\}$$

satisfies, in every dimension,

$$u_1 \geq \min \left\{ x_1 + \frac{\eta}{2}, 1 \right\}.$$

A coordinate $x_1 \leq 1/d$ jumps to at least $\eta/2$ unless $\eta \lesssim 1/d$.

Entropic right prox: multiplicative

$$\hat{u} = \operatorname{argmin}_{u \in \Delta_d} \left\{ \eta g(u) + D_h(x, u) \right\}$$

satisfies $\frac{x_1}{\hat{u}_1} = 1 - \eta + \eta \hat{u}_1$, so

$$x_1 \leq \hat{u}_1 \leq \frac{x_1}{1 - \eta}.$$

For $\eta = \frac{1}{2}$, mass $x_1 \leq \varepsilon$ stays $\hat{u}_1 \leq 2\varepsilon$!

Multiplicative control for piecewise-linear g

Lemma (KKT control for $g = \max\{a_1, \dots, a_K\}$)

For $0 < \eta < 1$ and $\hat{u} := \text{prox}_{\eta g}(x)$, there is $\lambda \in \Delta_d$ with $-\lambda \in \partial g(\hat{u})$, supported on the active coordinates at $\hat{u}$, such that, with $\hat{m} := \langle \lambda, \hat{u} \rangle$,

$$\frac{x_i}{\hat{u}_i} = 1 + \eta(\hat{m} - \lambda_i), \quad i = 1, \dots, d. \quad (\text{KKT})$$

Two consequences for $\eta = \frac{1}{2}$:

- Multiplicative stability:** $\hat{m} \geq 0, \lambda_i \leq 1 \implies \frac{x_i}{\hat{u}_i} \geq \frac{1}{2} \implies \hat{u} \leq 2x$ entrywise.
- Local smoothing loss:** if $\hat{u}_j \leq M$ on all active coordinates, then

$$\|x - \hat{u}\|_\infty \leq \frac{1}{2}M \text{ and}$$

$$g(x) - \frac{1}{2}M \leq S_{1/2}[g](x) \leq g(x).$$

The smoothed resisting oracle

“Layer 1 with smoothed answers; the price is $\delta > 2\varepsilon$.”

- **Round $t \leq T$:** choose j_t and g_t as in Layer 1; answer with $f_t := \frac{1}{2}S_{1/2}[g_t] \in \text{Ent}_d$.
- **At the approximate solution x_{T+1} :** choose j_{T+1} the same way; the final function is $f := \frac{1}{2}S_{1/2}[g_{T+1}] \in \text{Ent}_d$.
- **Constraints on the mass level ε and the offset step δ ($\hat{u}_t := \text{prox}_{g_t/2}(x_t)$):**

$$\underbrace{d - T \geq \frac{1}{\varepsilon}}_{\text{pigeonhole: } (x_t)_{j_t \leq \varepsilon}}$$

$$\underbrace{\delta > 2\varepsilon}_{\text{consistency; KKT (1): } (\hat{u}_t)_{j_t \leq 2(x_t)_{j_t \leq 2\varepsilon}}$$

$$\underbrace{2\varepsilon + \frac{3T}{2}\delta < \frac{1}{2(T+1)}}_{\text{gap} > 1/(4(T+1))}$$

Together they force $d - T > 2(T+1)(3T+2)$: **this construction needs $d \gtrsim T^2$.**

- **Choice:** $\varepsilon := \frac{1}{8(T+1)^2}$ and $\delta := \frac{8}{3}\varepsilon = \frac{1}{3(T+1)^2}$; all three hold when $d \geq 8(T+1)^2 + T$.

Verification (C): later atoms are invisible

Lemma (Positive margin at old prox points)

For all $t \leq T$ and every later atom
 $a \in \{a_{t+1}, \dots, a_{T+1}\}$:

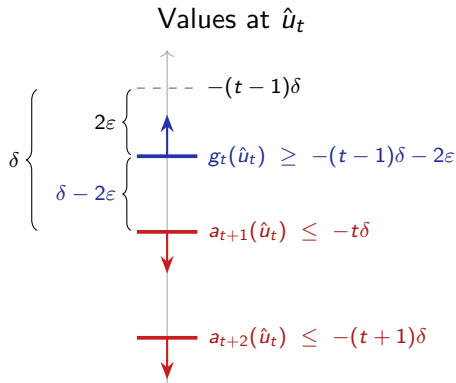
$$g_t(\hat{u}_t) - a(\hat{u}_t) \geq \delta - 2\varepsilon > 0.$$

- **Proof:** By KKT (1), $(\hat{u}_t)_{j_t} \leq 2\varepsilon$, hence

$$g_t(\hat{u}_t) \geq a_t(\hat{u}_t) \geq -2\varepsilon - (t-1)\delta,$$

whereas $a_{t+1}(\hat{u}_t), \dots, a_{T+1}(\hat{u}_t) \leq -t\delta$.

Chain: $g_{T+1} = g_t$ near $\hat{u}_t$ (1-Lipschitz) $\Rightarrow f = f_t$ near x_t (locality transfer) $\Rightarrow \mathcal{M}$ produces the same search points and approximate solution $x_1, \dots, x_{T+1}$ on f (induction on t).



Verification (G): the gap

- Final function $f = \frac{1}{2}S_{1/2}[g_{T+1}]$; by construction, $(x_{T+1})_{j_{T+1}} \leq \varepsilon$.

