Sharp Risk Bounds for Early-stopping in Gaussian Linear Regression

Tobias Wegel
ETH Zurich

Gil Kur
ETH Zurich

Patrick Rebeschini
University of Oxford

Abstract

We study early-stopped mirror descent (ESMD) for high-dimensional Gaussian linear regression over arbitrary convex bodies and design matrices, where the task is to minimize the in-sample mean squared error. Our main result shows that some of the sharpest risk bounds for the least squares estimator (LSE), based on the local Gaussian width, extend to ESMD. We derive sufficient conditions on the potential, expressed via the Minkowski functional, under which our result holds. These conditions allow us to construct new potentials and analyze existing ones. Our results then yield general sufficient conditions for minimax optimality of ESMD, provide a systematic comparison with the LSE, and establish the tightest known risk bound in the ℓ_1-constrained setting.[1]

1 INTRODUCTION

Regularization methods generally fall into two categories: *Explicit* regularization, where the learning objective is altered to reduce model complexity via constraints or penalty terms, and *implicit* regularization, where the optimization solver inherently controls model complexity via algorithmic primitives and parameter tuning. A commonly used implicit regularization technique is to stop iterative optimization algorithms before convergence. This is called *early-stopping* or *iterative regularization*, and has been investigated for different settings and algorithms such as linear and kernel regression with coordinate descent (Hastie et al., 2001; Efron et al., 2004; Rosset et al., 2004; Zhang and Yu, 2005), gradient descent (Ali et al., 2019; Bühlmann and Yu, 2003; Yao et al., 2007; Raskutti et al., 2011b; Bauer et al., 2007), primal-dual gradient methods (Molinari et al., 2021, 2024) and algorithms based on factorized parameterizations (Gunasekar et al., 2017; Li et al., 2018; Vaškevičius et al., 2019; Zhao et al., 2022). One benefit of iterative regularization is the simultaneous study of modeling and numerical aspects; often iterative regularization improves computational efficiency while retaining good statistical performance (Molinari et al., 2021; Yao et al., 2007).

A recurring approach to understanding iterative regularization methods is to tie them to "corresponding" explicit regularization methods or constraint geometries. Usually, the stopping time then takes the role of regularization strength, analogous to the (inverse) coefficient of the penalty in explicit regularization. For example, coordinate descent has been shown to be related to explicit ℓ_1-regularization (Hastie et al., 2001; Efron et al., 2004; Rosset et al., 2004; Zhang and Yu, 2005) and gradient descent has been shown to trace the path of ridge regularization (Bauer et al., 2007; Ali et al., 2019). However, most results on early-stopping fall into at least one of three categories: either the risk bounds are *un*localized (e.g., for online mirror descent (Shalev-Shwartz, 2007; Bach, 2024)), they only hold in the low-dimensional regime (e.g., Suggala et al. (2018)), or they use tools (such as spectral analysis) that only apply to specific geometries like ℓ_2 or Hilbert spaces (e.g., Wei et al. (2017); Ali et al. (2019)). In particular, a sharp localized analysis of early-stopping for general geometries in high dimensions seems to be lacking. Since the role of the geometry is particularly important to circumvent the curse of dimensionality, sharp localized risk bounds in this setting are of particular interest.

In this work, we address this gap in the high-dimensional linear regression setting using the framework introduced by Kanade et al. (2023). In this setting, we observe a design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ (with possibly $d \gg n$) and n random responses from the model

$$y = \mathbf{X}\alpha^\star + \xi \in \mathbb{R}^n, \tag{1}$$

with a linear ground-truth function parametrized by $\alpha^\star \in \mathbb{R}^d$ and additive i.i.d. Gaussian noise $\xi \sim \mathcal{N}(0, \mathbf{I}_n)$. We assume that the ground truth satisfies the shape constraint given by

$$\alpha^\star \in K_\tau := \tau K = \{\tau\alpha \mid \alpha \in K\} \subset \mathbb{R}^d, \tag{2}$$

where $\tau > 0$ is a "radius" and $K \subset \mathbb{R}^d$ is any convex body, that is, it is convex, compact and the origin is contained in its

[1] Please refer to arXiv for an updated version of this paper.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

interior. We define the empirical and in-sample prediction risk using squared loss, respectively, as

$$\widehat{\mathcal{R}}(\widehat{\alpha}) := \frac{1}{n}\left\|\mathbf{X}\widehat{\alpha} - y\right\|_2^2, \quad \mathcal{R}(\widehat{\alpha}) := \frac{1}{n}\left\|\mathbf{X}(\widehat{\alpha} - \alpha^\star)\right\|_2^2.$$

Under these assumptions, the aim of a predictor $\widehat{\alpha} \equiv \widehat{\alpha}(\mathbf{X}, y) \in \mathbb{R}^d$ is to achieve minimal in-sample risk using the observations $\mathbf{X}$ and y, as well as knowledge of the convex body K and the radius τ.[2] Notice that this setting can be viewed as a Gaussian sequence model over the convex constraints $\mathbf{X}K_\tau$, see Johnstone (2017) for a detailed account. Our results are restricted to in-sample prediction (also known as fixed design), see e.g., Wainwright (2019, Sec. 14.1) or Bach (2024, Chp. 3) for a discussion of the differences to random design.

Notation. The Bregman divergence of a strictly convex and differentiable function $\psi : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$D_\psi(\alpha, \alpha') = \psi(\alpha) - \psi(\alpha') - \left\langle \nabla\psi(\alpha'), \alpha - \alpha' \right\rangle$$

for $\alpha, \alpha' \in \mathbb{R}^d$. We denote the Bregman ball as $B_\psi(\alpha', r) = \left\{\alpha \in \mathbb{R}^d \;\middle|\; D_\psi(\alpha', \alpha) \leqslant r\right\}$ and ℓ_p-norm balls as $B_p^d = \{\alpha \in \mathbb{R}^d | \left\|\alpha\right\|_p \leqslant 1\}$. We write $a \lesssim b$ if there is a constant $C > 0$ such that $a \leqslant Cb$, $a \asymp b$ if $a \lesssim b \lesssim a$, and $a \wedge b = \min\{a, b\}$, $a \vee b = \max\{a, b\}$. We use Minkowsi sum notation throughout: for a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, sets $K, K' \subset \mathbb{R}^d$ and $v \in \mathbb{R}^d$ we write $\mathbf{X}K := \{\mathbf{X}\alpha \mid \alpha \in K\}$, $K + v := \{\alpha + v \mid \alpha \in K\}$ and $K + K' := \{\alpha + \alpha' : \alpha \in K, \alpha' \in K'\}$.

1.1 Local Gaussian width and the LSE

A natural and well-studied estimator in this setting is the constrained *Least Squares Estimator* (LSE), also known as the maximum likelihood estimator. It is defined by minimizing the empirical risk over the constraint set from (2),

$$\widehat{\alpha}_{\mathrm{LSE}} \in \operatorname*{arg\,min}_{\alpha \in K_\tau} \widehat{\mathcal{R}}(\alpha). \tag{3}$$

The predictions of the LSE on the sample $\mathbf{X}$ are given by the orthogonal projection of y onto $\mathbf{X}K_\tau = \{\mathbf{X}\alpha \mid \alpha \in K_\tau\}$, which is unique by the convexity of $\mathbf{X}K_\tau$. Hence, while the minimizer in Equation (3) is not necessarily unique (especially in the high-dimensional setting where $d > n$), its predictions on the sample are, and consequently, we do not need to distinguish between the minimizers any further.

The *Gaussian width* (cf. Vershynin (2018)) of a set $S \subset \mathbb{R}^n$ is defined with a Gaussian vector $\xi \sim \mathcal{N}(0, \mathbf{I}_d)$ as

$$w(K) := \mathbb{E}_\xi\left[\sup_{\theta \in K} \langle \xi, \alpha \rangle\right].$$

In his seminal paper, Chatterjee (2014) showed that the risk of the LSE concentrates sharply around a *critical radius* that maximizes a function of the *local Gaussian width*:

[2] The knowledge of τ is not always required, which we specify in those instances.

Definition 1. For $\alpha \in K \subset \mathbb{R}^d$, define the function $f_{\alpha,K} : [0, \infty) \to [-\infty, \infty)$ as

$$f_{\alpha,K}(r) := w\left((K - \alpha) \cap rB_2^d\right) - \frac{r^2}{2}.$$

The *critical* and the *stationary radius* of a set $K \subset \mathbb{R}^d$ around a point $\alpha \in K$ are defined, respectively, as

$$\begin{aligned} r_\star(\alpha, K) &:= \operatorname*{arg\,max}_{r \geqslant 0} f_{\alpha,K}(r), \\ r_0(\alpha, K) &:= \inf\left\{r > 0 \mid f_{\alpha,K}(r) \leqslant 0\right\}. \end{aligned} \tag{4}$$

We denote the maximal critical and stationary radii on K as

$$r_\star(K) := \sup_{\alpha \in K} r_\star(\alpha, K), \quad r_0(K) := \sup_{\alpha \in K} r_0(\alpha, K).$$

Specifically, in our notation, Chatterjee (2014) showed that $\mathcal{R}(\widehat{\alpha}_{\mathrm{LSE}})$ concentrates sharply around $r_\star^2(\mathbf{X}\alpha^\star, \mathbf{X}K_\tau)/n$, up to constant factors. Multiple other works such as Bellec (2016); Prasadan and Neykov (2024) bound $\mathcal{R}(\widehat{\alpha}_{\mathrm{LSE}})$ in terms of the stationary radius $r_0^2(\mathbf{X}\alpha^\star, \mathbf{X}K_\tau)$ instead, yielding tight leading constants. Importantly, the stationary radius $r_0(\alpha, K)$ can always be bounded by solving

$$w\left((K - \alpha) \cap rB_2^d\right) \leqslant \frac{r^2}{2} \tag{5}$$

for $r \geqslant 0$, where one trivial solution of (5) is always given by $r^2 = 2w(K)$. Moreover, the stationary radius is an upper bound on the critical radius (Chatterjee, 2014, Proposition 1.3). To summarize, if $r \geqslant 0$ is a solution to (5) for a convex body K and $\alpha \in K$, it holds that

$$r_\star^2(\alpha, K) \leqslant r_0^2(\alpha, K) \leqslant \min\left\{r^2, 2w(K)\right\}. \tag{6}$$

1.2 Overview of Contributions

We study *mirror descent* (Nemirovski and Yudin, 1984), which is a gradient-based iterative optimization method that generalizes gradient descent to different geometries.

Definition 2. Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable and have positive definite Hessian $\nabla^2\psi$ everywhere. Unconstrained *continuous-time mirror descent* using ψ and initialized at $\alpha_0 \equiv 0 \in \mathbb{R}^d$ is defined through the ODE

$$\frac{d}{dt}\alpha_t = -\left(\nabla^2\psi(\alpha_t)\right)^{-1} \nabla\widehat{\mathcal{R}}(\alpha_t). \tag{7}$$

Definition 3. Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be differentiable, strictly convex and let the gradient of ψ be surjective, i.e., $\left\{\nabla\psi(\alpha) \;\middle|\; \alpha \in \mathbb{R}^d\right\} = \mathbb{R}^d$. Unconstrained *discrete-time mirror descent* using ψ, initialization $\alpha_0 \equiv 0 \in \mathbb{R}^d$ and fixed step-size $\eta > 0$, is defined through the recursion

$$\nabla\psi(\alpha_{t+1}) = \nabla\psi(\alpha_t) - \eta\nabla\widehat{\mathcal{R}}(\alpha_t). \tag{8}$$

In both cases, the function ψ is called the *mirror map* or *potential* of the mirror descent algorithm. If there are multiple minimizers of the empirical risk, the potential determines which of them mirror descent converges to (Gunasekar et al., 2018). More generally, it determines the optimization path, see Appendix A.2 on page 14 for a visualization. We remark that continuous-time mirror descent is also referred to as Riemannian gradient flow (Gunasekar et al., 2021). Stopping mirror descent before convergence, that is, using $\alpha_{t^\star}$ for some $t^\star > 0$ as the estimator is called *Early-Stopped Mirror Descent* (ESMD).

So far, an analysis akin to the works of Chatterjee (2014) and Bellec (2016) has eluded early-stopped mirror descent. In this work, we close this gap and show that a risk bound almost identical to the one from Bellec (2016) also applies to ESMD, provided the potential is appropriately chosen based on K and τ. This bound holds for *any* convex constraints and *any* design matrix, and importantly, in the high-dimensional setting ($d \gg n$).

We summarize our main contributions below.

- We prove a tight bound on the in-sample risk of ESMD in terms of the stationary radius (Theorem 1). As a consequence, we provide sufficient conditions under which the worst-case risk of ESMD is bounded by that of the LSE (Corollary 1) and for minimax optimality (Corollary 2).
- Using the Minkowski functional of the convex body, we provide sufficient conditions on the optimization potential for our bound to apply (Assumption A). We use these conditions for developing new (and analyzing existing) potentials in several examples (Section 4).
- We apply our risk bounds to ℓ_p-norm balls with $p \in [1, 2)$ as well as general M-convex hulls and derive sharp statistical rates (Section 4). We accompany our bound for $p \in (1, 2)$ with a matching minimax lower bound for column-normalized fixed design matrices. For ℓ_1-constraints, our bound improves upon the best known bounds, demonstrating the benefits of our tight analysis.

2 RELATED WORKS

Constrained Least Squares. The statistical performance of the LSE under convex constraints is well-studied, for example, in Birgé and Massart (1993); Vershynin (2015); Bellec (2016); Plan et al. (2017); Kur et al. (2023). Tight estimates of the local Gaussian width of specific convex bodies are established, for instance, in Gordon et al. (2007); Bellec (2017). The minimax rates of the Gaussian sequence model under convex constraints are characterized exactly in Neykov (2022), and the minimax sub-optimality of the LSE for certain constraints has been described in Prasadan and Neykov (2024). When the convex body is an ℓ_1-norm ball, LSE recovers the LASSO (Tibshirani, 1996) in constrained form, which has been studied extensively in Candes and Tao (2005); Bunea et al. (2007); Ye and Zhang (2010); Bühlmann and van de Geer (2011); Pathak and Ma (2024).

Mirror Descent. Mirror descent as an optimization procedure is well-studied (Beck and Teboulle, 2003; Shalev-Shwartz, 2007; Agarwal et al., 2012). The implicit bias of mirror descent in the overparameterized regime is studied in Gunasekar et al. (2018); Sun et al. (2023). The generalization properties of *online* mirror descent have been studied extensively (Shalev-Shwartz, 2007; Orabona, 2023; Lattimore, 2024; Bach, 2024). Notable instances of this are Srebro et al. (2011); Levy and Duchi (2019); Gatmiry et al. (2024), who also relate the potential to the geometry of the constraint set, even showing a "universality" of online mirror descent. While these bounds can be minimax optimal, they are usually not *local* (with a few exceptions (Rakhlin et al., 2013)), or only give guarantees on averaged iterates.

Local risk bounds for early-stopping. The application of localized complexity measures to iterative regularization methods is scarce in the literature. While local Gaussian width appears in Wei et al. (2017), their results only apply to (unconstrained) RKHS. One could also obtain local risk bounds by directly tying mirror descent to explicit regularization paths, but known results either require strong convexity of the empirical risk which is necessarily violated in high dimensions (Suggala et al., 2018, Section 4), or apply only to the ℓ_2 geometry (Ali et al., 2019). In Kanade et al. (2023), a general analysis with offset Rademacher complexities is introduced; our work is based on their framework.

3 MAIN RESULTS

We now provide a list of sufficient conditions for the mirror descent potential based on the Minkowski functional of the convex body (Bonnesen and Fenchel, 1934).

Definition 4. For a convex body $K \subset \mathbb{R}^d$, the function $\varphi_K : \mathbb{R}^d \to \mathbb{R}$, defined as

$$\varphi_K(\alpha) := \inf \{\tau > 0 \mid \alpha \in \tau K\}$$

is called the *Minkowski functional* of K (also referred to as distance or gauge function).

The Minkowski functional is a norm if and only if K is centrally symmetric (i.e., $K = -K$). In that case, we use the notation $\|\cdot\|_K$ rather than $\varphi_K(\cdot)$. Any convex body $K \subset \mathbb{R}^d$ that contains the origin in its interior can be written as $K = \left\{\alpha \in \mathbb{R}^d \,\middle|\, \varphi_K(\alpha) \leqslant 1\right\}$. Noticably, convexity of K_τ and strong Lagrange duality imply that we can rewrite the LSE in unconstrained form as

$$\widehat{\alpha}_{\mathrm{LSE}} \in \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^d} \left\{\widehat{\mathcal{R}}(\alpha) + \lambda_n(\tau)\varphi_K(\alpha)\right\}$$

for some data-dependent, not necessarily computable regularization strength $\lambda_n(\tau) \geqslant 0$.

The conditions on the potential we formulate below then depend on whether discrete-time or continuous-time mirror descent is used.

Assumption A. The potential $\psi : \mathbb{R}^d \to [0, \infty)$ satisfies:

(I) For continuous-time MD, ψ is twice differentiable, and in discrete-time, ψ is differentiable. In both cases, the gradient of ψ vanishes at zero, that is, $\nabla\psi(0) = 0$.

(II) The square-root $\sqrt{\psi}$ is convex.

(III) In discrete time, ψ is ρ-strongly convex with respect to some norm, and in continuous time, it is strictly convex (cf. Definition 2).

(IV) There exist constants $c_l, c_u > 0$ independent of all other parameters, such that

$$\forall \alpha \in \mathbb{R}^d : \quad \varphi_K(\alpha) \leqslant c_l \sqrt{\psi(\alpha)},$$
$$\forall \alpha \in K_\tau : \quad \sqrt{\psi(\alpha)} \leqslant c_u \tau.$$

Throughout this paper, we denote the "approximation constant" $c_a = c_l \cdot c_u$. We remark that Assumption A is loosely connected to the notion of κ-regularity from Juditsky and Nemirovski (2008, Def. 2.1).

If the squared Minkowski functional φ_K^2 satisfies (I) and (III), and since it satisfies (II) and (IV) by definition, we could simply choose $\psi = \varphi_K^2$. For example, for any vector $v \in (1/2)B_2^d$ and $K = B_2^d - v$, it is easily verified that φ_K^2 is twice differentiable with vanishing gradient at zero (I) and φ_K^2 is $2/(1 + \|v\|_2)^2$-strongly convex (III). Notably, in this example φ_K is not a norm, as the convex body is not centrally symmetric.

However, in many interesting cases the squared Minkowski functional φ_K^2 is not smooth or not strongly convex (for example, the ℓ_1-norm), and we need to approximate it with a different function. This is possible for all K, as we now show in Lemma 1. Specifically, we can smoothen and "strongly convexify" φ_K^2: To that end, we denote the Moreau envelope (Moreau, 1965) of a closed and proper convex function f with $\lambda > 0$ as

$$(\mathcal{M}_\lambda f)(\alpha) = \inf_{\alpha' \in \mathbb{R}^d} \left\{ f(\alpha') + \frac{1}{2\lambda} \|\alpha - \alpha'\|_2^2 \right\}.$$

Lemma 1. *For any convex body $K \subset \mathbb{R}^d$ that contains the origin in its interior, there exists a potential ψ that satisfies both the continuous and discrete-time versions of Assumption A with approximation constant $c_a = 4$ and some $\rho > 0$. Furthermore, for $\rho = 2/(\max_{\alpha \in K} \|\alpha\|_2^2)$ and sufficiently small $\lambda > 0$ independent of τ, the potential*

$$\psi(\alpha) = (\mathcal{M}_\lambda \varphi_K^2)(\alpha) + \frac{\rho}{2} \|\alpha\|_2^2$$

satisfies the discrete-time version of Assumption A with approximation constant $c_a = 4$.

We prove Lemma 1 in Appendix B.2 using results from Planiden and Wang (2019). We note that in Lemma 1, the potentials do not require any knowledge of the radius. Beyond Moreau-smoothing, other smoothing methods, such as the Polar envelope (Friedlander et al., 2019), or infimal convolution smoothing (Beck and Teboulle, 2012) could be viable options instead. Moreover, there are applications where other potentials that are tailored to the convex body may be more suitable, as we will see in Section 4.

3.1 A Localized Gaussian Width Risk Bound

By Lemma 1, Assumption A is always non-vacuous, which leads us to the following theorem; our main result.

