On Traceability in ℓ_p Stochastic Convex Optimization

Sasha Voitovych[*] **Mahdi Haghifam**[†] **Idan Attias**[‡]

Gintare Karolina Dziugaite[§] **Roi Livni**[¶] **Daniel M. Roy**[‖]

Abstract

In this paper, we investigate the necessity of traceability for accurate learning in stochastic convex optimization (SCO) under ℓ_p geometries. Informally, we say a learning algorithm is *m-traceable* if, by analyzing its output, it is possible to identify at least m of its training samples. Our main results uncover a fundamental tradeoff between traceability and excess risk in SCO. For every $p \in [1, \infty)$, we establish the existence of an excess risk threshold below which every sample-efficient learner is traceable with the number of samples which is a *constant fraction* of its training sample. For $p \in [1, 2]$, this threshold coincides with the best excess risk of differentially private (DP) algorithms, i.e., above this threshold, there exist algorithms that are not traceable, which corresponds to a sharp phase transition. For $p \in (2, \infty)$, this threshold instead gives novel lower bounds for DP learning, partially closing an open problem in this setup. En route to establishing these results, we prove a sparse variant of the fingerprinting lemma, which is of independent interest to the community.

1 Introduction

Tracing or membership inference informally asks whether it is possible, using only the output of a learning algorithm, to distinguish samples in the training set from held-out samples. The existence of a tracer that identifies training examples reveals that the model has memorized specific examples rather than purely captured the underlying distribution [SSSS17; CCNSTT22]. In particular, understanding tracing has an important role in generalization theory, where an algorithm that is not traceable is known to generalize well beyond its training data [SZ20]. Tracing is also an important technical tool in, e.g., differential privacy (DP), where tracing attacks are the workhorse behind tight lower bounds for the risk-privacy trade-offs [BUV14]. From a privacy standpoint, even the leakage of a single example is viewed as catastrophic. However, from a generalization theory standpoint, we want to understand the exact relationship between an algorithm's generalization performance and the *number* of traceable examples.

To reason rigorously about tracing, following [DSSUV15], we define the problem of tracing as follows. Let $\mathcal{A}_n$ be a learning algorithm that, given a training set $S_n = (Z_1, \ldots, Z_n)$ of

[*]SV and MH are equal contribution authors.
[*]Institute for Data, Systems, and Society, Massachusetts Institute of Technology
[†]Khoury College of Computer Sciences, Northeastern University
[‡]University of Illinois at Chicago; Toyota Technological Institute at Chicago
[§]Google DeepMind
[¶]Department of Electrical Engineering, Tel Aviv University
[‖]Department of Statistical Sciences, University of Toronto; Vector Institute

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

n i.i.d. samples from some underlying distribution $\mathcal{D}$, outputs a learned model $\hat{\theta}$. Then, a *tracer* $\mathcal{T}$ is a hypothesis tester that, given the model $\hat{\theta}$ and a candidate point Z, outputs IN if it believes Z was in S_n, or OUT otherwise. Formally, for some small soundness parameter $\xi \in (0,1)$ and $m \leq n$, we require $\mathcal{T}$ to satisfy:

$$\Pr_{\substack{S_n \sim \mathcal{D}^{\otimes n} \\ Z \sim \mathrm{Unif}(S_n)}} [\mathcal{T}(\hat{\theta}, Z) = \mathrm{IN}] \geq \frac{m}{n} \qquad \Pr_{\substack{S_n \sim \mathcal{D}^{\otimes n} \\ Z \sim \mathcal{D} \perp\!\!\!\perp S_n}} [\mathcal{T}(\hat{\theta}, Z) = \mathrm{OUT}] \geq 1 - \xi$$

When such a tracer exists, we say that $\mathcal{A}_n$ is (ξ, m)-traceable. Equivalently, m is the expected number of samples in the training set S_n for which the tracer outputs IN, and we refer to m as *recall*. (See Definitions 2.3 and 2.4.)

A fundamental problem in learning theory is investigating how an algorithm's generalization ability interacts with the information it retains about the training samples (including, in our language, its traceability). The common wisdom is that any information about the training set in a learned model is in tension with generalization [XR17; BMNSY18; SZ20]. On the other hand, non-traceable algorithms, such as *differentially private* algorithms, are often unable to reach optimal excess risk. The central question we study in this paper is: *what is the exact tradeoff between the number of traceable examples and achievable excess risk?*

This question was considered, first, in the context of mean estimation of a d-dimensional vector, in the seminal work of [BUV14; DSSUV15]. It was also studied in the context of *Stochastic Convex Optimization* (SCO) [SSSS09] in the work of [ADHLR24]. The work of [DSSUV15] studied the tradeoff between excess risk and the number of traceable examples, when a mechanism publishes an estimate of the mean that is *accurate* in every coordinate (i.e., the output of the algorithm has error of α with respect to ℓ_∞ norm to the true mean). At a high level, they showed that for every algorithm that has accuracy better than that achievable by a private algorithm, $\Omega(1/\alpha^2)$ examples are traceable on some hard instance. Notice that for the task of mean estimation in ℓ_∞ norm, the statistical sample complexity is $\Theta(\log(d)/\alpha^2)$. Thus, for every algorithm, the preceding result only shows that it is possible to trace out a $1/\log(d)$ fraction of the input samples. In contrast, [ADHLR24] exhibited an SCO problem in ℓ_2 geometry for which a *constant* fraction of the training samples are traceable. An important open problem, then, is to further explore and understand in which setups we expect a constant fraction of the training sample to be traceable.