At x_{T+1} : f is large

- Last atom: $g_{T+1}(x_{T+1}) \geq -\varepsilon - T\delta$.
- Active coordinates at $\hat{u}_{T+1}$ have mass $\leq M := 2\varepsilon + T\delta$, so by KKT (2):

$$S_{1/2}[g_{T+1}](x_{T+1}) \geq -2\varepsilon - \frac{3T}{2}\delta.$$

At comparator $\bar{x}$: f is small

- Let $\bar{x} := \text{uniform on } \{j_1, \dots, j_{T+1}\}$.
- Since $a_t(\bar{x}) \leq -\frac{1}{T+1}$ and $S_{1/2}[g] \leq g$:

$$S_{1/2}[g_{T+1}](\bar{x}) \leq -\frac{1}{T+1}.$$

- Combine** with $\varepsilon = \frac{1}{8(T+1)^2}$ and $\delta = \frac{1}{3(T+1)^2}$:

$$f(x_{T+1}) - f^* \geq \frac{1}{2} \left[\frac{1}{T+1} - 2\varepsilon - \frac{3T}{2}\delta \right] = \frac{2T+3}{8(T+1)^2} > \frac{1}{4(T+1)}.$$



Extension: Quantum density matrices

- **Density matrices** $\mathfrak{D}_d$ with the **von Neumann entropy** $H(X) := \text{tr}(X \log X)$; its Bregman divergence D_H is the quantum relative entropy.
- **Quantum entropy-smooth class** $\text{Ent}_d^{\mathbb{H}}$: convex F with $D_F \leq D_H$.

Theorem (Aguirre & Ostrovskii 2026)

For $d \geq 8(T+1)^2 + T$, every deterministic T -step matrix algorithm on $\text{Ent}_d^{\mathbb{H}}$ suffers worst-case error $> 1/(4(T+1))$.

- **Reduction via diagonal restriction** $F(X) := f(\text{diag } X)$:
 - **Data-processing inequality (Lindblad 1975):**
 $D_h(\text{diag } X, \text{diag } U) \leq D_H(X, U) \implies F \in \text{Ent}_d^{\mathbb{H}}$.
 - **Oracle simulation:** $\nabla F(X) = \text{Diag}(\nabla f(\text{diag } X))$ reveals no off-diagonal information.

Summary and open problems

- **Main result:** On the entropy-smooth class Ent_d ($d \geq 8(T+1)^2 + T$),

$$\frac{1}{4(T+1)} < \text{Risk}_d(T) \leq \frac{\log d}{T}.$$

Mirror Descent is optimal up to $\log d$; no deterministic method attains $\text{polylog}(d)/T^{1+\alpha}$, $\alpha > 0$.

- **Why ℓ_1 accelerates but KL does not:** along FGM's coupling KL shrinks like θ , not θ^2 ; and **curvature explodes at the boundary**, where the right Bregman–Moreau envelope hides the atoms.
- **Open questions:**
 - Close the $\log d$ gap and extend to randomized methods.
 - Smaller dimensions $d \ll T^2$: does $\Omega(1/T)$ persist for $d \asymp T$?
 - **Logarithmic barrier** $\phi(x) = -\sum_i \log x_i$: D_ϕ is *not* jointly convex; it needs a new conjugate smoothing.

References

- J. M. Aguirre, D. M. Ostrovskii. Entropy-smooth convex optimization cannot be accelerated. *arXiv:2607.27476* (2026).
- H. H. Bauschke, J. Bolte, M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Math. Oper. Res.* (2017).
- H. Lu, R. M. Freund, Yu. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim.* (2018).
- R.-A. Dragomir, A. B. Taylor, A. d'Aspremont, J. Bolte. Optimal complexity and certification of Bregman first-order methods. *Math. Program.* (2022).
- R.-A. Dragomir. *Bregman Gradient Methods for Relatively Smooth Optimization*. PhD thesis, Toulouse (2021).
- C. Guzmán, A. S. Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *J. Complexity* (2015).
- A. Juditsky, A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, I: general purpose methods. In *Optimization for Machine Learning*, MIT Press (2011).
- Yu. Nesterov. Relative smoothness: new paradigm in convex optimization. Plenary talk, *EUSIPCO*, A Coruña (2019).
- C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.* (1948).
- R. E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inform. Theory* (1972).
- F. Hanzely, P. Richtárik, L. Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *Comput. Optim. Appl.* (2021).
- H. H. Bauschke, M. N. Dao, S. B. Lindstrom. Regularizing with Bregman–Moreau envelopes. *SIAM J. Optim.* (2018).
- H. H. Bauschke, J. M. Borwein. Joint and separate convexity of the Bregman distance. *Stud. Comput. Math.* 8 (2001).
- Yu. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.* (2005).
- M. I. Florea, Yu. Nesterov. An optimal lower bound for smooth convex functions. *Found. Comput. Math.* (2025).
- A. S. Nemirovski, D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley (1983).
- G. Lindblad. Completely positive maps and entropy inequalities. *Comm. Math. Phys.* (1975).
- S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Inform. Theory* (1972).
- K. He, J. Saunderson, H. Fawzi. A Bregman proximal perspective on classical and quantum Blahut–Arimoto algorithms. *IEEE Trans. Inform. Theory* (2024).

Thank you!

Questions & discussion

[arXiv:2607.27476](https://arxiv.org/abs/2607.27476)