Theorem 1. *Let $\alpha_0 = 0 \in \mathbb{R}^d$ and let $\{\alpha_t\}_{t \geqslant 0}$ be the continuous or discrete-time mirror descent updates on $\widehat{\mathcal{R}}$ using some ψ that satisfies Assumption A. In the discrete-time case, let ψ be ρ-strongly convex and $\widehat{\mathcal{R}}$ be β-smooth with respect to the same norm, and let the step-size satisfy $\eta \leqslant \frac{\rho}{\beta} \wedge \frac{D_\psi(\alpha^\star, 0)}{2}$. Then, for any $\varepsilon > 0$ and $T := c_u^2 \tau^2 / \varepsilon$ in continuous time and $T := \lceil 2c_u^2 \tau^2 / (\varepsilon \eta) \rceil$ in discrete time,*

$$\min_{0 \leqslant t \leqslant T} \mathcal{R}(\alpha_t) \leqslant \frac{2r_0^2(\mathbf{X}\alpha^\star, \mathbf{X}K_{3c_a\tau})}{n} + \frac{4\log(1/\delta)}{n} + \varepsilon \quad (9)$$

with probability at least $1 - \exp(-0.1n) - \delta$ over draws of the noise ξ. Moreover, in continuous time, we have that

$$\mathbb{E}_\xi \left[\min_{0 \leqslant t \leqslant T} \mathcal{R}(\alpha_t) \right] \leqslant \frac{2r_0^2(\mathbf{X}\alpha^\star, \mathbf{X}K_{3c_a\tau})}{n} + \frac{4}{n} + \varepsilon. \quad (10)$$

The proof can be found in Appendix B.3 and is outlined in Section 3.3. Throughout this paper, we choose $\varepsilon > 0$ to balance the right-hand side of (9), respectively (10), and we denote the oracle optimal stopping time as

$$t^\star := \arg\min_{0 \leqslant t \leqslant T} \mathcal{R}(\alpha_t).$$

We would like to stress that this stopping time can depend on the noise and the ground truth, and hence is not necessarily computable. However, $t^\star \leqslant T$ quantifies the maximal number of iterations necessary to achieve the statistical complexity. Note that in the case of discrete time, the strong convexity parameter does not influence the bound in (9), however, it impacts the bound on the stopping time.

Remark 1. It is easily shown (Appendix B.4) that for any convex body we can bound the stationary radius as

$$r_0^2(\mathbf{X}K) \leqslant 4 \operatorname{rk}(\mathbf{X}) \leqslant 4 \min\{n, d\}, \quad (11)$$

where $\operatorname{rk}(\mathbf{X})$ denotes the rank of $\mathbf{X}$. This is unsurprising since the unconstrained LSE is known to achieve the rate $\operatorname{rk}(\mathbf{X})/n$, which is the minimax risk without any shape constraints over $\mathbb{R}^d$ (Wainwright, 2019, Example 15.14).

3.2 A Few Consequences

Comparison with LSE. Theorem 1 immediately leads us to the following corollary connecting the in-sample risks of ESMD and the LSE, based on the results from Chatterjee (2014). For this, we must restrict ourselves to cases in which the following assumption holds.

Assumption B. There exists a constant $\mathcal{C} \geqslant 1$ such that

$$r_0^2(\mathbf{X}K_\tau) \leqslant \mathcal{C} \cdot r_\star^2(\mathbf{X}K_\tau),$$

implying $r_0^2(\mathbf{X}K_\tau) \asymp r_\star^2(\mathbf{X}K_\tau)$ by (6).

Assumption B is not very strong: it holds for Donsker classes, and many non-parametric classes (cf. (van de Geer, 2000)). But, importantly, it does not always hold; see Prasadan and Neykov (2024, Sec. 3.1.3) for a counterexample. Other notable examples of when it does not hold, appear in Kur et al. (2023), and see also Aolaritei et al. (2025).

We prove the following Corollary 1 in Appendix B.5.

Corollary 1. *In the setting of Theorem 1, if $r_\star(\mathbf{X}K_\tau) \geqslant (644/3)^2$ and Assumption B holds, it holds that*

$$\sup_{\alpha^\star \in K_\tau} \mathbb{E}_\xi \left[\mathcal{R}(\alpha_{t^\star})\right] \leqslant 84\mathcal{C}c_a \cdot \sup_{\alpha^\star \in K_\tau} \mathbb{E}_\xi \left[\mathcal{R}(\widehat{\alpha}_{\mathrm{LSE}})\right] \quad (12)$$

for continuous-time ESMD with large enough T.

It follows that if $\mathcal{C} \cdot c_a \lesssim 1$ and the LSE is minimax optimal, then ESMD with optimal stopping time is also minimax optimal. Whether this is the case depends highly on the convex body and the design matrix (Raskutti et al., 2011a; Kur et al., 2020). In Theorem 1 and Corollary 1 we did not optimize the constants, and tighter bounds (in terms of constants) could be derived using our arguments.

Minimax optimality. We can also directly derive a sufficient condition for minimax optimality. To that end, we define the maximum local entropy (e.g., from Neykov (2022)). Let $M(r, K)$ denote the packing number of K in ℓ_2-norm at radius r, and let $c^\star > 0$ be a sufficiently large absolute constant. The local entropy of a set $K \subset \mathbb{R}^d$ is defined as

$$M^{\mathrm{loc}}(r, K) = \sup_{\alpha \in K} M(r/c^\star, (K - \alpha) \cap rB_2^d).$$

We get a sufficient condition for minimax optimality.

Assumption C. It holds for all $r \lesssim \operatorname{diam}(\mathbf{X}K_\tau)$ that

$$\sup_{\alpha \in K_\tau} \frac{w(\mathbf{X}(K_\tau - \alpha) \cap rB_2^n)}{r} \leqslant \sqrt{\log M^{\mathrm{loc}}(r, \mathbf{X}K_\tau)}.$$

Examples that satisfy this condition are Donsker classes and set constrained models in general dimensions, see Han (2021); Kur et al. (2019) for discussions.

Corollary 2. *Consider the setting of Theorem 1. If Assumption C holds and $c_a \lesssim 1$, then continuous-time ESMD is minimax optimal (up to constant factors), that is,*

$$\sup_{\alpha^\star \in K_\tau} \mathbb{E}_\xi \left[\mathcal{R}(\alpha_{t^\star})\right] \lesssim \inf_{\widehat{\alpha}} \sup_{\alpha^\star \in K_\tau} \mathbb{E}_\xi \left[\mathcal{R}(\widehat{\alpha})\right].$$

We prove Corollary 2 in Appendix B.6 following Prasadan and Neykov (2024, Cor. 2.6). Corollary 2 yields, for example, that if $\mathbf{X}$ is the identity and K_τ is an ℓ_1- or ℓ_2-ball of arbitrary radius $\tau > 0$, ESMD is minimax rate optimal. We revisit the ℓ_1-constrained setting in Section 4.2.

Estimation. Finally, we would like to highlight that Theorem 1 can easily be used to derive bounds on the estimation risk whenever the design matrix has a vanishing kernel width, cf. Raskutti et al. (2011a). A matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has a vanishing kernel width with respect to f and K, if for all $\alpha \in K - K$

$$\frac{1}{n} \|\mathbf{X}\alpha\|_2^2 \geqslant \|\alpha\|_2^2 - f(K, n). \quad (13)$$

Corollary 3. *Under the conditions of Theorem 1 and if $\mathbf{X}$ has vanishing kernel width (13) with respect to $3c_a\tau K$ and $f(3c_a\tau K, n) \lesssim r_0^2(\mathbf{X}K_{3c_a\tau})/n$, the estimation error of ESMD is bounded for all $\alpha^\star \in K_\tau$ as*

$$\|\alpha_{t^\star} - \alpha^\star\|_2^2 \lesssim c_a \frac{r_0^2(\mathbf{X}K_\tau)}{n} + \frac{\log(1/\delta)}{n}$$

with probability at least $1 - \exp(-0.1n) - \delta$ over the noise.

We prove Corollary 3 in Appendix B.7. The additional assumption of vanishing kernel width is necessary because parameter estimation is ill-posed if the data matrix does not satisfy any regularity assumptions, especially in the high-dimensional regime where $d > n$.

3.3 Proof Outline of Theorem 1

Finally, we provide a short proof outline of Theorem 1; the full proof is in Appendix B.3. The first ingredient of the proof of Theorem 1 is to show that, under Assumption A, even without strong convexity of the potential, we have the following inclusion (cf. Figure 1)

$$\forall \alpha^\star \in K_\tau : \quad B_\psi(\alpha^\star, 2D_\psi(\alpha^\star, 0)) \subset 3c_a K_\tau. \quad (14)$$

This is useful, as Kanade et al. (2023) showed that optimally early-stopped mirror descent is contained in this Bregman ball while satisfying the so-called offset condition (Liang et al., 2015), which is defined as

$$\widehat{\mathcal{R}}(\alpha_{t^\star}) - \widehat{\mathcal{R}}(\alpha^\star) + \mathcal{R}(\alpha_{t^\star}) \leqslant \varepsilon. \quad (15)$$

The key step is then to show that using (14), we can relate the offset condition (15) to the stationary radius from Definition 1 using localization arguments akin to those in Bellec (2016). Specifically, we show that (15) and (14) imply that we can bound the in-sample risk with the supremum of a Gaussian process, that is,

$$\mathcal{R}(\alpha_{t^\star}) \leqslant \frac{1}{n}\left(\frac{Z_{r_0}}{r_0}\right)^2 + \varepsilon$$
$$\text{with} \quad Z_{r_0} = \sup_{\theta \in (\mathbf{X}(K_{3c_a\tau} - \alpha^\star) \cap r_0 B_2^n)} \langle \xi, \theta \rangle,$$

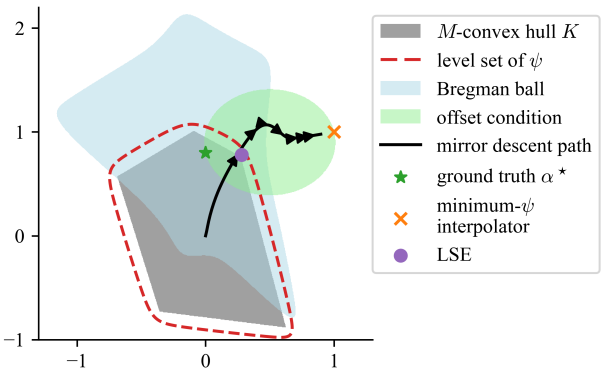


Figure 1: We plot an M-convex hull, a level set of the potential from (20) with $\gamma = 10$ and $\rho = 0.2$, the Bregman ball from (14), the set of points satisfying the offset condition (15), and the mirror descent path.

where we denote $r_0 = r_0(\mathbf{X}\alpha^\star, \mathbf{X}K_{3c_a\tau})$. We can bound Z_{r_0} using the concentration of the supremum of a Gaussian process to its expectation $\mathbb{E}[Z_{r_0}]$. By definition of the stationary radius, we have that $\mathbb{E}[Z_{r_0}] \leqslant r_0^2/2$, which we can plug in and putting things together yields (9).

4 APPLICATIONS

So far, we have only discussed the general case of arbitrary convex bodies and design matrices, which yields the full generality of our main results Theorem 1 and Corollaries 1 to 3. These results allow us to view Assumption A as a blueprint. For a given convex body, one can construct a potential that satisfies (I)-(IV). Using this potential, ESMD then enjoys the guarantees of Theorem 1. We now demonstrate this approach for specific choices of convex bodies and further assumptions on the design matrices. We analyze existing potentials and explicitly construct novel potentials.

Assumptions on the Design Matrix. For our main results, we made no assumptions on the design matrix. Now we consider two special cases in the applications. The first is when the design matrix is fixed and we only assume normalized columns. $\mathbf{X} \in \mathbb{R}^{n\times d}$ is said to be column-normalized, if its columns $\mathbf{X}_i, i \in [d]$ satisfy

$$\|\mathbf{X}_i\|_2 \leqslant \sqrt{n}. \tag{16}$$

Note that normalizing to $\sqrt{n}$ here is somewhat arbitrary, and can readily be changed. The second setting is that of Gaussian design: $\mathbf{X}$ is Gaussian, if the entries $\mathbf{X}_{ij}, i \in [n], j \in [d]$ of $\mathbf{X}$ are i.i.d. standard Gaussian, that is, $\mathbf{X}_{ij} \sim \mathcal{N}(0,1)$. Up to constants, Gaussianity implies having normalized columns (16) and vanishing kernel width (13) for some K with high probability; see Raskutti et al. (2011a, Sec. 3.2).

4.1 ℓ_p-norms with $p \in (1,2)$

The following result shows that ℓ_p-norms with $p \in (1,2)$ are regular enough such that we can simply choose ψ as the squared Minkowsi functional, i.e., as $\|\cdot\|_p^2$. The mirror descent algorithm associated with this potential is known as the p-norm algorithm (Shalev-Shwartz, 2007; Levy and Duchi, 2019; Orabona, 2023) and can be implemented very efficiently (Gentile, 2003). Orthogonally, it is worth pointing out that in the interpolation regime, when mirror descent with $\|\cdot\|_p^2$ is *not* early-stopped, it converges to the minimum ℓ_p-norm interpolator (Gunasekar et al., 2018). These predictors have been studied as part of the benign overfitting literature (Donhauser et al., 2022; Kur et al., 2024).

We now derive the rates from (9) explicitly for column-normalized and Gaussian design.

Proposition 2. *Let $p \in (1,2)$, $1/p + 1/q = 1$ and $K = B_p^d$. Then $\psi(\alpha) = \|\alpha\|_p^2$ satisfies the discrete-time version of Assumption A where the potential is $\rho = 2(p-1)$-strongly convex with respect to the ℓ_p-norm, and $c_a = 1$. If $\mathbf{X}$ is column normalized (16), optimally early-stopped mirror descent achieves for all $\alpha^\star \in \tau B_p^d$*

$$\mathcal{R}(\alpha_{t^\star}) \lesssim \frac{\mathrm{rk}(\mathbf{X})}{n} \wedge \frac{\tau}{\sqrt{n}} \begin{cases} \sqrt{\log d} & \text{if } p \leqslant 1 + \frac{1}{\log d}, \\ \sqrt{q} d^{1/q} & \text{if } p > 1 + \frac{1}{\log d} \end{cases}$$

with probability at least $0.99 - \exp(-0.1n)$ over the noise. If $\mathbf{X}$ is Gaussian, then for all $\alpha^\star \in \tau B_p^d$, ESMD achieves the same bound with $\mathrm{rk}(\mathbf{X})/n = 1$ and with probability at least $0.99 - 2\exp(-0.1n)$ jointly over draws of the design matrix $\mathbf{X}$ and the noise ξ.

We prove Proposition 2 in Appendix B.8. The strong convexity was proved, for example, in Shalev-Shwartz (2007); Ball et al. (1994). A tighter bound for specific scalings may be possible using the bounds on the localized Gaussian width from Gordon et al. (2007).

Minimax (Sub-)Optimality. When $\mathbf{X}$ is a scaled identity, the LSE is known to be *sub*-optimal under ℓ_p-norm constraints (Donoho and Johnstone, 1994; Johnstone, 2017; Prasadan and Neykov, 2024; Aolaritei et al., 2025), and so by Equation (6) we cannot generally hope for our bounds to prove minimax optimality of ESMD in that case.

Surprisingly, however, there seems to be no work (explicitly) establishing the minimax rate under the *worst-case* fixed and Gaussian design. We now show that the rate from Proposition 2 is optimal (up to p-dependent-factor) for a worst-case fixed design matrix that is column-normalized. To that end, we explicitly construct a column-normalized data matrix as a hard instance. This is similar in spirit to a line of research investigating particularly hard design matrices (Rigollet and Tsybakov, 2010; Zhang et al., 2017; Pathak and Ma, 2024; Foygel and Srebro, 2011; Dalalyan

et al., 2017). At the same time, we show that the rate from Proposition 2 is sub-optimal for the Gaussian design matrix.

Theorem 3. *Let $p \in [1+1/\log d, 2)$, $1/p+1/q=1$ and let $\tau = 1$ for simplicity. Assume that $n^{p/2} \leqslant d \leqslant n^{q/2}$. There exists a design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ that is column normalized (16), such that the minimax in-sample risk satisfies*

$$\inf_{\widehat{\alpha}} \sup_{\alpha^\star \in B_p^d} \mathbb{E}_\xi \left[\mathcal{R}(\widehat{\alpha})\right] \geqslant 1 \wedge c_p \frac{d^{1/q}}{\sqrt{n}},$$

where the infimum is taken over all estimators and $c_p > 0$ is a constant that may depend on p. If $\mathbf{X}$ is Gaussian, let $n(\log d)^{c_0} \leqslant d \leqslant n^{q/2}$ and $p \in [1 + c_1/\log\log d, 2)$ for some universal constant $c_0, c_1 > 0$. Then

$$\inf_{\widehat{\alpha}} \sup_{\alpha^\star \in B_p^d} \mathbb{E}_\xi \left[\mathcal{R}(\widehat{\alpha})\right] \asymp n^{p/2-1}(\log d)^{1-p/2} \vee \frac{d^{2/q}}{n}$$

with probability at least $1 - c_2\exp(-c_3 n)$ over draws of the matrix $\mathbf{X}$, where $c_2, c_3 > 0$ are some constants.

We prove Theorem 3 in Appendix B.9. Note that the upper bound from Proposition 2 and the lower bound from Theorem 3 essentially match for the fixed data matrix. In Appendix A.1, we present simulations of the LSE on the adversarial data matrix that we construct in the proof of the first lower bound, showing that it exhibits the rate $d^{1/q}/\sqrt{n}$.

4.2 ℓ_1-norm

When the convex body is an ℓ_1-norm ball, the corresponding LSE is the LASSO estimator in its constrained form (Tibshirani, 1996). It is known (Bellec, 2017, Thm. 7) that for all column-normalized design matrices (16), if $d \geqslant \tau\sqrt{n}$, the LSE achieves with high probability over the noise

$$\mathcal{R}(\widehat{\alpha}_{\mathrm{LSE}}) \lesssim \frac{\mathrm{rk}\,(\mathbf{X})}{n} \wedge \tau\sqrt{\frac{\log(ed/(\tau\sqrt{n}))}{n}}, \tag{17}$$

and there exists at least one column-normalized data matrix for which this is minimax optimal (Rigollet and Tsybakov, 2010, Thm. 5.3 and Eq. 5.25). If $d/(\tau\sqrt{n}) \asymp d^\kappa$ with some constant $\kappa > 0$ and the rank of the design matrix is large enough, the bound from (17) reduces to the rate $\tau\sqrt{\log(d)/n}$. In Raskutti et al. (2011a, Thm. 3), it is shown that if this scaling assumption holds and the design matrix has vanishing kernel width (13) with $f(\tau B_1^d, n) \lesssim \sqrt{\tau}\,(\log(d)/n)^{1/4}$ (implying it has rank n), the minimax lower bound matches the simpler bound. However, without the scaling assumption and if $\tau\sqrt{n} \geqslant e$, (17) is a stronger bound (Rigollet and Tsybakov, 2010, pgs. 16-17).

Kanade et al. (2023, Thm. 5) showed that when the design matrix is column-normalized (16), optimally early-stopped continuous-time mirror descent with the hyperbolic entropy from Ghai et al. (2020), defined as

$$\psi(\alpha) = \sum_{i=1}^d \alpha_i \operatorname{arcsinh}(\alpha_i/\gamma) - \sqrt{\alpha_i^2 + \gamma^2},$$

achieves for appropriate $\gamma > 0$ with high probability over draws of the noise

$$\mathcal{R}(\alpha_{t^\star}) \lesssim \tau\frac{\log^{3/2} d}{\sqrt{n}}. \tag{18}$$

As one can see, there is a gap from (18) to (17) of order (at least) $\log d$. The same potential has also been used for the related setting of sparse noisy phase retrieval (Wu and Rebeschini, 2022) and in the analysis of diagonal networks (Woodworth et al., 2020). And while a batch conversion of online mirror descent with the first potential from Table 1 is known to achieve the rate $\sqrt{\log(d)/n}$ (e.g., Bach (2024)), as discussed, there is still a gap to (17) and their proof technique does not apply to our definition of mirror descent.

Employing our results, together with results from Bellec (2017), we can improve upon (18) and fully close the gap between (17) and (18). Because $x \mapsto \|x\|_1^2$ is *not* differentiable, nor strictly convex, we cannot use it as a potential itself. However, we can use Assumption A to derive several alternatives. The next theorem provides a joint analysis of the potentials described in Table 1.

Theorem 4. *Suppose that $K = B_1^d$ and $d \geqslant \tau\sqrt{n}$. Then all potentials from Table 1 satisfy Assumption A with the specified constants, and if $\mathbf{X}$ is column-normalized (16), optimally stopped mirror descent using any of the potentials from Table 1 achieves for all $\alpha^\star \in \tau B_1^d$*

$$\mathcal{R}(\alpha_{t^\star}) \lesssim \left(\frac{\mathrm{rk}\,(\mathbf{X})}{n} \wedge c_a\tau\sqrt{\frac{\log(ed/(\tau\sqrt{n}))}{n}}\right) + \frac{\log(1/\delta)}{n}$$

with probability at least $1 - \exp(-0.1n) - \delta$ over the noise. If $\mathbf{X}$ is Gaussian, then optimal early-stopping achieves

$$\mathcal{R}(\alpha_{t^\star}) \lesssim 1 \wedge c_a\tau\sqrt{\frac{\log d}{n}}$$

with probability at least $0.99 - 2\exp(-0.1n)$ jointly over draws of the design matrix $\mathbf{X}$ and the noise ξ.

We prove Theorem 4 in Appendix B.10, where we also specify the constants of the first bound. As discussed above, because when $\mathbf{X}$ is Gaussian it has vanishing kernel width (Raskutti et al., 2011a, Proposition 1), both bounds are minimax optimal under weak scaling assumptions. Therefore, they cannot be improved upon beyond the constants and ESMD is minimax optimal, as is the LSE.

Notice how the third potential in Table 1 is an adjusted version of the hypentropy potential from Ghai et al. (2020). With only a few changes to the potential, we closed the logarithmic gap from (18) to (17); In particular, the key is that we *square* the potential. The sigmoidal example in Table 1 for continuous-time mirror descent, where strong convexity is not required, is a natural smooth approximation of the absolute value function from Schmidt et al. (2007). Some example paths using potentials from Table 1, and the risk along the optimization path, are plotted in Appendix A.2.