In this work, we investigate this question, of traceability, in the fundamental learning setup of *Stochastic Convex Optimization* for general ℓ_p geometries. We show that, in this general learning setup, when private learning is not possible, *there is no meaningful gap* between sample complexity and traceability. That is, in every geometry, there exist a hard problem for which every (sample-efficient) algorithm is traceable with a recall which is a constant fraction of its sample size. Due to connections between SCO and mean estimation problems, our results also extend to the latter settings; in particular, we close the $\log(d)$ gap in the setting of [DSSUV15] and show that optimal traceability is *dimension-dependent*.

SCO is an ideal testbed for this problem: (1) as in modern machine learning practices, first-order methods are known to achieve optimal sample-complexity rates in this setting [Fel16; HRS16; AKL21], and (2) within this framework, we can design provable methods that mitigate tracing, such as DP algorithms [CMS11; BST14; BFTG19]. Therefore, by studying the problem of traceability in SCO, we also deepen our understanding of the interaction between privacy risks and sample-optimal learning.

To present our results, we recall the basic setup of SCO. An SCO problem is characterized via a triple $\mathcal{P} = (\mathcal{Z}, \Theta, f)$, where $\mathcal{Z}$ is the data space, $\Theta \subset \mathbb{R}^d$ is the parameter space, which must be convex, and $f \colon \Theta \times \mathcal{Z} \to \mathbb{R}$ is a loss function such that $f(\cdot, z)$ is convex for all $z \in \mathcal{Z}$. In SCO, data points are drawn from an underlying distribution $\mathcal{D}$ over $\mathcal{Z}$, unknown to the learner. The objective of the learner is to minimize the *expected risk* based on observed samples. Then, a learning algorithm $\mathcal{A}_n \colon \mathcal{Z}^n \to \Theta$ receives a sample $S_n = (Z_1, \ldots, Z_n)$ of n data points from $\mathcal{Z}^n$ and returns a (perhaps randomized) output in Θ. Then, for $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$, expected risk is defined as $F_{\mathcal{D}}(\theta) := \mathbb{E}_{Z \sim \mathcal{D}} [f(Z, \theta)]$. For an SCO problem to be learnable, one often assumes that the loss function f is *Lipschitz* and the *diameter* of the space Θ is bounded, both of which can be measured w.r.t. different norms. These bounds govern the behavior of learnability, but they can be measured in different geometries. A

canonical class of SCO problems is induced by the ℓ_p norms, in which case we assume that Θ has bounded ℓ_p-diameter and $f(\cdot, z)$ is ℓ_p-Lipschitz, for a fixed $p \in [1, \infty]$.

1.1 Contributions

In this paper, we establish a fundamental tradeoff between traceability and excess risk for algorithms in the context of SCO in general geometries. Some settings in which tracing is not possible are already well-understood: in the excess risk regime where DP [DMNS06] is possible, no samples can be traced. Due to this observation, the problem of traceability is only meaningful outside the DP risk regime. More formally, let us define minimax statistical and DP excess risks in ℓ_p SCO. Specifically, for a family of ℓ_p Lipschitz problems $\mathcal{L}_p^d$, we let

$$\alpha_{\mathsf{stat}}(p, n) = \max_{\mathcal{P} \in \mathcal{L}_p^d} \min_{\mathcal{A}_n} \max_{\mathcal{D}} \left\{ \mathbb{E}\left[F_{\mathcal{D}}(\mathcal{A}_n(S_n))\right] - \inf_{\theta \in \Theta} F_{\mathcal{D}}(\theta) \right\} \quad \text{and} \tag{1}$$

$$\alpha_{\mathsf{DP}}(p, n) = \max_{\mathcal{P} \in \mathcal{L}_p^d} \min_{(\varepsilon,\delta)\text{-DP-}\mathcal{A}_n} \max_{\mathcal{D}} \left\{ \mathbb{E}\left[F_{\mathcal{D}}(\mathcal{A}_n(S_n))\right] - \inf_{\theta \in \Theta} F_{\mathcal{D}}(\theta) \right\}, \tag{2}$$

where, for concreteness, we take $\varepsilon = 0.01$ and $\delta = 1/n^2$ in the above.

Main Contribution: We show that every sample-efficient algorithm that achieves an excess risk outside the DP regime (that is, $\alpha = o(\alpha_{\mathsf{DP}})$) is traceable with recall proportional to the number of samples. The precise statement of our main contribution varies based on the geometry of SCO:

Tracing when $p = 1$. For the case $p = 1$, we show that any learner whose excess risk is better, by a small polynomial factor, than the best risk attainable by a DP algorithm with constant ε and $\delta = 1/n^2$ must be traceable. Moreover, we give an essentially optimal lower bound on the number of samples that can be traced. In more detail, we show that there exists an ℓ_1-SCO problem such that, if an algorithm achieves risk of

$$\alpha \lesssim \frac{\alpha_{\mathsf{DP}}}{d^{0.01} \log^2(n)},$$

then $\Omega(\log(d)/\alpha^2))$ of the training samples can be traced (see Theorem 2.6). We note that the choice of the constant 0.01 above is arbitrary. It is instructive to compare our results to [DSSUV15]. While the settings of mean estimation and SCO are generally different, our lower bound for ℓ_1 geometry also extends to mean estimation in ℓ_∞ norm (see Corollary 2.9 for a formal argument). In both settings, the sample complexity scales like $\log(d)/\alpha^2$, however, [DSSUV15] showed traceability of only $1/\log(d)$ fraction of the samples. On the other hand, in our work we show that there is no meaningful gap between sample complexity and traceability, and every sample-efficient algorithm outside the DP regime must memorize a constant fraction of its sample. Notably, our results also imply that traceability is *dimension-dependent* in this setup.