Table 1: Mirror descent potentials for linear regression over the ℓ_1-norm ball with which early-stopping is minimax optimal (Theorem 4). Note that the Moreau envelope is *not* the same as in Lemma 1, which would also be a valid potential for Theorem 4. Here we show $(\mathcal{M}_\lambda \|\cdot\|_1 + \frac{d\lambda}{2})^2 + \frac{\rho}{2}\|\cdot\|_2^2$, because it has a closed-form solution. See Figure 5 in Appendix A.2 for example optimization paths.

name	**potential** $\psi(\alpha)$	**strong convexity parameter** ρ	**approximation constant** c_a
squared ℓ_p-norm	$\|\alpha\|_p^2$ with $1 < p \leqslant 1 + 1/\log(d)$	$\rho = 2(p-1) \leqslant \frac{2}{\log d}$ with ℓ_p-norm	$c_a = e$
Moreau envelope (Huber loss)	$(\sum_{i=1}^d h(\alpha_i) + \tau)^2 + \|\alpha\|_2^2$ where $h(x) = \begin{cases} \frac{x^2 d}{4\tau} & \|x\| \leqslant \frac{2\tau}{d} \\ \|x\| - \frac{\tau}{d} & \|x\| > \frac{2\tau}{d} \end{cases}$	$\rho = 2$ with ℓ_2-norm	$c_a = \sqrt{5}$
adjusted hypentropy	$\left(\sum_{i=1}^d \frac{\left(\alpha_i \operatorname{arcsinh}(\alpha_i/\gamma) - \sqrt{\alpha_i^2+\gamma^2} + \gamma + 1\right)}{\operatorname{arcsinh}(\gamma^{-1})}\right)^2$ with $\gamma \leqslant \sinh(d/\tau)^{-1} \wedge (4\tau)^{-1} \wedge 2^{-1/2}$	$\rho = \operatorname{arcsinh}(\gamma^{-1})^{-2}$ $\leqslant (\tau/d)^2$ with ℓ_2-norm	$c_a = 3$
sigmoidal	$\left(\sum_{i=1}^d \frac{(\log(1+\exp(-\gamma\alpha_i)) + \log(1+\exp(\gamma\alpha_i)))}{\gamma}\right)^2$ with $\gamma \geqslant d\log(4)/\tau$	only valid for continuous time	$c_a = 2$

Computational-statistical trade-off. As we can see in Table 1, the strong convexity parameter ρ varies depending on the potential we use. For example, the strong convexity parameter $2(p-1) \leqslant 2/\log(d)$ of the squared ℓ_p-norm vanishes logarithmically as $d \to \infty$. However, the Moreau envelope-based potential showcases that this is not necessary. Importantly, the stopping time from Theorem 1 behaves as $T \asymp 1/\rho$, ignoring all other dependencies. Hence, whether (and how fast) the strong convexity parameter vanishes determines the bound on the computational cost of achieving minimax optimality with the given potentials. Thus, the squared ℓ_p-norm is not necessarily the best candidate. Generally, we can improve the constant c_a arbitrarily close to 1 from above by paying in a smaller ρ.

4.3 M-convex Hulls

Let the convex body be the convex hull of M points $k_i \in \mathbb{R}^d$ that contains the origin in its interior. Using the points to represent the convex body is known as the V-representation, but one can also transform this representation into a so-called H-representation, where the convex body is characterized as the set of solutions to the linear inequality $\mathbf{A}\alpha \leqslant \mathbf{1}_m$ with $\mathbf{A} \in \mathbb{R}^{m \times d}$ and $\mathbf{1}_m = (1, \dots, 1)^\top \in \mathbb{R}^m$, where $m \in \mathbb{N}$ is some finite number that can be bounded as $m \geqslant d+1$ and in terms of M. See Schrijver (1998) for more in-depth background. Given this representation, we can equivalently write the convex body as the solution to $\max_{i \in [m]} \langle a_i, \alpha \rangle \leqslant 1$, where a_i is the i-th row of $\mathbf{A}$. Hence, it is easy to see that

$$\varphi_K(\alpha) = \max_{i \in [m]} \langle a_i, \alpha \rangle . \tag{19}$$

Special cases of this are the ℓ_1-norm and the ℓ_∞-norm. We can approximate this Minkowski functional using a smoothing from Beck and Teboulle (2012, Example 4.5).

Proposition 5. *Suppose that K is an M-convex hull that contains the origin its interior, as described above, and that the Minkowski functional is in the form of (19) with $\sum_{i=1}^m a_i = 0$. Then ψ, defined as*

$$\psi(\alpha) = \frac{1}{\gamma^2} \log^2 \left(m \sum_{i=1}^m \exp(\gamma \langle a_i, \alpha \rangle) \right) + \frac{\rho}{2} \|\alpha\|_2^2 \tag{20}$$

with $\rho = 2/\max_{i \in [M]} \|k_i\|_2^2$ and $\gamma \geqslant 2\log(m)/\tau$ satisfies the continuous and discrete-time version of Assumption A with ρ for the ℓ_2-norm and $c_a = 3$. Moreover, if

$$\phi = \max_{i \in [M]} \frac{\|\mathbf{X} k_i\|_2}{\sqrt{n}}$$

and $M \geqslant \tau\phi\sqrt{n}$, optimally early-stopped mirror descent with the potential from (20) achieves for all $\alpha^\star \in K_\tau$

$$\mathcal{R}(\alpha_{t^\star}) \lesssim \frac{\operatorname{rk}(\mathbf{X})}{n} \wedge c_a \tau\phi \sqrt{\frac{\log\left(eM/(\tau\phi\sqrt{n})\right)}{n}}$$

with probability at least $0.99 - \exp(-0.1n)$.

Using the bound on localized Gaussian width of M-convex hulls from Bellec (2017), we prove Proposition 5 in Appendix B.11, where we also specify the constants. Noticably, when K is an ℓ_1-norm ball, we recover the bound from Theorem 4 with $M = 2d$, $m = 2^d$ and $\phi = 1$.

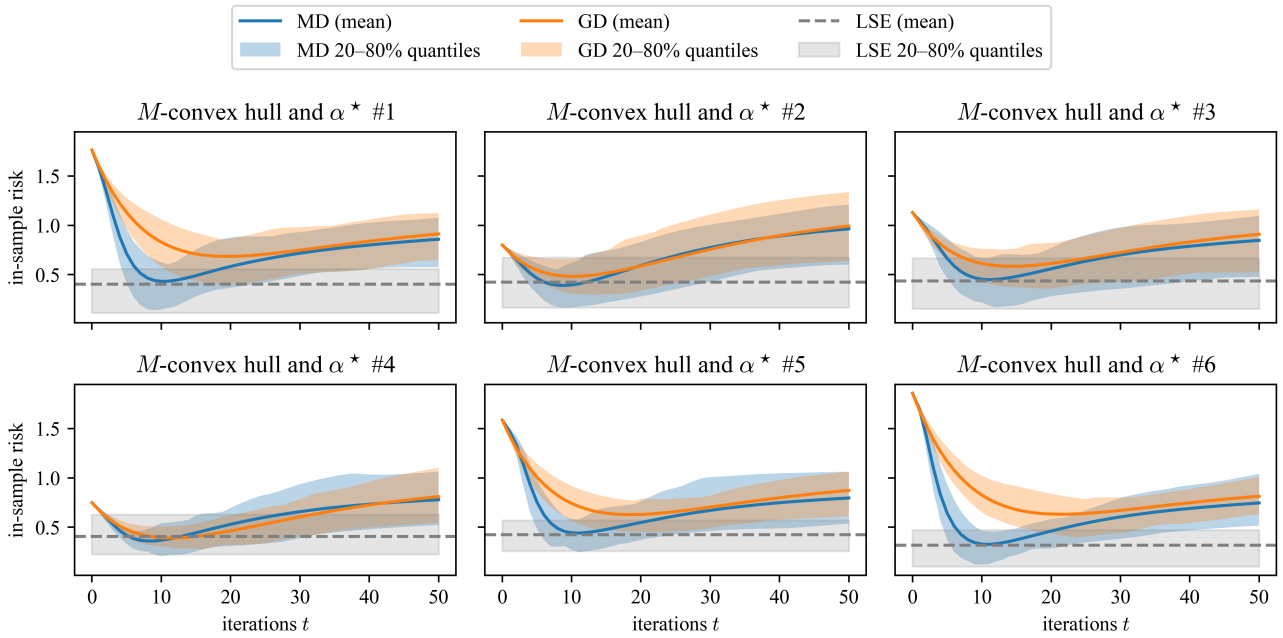


Figure 2: Risks of mirror descent (MD) with potential from (20), gradient descent (GD) and the LSE on six randomly generated convex bodies and ground truths. Using the potential clearly adapts ESMD to the geometry of the convex body.

Simulations. In Figure 2, we randomly generate six M-convex hulls in $\mathbb{R}^{10}$ with $m = 50$ points, and sample a ground truth within this convex body. We let $\mathbf{X}$ be the identity. Figure 2 shows how the mirror potential from Equation (20) correctly adapts to the geometry of the convex constraint and how early-stopped mirror descent performs better than early-stopped gradient descent. We repeat the experiment 30 times and plot mean, 20th and 80th percentile for different instances of the noise. Note that the best iterate has similar risk to the LSE, as described in Corollary 1.

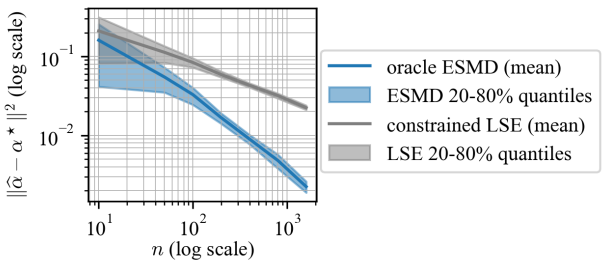


Figure 3: Risk as a funciton of n in the $\ell_{3/2}$-instance described in Section 5: ESMD outperforms the LSE.

5 DISCUSSION

Our main contribution is to show that the sharp Gaussian width analysis of the LSE from Bellec (2016) translates to early-stopped mirror descent, when the mirror potential is chosen to approximate the squared Minkowski functional of the convex constraints. This enables a general comparison to the LSE, a formulation of sufficient conditions for minimax optimality, and an improvement over the best known bounds for the ℓ_1-constrained case. Our results extend to general convex constraints and the high-dimensional setting, a regime that had evaded some of the previous analyses.

A caveat of our bound is that it cannot prove that ESMD achieves better rates than the LSE in settings where this may be expected. For instance, a setting where the LSE is expected to perform poorly (see Section 2.2 in Aolaritei et al. (2025)) is when we choose $\mathbf{X} = \mathbf{I}_n$, $K = B^d_{3/2}$ and $\alpha^\star = e_1$, and we scale the noise as $\xi \sim \mathcal{N}(0, n^{-4/3}\mathbf{I}_n)$. Then we can observe in Figure 3 that optimally stopped mirror descent with potential $\|\cdot\|^2_{3/2}$ outperforms the LSE. However, a fundamentally different proof technique is required to prove such a discrepancy, as the stationary radius in our bound for ESMD also bounds the risk of the LSE.

Moreover, our comparison between ESMD and the LSE only pertains to the maximal risk over the constraint set, and whether a similar comparison is possible point-wise remains an interesting problem.

Lastly, our analysis is specific to in-sample prediction. The results from Kanade et al. (2023) may also be applicable for showing analogous bounds for out-of-sample prediction, but this would require tight bounds on the offset Rademacher complexity of convex bodies and, to facilitate a comparison to the LSE, precise lower bounds for the LSE in out-of-sample prediction.

ACKNOWLEDGEMENTS

We thank Reese Pathak for helpful feedback and suggesting the experiment in Figure 3, as well as Tomas Vaškevičius and Fanny Yang for insightful discussions. Tobias Wegel was partially supported by SNF Grant 204439, and conducted some of this work while he was visiting the Simons Institute for the Theory of Computing. Gil Kur conducted part of this work during his visit to the IDEAL Institute, hosted by Lev Reyzin and supported by NSF ECCS-2217023. Patrick Rebeschini and Tobias Wegel were partially funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number EP/Y028333/1].

References

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 1370–1378, 2019.

Liviu Aolaritei, Michael I Jordan, Reese Pathak, and Annie Ulichney. Revisiting mean estimation over ℓ_p balls: Is the MLE optimal? *arXiv preprint arXiv:2506.10354*, 2025.

Francis Bach. *Learning theory from first principles*. MIT press, 2024.

Keith Ball, Eric A Carlen, and Elliott H Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones mathematicae*, 115(1):463–482, 1994.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Amir Beck and Marc Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012.

Pierre C. Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression, 2016.

Pierre C. Bellec. Localized Gaussian width of M-convex hulls with applications to Lasso and convex aggregation, 2017.

Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150, 1993.

Tommy Bonnesen and Werner Fenchel. *Theorie der konvexen Körper*. Springer Berlin Heidelberg, 1934.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Peter Bühlmann and Bin Yu. Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

Peter Bühlmann and Sara A. van de Geer. *Statistics for High-Dimensional Data*. Springer Berlin Heidelberg, 2011.

Emmanuel J. Candes and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Sourav Chatterjee. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42(6), 2014.

Arnak S. Dalalyan, Mohamed Hebiri, and Johannes Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23(1):552–581, 2017.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 2016.

Konstantin Donhauser, Nicolò Ruggeri, Stefan Stojanovic, and Fanny Yang. Fast rates for noisy interpolation require rethinking the effect of inductive bias. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 5397–5428, 2022.

David L Donoho and Iain M Johnstone. Minimax risk over lp-balls for lp-error. *Probability theory and related fields*, 99(2):277–303, 1994.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least Angle Regression. *The Annals of Statistics*, 32(2):407–451, 2004.

Rina Foygel and Nathan Srebro. Fast-rate and optimistic-rate error bounds for L1-regularized regression. *arXiv preprint arXiv:1108.0373*, 2011.

Michael P. Friedlander, Ives Macêdo, and Ting Kei Pong. Polar convolution. *SIAM Journal on Optimization*, 29(2): 1366–1391, 2019.

Khashayar Gatmiry, Jon Schneider, and Stefanie Jegelka. Computing optimal regularizers for online linear optimization. *arXiv preprint arXiv:2410.17336*, 2024.

Claudio Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299, 2003.

Udaya Ghai, Elad Hazan, and Yoram Singer. Exponentiated Gradient Meets Gradient Descent. *Proceedings of Machine Learning Research*, 117:386–407, 2020.

Yehoram Gordon, Alexander E Litvak, Shahar Mendelson, and Alain Pajor. Gaussian averages of interpolated bodies and applications to approximate reconstruction. *Journal of Approximation Theory*, 149(1):59–73, 2007.

Antonio José Guirao, Vicente Montesinos, and Václav Zizler. Renormings in Banach Spaces. *Monografie Matematyczne*, 75, 2022.

Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit Regularization in Matrix Factorization. In *Advances in Neural Information Processing Systems*, volume 30, pages 6151–6159, 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing Implicit Bias in Terms of Optimization Geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1832–1841, 2018.

Suriya Gunasekar, Blake E. Woodworth, and Nathan Srebro. Mirrorless Mirror Descent: A Natural Derivation of Mirror Descent. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 2305–2313, 2021.

Petr Hájek and Michal Johanis. *Smooth analysis in Banach spaces*, volume 19. Walter de Gruyter GmbH & Co KG, 2014.

Qiyang Han. Set structured global empirical risk minimizers are rate optimal in general dimensions. *The Annals of Statistics*, 49(5):2642–2671, 2021.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York Inc., 2001.

Iain M. Johnstone. Gaussian estimation: Sequence and wavelet models. Available online, 2017.

Anatoli Juditsky and Arkadii S. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

Varun Kanade, Patrick Rebeschini, and Tomas Vaškevičius. The statistical complexity of early-stopped mirror descent. *Information and Inference: A Journal of the IMA*, 12(4): 3010–3041, 2023.

Gil Kur, Yuval Dagan, and Alexander Rakhlin. Optimality of maximum likelihood for log-concave density estimation and bounded convex regression. *arXiv preprint arXiv:1903.05315*, 2019.

Gil Kur, Alexander Rakhlin, and Adityanand Guntuboyina. On suboptimality of least squares with application to estimation of convex bodies. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 2406–2424, 09–12 Jul 2020.

Gil Kur, Eli Putterman, and Alexander Rakhlin. On the variance, admissibility, and stability of empirical risk minimization. *Advances in Neural Information Processing Systems*, 36:37527–37539, 2023.

Gil Kur, Pedro Abdalla, Pierre Bizeul, and Fanny Yang. Minimum Norm Interpolation Meets The Local Theory of Banach Spaces. In *Forty-first International Conference on Machine Learning*, 2024.

Tor Lattimore. Bandit convex optimisation. *arXiv preprint arXiv:2402.06535*, 2024.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method II: complexity dependent error rates. *Journal of Machine Learning Research*, 18(146): 1–48, 2017.

Daniel Levy and John C. Duchi. Necessary and Sufficient Geometries for Gradient Methods. In *Advances in Neural Information Processing Systems*, volume 32, pages 11495–11505, 2019.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic Regularization in Over-parameterized Matrix Sensing and Neural Networks with Quadratic Activations. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 2–47, 2018.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset Rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285. PMLR, 2015.

Cesare Molinari, Mathurin Massias, Lorenzo Rosasco, and Silvia Villa. Iterative regularization for convex regularizers. In *International conference on artificial intelligence and statistics*, 2021.

Cesare Molinari, Mathurin Massias, Lorenzo Rosasco, and Silvia Villa. Iterative regularization for low complexity regularizers. *Numerische Mathematik*, 156:641–689, 4 2024.

Jean-Jacques Moreau. Proximité et dualité dans un espace Hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

Arkadi S. Nemirovski and David B. Yudin. *Problem complexity and method efficiency in optimization*. Chichester: Wiley, 1984.

Matey Neykov. On the minimax rate of the Gaussian sequence model under bounded convex constraints. *IEEE Transactions on Information Theory*, 69(2):1244–1260, 2022.

Kazimierz Nikodem and Zsolt Pales. Characterizations of inner product spaces by strongly convex functions. *Banach Journal of Mathematical Analysis*, 5(1):83 – 87, 2011.

Francesco Orabona. A modern introduction to online learning, 2023.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

Reese Pathak and Cong Ma. On the design-dependent suboptimality of the Lasso. *arXiv preprint arXiv:2402.00382*, 2024.

Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

Chayne Planiden and Xianfu Wang. Proximal mappings and Moreau envelopes of single-variable convex piecewise cubic functions and multivariable gauge functions. *Nonsmooth optimization and its applications*, pages 89–130, 2019.

Akshay Prasadan and Matey Neykov. Some facts about the optimality of the LSE in the Gaussian sequence model with convex constraint. *arXiv preprint arXiv:2406.05911*, 2024.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Localization and adaptation in online learning. In *Artificial Intelligence and Statistics*, pages 516–526. PMLR, 2013.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax Rates of Estimation for High-Dimensional Linear Regression Over ℓ_q-Balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011a.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping for non-parametric regression: An optimal data-dependent stopping rule. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1318–1325, 2011b.

Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation, 2010.

R. Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a Regularized Path to a Maximum Margin Classifier. *Journal of Machine Learning Research*, 5:941–973, 2004.

Mark Schmidt, Glenn Fung, and Rómer Rosales. Fast Optimization Methods for L1 Regularization: A Comparative Study and Two New Approaches. *Lecture Notes in Computer Science*, page 286–297, 2007.

Rolf Schneider. *Convex bodies: the Brunn–Minkowski theory*, volume 151. Cambridge University Press, 2013.

Alexander Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998.

Carsten Schütt. Entropy numbers of diagonal operators between symmetric Banach spaces. *Journal of Approximation Theory*, 40(2):121–128, 1984.

Shai Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. Hebrew University, 2007.

Nati Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. *Advances in neural information processing systems*, 24, 2011.

Arun Suggala, Adarsh Prasad, and Pradeep K. Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Haoyuan Sun, Khashayar Gatmiry, Kwangjun Ahn, and Navid Azizan. A unified approach to controlling implicit regularization via mirror descent. *Journal of Machine Learning Research*, 24(393):1–58, 2023.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Sara A. van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Tomas Vaškevičius, Varun Kanade, and Patrick Rebeschini. Implicit Regularization for Optimal Sparse Recovery. In *Advances in Neural Information Processing Systems*, volume 32, pages 2972–2983, 2019.

Roman Vershynin. Estimation in high dimensions: a geometric perspective. In *Sampling Theory, a Renaissance: Compressive Sensing and Other Developments*, pages 3–66. Springer, 2015.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Yuting Wei, Fanny Yang, and Martin J. Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. In *Advances in Neural Information Processing Systems*, volume 30, pages 6065–6075, 2017.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

Fan Wu and Patrick Rebeschini. Nearly minimax-optimal rates for noisy sparse phase retrieval via early-stopped mirror descent. *Information and Inference: A Journal of the IMA*, 12(2):633–713, 2022.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On Early Stopping in Gradient Descent Learning. *Constructive Approximation*, 26:289–315, 08 2007.