Tracing when $p \in (2, \infty)$. For ℓ_p SCO with $p \geq 2$, in Theorem 2.7, we show that for

$$\alpha \lesssim \frac{\underline{\alpha}_{\mathsf{DP}}}{\log(n)},$$

where $\underline{\alpha}_{\mathsf{DP}}$ is set as $\underline{\alpha}_{\mathsf{DP}} = \Theta\left(d/n^2\right)^{1/p}$ we can construct an SCO problem such that, if a learner achieves a risk of α, then $\Omega(1/\alpha^p)$ of its samples are traceable. Note that, the non-private sample complexity of learning for $p > 2$ is precisely $\Theta(1/\alpha^p)$ in the relevant parameter regime,[**] i.e., the number of traced out samples is of the order of the sample complexity. Note that, the optimal DP risk in this setup constitutes an open problem, and the quantity $\underline{\alpha}_{\mathsf{DP}}$ above need not be the optimal DP risk in this setting. Nevertheless, this quantity can be shown to be a *lower bound* on the optimal DP risk (in the regime $\varepsilon \in \Theta(1)$). We extend this result to other regimes of ε, and another important contribution of our work is proving such DP lower bounds.

[**]We point out that, in general, the sample complexity for $p \geq 2$ scales as $1/n^{1/p} \wedge d^{1/2-1/p}/\sqrt{n}$, however, $\alpha_{\mathsf{stat}} \ll \alpha_{\mathsf{DP}}$ only when $d \gtrsim n$. Thus, the question of traceability is only non-vacuous in the overparameterized regime.

p	Recall	Range of α	Sample complexity	Minimax DP rate	Refs.
1	$\frac{\log(d)}{\alpha^2}$	$\left(\sqrt{\frac{\log(d)}{n}},\ \frac{d^{0.49}}{n\sqrt{\log(1/\xi)}}\right)$	$\frac{\log(d)}{\alpha^2}$	$\sqrt{\frac{\log(d)}{n}} + \frac{\sqrt{d}}{\varepsilon n}$	Thm. 2.6
$(1,2]$	$\frac{1}{\alpha^2}$	$\left(\sqrt{\frac{1}{n}},\ \frac{\sqrt{d}}{n\sqrt{\log(1/\xi)}}\right)$	$\frac{1}{\alpha^2}$	$\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\varepsilon n}$	Thm. 2.5
$[2,\infty)$	$\frac{1}{\alpha^p}$	$\left(\min\left\{\frac{1}{n^{1/p}}, \frac{d^{1/2-1/p}}{\sqrt{n}}\right\},\ \left(\frac{d}{n^2\log(1/\xi)}\right)^{1/p}\right)$	$\frac{1}{\alpha^p}$ †	Open	Thm. 2.7

Table 1: Summary of traceability results. All results are stated up to constants. The sample complexity bounds are implied by Theorem A.1. Minimax DP rates are known due to [BFTG19; BGN21; AFKT21; GLLST23] and are displayed up to log factors and with $\delta = 1/n^2$. (†) Although, in general, the sample complexity in this setting is a minimum of two terms, within the stated range of α, the term $1/\alpha^p$ dominates.

In particular, we provide an improved lower bound on DP-SCO under ℓ_p geometries for $p > 2$ in the high dimensional regime, i.e., $d \geq \varepsilon n$, which is arguably the most interesting regime as it is more relevant for the modern ML applications. Specifically, we show, in Theorem 2.8, that for all $\varepsilon < 1$ and small δ we can construct a problem such that for every (ε, δ)-DP algorithm, $\mathcal{A}_n$, there exists a data distribution such that:

$$\mathbb{E}_{S_n \sim \mathcal{D}^{\otimes n}, \hat{\theta} \sim \mathcal{A}_n(S_n)} \left[F_{\mathcal{D}}(\hat{\theta}) \right] - \inf_{\theta \in \Theta} F_{\mathcal{D}}(\theta) \gtrsim \left(\frac{d}{n^2 \varepsilon^2} \right)^{1/p}.$$

In particular, the above implies that when $d \geq \varepsilon n$, the risk due to privacy dominates the statistical risk. This result improves upon all previous best bounds in the literature when $d \geq \varepsilon n$ [ABGMU22; LLL24]. In particular, Theorem 3.1 of [LLL24] gives a lower bound $\sqrt{d/n^2\varepsilon^2}$, which is weaker than our lower bound for every $p > 2$. Corollary 4 of [ABGMU22] gives a lower bound of $\min\left\{\left(\frac{1}{\varepsilon n}\right)^{\frac{1}{p}}, \frac{d^{1-1/p}}{\varepsilon n}\right\}$ which is weaker than our lower bound for $d \geq \varepsilon n$.

Tracing when $p \in (1,2]$. For each $p \in (1,2]$, we show that there exists an ℓ_p SCO problem such that, if an algorithm achieves excess risk of $\alpha \lesssim \frac{\alpha_{\mathrm{DP}}}{\log^2(n)}$, then $\Omega(1/\alpha^2)$ of its samples can be traced (see Theorem 2.5). This result uncovers a fundamental dichotomy between traceability and privacy in ℓ_p SCO. It is known that $p \in (1,2]$, $\Theta(1/\alpha^2)$ is precisely the sample complexity of learning ℓ_p-Lipschitz problems [AWBR09]. We note that α_{DP} for $p = 2$ is known due to [BST14]. However, as we discuss in Appendix B.1, combining [BST14] with the tracing results of [DSSUV15] does not give the optimal tracing of $\Theta(1/\alpha^2)$ samples.