Fei Ye and Cun-Hui Zhang. Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, 11(114):3519–3540, 2010.

Yao-Liang Yu. The strong convexity of von neumann's entropy. *unpublished note*, 2015.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4), 2005.

Yuchen Zhang, Martin J. Wainwright, and Michael I. Jordan. Optimal prediction for sparse linear models? Lower bounds for coordinate-separable M-estimators. *Electronic Journal of Statistics*, 11(1):752 – 799, 2017.

Peng Zhao, Yun Yang, and Qiao-Chu He. High-dimensional linear regression via implicit regularization. *Biometrika*, 109(4):1033–1046, 2022.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes (Section 1)
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes (Section 3)
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. No, the paper is mostly theoretical and the simulations are only small-scale, synthetic, and to support the intuition of the theoretical results.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes (Section 3)
   (b) Complete proofs of all theoretical results. Yes (Appendix B)
   (c) Clear explanations of any assumptions. Yes (Section 3)

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes (Appendix A)
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes (Appendix A)
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes (Appendix A)
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). No, the simulations are only small-scale and run within minutes on standard laptops.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable, we are not using or releasing existing assets.
   (b) The license information of the assets, if applicable. Not Applicable, we are not using or releasing existing assets.
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable, we are not using or releasing existing assets.
   (d) Information about consent from data providers/curators. Not Applicable, we are not using or releasing existing assets.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable, we are not using or releasing existing assets.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

Supplementary Material: Sharp Risk Bounds for Early-stopping in Gaussian Linear Regression

A SIMULATIONS

A.1 A Sanity Check for the Hard Design Matrix for ℓ_p-Constraints

In this section, we present a simple simulation using the LSE on the adversarial data matrix constructed in the proof of Theorem 3. We do this as a sanity check of our construction and the rate $d^{1/q}/\sqrt{n}$: Since we know that the LSE also enjoys the upper bound of order $d^{1/q}/\sqrt{n}$ from Proposition 2 (e.g., due to Bellec (2016)), and according to Theorem 3 on the adversarial matrix we have

$$\sup_{\alpha^\star \in B_p^d} \mathbb{E}_\xi \mathcal{R}(\widehat{\alpha}_{\mathrm{LSE}}) \gtrsim \frac{d^{1/q}}{\sqrt{n}},$$

there should be a ground-truth $\alpha^\star \in B_p^d$ attaining the supremum, such that $\mathbb{E}_\xi \mathcal{R}(\widehat{\alpha}_{\mathrm{LSE}}) \asymp d^{1/q}/\sqrt{n}$.

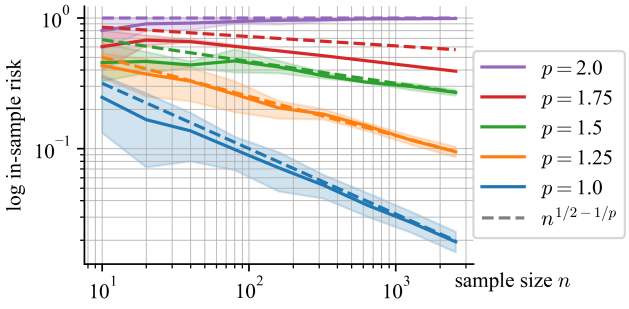


Figure 4: Simulation results for Theorem 3. For each p, we plot the log average in-sample risk over 30 experiments in solid, and the log 20th and 80th quantile as the shaded regions. The dashed lines show $(1/2 - 1/p)\log n$ for each p.

We do not prove it, but the following simulations suggest that—at least approximately—the LSE exhibits the rate $d^{1/q}/\sqrt{n}$ at the ground truth $\alpha^\star = 0$.[3] We let $d = n$ vary over $\{10 \cdot 2^i : i \in \{0, \dots, 8\}\}$, as well as p over $\{1, 1.25, 1.5, 1.75, 2\}$. We take $\mathbf{X}$ to be the data matrix from the proof of Theorem 3, as defined in Equation (26) of Appendix B.9, and $\alpha^\star = 0$. Thus, the rate from Theorem 3 is given by $\sim d^{1/q}/\sqrt{n} = n^{1-1/p-1/2} = n^{1/2-1/p}$.

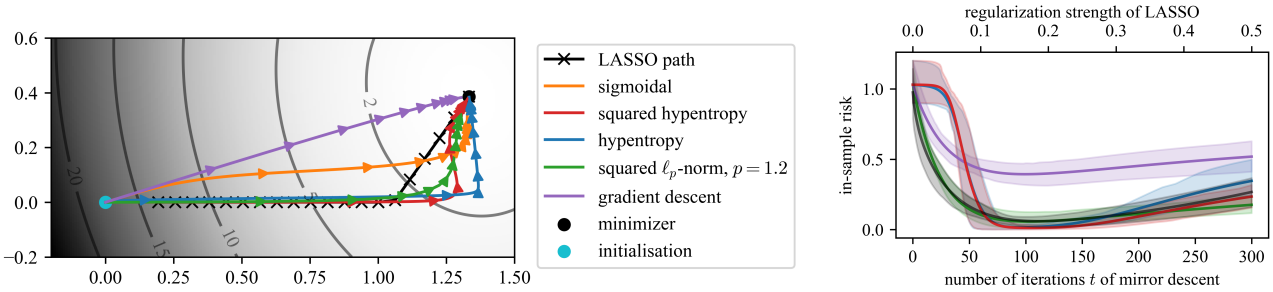
For each (n, p), we then repeat the following experiment 30 times. We sample $y = \mathbf{X}\alpha^\star + \xi = \xi \sim \mathcal{N}_n(0, \mathbf{I}_n)$ and compute the estimator $\widehat{\alpha}_{\mathrm{LSE}} \in \arg\min_{\alpha \in B_p^d} \widehat{\mathcal{R}}(\alpha)$ using the Python library `cvxpy` (Diamond and Boyd, 2016), which is designed for convex optimization.

Figure 4 shows the results. We plot the sample size n against the average in-sample prediction risk $\mathcal{R}(\widehat{\alpha}_{\mathrm{LSE}})$ in log-scale (solid lines), as well as the (logarithm of) the 20th and 80th quantiles (shaded regions) for every pair (n, p). Because $\mathcal{R}(\widehat{\alpha}_{\mathrm{LSE}}) \sim n^{1/2-1/p}$, we should see a linear dependence with slope $1/2 - 1/p$. Therefore, we also plot the corresponding lines (dashed lines). And indeed, comparing the risks achieved by the LSE to the rate $n^{1/2-1/p}$, we see that the behavior matches. This confirms our construction and Theorem 3 according to this sanity check.

A.2 Risk Along ℓ_1 Optimization Paths

We plot some example optimization paths of some potentials from Section 4.2 respectively Table 1 in Figure 5. The squared hypentropy "fixes" issues arising for previously used hypentropy from Ghai et al. (2020). In Figure 5(a) we plot paths on one data instance for a two-dimensional problem. Note that the paths can (somewhat) deviate from the LASSO path, but early-stopping still achieves minimax rates (Section 4.2). In Figure 5(b), we take $\mathbf{X}$ to be Gaussian for $n = d = 100$ and $\alpha^\star$ to be 1-sparse. We repeat the experiment 50 times and plot mean, 10th and 90th percentile.

[3] Note that the LSE achieving its worst-case risk at the origin is not obvious a priori (Aolaritei et al., 2025).



(a) Optimization and regularization paths.

(b) Risk along mirror descent paths.

Figure 5: Optimization paths of different known and new potentials from this work for regression over ℓ_1-balls.

A.3 Implementation Details

Finally, we just note that for all potentials, even those for which the (inverse) gradient does not have a closed-form solution, we can implement mirror descent using the the following expression which is equivalent to Definition 3;

$$\alpha_{t+1} \in \arg\min_{\alpha \in \mathbb{R}^d} \left(\left\langle \nabla \widehat{\mathcal{R}}(\alpha_t), \alpha - \alpha_t \right\rangle + \frac{1}{\eta} D_\psi(\alpha, \alpha_t) \right).$$

We implement all mirror descent experiments with this approach using `cvxpy` (Diamond and Boyd, 2016). In particular, Figures 1 to 3 and 5 are created using mirror descent updates that are implemented in the same way.

B PROOFS

B.1 Preliminaries

We first show this preliminary result that we use in multiple proofs.

Lemma 2. *For any convex body K that contains 0 in its interior, and any $\tau \geqslant 1$, it holds that*

$$r_0^2(\tau K) := \sup_{\alpha^\star \in \tau K} r_0^2(\alpha^\star, \tau K) \leqslant \tau \cdot \sup_{\alpha^\star \in K} r_0^2(\alpha^\star, K) =: \tau \cdot r_0^2(K). \tag{21}$$

Proof. Note that, by definition, $r_0(K)$ is bounded by any $r \geqslant 0$ that satisfies for all $\alpha^\star \in K$ (cf. (5))

$$w\left((K - \alpha^\star) \cap r B_2\right) \leqslant \frac{r^2}{2}.$$

Take any $\alpha^\star \in \tau K$ and define $R = \sqrt{\tau} \cdot r_0(K)$. We have that

$$\begin{aligned} w\left((\tau K - \alpha^\star) \cap R B_2\right) &= w\left(\tau\left((K - \alpha^\star/\tau) \cap (R/\tau) B_2\right)\right) \\ &= \tau \cdot w\left((K - \alpha^\star/\tau) \cap (R/\tau) B_2\right) \\ &= \tau \cdot w\left((K - \alpha^\star/\tau) \cap (r_0(K)/\sqrt{\tau}) B_2\right) \\ &\leqslant \tau \cdot w\left((K - \alpha^\star/\tau) \cap r_0(K) B_2\right) \\ &\leqslant \tau \cdot \frac{r_0^2(K)}{2} = \frac{R^2}{2}. \end{aligned}$$

Consequently, $r_0^2(\alpha^\star, \tau K) \leqslant R^2 = \tau r_0^2(K)$ for every $\alpha^\star \in \tau K$, and hence we have that $\sup_{\alpha^\star \in \tau K} r_0^2(\alpha^\star, \tau K) \leqslant \tau \cdot \sup_{\alpha^\star \in K} r_0^2(\alpha^\star, K)$, or in short, $r_0^2(\tau K) \leqslant \tau \cdot r_0^2(K)$. □

B.2 Proof of Lemma 1

Let K be any convex body with zero contained in its interior. It is known that for every $\delta > 0$, there exists a convex body K_δ such that $K \subset K_\delta \subset (1 + \delta) K$, and the corresponding Minkowski functional squared $\varphi_\delta^2 := \varphi_{K_\delta}^2$ is twice continuously

differentiable on $\mathbb{R}^d$ with $\nabla\varphi_\delta^2(0) = 0$, see for example Bonnesen and Fenchel (1934, Section 27), Schneider (2013), Guirao et al. (2022, Chapter 13) and Hájek and Johanis (2014, Chapter 7, §9). For this convex body, it clearly holds that $\frac{1}{1+\delta}\varphi_K \leqslant \varphi_\delta \leqslant \varphi_K$.

Furthermore, since φ_δ^2 is a convex and twice continuously differentiable function, the function ψ given by

$$\psi(\alpha) = \varphi_\delta^2(\alpha) + \frac{\rho}{2}\|\alpha\|_2^2$$

with $\rho = 2/(\max_{\alpha\in K}\|\alpha\|_2^2)$ satisfies

(I). ψ is twice continuously differentiable and the gradient satisfies $\nabla\psi(0) = 0$. Therefore, (I) is satisfied.

(II). The function $\alpha \mapsto \sqrt{\psi(\alpha)} = \sqrt{\varphi_\delta^2(\alpha) + \frac{\rho}{2}\|\alpha\|_2^2}$ is convex, as we can write it as

$$\sqrt{\psi(\alpha)} = \left\| \left[\varphi_\delta(\alpha), \sqrt{\frac{\rho}{2}}\|\alpha\|_2 \right]^\top \right\|_2$$

which is the composition of convex functions and a non-decreasing one, making $\sqrt{\psi}$ convex. Therefore, (II) is satisfied.

(III). ψ is ρ-strongly convex with respect to the ℓ_2-norm (III). This is due to the well-known fact (Nikodem and Pales, 2011) that, because $\|\cdot\|_2$ is induced by an inner product, ψ is ρ-strongly convex with respect to $\|\cdot\|_2$ if and only if $\psi - \frac{\rho}{2}\|\cdot\|_2^2 = \varphi_\delta^2$ is convex, which clearly holds since φ_δ is a Minkowski functional.

(IV). We have that for all $\alpha \in \mathbb{R}^d$ and $\alpha' \in \tau K$

$$\sqrt{\psi(\alpha)} = \sqrt{\varphi_\delta^2(\alpha) + \frac{\rho}{2}\|\alpha\|_2^2} \geqslant \varphi_\delta(\alpha) \geqslant \frac{1}{1+\delta}\varphi_K(\alpha) =: \frac{1}{c_l}\varphi_K(\alpha)$$
$$\sqrt{\psi(\alpha')} = \sqrt{\varphi_\delta^2(\alpha') + \frac{\rho}{2}\|\alpha'\|_2^2} \leqslant \sqrt{\varphi_K^2(\alpha') + \frac{1}{\max_{\alpha\in K}\|\alpha\|_2^2}\|\alpha'\|_2^2} \leqslant \sqrt{2}\tau =: c_u\tau.$$

Hence (IV) is satisfied.

For discrete time, we only need the potential to be differentiable (not twice differentiable), and hence we can also use the Moreau envelope of φ_K^2, defined as

$$\mathcal{M}_\lambda\varphi_K^2(\alpha) := \inf_{\alpha'}\left\{\varphi_K^2(\alpha') + \frac{1}{2\lambda}\|\alpha' - \alpha\|_2^2\right\}$$

and let $\psi(\alpha) = \mathcal{M}_\lambda\varphi_K^2(\alpha) + \frac{\rho}{2}\|\alpha\|_2^2$ with $\rho = 2/(\max_{\alpha\in K}\|\alpha\|_2^2)$. We can verify all necessary properties.

(I). ψ is continuously differentiable with gradient (Rockafellar and Wets, 2009, Theorem 2.26)

$$\nabla\psi(\alpha) = \nabla\mathcal{M}_\lambda\varphi_K^2(\alpha) + \rho\alpha = \frac{1}{\lambda}(\alpha - P_\lambda\varphi_K^2(\alpha)) + \rho\alpha \quad \text{with} \quad P_\lambda\varphi_K^2(\alpha) = \arg\min_{\alpha'}\left\{\varphi_K^2(\alpha') + \frac{1}{2\lambda}\|\alpha' - \alpha\|_2^2\right\},$$

so that $\nabla\psi(0) = 0$, cf. Planiden and Wang (2019, Theorem 5.6), and thus is satisfies (I).

(II). $\sqrt{\psi}$ is convex by Planiden and Wang (2019, Theorem 5.6) and the same argument as above, hence it satisfies (II).

(III). ψ is ρ-strongly convex with respect to the ℓ_2-norm (III) by the same argument as above, noting that the squared Moreau-envelope is convex (Planiden and Wang, 2019, Theorem 5.6).

(IV). To verify (IV), we use the following fact that $\sqrt{\mathcal{M}_\lambda \varphi_K^2} \leqslant \varphi_K$, as well as the pointwise convergence $\lim_{\lambda\to 0}\sqrt{\mathcal{M}_\lambda \varphi_K^2}(\alpha) = \varphi_K(\alpha)$ for all $\alpha \in \mathbb{R}^d$ (Planiden and Wang, 2019, Theorem 5.6). As $\sqrt{\mathcal{M}_\lambda \varphi_K^2}$ is continuous and $\left\{\alpha \in \mathbb{R}^d \,\middle|\, \varphi_K(\alpha) = 1\right\}$ is compact, by Dini's theorem this pointwise convergence implies uniform convergence, i.e., for the function $\zeta(\lambda) = \min_{\varphi_K(\alpha)=1} \sqrt{\mathcal{M}_\lambda \varphi_K^2}(\alpha) \leqslant 1$ that $\lim_{\lambda\to 0}\zeta(\lambda) = 1$. Therefore, there exists a λ_0 such that for all $\lambda \leqslant \lambda_0$ it holds $\zeta(\lambda) \geqslant 1/2$ and hence also $\sqrt{\mathcal{M}_\lambda \varphi_K^2(\alpha)} \geqslant \varphi_K(\alpha)/2$ for all α with $\varphi_K(\alpha) = 1$. Because $\sqrt{\mathcal{M}_\lambda \varphi_K^2}$ is another Minkowski functional (Planiden and Wang, 2019, Theorem 5.6) and hence positive homogeneous, the same holds on $\mathbb{R}^d$.

Therefore, for $\rho = 2/(\max_{\alpha\in K}\|\alpha\|_2^2)$ and all $\alpha \in \mathbb{R}^d$, $\alpha' \in \tau K$ it holds

$$\sqrt{\psi(\alpha)} = \sqrt{\mathcal{M}_\lambda \varphi_K^2(\alpha) + \frac{\rho}{2}\|\alpha\|_2^2} \geqslant \sqrt{\mathcal{M}_\lambda \varphi_K^2(\alpha)} \geqslant \frac{1}{2}\varphi_K(\alpha) = \frac{1}{c_l}\cdot\varphi_K(\alpha)$$
$$\sqrt{\psi(\alpha')} = \sqrt{\mathcal{M}_\lambda \varphi_K^2(\alpha') + \frac{\rho}{2}\|\alpha'\|_2^2} \leqslant \sqrt{\varphi_K^2(\alpha') + \frac{1}{\max_{\alpha\in K}\|\alpha\|_2^2}\|\alpha'\|_2^2} \leqslant \sqrt{2}\tau =: c_u\tau$$

This concludes the proof.

B.3 Proof of Theorem 1

We begin by proving this auxiliary statement: Let Assumption A hold. Then

$$B_\psi(\alpha^\star, 2D_\psi(\alpha^\star, 0)) \subset K_{3c_a\tau}. \tag{22}$$

The proof is based on the following two inequalities. We denote $f = \sqrt{\psi}$ so that $\psi(\alpha) = f^2(\alpha)$. Hence, for $f^2(\alpha) > 0$,

$$\begin{aligned}
2D_\psi(\alpha^\star, 0) &= 2\left(f^2(\alpha^\star) - f^2(0) - \left\langle \nabla f^2(0), \alpha^\star - 0\right\rangle\right) \\
&\leqslant 2f^2(\alpha^\star) && ((I): \nabla f^2(0) = 0) \\
D_\psi(\alpha^\star, \alpha) &= f^2(\alpha^\star) - f^2(\alpha) - \left\langle \nabla f^2(\alpha), \alpha^\star - \alpha\right\rangle \\
&= f^2(\alpha^\star) - f^2(\alpha) - 2f(\alpha)\left\langle \nabla f(\alpha), \alpha^\star - \alpha\right\rangle && \text{(chain rule)} \\
&\geqslant f^2(\alpha^\star) - f^2(\alpha) - 2f(\alpha)(f(\alpha^\star) - f(\alpha)) && ((II): \text{convexity of } f) \\
&= f^2(\alpha^\star) + f^2(\alpha) - 2f(\alpha)f(\alpha^\star) \\
&= (f(\alpha^\star) - f(\alpha))^2
\end{aligned}$$

Since for every $\alpha \in B_\psi(\alpha^\star, 2D_\psi(\alpha^\star, 0))$ we have by definition that $D_\psi(\alpha^\star, \alpha) \leqslant 2D_\psi(\alpha^\star, 0)$, the following inequality holds too;

$$(f(\alpha^\star) - f(\alpha))^2 \leqslant 2f^2(\alpha^\star) \implies f(\alpha) \leqslant (1+\sqrt{2})f(\alpha^\star) \leqslant 3f(\alpha^\star).$$

Therefore, by (IV) we have that

$$\varphi_K(\alpha) \leqslant c_l f(\alpha) \leqslant 3c_l f(\alpha^\star) \leqslant 3c_l c_u \tau = 3c_a\tau$$

and hence $B_\psi(\alpha^\star, 2D_\psi(\alpha^\star, 0)) \subset 3c_a K_\tau$, which finishes the proof of this first auxiliary part. Notice how we could improve the constant 3 arbitrarily close to 2 in the above argument.