1.1.1 Traceability beyond SCO: PAC Learning

A natural question is whether a similar phenomenon holds true for other learning setups. Consider the setting of binary classification PAC learning. We show that, for every class with VC dimension bounded by d_{vc}, the recall of every tracer is in $O(d_{\mathsf{vc}} \log^2(n))$, i.e., it is at most a small fraction of the training sample provided $n \gg d_{\mathsf{vc}}$. Since many such classes, including the class of thresholds, are not privately learnable [BNSV15; ALMM19; BLM20], the sharp transitions between privacy and traceability does not hold in PAC classification. We also point out that for the class of thresholds, we can remove the $\log^2(n)$ factor from the recall upper bound. See Appendix H.

1.2 Technical contributions

Our technical contributions are elaborated on in Section 2.3. In essence, our technical novelties are twofold. First, we present a novel *sparse fingerprinting lemma* that, intuitively, shows that learners over sparse domains must be correlated to their samples. The key novelty of this result is that the *correlation is inversely proportional to the sparsity parameter.* This feature is not present in prior work, since fingerprinting lemmas are most often applied for learners/estimators over a hypercube domain.

Second, armed with this new fingerprinting result, we present a generic conversion result using a notion of a *subgaussian trace value*, which converts any lower bound on *correlation*

with the samples into a *number* of samples that can be traced. While it is well-appreciated by prior work that, conceptually, a fingerprinting lower bound implies a traceability lower bound, proving results for our setting of SCO involves complicated sparse domains embedded into ℓ_p balls. This makes it more technically challenging to prove the necessary concentration phenomena holds for a tracer over the corresponding domain, which motivates us to restrict our attention to tracers that induce a *subgaussian process* over the domain.

1.3 Related Work

Our work is most similar in spirit to [DSSUV15; ADHLR24]. Our work builds on top of these results on a number of fronts. A key distinct aspect of our approach is the difference in the structure of hard problems and the new sparse fingerprinting lemma. Also, our generic traceability theory of *subgussian trace value* (Section 2.3) provides an abstract treatment of the approach in [DSSUV15]. Our approach allows to seamlessly convert fingerprinting lemmas into traceability results and even non-private sample complexity lower bound.

Our work also makes progress towards closing the gap regarding the optimal excess error for ℓ_p DP-SCO for $p > 2$. The best known upper bounds for DP-SCO in ℓ_p geometry for $p > 2$ are due to [BGN21; GLLST23], and Theorem 2.8 is the best lower bound.

To put our *sparse fingerprinting lemma* into the context of prior work, it can be seen to generalize the results of [SU17] to sparse sets. Another "sparse fingerprinting lemma" in the literature is given by [CWZ23]. Our results are distinct by the way sparsity enters the lemmas: in [CWZ23] the *mean vector* is sparse (and data is dense), and in our case, the mean is dense and the *data vectors* are sparse. The proof techniques also differ substantially. Our sparse fingerprinting lemma is also an example of a fingerprinting lemma for the setting where the coordinates of the data vector are not independent, similar to [KMS22; LT24]. Additional related work is discussed in Appendix B.

2 Problem Setup and Main Results

We begin by some definitions. For a (measurable) space $\mathcal{R}$, $\mathcal{M}_1(\mathcal{R})$ denotes the set of all probability measures on $\mathcal{R}$. In SCO, an α-learner is defined to be a learner whose expected *excess risk* is bounded by α. A formal definition is given below.

Definition 2.1 (α-learner)**.** Fix $\alpha > 0$, $n \in \mathbb{N}$ and SCO problem $(\Theta, \mathcal{Z}, f)$. We say $\mathcal{A}_n : \mathcal{Z}^n \to \mathcal{M}_1(\Theta)$ is an α-learner for $(\Theta, \mathcal{Z}, f)$ iff for every $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$, we have $\mathbb{E}_{S_n \sim \mathcal{D}^{\otimes n}, \hat{\theta} \sim \mathcal{A}_n(S_n)} \left[F_{\mathcal{D}}(\hat{\theta}) \right] - \inf_{\theta \in \Theta} F_{\mathcal{D}}(\theta) \leq \alpha$.

In our work, we focus on learning *Lipschitz-bounded* families of problems, which are defined below. For every $p \in [1, \infty]$, let $\mathcal{B}_p(r) = \{\theta \in \mathbb{R}^d : \|\theta\|_p \leq r\}$ be the unit ball in ℓ_p norm.

Definition 2.2 (Lipschitz-bounded problems)**.** Fix $p \in [1, \infty]$, and let $d < \infty$ be a natural number. We let $\mathcal{L}_p^d$ denote the set of all ℓ_p-Lipschitz-bounded SCO problems in d dimensions. Namely, $\mathcal{P} = (\Theta, \mathcal{Z}, f) \in \mathcal{L}_p^d$ iff (i) $\Theta \subset \mathcal{B}_p(1)$, and (ii) for every $\theta_1, \theta_2 \in \Theta$ and $z \in \mathcal{Z}$, we have $|f(z, \theta_1) - f(z, \theta_2)| \leq \|\theta_1 - \theta_2\|_p$.

2.1 Tracing

The key notion we study here is *tracing*, and we next introduce our framework for traceability. We consider families of tracers that assign each candidate point a real-valued score capturing how likely it is to have been seen during training. Then, the tracer converts these scores into binary In or Out decisions by thresholding the score. Intuitively the score corresponds to the likelihood of the event that the learner saw a data point during training.

Definition 2.3 (Tracer)**.** Fix data space $\mathcal{Z}$ and parameter space Θ. A tracer's strategy is a tuple of $\mathcal{T} = (\phi, \mathcal{D})$ where $\phi : \Theta \times \mathcal{Z} \to \mathbb{R}$ and $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$.