Define the event

$$A_1 := \left\{\widehat{\mathcal{R}}(\alpha^\star) \leqslant 2\right\}.$$

This event holds with high probability, as $\widehat{\mathcal{R}}(\alpha^\star) = \|\xi\|_2^2/n$, and thus

$$\begin{aligned}
\mathbb{P}(A_1) &= \mathbb{P}\left(\|\xi\|_2^2/n \leqslant 2\right) = \mathbb{P}\left(\|\xi\|_2^2 \leqslant 2n\right) \\
&\geqslant 1 - \left(\frac{2n}{n}\exp\left(1 - \frac{2n}{n}\right)\right)^{n/2} \\
&= 1 - \exp\left(\left(\log(2) + 1 - 2\right)n/2\right) \\
&\geqslant 1 - \exp\left(-0.1n\right)
\end{aligned}$$

where we use Chernoff's bound for χ^2_n-distributed random variables in the first inequality. Conditioned on A_1, we have by Kanade et al. (2023, Theorems 2 and 3) that for

$$\eta \leqslant \frac{\rho}{\beta} \wedge \frac{D_\psi(\alpha^\star, 0)}{2}$$

(the latter of which implies $\eta\widehat{\mathcal{R}}(\alpha^\star) \leqslant D_\psi(\alpha^\star, 0)$) and any $\varepsilon > 0$ we may choose later, there exists a stopping time

$$t^\star \leqslant \left\lceil \frac{D_\psi(\alpha^\star, 0) + \eta\widehat{\mathcal{R}}(\alpha^\star)}{\eta\varepsilon} \right\rceil \leqslant \left\lceil \frac{2D_\psi(\alpha^\star, 0)}{\eta\varepsilon} \right\rceil \leqslant \left\lceil \frac{2c_u^2\tau^2}{\eta\varepsilon} \right\rceil = T \quad \text{or} \quad t^\star \leqslant \frac{D_\psi(\alpha^\star, 0)}{\varepsilon} \leqslant \frac{c_u^2\tau^2}{\varepsilon} = T$$

for the discrete and continuous-time cases respectively, such that the following two hold:

1. For all $t \leqslant t^\star$, it holds $\alpha_t \in B_\psi(\alpha^\star, D_\psi(\alpha^\star, 0) + \eta\widehat{\mathcal{R}}(\alpha^\star))$.
2. $\widehat{\mathcal{R}}(\alpha_{t^\star}) - \widehat{\mathcal{R}}(\alpha^\star) + \mathcal{R}(\alpha_{t^\star}) \leqslant \varepsilon$, which is called the offset condition with parameter $\varepsilon > 0$.

From the choice of η, on the event A_1, we get $\alpha_t \in B_\psi(\alpha^\star, 2D_\psi(\alpha^\star, 0))$. As we have shown in the auxiliary part (22), this implies $\alpha_t \in 3c_aK_\tau$ for all $t \leqslant t^\star$. We now combine this with the second part in a localization argument.

By rearranging the offset condition, we retrieve the inequality

$$\begin{aligned}
\mathcal{R}(\alpha_{t^\star}) &\leqslant \widehat{\mathcal{R}}(\alpha^\star) - \widehat{\mathcal{R}}(\alpha_{t^\star}) + \varepsilon \\
&= \frac{1}{n}\left\|y - \mathbf{X}\alpha^\star\right\|_2^2 - \frac{1}{n}\left\|y - \mathbf{X}\alpha_{t^\star}\right\|_2^2 + \varepsilon \\
&= \frac{1}{n}\left\|\xi\right\|_2^2 - \frac{1}{n}\left\|\mathbf{X}(\alpha^\star - \alpha_{t^\star})\right\|_2^2 - \frac{2}{n}\left\langle \xi, \mathbf{X}(\alpha^\star - \alpha_{t^\star})\right\rangle - \frac{1}{n}\left\|\xi\right\|_2^2 + \varepsilon \\
&= -\frac{1}{n}\left\|\mathbf{X}(\alpha^\star - \alpha_{t^\star})\right\|_2^2 - \frac{2}{n}\left\langle \xi, \mathbf{X}(\alpha^\star - \alpha_{t^\star})\right\rangle + \varepsilon, \\
&= \frac{2}{n}\left\langle \xi, \mathbf{X}(\alpha_{t^\star} - \alpha^\star)\right\rangle - \mathcal{R}(\alpha_{t^\star}) + \varepsilon. \qquad (23)
\end{aligned}$$

We *could* use this to get a unlocalized uniform bound, by bounding this as $\mathcal{R}(\alpha_{t^\star}) \leqslant \sup_{\alpha \in K_{3c_a\tau}} \frac{1}{n}\left\langle \xi, \mathbf{X}(\alpha^\star - \alpha)\right\rangle + \varepsilon/2$. However, we can improve upon this using a localization argument. We follow an argument akin to Bellec (2016, Theorem 2.3) where it was used for the LSE. Define the random variable

$$Z_r := \sup_{\theta \in (\mathbf{X}K_{3c_a\tau} - \mathbf{X}\alpha^\star) \cap rB_2^n} \left\langle \xi, \theta \right\rangle$$

which is the supremum of a Gaussian process indexed by $\theta \in (\mathbf{X}K_{3c_a\tau} - \mathbf{X}\alpha^\star) \cap rB_2^n$. By concentration of the supremum of Gaussian processes (Boucheron et al., 2013, Theorem 5.8), for any $r > 0$, the event

$$A_2 = \left\{ Z_r \leqslant \mathbb{E}\left[Z_r\right] + r\sqrt{2\log(1/\delta)} \right\}$$

has probability at least $1 - \delta$. Recall Definition 1, in particular, of the stationary radius

$$r_0(\mathbf{X}\alpha^\star, \mathbf{X}K_{3c_a\tau}) = \inf\left\{ r \geqslant 0 \;\middle|\; w\left((\mathbf{X}K_{3c_a\tau} - \mathbf{X}\alpha^\star) \cap rB_2^n\right) - \frac{r^2}{2} \leqslant 0 \right\}$$

and denote $r_0 = r_0(\mathbf{X}\alpha^\star, \mathbf{X}K_{3c_a\tau})$. By definition, we have that

$$\mathbb{E}\left[Z_{r_0}\right] = w\left((\mathbf{X}K_{3c_a\tau} - \mathbf{X}\alpha^\star) \cap r_0B_2^n\right) \leqslant \frac{r_0^2}{2}$$

and hence, conditioned on A_2 with $r = r_0$, we have

$$Z_{r_0} \leqslant \frac{r_0^2}{2} + r_0\sqrt{2\log(1/\delta)}.$$

We now make a case distinction between $\mathcal{R}(\alpha_{t^\star}) \leqslant r_0^2/n$ and $\mathcal{R}(\alpha_{t^\star}) > r_0^2/n$ conditioned on A_2. In the first case, we get trivially

$$\mathcal{R}(\alpha_{t^\star}) \leqslant \frac{r_0^2}{n} \leqslant \frac{1}{n}\left(r_0 + \sqrt{2\log(1/\delta)}\right)^2 + \varepsilon.$$

In the second case when $\mathcal{R}(\alpha_{t^\star}) = \frac{1}{n}\|\mathbf{X}(\alpha_{t^\star} - \alpha^\star)\|_2^2 > r_0^2/n$, we define $\lambda = r_0/\|\mathbf{X}(\alpha_{t^\star} - \alpha^\star)\|_2 \in (0,1)$ and $v = \lambda\mathbf{X}\alpha_{t^\star} + (1-\lambda)\mathbf{X}\alpha^\star \in \mathbf{X}K_{3c_a\tau}$. Note $(v - \mathbf{X}\alpha^\star)/\lambda = \mathbf{X}\alpha_{t^\star} - \mathbf{X}\alpha^\star$ as well as $\mathcal{R}(\alpha_{t^\star}) = \frac{1}{n}\|\mathbf{X}(\alpha^\star - \alpha_{t^\star})\|_2^2 = r_0^2/(n\lambda^2)$, and hence bounding Equation (23) yields with probability at least $1-\delta$

$$\begin{aligned}
\frac{2}{n}\langle \xi, \mathbf{X}(\alpha_{t^\star} - \alpha^\star)\rangle - \mathcal{R}(\alpha_{t^\star}) + \varepsilon &= \frac{1}{n}\left(\frac{2}{\lambda}\langle \xi, v - \mathbf{X}\alpha^\star\rangle - \frac{r_0^2}{\lambda^2}\right) + \varepsilon \\
&\leqslant \frac{1}{n}\left(\frac{2}{\lambda}Z_{r_0} - \frac{r_0^2}{\lambda^2}\right) + \varepsilon \\
&= \frac{1}{n}\left(\frac{2r_0}{\lambda}\frac{Z_{r_0}}{r_0} - \frac{r_0^2}{\lambda^2}\right) + \varepsilon \\
&\leqslant \frac{1}{n}\left(\frac{Z_{r_0}}{r_0}\right)^2 + \varepsilon \qquad \text{(since } 2ab - b^2 \leqslant a^2\text{)} \\
&\leqslant \frac{1}{n}\left(\frac{r_0}{2} + \sqrt{2\log(1/\delta)}\right)^2 + \varepsilon.
\end{aligned}$$

Hence, either way, we have that

$$\mathcal{R}(\alpha_{t^\star}) = \frac{1}{n}\|\mathbf{X}(\alpha_{t^\star} - \alpha^\star)\|_2^2 \leqslant \frac{1}{n}\left(r_0 + \sqrt{2\log(1/\delta)}\right)^2 + \varepsilon \leqslant \frac{2r_0^2 + 4\log(1/\delta)}{n} + \varepsilon.$$

Taking the union bound on A_1 and A_2, we get that with probability at least $1 - \exp(-0.1n) - \delta$, it holds

$$\mathcal{R}(\alpha_{t^\star}) \leqslant \frac{2r_0^2}{n} + \frac{4\log(1/\delta)}{n} + \varepsilon$$

where $r_0 = r_0(\mathbf{X}\alpha^\star, \mathbf{X}K_{3c_a\tau})$. This concludes the proof of the high probability bound.

The in-expectation bound for continuous time mirror descent follows by observing that we do not require the event A_1 to hold, and so the high probability bound holds with probability $1-\delta$ (in fact, one could improve its constants, which we neglect here for simplicity). In particular, denoting $z = \frac{2r_0^2}{n} + \varepsilon$ we have that

$$\mathcal{R}(\alpha_{t^\star}) \leqslant z + \frac{4\log(1/\delta)}{n}.$$

From tail integration, we then obtain

$$\begin{aligned}
\mathbb{E}_\xi\left[\mathcal{R}(\alpha_{t^\star})\right] &\leqslant \int_0^\infty \mathbb{P}\left(\mathcal{R}(\alpha_{t^\star}) > x\right) dx \\
&= \int_0^z \mathbb{P}\left(\mathcal{R}(\alpha_{t^\star}) > x\right) dx + \int_z^\infty \mathbb{P}\left(\mathcal{R}(\alpha_{t^\star}) > x\right) dx \\
&\leqslant z + \int_z^\infty \exp\left(-\frac{n}{4}(x - z)\right) dx \\
&= z + \frac{4}{n},
\end{aligned}$$

concluding the in-expectation bound and hence the proof.

B.4 Proof of Remark 1

An argument akin to the proof of Theorem 7 in Bellec (2017) shows that if we let $r = 2\sqrt{\mathrm{rk}(\mathbf{X})}$ and we denote the orthogonal projection onto the column space of $\mathbf{X}$ as $\Pi_{\mathbf{X}}$, we have $\mathbb{E}\left[\|\Pi_{\mathbf{X}}\xi\|_2\right] \leqslant \sqrt{\mathrm{rk}(\mathbf{X})}$, and hence for every $\alpha^\star \in K$ we have

$$w((\mathbf{X}K - \mathbf{X}\alpha^\star) \cap rB_2^d) = \mathbb{E}\left[\sup_{\alpha \in (K-\alpha^\star), \|\mathbf{X}\alpha\|_2 \leqslant r} \langle \xi, \mathbf{X}\alpha\rangle\right] \leqslant r\mathbb{E}\left[\|\Pi_{\mathbf{X}}\xi\|_2\right] \leqslant r\sqrt{\mathrm{rk}(\mathbf{X})} = \frac{r^2}{2}$$

implying by (5) that $r_0^2(\mathbf{X}K) \leqslant 4\,\mathrm{rk}\,(\mathbf{X})$.

B.5 Proof of Corollary 1

We begin by quantifying the constant C in Chatterjee (2014, Corollary 1.2): Temporarily denote $r_\star := r_\star(\mathbf{X}\alpha^\star, \mathbf{X}K_\tau)$. By Theorem 1.1 in Chatterjee (2014), we know that

$$\mathbb{P}\left(\left|\|\mathbf{X}\left(\widehat{\alpha}_{\mathrm{LSE}} - \alpha^\star\right)\|_2 - r_\star\right| \geqslant x\sqrt{r_\star}\right) \leqslant 3\exp\left(-\frac{x^4}{32(1+\frac{x}{\sqrt{r_\star}})^2}\right) \leqslant 3\exp\left(-\frac{x^4}{32(1+x)^2}\right),$$

where the second inequality is true if $r_\star \geqslant 1$. From tail integration we get

$$\begin{aligned}
\mathbb{E}_\xi\left(\|\mathbf{X}(\widehat{\alpha}_{\mathrm{LSE}} - \alpha^\star)\|_2 - r_\star\right)^2 &= \int_0^\infty \mathbb{P}\left(\left(\|\mathbf{X}(\widehat{\alpha}_{\mathrm{LSE}} - \alpha^\star)\|_2 - r_\star\right)^2 \geqslant u\right) du \\
&= 2r_\star \int_0^\infty x\mathbb{P}\left(\left|\|\mathbf{X}\left(\widehat{\alpha}_{\mathrm{LSE}} - \alpha^\star\right)\|_2 - r_\star\right| \geqslant x\sqrt{r_\star}\right) dx \\
&\leqslant 6r_\star \int_0^\infty x\exp\left(-\frac{x^4}{32(1+x)^2}\right) dx \\
&\leqslant 125 r_\star
\end{aligned}$$

and similarly, without the square,

$$\mathbb{E}_\xi\left|\|\mathbf{X}(\widehat{\alpha}_{\mathrm{LSE}} - \alpha^\star)\|_2 - r_\star\right| \leqslant 3\sqrt{r_\star}\int_0^\infty \exp\left(-\frac{x^4}{32(1+x)^2}\right) dx \leqslant 18\sqrt{r_\star}.$$

Combining the two, we get

$$\begin{aligned}
\left|\mathbb{E}_\xi\,\|\mathbf{X}(\widehat{\alpha}_{\mathrm{LSE}} - \alpha^\star)\|_2^2 - r_\star^2\right| &= \left|2r_\star\mathbb{E}_\xi(\|\mathbf{X}(\widehat{\alpha}_{\mathrm{LSE}} - \alpha^\star)\|_2 - r_\star) + \mathbb{E}_\xi(\|\mathbf{X}(\widehat{\alpha}_{\mathrm{LSE}} - \alpha^\star)\|_2 - r_\star)^2\right| \\
&\leqslant 2\cdot 18\cdot r_\star^{3/2} + 125 r_\star \leqslant 161 r_\star^{3/2}.
\end{aligned}$$

Consequently, for every $\alpha^\star \in K_\tau$ with $r_\star(\mathbf{X}\alpha^\star, \mathbf{X}K_\tau) \geqslant (644/3)^2$, it holds $\mathbb{E}_\xi\mathcal{R}(\widehat{\alpha}_{\mathrm{LSE}}) \geqslant \frac{1}{4}\frac{r_\star^2}{n}$, and in particular, if $r_\star(\mathbf{X}K_\tau) \geqslant (644/3)^2$, then

$$\sup_{\alpha^\star\in K_\tau} \mathbb{E}_\xi\mathcal{R}(\widehat{\alpha}_{\mathrm{LSE}}) \geqslant \frac{r_\star^2(\mathbf{X}K_\tau)}{4n}. \tag{24}$$

Moreover, from Theorem 1 we know that for $r_0(\mathbf{X}\alpha^\star, \mathbf{X}K_{3c_a\tau}) \geqslant 1$ and $\varepsilon = r_0(\mathbf{X}\alpha^\star, \mathbf{X}K_{3c_a\tau})/n$, in continuous time,

$$\mathbb{E}_\xi\mathcal{R}(\alpha_{t^\star}) \leqslant \frac{7r_0^2(\mathbf{X}\alpha^\star, \mathbf{X}K_{3c_a\tau})}{n}.$$

Hence, combining the two equations and using Equation (21) yields

$$\begin{aligned}
\sup_{\alpha^\star\in K_\tau} \mathbb{E}_\xi\mathcal{R}(\alpha_{t^\star}) &\leqslant \sup_{\alpha^\star\in K_\tau} \frac{7r_0^2(\mathbf{X}\alpha^\star, \mathbf{X}K_{3c_a\tau})}{n} \\
&\leqslant 21 c_a \frac{r_0^2(\mathbf{X}K_\tau)}{n} && \text{(from (21))} \\
&\leqslant 21\mathcal{C}c_a \cdot \frac{r_\star^2(\mathbf{X}K_\tau)}{n} && \text{(by Assumption B)} \\
&\leqslant 84\mathcal{C}c_a \cdot \sup_{\alpha^\star\in K_\tau} \mathbb{E}_\xi\mathcal{R}(\widehat{\alpha}_{\mathrm{LSE}}), && \text{(from (24))}
\end{aligned}$$

which concludes the proof.

B.6 Proof of Corollary 2

The proof of Corollary 2 is analgous to the proof of Prasadan and Neykov (2024, Corollary 2.6), where we can replace the upper bound on the LSE from their Proposition 2.4 with our bound on continuous-time ESMD from Theorem 1 and use Equation (21).

Specifically, from Theorem 1, applying Equation (21) and $c_a \lesssim 1$, we know that

$$\sup_{\alpha^\star \in K_\tau} \mathbb{E}_\xi \mathcal{R}(\alpha_{t^\star}) \lesssim \frac{r_0^2(\mathbf{X}K_{3c_a\tau})}{n} \lesssim \frac{r_0^2(\mathbf{X}K_\tau)}{n}.$$

Also note that in the proof of Theorem 1 (Appendix B.3), we showed that $\alpha_{t^\star} \in 3c_aK_\tau$, and so $\mathbb{E}_\xi \mathcal{R}(\alpha_{t^\star}) \lesssim (\operatorname{diam}(\mathbf{X}K_\tau))^2/n$.

Denote now temporarily

$$\overline{\mathcal{R}} := \sup_{\alpha^\star \in K_\tau} \mathbb{E}_\xi \mathcal{R}(\alpha_{t^\star}).$$

Now, for every $\alpha^\star \in K_\tau$, $r \mapsto \sup_{\alpha^\star \in K_\tau} w(\mathbf{X}(K_\tau - \alpha^\star) \cap rB_2^n)/r$ is non-increasing (Prasadan and Neykov, 2024, page 7). Denoting $r_0 = r_0(\mathbf{X}K_\tau)$, we get that for some constant $c_1 > 0$, and some $r < r_0$

$$\begin{aligned}
c_1\sqrt{n\overline{\mathcal{R}}} \leqslant r &\leqslant 2 \sup_{\alpha^\star \in K_\tau} \frac{w(\mathbf{X}(K_\tau - \alpha^\star) \cap rB_2^n)}{r} && \text{(by Definition 1 and } r < r_0\text{)} \\
&\leqslant 2\frac{\sup_{\alpha^\star \in K_\tau} w(\mathbf{X}(K_\tau - \alpha^\star) \cap c_1\sqrt{n\overline{\mathcal{R}}}B_2^n)}{c_1\sqrt{n\overline{\mathcal{R}}}} && \text{(non-increasing)} \\
&\leqslant 2\max\left\{1, \frac{1}{c_1}\right\} \frac{\sup_{\alpha^\star \in K_\tau} w(\mathbf{X}(K_\tau - \alpha^\star) \cap \sqrt{n\overline{\mathcal{R}}}B_2^n)}{\sqrt{n\overline{\mathcal{R}}}} \\
&\lesssim \sqrt{\log M^{\mathrm{loc}}\left(\sqrt{n\overline{\mathcal{R}}}, \mathbf{X}K_\tau\right)} && \text{(by Assumption C)}
\end{aligned}$$

Hence, $\sqrt{n\overline{\mathcal{R}}} \leqslant \sup\left\{r > 0 : r \lesssim \sqrt{\log M^{\mathrm{loc}}(r, \mathbf{X}K_\tau)}\right\}$, which by the main result of Neykov (2022) yields minimax optimality of ESMD.

B.7 Proof of Corollary 3

Consider the setting of Theorem 1, and assume that $\mathbf{X}$ has vanishing kernel width satisfying

$$\frac{1}{n}\|\mathbf{X}\alpha\|_2^2 \geqslant \|\alpha\|_2^2 - f(3c_a\tau K, n) \quad \text{for all } \alpha \in 3c_a\tau(K - K).$$

Per assumption, we have that $f(K_{3c_a\tau}, n) \lesssim r_0^2(\mathbf{X}K_{3c_a\tau})/n$, which combined with the bound from Theorem 1 and the fact that $\alpha_{t^\star} \in 3c_a\tau K$ (see the proof of Theorem 1) yields for all $\alpha^\star \in K_\tau$

$$\|\alpha^\star - \alpha_{t^\star}\|_2^2 \leqslant \frac{1}{n}\|\mathbf{X}(\alpha_{t^\star} - \alpha^\star)\|_2^2 + f(3c_a\tau K, n) \lesssim \frac{r_0^2}{n} + \frac{\log(1/\delta)}{n} + \frac{r_0^2}{n} \asymp \frac{r_0^2}{n} + \frac{\log(1/\delta)}{n},$$

where we denoted $r_0 = r_0(\mathbf{X}K_{3c_a\tau})$. Applying Equation (21) concludes the proof of the corollary.

B.8 Proof of Proposition 2

Let $p \in (1, 2)$ and $\varphi_K = \|\cdot\|_p$. We first show that $\psi = \|\cdot\|_p^2$ satisfies the discrete-time version of Assumption A.