Definition 2.4 ((ξ, m)-traceability)**.** Let $n \in \mathbb{N}$, $\xi \in (0, 1)$, and $m \in \mathbb{N}$. We say a learning algorithm $\mathcal{A}_n$ is (ξ, m)-traceable if there exists a tracer $(\phi, \mathcal{D})$ and $\lambda \in \mathbb{R}$ such that, if $(Z_0, Z_1, \ldots, Z_n) \sim \mathcal{D}^{\otimes (n+1)}$ and $\hat{\theta} \sim \mathcal{A}_n(Z_1, \ldots, Z_n)$, we have (i) Soundness: $\Pr\left(\phi(\hat{\theta}, Z_0) \geq \lambda\right) \leq \xi$, and (ii) Recall: $\mathbb{E}\left[\left|\{i \in [n] : \phi(\hat{\theta}, Z_i) \geq \lambda\}\right|\right] \geq m$.

2.2 Main Results

2.2.1 Traceability of α-Learners

In this section, we discuss our traceability results for accurate learners in ℓ_p geometries. First, we will state a result that applies to $p \in [1, 2)$, and then present its slight refinement for $p = 1$. We will then present our result for $p \geq 2$. See Appendices F.1 to F.3 for proofs.

Theorem 2.5. *There exists a universal constant $c > 0$ such that, for all $p \in [1, 2)$, if d, n, $\xi \in (0, 1/e)$, and $\alpha > 0$ are such that*

$$\frac{c}{\sqrt{n}} \leq \alpha \leq \min\left\{c \cdot \sqrt{\frac{d}{n^2 \log(1/\xi)}},\ \frac{1}{6}\right\}, \tag{3}$$

then there exist an ℓ_p SCO problem that every α-learner is (ξ, m)-traceable with $m \in \Omega\left(\alpha^{-2}\right)$.

Note that the upper bound on α in Equation (3) is precisely the optimal DP excess risk for $\varepsilon \in \Theta(1)$ and $p \in [1, 2]$ [AFKT21; BGN21], and the lower bound is precisely the optimal non-private risk (except $p = 1$; see Theorem A.1). Moreover, for $p \in (1, 2]$, the lower bound on m *exactly* matches the statistical sample complexity.

As mentioned above, for $p = 1$, the lower bound on recall in Theorem 2.5 is less than sample complexity by a factor of $\log(d)$. This prompts us to establish the following refinement

Theorem 2.6. *There exists a universal constant $c > 0$ such that, if d is large enough and n, $\xi \in (0, 1/e)$, and $\alpha > 0$ are such that*

$$c \cdot \sqrt{\frac{\log(d)}{n}} \leq \alpha \leq \min\left\{c \cdot \frac{d^{0.49}}{n\sqrt{\log(1/\xi)}},\ \frac{1}{8}\right\}, \text{ then} \tag{4}$$

there exists a ℓ_1 SCO problem that every α-learner is (ξ, m)-traceable with $m \in \Omega\left(\log(d)/\alpha^2\right)$.

Note that the upper bound in Equation (4) is slightly stronger than in Equation (3); however, the lower bound on recall now matches the sample complexity of learning in ℓ_1 geometry. We now present a result for $p \geq 2$.

Theorem 2.7. *There exists a universal constant $c > 0$ such that, for all $p \in [2, \infty)$, if d, n, $\xi \in (0, 1/e)$, and $\alpha > 0$ are such that*

$$\frac{1}{6} \cdot \min\left\{\frac{1}{n^{1/p}},\ \frac{d^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{n}}\right\} \leq \alpha \leq \min\left\{c \cdot \left(\frac{d}{n^2 \log(1/\xi)}\right)^{1/p},\ \frac{1}{6}\right\}, \text{ then} \tag{5}$$

there exist an ℓ_p SCO problem such that every α-learner is (ξ, m)-traceable with $m \in \Omega\left(1/(6\alpha)^p\right)$.

For $p \in (2, \infty)$, our results have a different implication, showing that all sufficiently accurate learners need to memorize a number of samples on the order of the sample complexity. However, in this case, the upper bound in Equation (5) need not be the optimal DP risk. Instead, it provides a *lower bound* on the optimal DP risk, as we will see next.

2.2.2 Improved DP-SCO Lower Bound for $p > 2$

Theorem 2.8. *Let $p \in [2, \infty)$. There exist a universal constant $c > 0$ and an ℓ_p SCO problem $\mathcal{P} = (\Theta, \mathcal{Z}, f)$ such that every (ε, δ)-DP learner of $\mathcal{P}$ with $\varepsilon \leq 1$ and $\delta \leq c/n$ satisfies,*

$$\alpha \geq c \cdot \min\left\{\left(\frac{d}{\varepsilon^2 n^2}\right)^{\frac{1}{p}}, \frac{d^{1-1/p}}{\varepsilon n}, 1\right\}.$$

2.2.3 Consequences for mean estimation

Consider the setting of mean estimation in ℓ_∞ norm as in [DSSUV15]. Our results in Theorem 2.6 extend almost verbatim to this setting.

Corollary 2.9. *Let $\mathcal{Z} = \{\pm 1\}^d$, and suppose an estimator is given such that, given access to i.i.d. samples $Z_1, \ldots, Z_n \in \mathcal{Z}$, outputs $\hat{\mu}$ with $\mathbb{E}\left\|\hat{\mu} - \mathbb{E}[Z_1]\right\|_\infty \leq \alpha/2$. Then, there exists a universal constant $c > 0$ such that, if d is large enough and n, $\xi \in (0, 1/e)$, and $\alpha > 0$ satisfy Equation (4), then the estimator $\hat{\mu}$ is (ξ, m)-traceable with $m \in \Omega\left(\log(d)/\alpha^2\right)$.*

2.3 Roadmap of the proof

Our proofs rely on introducing two key technical elements that allow us to generalize tracing techniques to general ℓ_p setups. The first element is a generic conversion result involving a complexity notion which we term *the subgaussian trace value* of a problem. As we show in the proof of Theorem 2.8, we can use the subgaussian trace value to prove traceability results over general domains, establish DP sample complexity lower bounds, and, even to recover *non-private* sample complexity lower bounds. While the connection between the first two aspects is well-known [FS17], we find the ability of trace value to recover *non-private* lower bounds surprising.

Our second technical contribution concerns techniques for lower bounding the subgaussian trace value, which we accomplish through several novel *fingerprinting lemmas*. Previous works used the standard fingerprinting lemma, where the learner observes points on a hypercube, to lower bound DP and traceability in ℓ_2 geometry. However, when moving to general ℓ_p geometries, this setup no longer captures the hardest settings to learn. For instance, for $p > 2$, canonical instances of hard problems involve data drawn from sparse sets [AWBR09]. We thus prove new fingerprinting lemmas that enable us to leverage our framework in such settings. These fingerprinting lemmas are then applied to carefully constructed instances of hard problems, and we show that every accurate learner of these problems is traceable.

3 General framework: subgaussian trace value

We next describe more formally the framework of subgaussian tracers. For a random variable X, the subgaussian norm of X is the quantity $\|X\|_{\psi_2} := \inf\{t \colon \mathbb{E}\left[\exp(X^2/t^2)\right] \leq 2\}$ [Ver18]. We use the following definition of a subgaussian process:

Definition 3.1 (Subgaussian process)**.** We call an indexed collection of random variables $\{X_\theta\}$ a σ-subgaussian process w.r.t a metric space $(\Theta, \|\cdot\|)$ if for every $\theta, \theta' \in \Theta$, we have (i) $\|X_\theta - X_{\theta'}\|_{\psi_2} \leq \sigma \|\theta - \theta'\|$, and (ii) $\|X_\theta\|_{\psi_2} \leq \sigma \operatorname{diam}_{\|\cdot\|}(\Theta)$.

For origin symmetric convex body Θ, let $\|\cdot\|_\Theta$ denote the Minkowski norm w.r.t. Θ, that is $\|x\|_\Theta := \inf\{\lambda > 0 \colon x \in \lambda\Theta\}$. If Θ is not convex or not origin symmetric, we let $\|\cdot\|_\Theta$ be the Minkowski norm w.r.t. convex hull of $(\Theta \cup -\Theta)$. Note that $\|\cdot\|_\Theta$ is the minimal norm to contain Θ in its unit ball.

Definition 3.2 (Subgaussian tracer)**.** Fix $\kappa \in \mathbb{R}$ to be a constant, and let Θ be a convex body. We let $\mathfrak{T}_\kappa$ be the class of subgaussian tracers at scale $\kappa > 0$, that is, a tracer $(\phi, \mathcal{D}) \in \mathfrak{T}_\kappa$ iff

(i) $\{\phi(\theta, Z)\}_{\theta \in \Theta}$ where $Z \sim \mathcal{D}$ is a 1-subgaussian process w.r.t. $(\Theta, \|\cdot\|_\Theta)$.

(ii) $|\phi(\theta, z)| \leq \kappa$ for all $\theta \in \Theta$ and $z \in \mathcal{Z}$.

Definition 3.3 (Subgaussian trace value)**.** Fix $n \in \mathbb{N}$, $\alpha \in [0, 1]$, and $\kappa \in \mathbb{R}$. Consider an arbitrary SCO problem $\mathcal{P} = (\Theta, \mathcal{Z}, f)$. Let $\mathfrak{T}_\kappa$ be as in Definition 3.2. Then, we define the subgaussian trace value of problem $\mathcal{P}$ by

$$\operatorname{Tr}_\kappa(\mathcal{P}; n, \alpha) = \inf_{\alpha\text{-learner}\mathcal{A}_n} \sup_{\mathcal{T} = (\phi, \mathcal{D}) \in \mathfrak{T}_\kappa} \mathbb{E}_{S_n = (Z_1, \ldots, Z_n) \sim \mathcal{D}^{\otimes n}, \hat{\theta} \sim \mathcal{A}_n(S_n)} \left[\frac{1}{n} \sum_{i \in [n]} \phi(\hat{\theta}, Z_i)\right].$$

where the inf is taken over all $\mathcal{A}_n$ that achieve excess risk $\leq \alpha$ on $\mathcal{P}$ with n samples.

Traceability via subgaussian trace value. The subgaussian trace value characterizes the average score the pair $(\phi, \mathcal{D})$ assigns to the data points in the training set. However, the definition of recall in Definition 2.3 requires characterizing the *number* of samples in the training set that takes a large value. The former can be converted into the latter, provided the sum of squared scores of samples is not too large. A formal statement, which is a

consequence of Paley–Zygmund inequality, can be found in Lemma A.11. In the next lemma, we show how to control the sum of squares of the $\phi(\hat{\theta}, Z_i)$ using the subgaussian assumption.

Lemma 3.4. *Fix $n, d \in \mathbb{N}$. Suppose $\Theta \subset \mathbb{R}^d$ is a subset of a unit ball in some norm $\|\cdot\|$. Let $\phi\colon \Theta \times \mathcal{Z} \to \mathbb{R}$ and $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ be such that, as $Z \sim \mathcal{D}$, $\{\phi(\theta, Z)\}$ is a σ-subgaussian process w.r.t. $(\Theta, \|\cdot\|)$. Let $(Z_1, \dots, Z_n) \sim \mathcal{D}^{\otimes n}$. Then, there is a constant $C > 0$, such that*

$$\Pr\left[\sup_{\theta\in\Theta}\sqrt{\sum_{i=1}^{n}\left[\phi(\theta, Z_i)\right]^2} \le C\sigma\left(\sqrt{n} + \sqrt{d} + t\right)\right] \ge 1 - 4\exp(-t^2), \quad \forall t \ge 0.$$

Equipped with this lemma, in the next theorem, we show that if, the subgaussian trace value of a problem is large, then every α-learner is traceable.