(I). First, $\psi(\cdot) = \|\cdot\|_p^2$ with $p \in (1, 2)$ is differentiable (Lemma 17 Shalev-Shwartz, 2007; Juditsky and Nemirovski, 2008, Example 3.2) with

$$\nabla\psi(\alpha) = \begin{cases} 2\frac{\operatorname{sign}(\alpha_i)|\alpha_i|^{p-1}}{\|\alpha\|_p^{p-2}} & \alpha \neq 0 \\ 0 & \alpha = 0 \end{cases} \quad \text{and} \quad (\nabla\psi)^{-1}(\alpha) = \begin{cases} \frac{\operatorname{sign}(\alpha_i)|\alpha_i|^{q-1}}{2\|\alpha\|_q^{q-2}} & \alpha \neq 0 \\ 0 & \alpha = 0 \end{cases} \quad \text{with} \quad \frac{1}{p} + \frac{1}{q} = 1,$$

see for example Gentile (2003, Lemma 1). Hence, $\nabla\psi(0) = 0$.

(II). $\sqrt{\psi(\cdot)} = \|\cdot\|_p$ is convex, as it is a norm.

(III). ψ is $2(p-1)$-strongly convex with respect to the ℓ_p-norm by Shalev-Shwartz (2007, Lemma 17).

(IV) This clearly holds with $c_u = c_l = 1$.

We bound $r_0(\mathbf{X}K_{3\tau}) = r_0(3\tau\mathbf{X}B_p^d)$ under column-normalized design (16). We can use $r_0^2(3\tau\mathbf{X}B_p^d) \leqslant 6\tau w(\mathbf{X}B_p^d)$ and then further bound $w(\mathbf{X}B_p^d)$. We can apply Hölder's inequality with $1/p + 1/q = 1$ and get, if $p > 1 + 1/\log d$

$$w(\mathbf{X}B_p^d) = \mathbb{E}\left[\sup_{\alpha\in B_p^d}\langle\xi, \mathbf{X}\alpha\rangle\right] \leqslant \mathbb{E}\left[\sup_{\alpha\in B_p^d}\left\|\mathbf{X}^\top\xi\right\|_q\|\alpha\|_p\right] \leqslant \mathbb{E}\left[\left\|\mathbf{X}^\top\xi\right\|_q\right] \leqslant \left(\mathbb{E}\left[\left\|\mathbf{X}^\top\xi\right\|_q^q\right]\right)^{1/q}.$$

By column normalization, and since $|\langle\xi, \mathbf{X}_j\rangle| \overset{d}{=} \|\mathbf{X}_j\|_2 |Z|$ for $Z \sim \mathcal{N}(0,1)$, we get that

$$\left(\mathbb{E}\left[\left\|\mathbf{X}^\top\xi\right\|_q^q\right]\right)^{1/q} = \left(\mathbb{E}|Z|^q\right)^{1/q}\left(\sum_{j=1}^d\|\mathbf{X}_j\|_2^q\right)^{1/q} \asymp \sqrt{q}\sqrt{n}d^{1/q}.$$

On the other hand, if $p \leqslant 1 + 1/\log d$ and $d \geqslant 2$, we have that

$$w(\mathbf{X}B_p^d) = \mathbb{E}\left[\sup_{\alpha\in B_p^d}\langle\xi, \mathbf{X}\alpha\rangle\right] \leqslant \mathbb{E}\left[\sup_{\alpha\in B_p^d}\left\|\mathbf{X}^\top\xi\right\|_q\|\alpha\|_p\right] \leqslant d^{1/q}\mathbb{E}\left[\left\|\mathbf{X}^\top\xi\right\|_\infty\right] \leqslant d^{1/q}2\sqrt{n\log d} \leqslant 2e\sqrt{n\log d}.$$

where in the second last inequality we used that since $\mathbf{X}^\top\xi \sim \mathcal{N}(0, \mathbf{X}^\top\mathbf{X})$, and $\mathbf{X}$ is column-normalized, we know that $(\mathbf{X}^\top\xi)_i, i \in [d]$ is Gaussian with variance n, so that we may apply standard sub-Gaussian concentration, such as Vershynin (2018, Exercise 2.5.10) or Wainwright (2019, Exercise 2.12), and in the last inequality we used that for $p \leqslant 1 + 1/\log d$ it holds that $d^{1/q} = d^{1-1/p} \leqslant d^{1/\log d} \leqslant e$. Hence,

$$r_0^2(3\tau\mathbf{X}B_p^d) \lesssim \begin{cases}\tau\sqrt{n\log d} & \text{if } p \leqslant 1 + 1/\log d\\ \tau\sqrt{q}\sqrt{n}d^{1/q} & \text{if } p > 1 + 1/\log d\end{cases} \tag{25}$$

and from Theorem 1 and Remark 1 we get that

$$\begin{aligned}\mathcal{R}(\alpha_{t^\star}) &\leqslant \frac{4r_0^2(\mathbf{X}K_{3\tau})}{n} + \frac{8\log(1/\delta)}{n}\\ &\lesssim \frac{\mathrm{rk}\,(\mathbf{X})}{n} \wedge \frac{\tau}{\sqrt{n}}\begin{cases}\sqrt{\log d} & \text{if } 1 < p \leqslant 1 + \frac{1}{\log d},\\ \sqrt{q}d^{1/q} & \text{if } 1 + \frac{1}{\log d} < p < 2.\end{cases}\end{aligned}$$

To bound $r_0(3\tau\mathbf{X}B_p^d)$ in under Gaussian design, we use Vershynin (2018, Exercise 7.5.4) and get

$$r_0^2(\mathbf{X}B_p^d) \leqslant 2w(\mathbf{X}B_p^d) \leqslant 2\|\mathbf{X}\|_2 w(B_p^d).$$

By Wainwright (2019, Theorem 6.1) we have that

$$\mathbb{P}\left(\|\mathbf{X}\|_2 \leqslant 3\sqrt{n}\right) \geqslant 1 - \exp\left(-n/2\right)$$

and for $1/q + 1/p = 1$ it is easily verified that, because for small enough p we have $B_1^d \subset B_p^d \subset eB_1^d$ (Lecué and Mendelson, 2017; Vershynin, 2018), the Gaussian width is bounded as

$$w(B_p^d) \lesssim \begin{cases}\sqrt{\log d} & \text{if } 1 < p \leqslant 1 + \frac{1}{\log d},\\ \sqrt{q}d^{1/q} & \text{if } 1 + \frac{1}{\log d} < p < 2,\end{cases}$$

so that with probability $1 - \exp(-n/2)$ over draws of $\mathbf{X}$ the bound Equation (25) holds again. This concludes the proof.

B.9 Proof of Theorem 3

Let $p \in (1,2)$ and $1/p + 1/q = 1$. We introduce some further notation just for this proof. We use $N(\varepsilon, S)$, $M(\varepsilon, S)$ to denote the ε-covering number and, respectively, the ε-packing number of a set S with respect to the ℓ_2-norm (Wainwright, 2019, Definitions 5.1 and 5.4). $\log N(\varepsilon, S)$ is commonly referred to as the metric entropy of S. We assume basic knowledge about covering and packing numbers that can be found, e.g., in Chapter 5 in Wainwright (2019).

Fixed design. We will construct a hard design matrix $\mathbf{X}$ and reducing the problem to a Gaussian sequence model over a convex constraint set in $\mathbb{R}^n$.

Let $m \in \mathbb{N}$ and $k \in \mathbb{N}$ be defined as

$$m = \left\lfloor \frac{d^{1/p}}{\sqrt{n}} \right\rfloor \asymp \frac{d^{1/p}}{\sqrt{n}} \qquad \text{and} \qquad k = \left\lfloor \sqrt{n} d^{1/q} \right\rfloor \asymp \sqrt{n} d^{1/q}.$$

where we may write "$\asymp$" in the first case, because we assume that $d^{1/p} \geqslant \sqrt{n}$. We note that:

1. By definition of k and m, it holds that $k \cdot m \leqslant d^{1/p} d^{1/q} = d$.
2. By definition of k, it holds that $k \leqslant n$ if $\sqrt{n} \geqslant d^{1/q}$, which we assumed.

Define $\mathbf{1}_m = (1, \dots, 1)^\top \in \mathbb{R}^m$ as well as the all-zeros matrix $\mathbf{0}_{n \times d} \in \mathbb{R}^{n \times d}$. We define the data matrix

$$\mathbf{X} = \sqrt{n} \begin{pmatrix} \mathbf{A}_{k \times km} & \mathbf{0}_{k \times (d-km)} \\ \mathbf{0}_{(n-k) \times km} & \mathbf{0}_{(n-k) \times (d-km)} \end{pmatrix} \in \mathbb{R}^{n \times d} \quad \text{with} \quad \mathbf{A}_{k \times km} = \begin{pmatrix} \mathbf{1}_m^\top & \mathbf{0}_{1 \times m} & \dots & \mathbf{0}_{1 \times m} \\ \mathbf{0}_{1 \times m} & \mathbf{1}_m^\top & \dots & \mathbf{0}_{1 \times m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{1 \times m} & \mathbf{0}_{1 \times m} & \dots & \mathbf{1}_m^\top \end{pmatrix} \in \mathbb{R}^{k \times km}. \tag{26}$$

Note that the columns of $\mathbf{X}$ are normalised to $\|\mathbf{X}_i\|_2 \leqslant \sqrt{n}$.

As a first step, we reduce our problem to a more well-studied Gaussian sequence model. Define $t := \sqrt{n} m^{1/q}$ and $x_i, v_i \in \mathbb{R}^d, i \in [k]$ as

$$x_i = \sqrt{n} (\mathbf{0}_{1 \times m(i-1)}, \mathbf{1}_m^\top, \mathbf{0}_{1 \times (d-mi)})^\top \quad \text{and} \quad v_i = \frac{m^{-1/p}}{\sqrt{n}} x_i,$$

where x_i is the vector corresponding to the i-th row of $\mathbf{X}$. Note that $\|v_i\|_p = \left\| \frac{m^{-1/p}}{\sqrt{n}} x_i \right\|_p = m^{-1/p} m^{1/p} = 1$ and thus $v_i \in B_p^d$. Denoting the i-th standard-basis vector in $\mathbb{R}^n$ as $\mathbf{e}_i$, we have that

$$\mathbf{X} v_i = \frac{m^{-1/p}}{\sqrt{n}} \mathbf{X} x_i = \sqrt{n} m^{1-1/p} \mathbf{e}_i = \sqrt{n} m^{1/q} \mathbf{e}_i = t \mathbf{e}_i$$

for all $i \in [k]$. Taking the p-convex hull of the v_i amounts to choosing any $\delta_i, i \in [k]$ such that $\sum_{i=1}^k |\delta_i|^p \leqslant 1$ and letting $v_\delta = \sum_{i=1}^k \delta_i v_i$. It follows that $v_\delta \in B_p^d$ and

$$\mathbf{X} v_\delta = \sum_{i=1}^k \delta_i \mathbf{X} v_i = t \sum_{i=1}^k \delta_i \mathbf{e}_i.$$

Because $B_p^k \times \{0\}^{n-k}$ can be decomposed into $\sum_{i=1}^k \delta_i \mathbf{e}_i$, we have that

$$S_t := t B_p^k \times \{0\}^{n-k} \subset \mathbf{X} B_p^d \subset \mathbb{R}^n.$$

Therefore, since $y = \mathbf{X}\alpha^\star + \xi$ with $\xi \sim \mathcal{N}(0, \mathbf{I}_n)$ and $S_t \subset \mathbf{X} B_p^d$, the minimax rate for estimation in the Gaussian sequence model $y = \theta^\star + \xi$, with $\theta^\star \in S_t$ is a lower bound for our original problem, that is,

$$\inf_{\widehat{\alpha}} \sup_{\alpha^\star \in B_p^d} \mathbb{E}\left[\|\mathbf{X}(\alpha^\star - \widehat{\alpha})\|_2^2 \right] \geqslant \inf_{\widehat{\theta}} \sup_{\theta^\star \in S_t} \mathbb{E}\left[\|\theta^\star - \widehat{\theta}\|_2^2 \right]. \tag{27}$$

We may now use the following result from Johnstone (2017, Theorem 11.7), see also Aolaritei et al. (2025, page 6): Let $p \in (1,2)$ and consider the Gaussian sequence model $y = \theta^\star + \xi$ where $\xi \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$. Then, if $\sigma^2 \in [\frac{1}{d^{2/p}}, \frac{1}{1+\log d}]$, the minimax rate over B_p^d is given by

$$\mathfrak{M}(B_p^d, \sigma) := \inf_{\widehat{\theta}} \sup_{\theta^\star \in B_p^d} \mathbb{E}_\xi \left[\|\widehat{\theta}(y) - \theta^\star\|_2^2 \right] \asymp (\sigma^2 \log(ed\sigma^p))^{1-p/2}.$$

To apply this result to Equation (27), we use the following identity

$$\mathfrak{M}(\tau B_p^d, \sigma) = \tau^2 \cdot \mathfrak{M}\left(B_p^d, \frac{\sigma}{\tau}\right),$$

which follows from simple rescaling arguments. Moreover, we rely on the fact that if a convex body K (here S_t) is contained in a linear subspace, then the minimax risk of the Gaussian sequence model over $K \subset \mathbb{R}^d$ coincides with that of the model restricted to the subspace, since the orthogonal projection onto the subspace constitutes a sufficient statistic for the parameter. In particular, we can bound Equation (27) as

$$\inf_{\widehat{\theta}} \sup_{\theta^\star \in S_t} \mathbb{E}\left[\|\theta^\star - \widehat{\theta}\|_2^2 \right] \geqslant \mathfrak{M}(tB_p^k, 1) = t^2 \cdot \mathfrak{M}(B_p^k, 1/t) = t^p \left(\log\left(ek/t^p\right)\right)^{1-p/2}$$

where the last equality holds because

$$t^p = \left(\sqrt{n} m^{1/q}\right)^p \asymp_p \left(\sqrt{n}\left(\frac{d^{1/p}}{\sqrt{n}}\right)^{1/q}\right)^p = (\sqrt{n})^{p-p/q} d^{1/q} = \sqrt{n} d^{1/q} = k \quad \Longrightarrow \quad \left(\log\left(ek/t^p\right)\right)^{1-p/2} \asymp_p 1,$$

and in particular, $1/t^2 \asymp_p 1/k^{2/p} \in [\frac{1}{k^{2/p}}, \frac{1}{1+\log k}]$. Plugging this into the lower bound and dividing by n proves that the minimax rate on this data matrix is lower bounded by $c_p d^{1/q}/\sqrt{n}$, which concludes the first part.

Gaussian design. Let $p \in [1 + c_1/\log\log d, 2)$, $1/p + 1/q = 1$ and $n(\log d)^{c_0} \leqslant d \leqslant n^{q/2}$. Suppose $\mathbf{X} \in \mathbb{R}^{n \times d}$ is a standard Gaussian matrix.

We begin by noting that for some constant $c > 0$, the event

$$\mathcal{E}_1 = \left\{ \sqrt{d} - c\sqrt{n} \leqslant \sigma_{\min}(\mathbf{X}) \leqslant \sigma_{\max}(\mathbf{X}) \leqslant \sqrt{d} + c\sqrt{n} \right\}$$

has probability $\mathbb{P}(\mathcal{E}_1) \geqslant 1 - 2\exp(-n)$ (Vershynin, 2018, Theorem 4.6.1). Since $d \geqslant n(\log d)^{c_0}$ we have that $\sigma_{\min}(\mathbf{X}) \asymp \sqrt{d}$ with probability at least $1 - 2\exp(-n)$. Since $\|\cdot\|_p \leqslant d^{1/p-1/2} \|\cdot\|_2$, we know that $d^{1/2-1/p} B_2^d \subset B_p^d$, and hence on $\mathcal{E}_1$ that for some constant $c > 0$

$$c\sqrt{d} d^{1/2-1/p} B_2^n = c d^{1/q} B_2^n \subset \mathbf{X} B_p^d.$$

Following Neykov (2022), conditioned on $\mathbf{X}$, we need to solve the stationary condition of

$$\log M(\varepsilon, \mathbf{X}B_p^d \cap C\varepsilon B_2^n) \asymp \log M(\varepsilon, \mathbf{X}B_p^d) \asymp \varepsilon^2.$$

to determine the minimax rate given $\mathbf{X}$, where the first "$\asymp$" holds by a pigeon-hole argument. We solve this by first noting that on the event $\mathcal{E}_1$, the solution must satisfy $\varepsilon \gtrsim d^{1/q}$, since on $\mathcal{E}_1$ and if $\varepsilon \asymp d^{1/q}$ we have

$$\log M(\varepsilon, \mathbf{X}B_p^d) \geqslant \log M(\varepsilon, c\varepsilon B_2^n) \gtrsim n \geqslant d^{2/q} \asymp \varepsilon^2.$$

Thus, all we need is to estimate $\log M(\varepsilon, \mathbf{X}B_p^d)$ when $\varepsilon \gtrsim d^{1/q}$. This was done (implicitly) in Kur et al. (2024), when $p \geqslant 1 + c_1/\log\log(d)$, $d \gtrsim n(\log d)^{c_0}$ and n is sufficiently large. Specifically, it was shown that if $\varepsilon \gtrsim d^{1/q}$, the event

$$\mathcal{E}_2 = \left\{ \log M(\varepsilon, \mathbf{X}B_p^d) \asymp \log M(\varepsilon/\sqrt{n}, B_p^d) \right\} \tag{28}$$

has probability at least $1 - C\exp(-cn)$. We outline the proof of (28) after finishing the main proof. Using (28), for $\varepsilon \gtrsim d^{1/q}$ we can solve the inequality

$$\log M(\varepsilon/\sqrt{n}, B_p^d) \asymp \left(\frac{\varepsilon}{\sqrt{n}}\right)^{-\frac{2p}{2-p}} \log\left(d\left(\frac{\varepsilon}{\sqrt{n}}\right)^{\frac{2p}{2-p}}\right) \gtrsim \varepsilon^2,$$

where we used (29) from Schütt (1984):

$$\log N(\varepsilon, B_p^k) \asymp_p \begin{cases} \varepsilon^{-\frac{2p}{2-p}} \log\left(k\varepsilon^{\frac{2p}{2-p}}\right) & \text{if } \varepsilon \gtrsim k^{1/2-1/p}, \\ k \log\left(k^{-1}\varepsilon^{-\frac{2p}{2-p}}\right) & \text{if } \varepsilon \lesssim k^{1/2-1/p}, \end{cases} \tag{29}$$

We can see that choosing $\varepsilon^2 = n^{p/2}(\log d)^{1-p/2}$ satisfies this inequality. It follows from dividing by n and taking the union bound over $\mathcal{E}_1, \mathcal{E}_2$ that with probability $1 - c_2\exp(-c_3 n)$ over draws of the data matrix $\mathbf{X}$, we have

$$\inf_{\widehat{\alpha}} \sup_{\alpha^\star \in B_p^d} \mathbb{E}_\xi \mathcal{R}(\widehat{\alpha}) \asymp n^{p/2-1}(\log d)^{1-p/2} \vee \frac{d^{2/q}}{n},$$

which finishes the proof.

Proof of (28). We outline the proof of (28) based on arguments from Kur et al. (2024) here again for completeness. Define the i-th dyadic entropy number of any set S as

$$e_i(S) = \inf\left\{\varepsilon > 0 : \log_2 N(\varepsilon, S) \leqslant i - 1\right\}.$$

By definition, we have that $e_i(S) \leqslant \varepsilon$, if and only if $\log N(\varepsilon, S) \leqslant i$. We know by Schütt's Theorem (Schütt, 1984) that the i-th dyadic entropy number of B_p^k is given by

$$e_i(B_p^k) \asymp \begin{cases} 1 & \text{if } 1 \leqslant i \leqslant \log k, \\ \left(\frac{\log(ek/i)}{i}\right)^{1/p-1/2} & \text{if } \log k \leqslant i \leqslant k, \\ 2^{-i/k}k^{1/2-1/p} & \text{if } i \geqslant k. \end{cases} \tag{30}$$

For $i \in [d]$ with $\log d \leqslant i \leqslant d$, define $\Sigma_i = \{\sigma \subset [d] \mid |\sigma| \asymp i/\log(ed/i)\}$ and $\varepsilon_i := e_i(B_p^d) \asymp (\log(ed/i)/i)^{1/p-1/2}$ (recall (30)), and define the sets

$$\mathcal{V}_i = \bigcup_{\sigma \in \Sigma_i} A_\sigma \quad \text{where} \quad A_\sigma = \frac{\varepsilon_i}{\sqrt{|\sigma|}} \cdot B_\infty^\sigma = \frac{\varepsilon_i}{\sqrt{|\sigma|}} \left\{v \in \mathbb{R}^d \;\middle|\; v_i = 0 \text{ if } i \notin \sigma, \|v\|_\infty \leqslant 1\right\}.$$

Then it is easy to see that $\mathcal{V}_i \subset B_p^d$, as for all $v \in \mathcal{V}_i$ we have

$$\|v\|_p \leqslant |\sigma|^{1/p} \|v\|_\infty \leqslant |\sigma|^{1/p-1/2} \varepsilon_i \lesssim \left(\frac{i}{\log(ed/i)}\right)^{1/p-1/2} \left(\frac{\log(ed/i)}{i}\right)^{1/p-1/2} \leqslant 1.$$

Also note that by the same calculation, for $v \in \mathcal{V}_i$, it holds $\|v_i\|_2 \leqslant \varepsilon_i$.

Let $\mathcal{N}_i$ be an ε_i-covering of $\mathcal{V}_i$ and note that (Kur et al., 2024, underneath Equation (25)) we get $\log|\mathcal{N}_i| \lesssim i$, meaning it has the same metric entropy for ε_i as B_p^d. Following Schütt (1984); Kur et al. (2024), one can see that for $\mathcal{I} := \{2^0, 2^1, 2^2, \ldots, c \cdot d\}$, every $v \in B_p^d$ can be written as

$$v = v_0 + \sum_{i \in \mathcal{I}} \delta_i v_i \qquad \text{where} \qquad \|v_0\|_2 \leqslant \varepsilon_d \lesssim d^{1/q-1/2}$$

and $v_i \in \mathcal{N}_i$, $\|v_i\|_p \asymp 1$, $\sum_{i\in\mathcal{I}} \delta_i^p \leqslant 1$. Consider the partition of $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$ with $\mathcal{I}_1 = \{2^0, 2^1, 2^2, \ldots, cn\}$, $\mathcal{I}_2 = \{n, 2^1 n, 2^2 n, \ldots, c \cdot d\}$ (where w.l.o.g., we assume these terms are powers of two). By triangle inequality, it follows that we can bound the norm of $\mathbf{X}v$ for any $v \in B_p^d$ as

$$\|\mathbf{X}v\|_2 \leqslant \|\mathbf{X}v_0\|_2 + \left\|\sum_{i\in\mathcal{I}_1} \delta_i \mathbf{X} v_i\right\|_2 + \left\|\sum_{i\in\mathcal{I}_2} \delta_i \mathbf{X} v_i\right\|_2.$$

We now bound each term separately.

To that end, recall that by the Johnson-Lindenstrauss Lemma (Vershynin, 2018, Theorem 5.3.1 and Exercise 5.3.3), we know that for $\delta \in (0,1)$ and any finite set $\mathcal{N} \subset \mathbb{R}^d$ with $n \geqslant (C/\delta^2)\log|\mathcal{N}|$, it holds that

$$\forall v, w \in \mathcal{N}: \quad (1-\delta)\cdot\|v-w\|_2 \leqslant \frac{1}{\sqrt{n}}\|\mathbf{X}(v-w)\|_2 \leqslant (1+\delta)\cdot\|v-w\|_2 \tag{31}$$

with probability of at least $1 - 2\exp(-cn\delta^2)$.

1. Note that as the maximal singular value satisfies $\sigma_{\max}(\mathbf{X}) \lesssim \sqrt{d}$ on the event $\mathcal{E}_1$, we get that

$$\|\mathbf{X}v_0\|_2 \lesssim \sqrt{d} \cdot \varepsilon_d \lesssim d^{1/q}.$$

2. By (31), since $\log \left|\bigcup_{i\in\mathcal{I}_1} \mathcal{N}_i\right| \lesssim n$, we obtain for $\mathcal{N} = \bigcup_{i\in\mathcal{I}_1} \mathcal{N}_i$ that

$$\mathcal{E}_3 = \left\{\forall v, w \in \mathcal{N} : \ \|\mathbf{X}(v-w)\|_2 \asymp \sqrt{n} \cdot \|v-w\|_2\right\},$$

holds with probability at least $1 - 2\exp(-cn)$. Clearly, on $\mathcal{E}_3$, we also have that $\|\mathbf{X}v\|_2 \asymp \sqrt{n} \cdot \|v\|_2$ for all $v \in \mathcal{N}$.

3. Again, by (31) with $\delta \asymp \sqrt{i/n}$ so that $\log|\mathcal{N}_i| \lesssim i \asymp \delta^2 n$, the event

$$\mathcal{E}_4 = \left\{\forall i \in \mathcal{I}_2 : \ \sup_{v\in\mathcal{N}_i} \|\mathbf{X}v\|_2 \lesssim \sqrt{n} \cdot \sqrt{i/n} \cdot \|v\|_2 \leqslant \sqrt{n} \cdot \sqrt{i/n} \cdot \varepsilon_i = \sqrt{i} \cdot \varepsilon_i\right\}.$$

holds with probability at least $1 - 2\sum_{i\in\mathcal{I}_2} \exp(-ci) \geqslant 1 - C\exp(cn)$. Hence, by triangle and Hölder's inequalities, we obtain that

$$\left\|\sum_{i\in\mathcal{I}_2} \delta_i \mathbf{X} v_i\right\|_2 \leqslant \sum_{i\in\mathcal{I}_2} \sqrt{i}\,\varepsilon_i \cdot \delta_i \lesssim \left(\sum_{i\in\mathcal{I}_2} \delta_i^p\right)^{1/p} \cdot \left(\sum_{i\in\mathcal{I}_2} (\sqrt{i}\,\varepsilon_i)^q\right)^{1/q} \lesssim \left(\sum_{i\in\mathcal{I}_2} (\sqrt{i}\,\varepsilon_i)^q\right)^{1/q}.$$

Plugging in $\varepsilon_i \lesssim \log(d/i)^{1/p-1/2} i^{1/2-1/p}$ from (30) we get that this is bounded by

$$\left(\sum_{i\in\mathcal{I}_2} (\sqrt{i}\,\varepsilon_i)^q\right)^{1/q} \lesssim \left(\sum_{i\in\mathcal{I}_2} \left(\log(d/i)^{1/p-1/2} i^{1/q}\right)^q\right)^{1/q} \lesssim \log(d)^{2/q} \cdot d^{1/q} \lesssim d^{1/q}$$

where in the last inequality we used that $p \geqslant 1 + C/\log\log(d)$.

Note that if $\varepsilon \gtrsim d^{1/q}$, we have that $\log M(\varepsilon, \mathbf{X}B_p^d) \asymp \log M(\varepsilon, \mathbf{X}B_p^d \backslash d^{1/q} B_2^n)$. We just showed that on the intersection of events $\mathcal{E}_1, \mathcal{E}_3, \mathcal{E}_4$ we have that if $\|\mathbf{X}v\|_2 \geqslant d^{1/q}$, it holds that $d^{1/q} \leqslant \|\mathbf{X}v\|_2 \lesssim \sqrt{n}\left\|\sum_{i\in\mathcal{I}_1} \delta_i v_i\right\|_2 + d^{1/q}$, and hence $\left\|\sum_{i\in\mathcal{I}_1} \delta_i v_i\right\|_2 \gtrsim d^{1/q}/\sqrt{n}$. Therefore, the metric entropy beyond the threshold $d^{1/q}$ only depends on the i-th dyadic numbers of B_p^d with $i \in \mathcal{I}_1$. On this set, we already showed that the Johnson-Lindenstrauss Lemma applies (i.e., $\mathcal{E}_3$), and hence

$$\log M(\varepsilon, \mathbf{X}B_p^d) \asymp \log M(\varepsilon/\sqrt{n}, B_p^d)$$

with probability at least $1 - C\exp(-cn)$, which is (28). □

B.10 Proof of Theorem 4

We split the proof of Theorem 4 into several auxiliary lemmas for each potential, which we prove in Appendix B.12. Specifically, we show for each potential from Table 1 that it satisfies Assumption A with the parameters stated in Table 1. Lemma 3 provides the details for the squared ℓ_p-norm, Lemma 4 for the Moreau envelope-based potential, Lemma 5 for the adjusted hypentropy and Lemma 6 for the sigmoidal potential. We prove Lemmas 3 to 6 in Appendix B.12.

Lemma 3. *Suppose that $\varphi_K = \|\cdot\|_1$ and $\psi = \|\cdot\|_p^2$ with $1 < p \leqslant 1 + 1/\log(d)$. Then the discrete time versions of Assumption A are satisfied with constants $c_l = e, c_u = 1$ and $\rho = 2(p-1)$ w.r.t. the ℓ_p-norm.*

Lemma 4. *Suppose that $\varphi_K = \|\cdot\|_1$ and*

$$\psi(\alpha) = \left(\sum_{i=1}^d h_\lambda(\alpha_i) + \frac{d\lambda}{2}\right)^2 + \frac{\rho}{2}\|\alpha\|_2^2 \quad \text{with} \quad h_\lambda(x) = \begin{cases} \frac{x^2}{2\lambda} & |x| \leqslant \lambda \\ |x| - \frac{\lambda}{2} & |x| > \lambda \end{cases}$$

with $\lambda \leqslant 2\tau/d$ and $\rho = 2$. Then the discrete-time version of Assumption A is satisfied with $\rho = 2$ with respect to the ℓ_2-norm and constants $c_l = 1, c_u = \sqrt{5}$.

Lemma 5. *Suppose that* $\varphi_K = \|\cdot\|_1$ *and*

$$\psi(\alpha) = \left(\sum_{i=1}^{d} \frac{\left(\alpha_i \operatorname{arcsinh}(\alpha_i/\gamma) - \sqrt{\alpha_i^2 + \gamma^2} + \gamma + 1 \right)}{\operatorname{arcsinh}(\gamma^{-1})} \right)^2,$$

with $\gamma \leqslant \sinh(d/\tau)^{-1} \wedge (4\tau)^{-1} \wedge 2^{-1/2}$. *Then it satisfies both the continuous and discrete-time versions of Assumption A with constants* $c_l = 1, c_u = 3$ *and* $\rho = \operatorname{arcsinh}(\gamma^{-1})^{-2}$ *with respect to both the* ℓ_1 *and* ℓ_2*-norm.*

Lemma 6. *Suppose that* $\varphi_K = \|\cdot\|_1$, *and*

$$\psi(\alpha) = \left(\sum_{i=1}^{d} \frac{1}{\gamma} \left(\log\left(1 + \exp\left(-\gamma\alpha_i\right)\right) + \log\left(1 + \exp\left(\gamma\alpha_i\right)\right) \right) \right)^2$$

with $\gamma \geqslant d\log(4)/\tau$. *Then* ψ *satisfies the continous-time version of Assumption A with constants* $c_l = 1, c_u = 2$.

The proofs of Lemmas 3 to 6 can be found in Appendix B.12.

Therefore, we may invoke Theorem 1, and the result follows by directly bounding the stationary radius: Suppose that $K = B_1^d$. If $\mathbf{X}$ is column-normalized (16) and $d \geqslant \tau\sqrt{n}$, then

$$\frac{r_0^2(\mathbf{X}K_{3c_a\tau})}{n} \leqslant 4\frac{\operatorname{rk}(\mathbf{X})}{n} \wedge 186c_a\tau\sqrt{\frac{\log(2ed/(\tau\sqrt{n}))}{n}}.$$

This follows directly from the derivations in Bellec (2017, Theorem 7) or from Gordon et al. (2007). Plugging this into (9) from Theorem 1 we get that with probability at least $1 - \exp(-0.1n) - \delta$ it holds

$$\begin{aligned}
\mathcal{R}(\alpha_{t^\star}) &\leqslant \frac{4r_0^2(\mathbf{X}K_{3c_a\tau})}{n} + \frac{8\log(1/\delta)}{n} \\
&\leqslant 4\left(4\frac{\operatorname{rk}(\mathbf{X})}{n} \wedge 186c_a\tau\sqrt{\frac{\log(2ed/(\tau\sqrt{n}))}{n}} \right) + \frac{8\log(1/\delta)}{n} \\
&\leqslant \left(16\frac{\operatorname{rk}(\mathbf{X})}{n} \wedge 744c_a\tau\sqrt{\frac{\log(2ed/(\tau\sqrt{n}))}{n}} \right) + \frac{8\log(1/\delta)}{n}
\end{aligned}$$

To bound $r_0(3c_a\tau\mathbf{X}B_1^d)$ in under Gaussian design, we use Vershynin (2018, Examples 7.5.4 and 7.5.11) which yield that

$$r_0^2(\mathbf{X}B_1^d) \leqslant 2w(\mathbf{X}B_1^d) \leqslant 2\|\mathbf{X}\|_2 w(B_1^d) \lesssim \|\mathbf{X}\|_2 \sqrt{\log d}.$$

By Wainwright (2019, Theorem 6.1) we have that

$$\mathbb{P}\left(\|\mathbf{X}\|_2 \leqslant 3\sqrt{n} \right) \geqslant 1 - \exp(-n/2)$$

so that with probability $1 - \exp(-n/2)$ it holds $r_0^2(3c_a\tau\mathbf{X}B_1^d) \lesssim c_a\tau\sqrt{n\log d}$. Combining this with Theorem 1 and Remark 1, and noting that $\operatorname{rk}(\mathbf{X}) = n$ with probability 1, we get from a union bound that with probability $1 - 2\exp(-0.1n) - \delta$ that

$$\begin{aligned}
\mathcal{R}(\alpha_{t^\star}) &\leqslant \frac{4r_0^2(3c_a\tau\mathbf{X}B_1^d)}{n} + \frac{8\log(1/\delta)}{n} \\
&\lesssim \left(1 \wedge c_a\tau\sqrt{\frac{\log d}{n}} \right) + \frac{\log(1/\delta)}{n}
\end{aligned}$$

Choosing $\delta = 0.01$ so that $1 - 2\exp(-0.1n) - \delta = 0.99 - 2\exp(-0.1n)$, we get that

$$\mathcal{R}(\alpha_{t^\star}) \lesssim 1 \wedge c_a\tau\sqrt{\frac{\log d}{n}}.$$

This concludes the proof.

Remark 2. As a side-remark, we also provide a direct proof that if $\mathbf{X}$ is column-normalized (16), then

$$\frac{r_0^2(\mathbf{X}K_{3c_a\tau})}{n} \leqslant 4\frac{\mathrm{rk}\,(\mathbf{X})}{n} \wedge 12c_a\tau\sqrt{\frac{\log d}{n}}.$$

Note that the bound in terms of the rank of $\mathbf{X}$ follows from Remark 1 and Appendix B.4. We can use $r_0^2(\mathbf{X}K_\tau) \leqslant 2w(\mathbf{X}K_\tau)$ and to further bound $w(\mathbf{X}K_\tau)$, we can apply Hölder's inequality and get for $d \geqslant 2$

$$w(\mathbf{X}K_\tau) = \mathbb{E}\left[\sup_{\alpha\in K_\tau}\langle\xi, \mathbf{X}\alpha\rangle\right] \leqslant \mathbb{E}\left[\left\|\mathbf{X}^\top\xi\right\|_\infty\right]\sup_{\alpha\in K_\tau}\|\alpha\|_1 = \tau\mathbb{E}\left[\left\|\mathbf{X}^\top\xi\right\|_\infty\right] \leqslant 2\tau\sqrt{n\log d},$$

where in the last step we used that $\mathbf{X}^\top\xi \sim \mathcal{N}(0, \mathbf{X}^\top\mathbf{X})$, and $\mathbf{X}$ is column-normalized (16), which implies that $(\mathbf{X}^\top\xi)_i, i \in [d]$ is Gaussian with variance n, so that we may use standard sub-Gaussian concentration (Wainwright, 2019, Exercise 2.12), which yields the second upper bound. Plugging the second bound into (9) from Theorem 1 we get that with probability at least $1 - \exp(-0.1n) - \delta$ it holds

$$\begin{aligned}
\mathcal{R}(\alpha_{t^\star}) &\leqslant \frac{4r_0^2(\mathbf{X}K_{3c_a\tau})}{n} + \frac{8\log(1/\delta)}{n} \\
&\leqslant 4\left(4\frac{\mathrm{rk}\,(\mathbf{X})}{n} \wedge 12c_a\tau\sqrt{\frac{\log d}{n}}\right) + \frac{8\log(1/\delta)}{n} \\
&\leqslant \left(16\frac{\mathrm{rk}\,(\mathbf{X})}{n} \wedge 48c_a\tau\sqrt{\frac{\log d}{n}}\right) + \frac{8\log(1/\delta)}{n}.
\end{aligned}$$

B.11 Proof of Proposition 5

We check that Assumption A is satisfied. We use the results from Beck and Teboulle (2012, Example 4.5).

(I). The differentiability of ψ is clear, with gradient and Hessian of ψ given by

$$\begin{aligned}
\nabla\psi(x) &= \frac{2}{\gamma}\log\left(m\sum_{i=1}^m \exp(\gamma\langle a_i, x\rangle)\right)\mu(x) + \rho x \\
\nabla^2\psi(x) &= 2\log\left(m\sum_{i=1}^m \exp(\gamma\langle a_i, x\rangle)\right)\left(\sum_{i=1}^m p_i(x)a_ia_i^\top - \mu(x)\mu(x)^\top\right) + 2\mu(x)\mu(x)^\top + \rho\mathbf{I}_d
\end{aligned}$$

where we defined the quantities

$$p_i(x) = \frac{\exp(\gamma\langle a_i, x\rangle)}{\sum_{j=1}^m \exp(\gamma\langle a_j, x\rangle)} \quad \text{and} \quad \mu(x) = \sum_{i=1}^m p_i(x)a_i$$

Therefore, since $p_i(0) = 1/m$ we have $\mu(0) = \frac{1}{m}\sum_{i=1}^m a_i$ which we assumed to be zero, and hence $\nabla\psi(0) = 0$.

(II). The convexity of $\sqrt{\psi}$ follows as previously because $\sqrt{\psi}$ is the Euclidean norm of $\sqrt{\frac{\rho}{2}}\|x\|_2$ and $\frac{1}{\gamma}\log\left(m\sum_{i=1}^m \exp\left(\gamma\langle a_i, x\rangle\right)\right)$, of which the Hessian is given by

$$\left(\sum_{i=1}^m p_i(x)a_ia_i^\top - \mu(x)\mu(x)^\top\right)$$

which is clearly positive semi-definite for every $x \in \mathbb{R}^d$ (as it is the covariance matrix of the discrete distribution over a_i with probabilities $p_i(x)$) and thus convex.

(III). The ρ-strong convexity follows as previously because

$$x \mapsto \psi(x) - \frac{\rho}{2}\|x\|_2^2 = \frac{1}{\gamma^2}\log^2\left(m\sum_{i=1}^m \exp\left(\gamma\langle a_i, x\rangle\right)\right)$$

is convex (which can be seen by the Hessian above).

(IV). By Beck and Teboulle (2012, Example 4.5), we know that for all $x \in \mathbb{R}^d$

$$\varphi_K(x) - \frac{1}{\gamma}\log m \leqslant \frac{1}{\gamma}\log\left(\sum_{i=1}^m \exp\left(\gamma\langle a_i, x\rangle\right)\right) \leqslant \varphi_K(x) + \frac{1}{\gamma}\log m.$$

It follows that

$$\varphi_K(x) \leqslant \frac{1}{\gamma}\log\left(m\sum_{i=1}^m \exp\left(\gamma\langle a_i, x\rangle\right)\right) \leqslant \varphi_K(x) + \frac{2}{\gamma}\log m.$$

and therefore, for all $\alpha \in \mathbb{R}^d$ and $\alpha' \in \tau K$ with $\rho = 2/\max_{i\in[M]}\|k_i\|_2^2$ and $\gamma \geqslant 2\log(m)/\tau$

$$\varphi_K(\alpha) \leqslant \frac{1}{\gamma}\log\left(m\sum_{i=1}^m \exp\left(\gamma\langle a_i, \alpha\rangle\right)\right) \leqslant c_l\sqrt{\psi(\alpha)}$$
$$\sqrt{\psi(\alpha')} = \sqrt{\frac{1}{\gamma^2}\log^2\left(m\sum_{i=1}^m \exp\left(\gamma\langle a_i, \alpha'\rangle\right)\right) + \frac{\rho}{2}\|\alpha'\|_2^2} \leqslant \varphi_K(\alpha') + \frac{2}{\gamma}\log m + \tau \leqslant 3\tau = c_u\tau$$

with $c_l = 1$, $c_u = 3$.

Using Bellec (2017, Proposition 5) we get that if $M \geqslant \tau\phi\sqrt{n}$, then

$$r_0^2(\mathbf{X}K_\tau) \leqslant 31\tau\phi\sqrt{n}\sqrt{\log\left(\frac{eM}{\tau\phi\sqrt{n}}\right)}.$$

Hence, from Theorem 1 and Remark 1, we know that with probability $1 - \exp(-0.1n) - \delta$ it holds that

$$\begin{aligned}\mathcal{R}(\alpha_{t^\star}) &\leqslant \frac{4r_0^2(\mathbf{X}K_{3c_a\tau})}{n} + \frac{8\log(1/\delta)}{n}\\ &\leqslant 4\left(\frac{4\,\mathrm{rk}\,(\mathbf{X})}{n} \wedge 93c_a\tau\phi\sqrt{\frac{\log\left(eM/(\tau\phi\sqrt{n})\right)}{n}}\right) + \frac{8\log(1/\delta)}{n}\\ &= \left(\frac{16\,\mathrm{rk}\,(\mathbf{X})}{n} \wedge 372c_a\tau\phi\sqrt{\frac{\log\left(eM/(\tau\phi\sqrt{n})\right)}{n}}\right) + \frac{8\log(1/\delta)}{n}.\end{aligned}$$

Choosing $\delta = 0.01$ yields the result. This concludes the proof.

B.12 Proofs of the Auxiliary Lemmata for Theorem 4

B.12.1 Proof of Lemma 3

We check each condition from Assumption A. We already proved (I),(II) and (III) in the proof of Proposition 2.