Theorem 3.5. *Fix $n \in \mathbb{N}$, $d \in \mathbb{N}$, $\kappa > 0$ and $\alpha \in [0, 1]$. Consider an arbitrary SCO problem $\mathcal{P} = (\Theta, \mathcal{Z}, f)$. Let $T = \mathrm{Tr}_\kappa(\mathcal{P}; n, \alpha)$ be the subgaussian trace value of $\mathcal{P}$. Then, for some constant $c > 0$, every α-learner $\mathcal{A}_n$ is (ξ, m)-traceable with*

$$\xi = \exp(-cT^2), \quad m = c\left[\frac{n^2T^2}{n+d} - \frac{16\kappa^2 n}{\exp(n+d)}\right].$$

Privacy lower bounds via subgaussian trace value. In the next theorem, we show that the notion of subgaussian trace value directly lower bounds the best privacy parameters achievable by a DP algorithm. The proof is based on [FS17].

Theorem 3.6. *There exists a universal constant $c > 0$, such that the following holds. Fix $p \in [1, \infty)$, $n \in \mathbb{N}$, $d \in \mathbb{N}$, $\alpha \in [0, 1]$, $\kappa > 0$ $\varepsilon > 0$, and $\delta \in [0, 1]$. Consider an arbitrary SCO problem $\mathcal{P} = (\Theta, \mathcal{Z}, f)$ in $\mathbb{R}^d$. Let $T = \mathrm{Tr}_\kappa(\mathcal{P}; n, \alpha)$ be the subgaussian trace value of problem $\mathcal{P}$. Then, for every (ε, δ)-DP α-learner $\mathcal{A}_n$, we have $\exp(\varepsilon) - 1 \ge c\,(T - 2\delta\kappa)$.*

Non-private sample complexity via subgaussian trace value. Surprisingly, if we directly use *subgaussian trace value*, we can recover optimal sample complexity bounds *for all $p \in [1, \infty)$ and all regimes of (d, α)*, thus unifying traceability with private and non-private sample complexity lower bounds. While we detail the argument formally in Appendix G, we consider here a helpful example of ℓ_2 geometry. First, it can be shown that we always have $\mathrm{Tr}(\mathcal{P}; n, \alpha) \lesssim \sqrt{d/n}$ for arbitrary problem $\mathcal{P}$ (see Proposition G.1). Also, we will later show that, for every $\alpha > 0$, there exist an ℓ_2 problem $\mathcal{P}$ with $\mathrm{Tr}(\mathcal{P}; n, \alpha) \gtrsim \sqrt{d}/n\alpha$ (see Theorem 5.1). Combining these two inequalities gives $n \gtrsim 1/\alpha^2$, which is optimal.

4 The sparse fingerprinting lemma

By introducing the notion of subgaussian trace value, we have reduced the problems of traceability and privacy lower bounds to the question of lower bounding the subgaussian trace value. Now, we discuss the techniques to lower bound subgaussian trace value. The proofs can be found in Appendix D. Due to space limitation, we only discuss the details for the case of $p > 1$ and present the details of $p = 1$ in Appendix E.2.

For ℓ_2 geometry, one can lower bound subgaussian trace value using the classical fingerprinting lemma in [DSSUV15]. While this strategy leads to traceability results in ℓ_2 geometry, examples of hard problems for ℓ_p geometry with $p > 2$ are those with *sparse* sets $\mathcal{Z}$ (e.g., as in [AWBR09]). This motivates us to prove the following *sparse fingerprinting lemma*, which is another important contribution of our work. For a vector $x \in \mathbb{R}^d$, let $\mathrm{supp}(x)$ be the set of its non-zero coordinates and denote $\|x\|_0 = |\,\mathrm{supp}(x)|$.

Definition 4.1 (Sparse distributions family)**.** Fix $d \in \mathbb{N}$, $k \in [d]$ and $\mu \in [-k/d, k/d]^d$. Consider the mixture distribution on $\mathcal{Z}_k = \{z \in \{0, \pm 1\}^d : \|z\|_0 = k\}$ given by, for all $z \in \mathcal{Z}_k$, $\mathcal{D}_{\mu,k}(z) = \mathbb{E}_{J\sim\mathsf{unif}\left(\binom{[d]}{k}\right)}\left[P_{\mu,k,J}(z)\right]$, where

$$P_{\mu,k,J}(z) = \mathbb{1}(\mathrm{supp}(z) = J) \cdot \prod_{j\in J}\left(\frac{1 + (d/k)\cdot\mu^j z^j}{2}\right).$$

Note that, in particular, $\mathbb{E}_{Z\sim\mathcal{D}_{\mu,k}}[Z] = \mu$. Intuitively, one can think of *sampling* from $\mathcal{D}_{\mu,k}$ using the following procedure: (i) sample the support coordinates $J \sim \mathsf{unif}\binom{[d]}{k}$, (ii) for each $j \in J$, sample Z^j from $\{\pm 1\}$ with mean $\frac{d}{k}\mu^j$ independently, (iii) for each $j \notin J$, set $Z^j = 0$.