(IV). We use the norm inequalities $\|\cdot\|_p \leqslant \|\cdot\|_1 \leqslant d^{1-1/p}\|\cdot\|_p$ and the fact that if $1 < p \leqslant 1 + 1/\log(d)$

$$d^{1-1/p} \leqslant d^{\frac{1}{1+\log d}} = \exp\left(\frac{\log d}{1+\log d}\right) \leqslant \exp(1) = e.$$

which implies that for all $\alpha \in \mathbb{R}^d$ and $\alpha' \in \tau B_1^d$

$$\varphi_K(\alpha) = \|\alpha\|_1 \leqslant e\|\alpha\|_p = c_l\sqrt{\psi(\alpha)}$$
$$\sqrt{\psi(\alpha')} = \|\alpha'\|_p \leqslant \|\alpha'\|_1 \leqslant c_u\tau$$

with $c_l = e, c_u = 1$. This concludes the proof.

B.12.2 Proof of Lemma 4

We first note that the potential in Lemma 4 is indeed given by the Moreau envelope, since by, e.g., Beck and Teboulle (2012, Example 4.2) it holds

$$(\mathcal{M}_\lambda \|\cdot\|_1)(\alpha) = \sum_{i=1}^d h_\lambda(\alpha_i) \quad \text{with} \quad h_\lambda(x) = \begin{cases} \frac{x^2}{2\lambda} & |x| \leqslant \lambda, \\ |x| - \frac{\lambda}{2} & |x| > \lambda. \end{cases}$$

It follows that $\mathcal{M}_\lambda \|\cdot\|_1 + \frac{d\lambda}{2} \geqslant \|\cdot\|_1 \geqslant 0$. We check that Assumption A is indeed satisfied.

(I). First note that the function is differentiable, as Moreau envelopes are continuously differentiable (Parikh and Boyd, 2014, §3.1) or (Beck and Teboulle, 2012, §4.2). Moreover, a calculation shows that

$$\nabla\psi(\alpha) = 2\left(\sum_{i=1}^d h_\lambda(x) + \frac{d\lambda}{2}\right)\left(h'_\lambda(\alpha_1), \dots, h'_\lambda(\alpha_d)\right) + \rho\alpha \quad \text{with} \quad h'_\lambda(x) = \begin{cases} \frac{x}{\lambda} & |x| \leqslant \lambda \\ \operatorname{sign}(x) & |x| \geqslant \lambda \end{cases}$$

and so $\nabla\psi(0) = 0$.

(II). The Moreau envelope $\mathcal{M}_\lambda \|\cdot\|_1 + \frac{d\lambda}{2}$ is convex (Beck and Teboulle, 2012, §4.2) and non-negative, so that

$$\sqrt{\psi(\alpha)} = \sqrt{\left((\mathcal{M}_\lambda \|\cdot\|_1)(\alpha) + \frac{d\lambda}{2}\right)^2 + \frac{\rho}{2}\|\alpha\|_2^2}$$

is convex, as it is the Euclidean norm of two non-negative convex functions.

(III). E.g., by Nikodem and Pales (2011), the ρ-strong convexity of ψ with respect to ℓ_2-norm holds if and only if

$$\psi - \frac{\rho}{2}\|\alpha\|_2^2 = \left((\mathcal{M}_\lambda \|\cdot\|_1)(\alpha) + \frac{d\lambda}{2}\right)^2$$

is convex. Since $\mathcal{M}_\lambda \|\cdot\|_1$ is convex (Beck and Teboulle, 2012, §4.2), $\mathcal{M}_\lambda \|\cdot\|_1 + \frac{d\lambda}{2} \geqslant 0$ and squaring is non-decreasing, this is the case.

(IV). Finally, we have that for all $\alpha \in \mathbb{R}^d$, $\alpha' \in \tau B_1^d$ and $\rho = 2$, $\lambda \leqslant 2\tau/d$ that

$$\varphi_K(\alpha) = \|\alpha\|_1 \leqslant (\mathcal{M}_\lambda \|\cdot\|_1)(\alpha) + \frac{d\lambda}{2} \leqslant \sqrt{\left((\mathcal{M}_\lambda \|\cdot\|_1)(\alpha) + \frac{d\lambda}{2}\right)^2 + \frac{\rho}{2}\|\alpha\|_2^2} = c_l\sqrt{\psi(\alpha)},$$

$$\sqrt{\psi(\alpha')} = \sqrt{\left((\mathcal{M}_\lambda \|\cdot\|_1)(\alpha') + \frac{d\lambda}{2}\right)^2 + \frac{\rho}{2}\|\alpha'\|_2^2} \leqslant \sqrt{(\|\alpha'\|_1 + \tau)^2 + \|\alpha'\|_2^2} \leqslant \sqrt{5}\tau = c_u\tau,$$

with $c_l = 1, c_u = \sqrt{5}$. Here we used that for any Moreau envelope, $\mathcal{M}_\lambda f \leqslant f$ by definition.

B.12.3 Proof of Lemma 5

Recall that we defined the function

$$\begin{aligned}\psi(\alpha) &= \left(\operatorname{arcsinh}(\gamma^{-1})^{-1}\sum_{i=1}^d \alpha_i \operatorname{arcsinh}(\alpha_i/\gamma) - \sqrt{\alpha_i^2 + \gamma^2} + \gamma + 1\right)^2 \\ &= \left(\sum_{i=1}^d g_\gamma(\alpha_i)\right)^2 \\ &= f^2(\alpha)\end{aligned}$$

where we now introduced the new notation

$$f = \sqrt{\psi} \quad \text{and} \quad g_\gamma(x) := \operatorname{arcsinh}\left(\gamma^{-1}\right)^{-1}\left(x\operatorname{arcsinh}(x/\gamma) - \sqrt{x^2+\gamma^2} + \gamma + 1\right).$$

We prove that ψ satisfies Assumption A.

(I). ψ is twice continuously differentiable with gradient

$$\nabla\psi(\alpha) = 2\,\mathrm{arcsinh}(\gamma^{-1})^{-2} f(\alpha) \begin{pmatrix} \mathrm{arcsinh}(\alpha_1/\gamma) \\ \vdots \\ \mathrm{arcsinh}(\alpha_d/\gamma) \end{pmatrix} \implies \nabla\psi(0) = 0$$

This follows by straight-forward calculations (Ghai et al., 2020).

(II). $\sqrt{\psi} = f$ is (strictly) convex, as a simple calculation (Ghai et al., 2020) shows that the Hessian of f is given by

$$\nabla^2 f(x) = \mathrm{arcsinh}(\gamma^{-1})^{-1} \mathrm{diag}\left(\frac{1}{\sqrt{x_1^2+\gamma^2}} \dots \frac{1}{\sqrt{x_d^2+\gamma^2}} \right)$$

which is positive definite everywhere.

To proceed, we state two useful facts.

- *Fact 1: For all $\gamma > 0$ and $\alpha \in \mathbb{R}^d$, $f(\alpha) \geqslant \|\alpha\|_1 \vee d\,\mathrm{arcsinh}(\gamma^{-1})^{-1}$.*
  We first show that $f(\alpha) \geqslant \|\alpha\|_1$. It suffices to show that $h_\gamma(x) = g_\gamma(x) - |x| \geqslant 0$ for all $x \in \mathbb{R}$. First of all, $h_\gamma(0) = \mathrm{arcsinh}\left(\gamma^{-1}\right)^{-1} > 0$ for all $\gamma > 0$, and due to the symmetry of h_γ, it therefore suffices to show that $h_\gamma(x) \geqslant 0$ for all $x \in (0,\infty)$. On this interval, $x \mapsto h_\gamma(x)$ is twice differentiable with
  $$\frac{dh_\gamma}{dx}(x) = \mathrm{arcsinh}\left(\gamma^{-1}\right)^{-1} \mathrm{arcsinh}\left(x/\gamma\right) - 1 \quad \text{and} \quad \frac{d^2h_\gamma}{(dx)^2}(x) = \frac{\mathrm{arcsinh}\left(\gamma^{-1}\right)^{-1}}{\sqrt{x^2+\gamma^2}}.$$
  Since $\frac{d^2h_\gamma}{(dx)^2} > 0$ for all $x \in (0,\infty)$, we know that h_γ is strictly convex on $(0,\infty)$. Therefore, h_γ has a unique global minimum on $(0,\infty)$, as $\lim_{x\to\infty} h_\gamma(x) = \infty$ and $\frac{dh_\gamma}{dx}(x) < 0$ for small enough $x > 0$. The first-order condition of optimality $\frac{dh_\gamma}{dx}(x^\star) = 0$ yields $x^\star = 1$, and thus we know that for all $x \in (0,\infty)$
  $$\begin{aligned} h_\gamma(x) &\geqslant h_\gamma(x^\star) \\ &= \mathrm{arcsinh}\left(\gamma^{-1}\right)^{-1} \left(\mathrm{arcsinh}\left(1/\gamma\right) - \sqrt{1^2+\gamma^2} + \gamma + 1\right) - 1 \\ &= \frac{\gamma + 1 - \sqrt{1+\gamma^2}}{\mathrm{arcsinh}(1/\gamma)} > 0. \end{aligned}$$
  hence, $f(\alpha) \geqslant \|\alpha\|_1$. Now, we can also see that for all $x \in \mathbb{R}$
  $$\frac{dg_\gamma}{dx}(x) = \mathrm{arcsinh}\left(\gamma^{-1}\right)^{-1} \mathrm{arcsinh}\left(x/\gamma\right) \quad \text{and} \quad \frac{d^2h_\gamma}{(dx)^2}(x) = \frac{\mathrm{arcsinh}\left(\gamma^{-1}\right)^{-1}}{\sqrt{x^2+\gamma^2}}.$$
  Hence, g_γ has a unique global minimum at $x^\star = 0$, and thus for all $x \in \mathbb{R}$, $g_\gamma(x) \geqslant g_\gamma(0) = \mathrm{arcsinh}(\gamma^{-1})^{-1}$, and thus $f(\alpha) = \sum_{i=1}^d g_\gamma(\alpha_i) \geqslant d\,\mathrm{arcsinh}(\gamma^{-1})^{-1}$.
- *Fact 2: $\gamma\,\mathrm{arcsinh}(1/\gamma) \leqslant 1$ for all $\gamma > 0$.*
  This follows by observing that $\frac{d^2}{(dx)^2} x\,\mathrm{arcsinh}(1/x) = -\sqrt{1/x^2+1}/(x^2+1)^2 < 0$, which means its concave, and $\frac{d}{dx} x\,\mathrm{arcsinh}(1/x) = \mathrm{arcsinh}(1/x) - 1/(x\sqrt{1/x^2+1}) > 0$, which means it has no maximizer, and
  $$\lim_{x\to 0} x\,\mathrm{arcsinh}(1/x) = 0, \quad \lim_{x\to\infty} x\,\mathrm{arcsinh}(1/x) = 1.$$

(III). We follow similar arguments used in Ghai et al. (2020) and will apply Yu (2015, Theorem 3) to show that ψ is ρ_γ-strongly convex on $\mathbb{R}^d$ with respect to $\|\cdot\|_1$ and $\|\cdot\|_2$, where $\rho_\gamma = \mathrm{arcsinh}(\gamma^{-1})^{-2}$.

First note that the Hessian of ψ is given by

$$\nabla^2\psi(x) = 2\,\mathrm{arcsinh}(\gamma^{-1})^{-2} \left[\begin{pmatrix} \mathrm{arcsinh}(x_1/\gamma) \\ \vdots \\ \mathrm{arcsinh}(x_d/\gamma) \end{pmatrix} \begin{pmatrix} \mathrm{arcsinh}(x_1/\gamma) \\ \vdots \\ \mathrm{arcsinh}(x_d/\gamma) \end{pmatrix}^\top + f(x) \begin{pmatrix} \frac{1}{\sqrt{x_1^2+\gamma^2}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{x_d^2+\gamma^2}} \end{pmatrix} \right]$$

and hence we have that for any $y \in \mathbb{R}^d$

$$y^\top \nabla^2 \psi(x) y \geqslant 2 \operatorname{arcsinh}(\gamma^{-1})^{-2} f(x) \sum_{i=1}^d \frac{y_i^2}{\sqrt{x_i^2+\gamma^2}}.$$

Using Fact 1 that $f(x) \geqslant \|x\|_1 \vee d \operatorname{arcsinh}(\gamma^{-1})^{-1}$, we can bound

$$\begin{aligned}
\inf_{x,\|y\|_1=1} f(x) \sum_{i=1}^d \frac{y_i^2}{\sqrt{x_i^2+\gamma^2}} &= \inf_{x,\|y\|_1=1} \frac{f(x)}{\sum_{i=1}^d \sqrt{x_i^2+\gamma^2}} \left(\sum_{i=1}^d \frac{y_i^2}{\sqrt{x_i^2+\gamma^2}}\right)\left(\sum_{i=1}^d \sqrt{x_i^2+\gamma^2}\right) \\
&\geqslant \inf_{x,\|y\|_1=1} \frac{f(x)}{\sum_{i=1}^d \sqrt{x_i^2+\gamma^2}} \left(\sum_{i=1}^d \sqrt{y_i^2}\right) \\
&\geqslant \inf_x \frac{\|x\|_1 \vee d \operatorname{arcsinh}(\gamma^{-1})^{-1}}{\sum_{i=1}^d \sqrt{x_i^2+\gamma^2}} \\
&\geqslant \inf_x \frac{\|x\|_1 \vee d \operatorname{arcsinh}(\gamma^{-1})^{-1}}{\|x\|_1 + d\gamma} \\
&= \frac{d \operatorname{arcsinh}(\gamma^{-1})^{-1}}{d \operatorname{arcsinh}(\gamma^{-1})^{-1} + d\gamma} = \frac{1}{1+\gamma \operatorname{arcsinh}(\gamma^{-1})} \geqslant \frac{1}{2}
\end{aligned}$$

where we used Fact 2 that $\gamma \operatorname{arcsinh}(1/\gamma) \leqslant 1$ for all $\gamma > 0$. Therefore, $\inf_{x,\|y\|_1=1} y^\top \nabla^2 \psi(x) y \geqslant \operatorname{arcsinh}(\gamma^{-1})^{-2}$ which implies ρ_γ-strong convexity with respect to the ℓ_1-norm (Yu, 2015, Theorem 3). With a similar argument, we have that

$$\begin{aligned}
\inf_{x,\|y\|_2=1} f(x) \sum_{i=1}^d \frac{y_i^2}{\sqrt{x_i^2+\gamma^2}} &\geqslant \inf_x \frac{f(x)}{\max_i \sqrt{x_i^2+\gamma^2}} \\
&\geqslant \inf_x \frac{\|x\|_1 \vee d \operatorname{arcsinh}(\gamma^{-1})^{-1}}{\|x\|_1 + \gamma} \\
&= \frac{1}{1+\gamma \operatorname{arcsinh}(\gamma^{-1})/d} \geqslant \frac{1}{2}
\end{aligned}$$

and therefore $\inf_{x,\|y\|_2=1} y^\top \nabla^2 \psi(x) y \geqslant \operatorname{arcsinh}(\gamma^{-1})^{-2}$.

(IV). From Fact 1 it immediately follows that for any $\alpha \in \mathbb{R}^d$ it holds

$$\varphi_K(\alpha) = \|\alpha\|_1 \leqslant f(\alpha) = \sqrt{\psi(\alpha)}$$

so that $c_l = 1$.

It remains to show the upper bound. For $\alpha' \in \tau B_1^d$ we have that $|\alpha_i'| \leqslant \tau$ for all i. Hence,

$$\begin{aligned}
\sqrt{\psi(\alpha')} = f(\alpha') &= \sum_{i=1}^d \operatorname{arcsinh}\left(\gamma^{-1}\right)^{-1} \left(\alpha_i' \operatorname{arcsinh}\left(\alpha_i'/\gamma\right) - \sqrt{(\alpha_i')^2+\gamma^2} + \gamma + 1\right) \\
&\leqslant \sum_{i=1}^d \operatorname{arcsinh}\left(\gamma^{-1}\right)^{-1} \left(\alpha_i' \operatorname{arcsinh}\left(\alpha_i'/\gamma\right) + 1\right) \\
&= \sum_{i=1}^d \operatorname{arcsinh}\left(\gamma^{-1}\right)^{-1} \left(|\alpha_i'| \log\left(\frac{1}{\gamma}\left(\sqrt{(\alpha_i')^2+\gamma^2} + |\alpha_i'|\right)\right) + 1\right) \\
&\leqslant \sum_{i=1}^d \operatorname{arcsinh}\left(\gamma^{-1}\right)^{-1} \left(|\alpha_i'| \log\left(\frac{1}{\gamma}\left(\sqrt{\tau^2+\gamma^2} + \tau\right)\right) + 1\right) \\
&= \|\alpha'\|_1 \operatorname{arcsinh}\left(\gamma^{-1}\right)^{-1} \log\left(\frac{1}{\gamma}\left(\sqrt{\tau^2+\gamma^2} + \tau\right)\right) + d \operatorname{arcsinh}\left(\gamma^{-1}\right)^{-1} \\
&\leqslant \tau \log\left(\gamma^{-1}\right)^{-1} \log\left(\frac{2\tau}{\gamma} + 1\right) + d \operatorname{arcsinh}\left(\gamma^{-1}\right)^{-1}
\end{aligned}$$

For the first term, we notice

$$\gamma \leqslant \frac{1}{\sqrt{2}} \wedge \frac{1}{4\tau} \implies \frac{2\tau}{\gamma} + 1 \leqslant \frac{1}{\gamma^2} \implies \log\left(\frac{2\tau}{\gamma} + 1\right) \leqslant 2\log(\gamma^{-1})$$

so we can bound the first term as 2τ. The second term can be bounded as

$$\gamma \leqslant \sinh(d/\tau)^{-1} \implies d\,\mathrm{arcsinh}(\gamma^{-1})^{-1} \leqslant d\,\mathrm{arcsinh}(\sinh(d/\tau))^{-1} = \tau$$

So overall, we get that $\sqrt{\psi(\alpha')} = f(\alpha') \leqslant 3\tau$, so we may set $c_u = 3$. This concludes the proof.

B.12.4 Proof of Lemma 6

We show that Assumption A holds by going through each point. We define the function

$$g_\gamma(x) = \frac{1}{\gamma}\left(\log(1+\exp(-\gamma x)) + \log(1+\exp(\gamma x))\right)$$

so that $\psi(\alpha) = \left(\sum_{i=1}^d g_\gamma(\alpha_i)\right)^2$.

(I). The fact that the potential is twice differentiable is clear, and the gradient is given by (Schmidt et al., 2007)

$$\nabla\psi(\alpha) = 2\left(\sum_{i=1}^d g_\gamma(\alpha_i)\right)\left(g_\gamma'(\alpha_1),\dots,g_\gamma'(\alpha_d)\right)^\top \quad \text{with} \quad g_\gamma'(x) = \frac{1}{1+\exp\left(-\gamma x\right)} - \frac{1}{1+\exp\left(\gamma x\right)}$$

so that $\nabla\psi(0) = 0$.

(II). We have by Schmidt et al. (2007) that

$$\nabla^2\sqrt{\psi(\alpha)} = \mathrm{diag}\left(\frac{d^2}{(dx)^2}g_\gamma(\alpha_1),\dots,\frac{d^2}{(dx)^2}g_\gamma(\alpha_d)\right) \quad \text{with} \quad \frac{d^2}{(dx)^2}g_\gamma(x) = \frac{2\gamma\exp(\gamma x)}{(1+\exp(\gamma x))^2},$$

and so all its eigenvalues (strictly) greater than zero everywhere, implying that $\sqrt{\psi}$ is convex.

(III). Strict convexity of ψ follows analogously, as

$$\nabla^2\psi(\alpha) = 2\left(\nabla\sqrt{\psi(\alpha)}\nabla\sqrt{\psi(\alpha)}^\top + \sqrt{\psi(\alpha)}\nabla^2\sqrt{\psi(\alpha)}\right)$$

also has all its eigenvalues strictly greater than zero, implying that ψ is strictly convex.

(IV). We can easily show that this approximation satisfies the conditions. First, notice that it holds $g_\gamma(0) = \log(4)/\gamma$, and for $x > 0$

$$\frac{d}{dx}\left(g_\gamma(x) - |x|\right) = \frac{1}{1+\exp\left(-\gamma x\right)} - \frac{1}{1+\exp\left(\gamma x\right)} - 1 < 0$$

and by symmetry of g_γ for $x < 0$ we have $\frac{d}{dx}\left(g_\gamma(x) - |x|\right) > 0$. Combining this with $\lim_{x\to\infty} g_\gamma(x) - |x| = 0$ and $\lim_{x\to-\infty} g_\gamma(x) - |x| = 0$ yields $g_\gamma(x) \in [|x|, |x| + \log(4)/\gamma]$ for all $x \in \mathbb{R}$. Therefore, for all $\alpha \in \mathbb{R}^d$ and all $\alpha' \in \tau B_1^d$, we have for $\gamma \geqslant d\log(4)/\tau$

$$\varphi_K(\alpha) = \|\alpha\|_1 \leqslant \sum_{i=1}^d g_\gamma(\alpha_i) = c_l\sqrt{\psi}$$

$$\sqrt{\psi(\alpha')} = \sum_{i=1}^d g_\gamma(\alpha_i') \leqslant \sum_{i=1}^d \left(|\alpha_i'| + \frac{\log 4}{\gamma}\right) \leqslant \|\alpha'\|_1 + \frac{d\log 4}{\gamma} \leqslant c_u\tau$$

with $c_l = 1$ and $c_u = 2$. This concludes the proof.