With this distribution family at hand, we may state the sparse fingerprinting lemma. For $x, y \in \mathbb{R}^d$ and a subset $R \subseteq [d]$ of coordinates, we use $\langle \cdot, \cdot \rangle_S$ to denote the inner product $\langle x, y \rangle_R := \sum_{i \in R} x_i y_i$. Also, for $\alpha, \beta, \gamma > 0$, let $\mathsf{s\text{-}beta}_{[-\gamma,\gamma]}(\alpha, \beta)$ be the *symmetric* beta-distribution, i.e., beta distribution with parameters α, β scaled and shifted to have support $[-\gamma, \gamma]$ (see Definition A.13).

Lemma 4.2 (Sparse fingerprinting)**.** *Fix $d, n \in \mathbb{N}$ and let $k \in [d]$. For each $\mu \in [-k/d, k/d]^d$, let $\mathcal{Z}_k$ and $\mathcal{D}_{\mu,k}$ be as in Definition 4.1 . Let $\pi = \mathsf{s\text{-}beta}_{[-k/d,k/d]}(\beta, \beta)^{\otimes d}$ be a prior and set*

$$\phi_\mu(\theta, Z) := \left\langle \theta, \left(Z - \frac{d}{k}\mu \right) \right\rangle_{\mathrm{supp}(Z)}.$$

Then, for every learning algorithm $\mathcal{A}_n : \mathcal{Z}^n \to \mathcal{M}_1(\mathbb{R}^d)$ with sample $S_n = (Z_1, \ldots, Z_n)$,

$$\mathbb{E}_{\mu\sim\pi}\mathbb{E}_{S_n\sim\mathcal{D}_{\mu,k}^{\otimes n},\hat{\theta}\sim\mathcal{A}_n(S_n)}\left[\sum_{i=1}^n \phi_\mu(\hat{\theta}, Z_i)\right] = \frac{2\beta d}{k}\mathbb{E}_{\mu\sim\pi}\left\langle \mu, \mathbb{E}_{S_n\sim\mathcal{D}_{\mu,k}^{\otimes n},\hat{\theta}\sim\mathcal{A}_n(S_n)}[\hat{\theta}]\right\rangle.$$

The key novelty of this lemma is that it provides a way to study the correlation between a learner's output and training samples on sparse sets $\mathcal{Z}_k$. An important and distinctive feature of this result is that the right-hand side scales by a factor of d/k, highlighting the fact that sparse problems correspond to greater subgaussian trace values. Intuitively, this stems from the fact that each coordinate is *seen* fewer times by the learning algorithm, meaning it must retain more information from each training sample in order to learn accurately. Additionally, for the special case $k = d$, the result precisely recovers the fingerprinting lemma from [SU17].

5 Final steps: bounding the subgaussian trace value for hard problems

Finally, we go over the construction of hard problems. To illustrate the difficulty of problem constructions, we give an example of a problem that requires many samples to learn but nevertheless is not traceable. Consider learning over ℓ_1 ball with linear loss. Let

$$\Theta = \mathcal{B}_1(1), \quad \mathcal{Z} = \{\pm 1\}^d, \quad f(\theta, Z) = -\langle \theta, Z \rangle. \tag{6}$$

Consider a difficult set of distributions $\{\mathcal{D}_i\}_{i=1}^d$ where $\mathcal{D}_i$ is a product distribution on $\mathcal{Z}$ and has mean α on coordinate i and mean zero on all other coordinates. It can be shown this problem requires $\Theta(\log(d)/\alpha^2)$ samples to learn up to risk of $\alpha/3$, and ERM is an optimal learner. However, after seeing $\Theta(\log(d)/\alpha^2)$ samples from $\mathcal{D}_i$, the ERM takes the value $\hat{\theta} = e_i$ w.h.p., which is also the *population* risk minimizer. In other words, it becomes impossible to trace out any specific samples on which $\hat{\theta}$ was trained.

Generic construction for $p \in (1, \infty)$. As mentioned above, to obtain optimal results for $p > 2$, problems constructed need to be sparse, and the main subtlety in our constructions is choosing the sparsity parameter. For some $k \in [d]$ to be chosen later, consider the following ℓ_p-Lipschitz problem $\mathcal{P}_{k,p}$.

$$\Theta = \mathcal{B}_\infty(d^{-1/p}), \quad \mathcal{Z} = \{z \in \{0, \pm 1\}^d \colon \|z\|_0 = k\}, \quad f(\theta, z) = -k^{-1/q}\langle \theta, z \rangle. \tag{7}$$

Here, the parameter space Θ is the largest ℓ_∞ ball inscribed into the unit ℓ_p ball, and q is the Hölder conjugate of p, i.e., $\frac{1}{p} + \frac{1}{q} = 1$. The next step is to show that α-learners for the above problem must be correlated with the mean of the unknown data distribution. Let $\mathcal{D}$ be a distribution with mean μ, and suppose $\mathcal{A}_n$ is an α-learner for Equation (7). Then,

$$\mathbb{E}_{S_n\sim\mathcal{D}^{\otimes n},\hat{\theta}\sim\mathcal{A}_n(S_n)}\left[\left\langle \mu, \hat{\theta}\right\rangle\right] \geq \sup_{\theta\in\Theta}\langle \mu, \theta \rangle - k^{1/q}\cdot\alpha = d^{-1/p}\|\mu\|_1 - k^{1/q}\alpha.$$

Now, we apply the sparse fingerprinting lemma (Lemma 4.2). A key step is choosing the scale $\beta \geq 1$ of the beta-prior. On the one hand, β should be small enough to guarantee $\mathbb{E}d^{-1/p}\|\mu\|_1 > k^{1/q}\alpha$, so that the above lower bound is non-vacuous. On the other hand, taking β too small decreases the sample complexity of learning the problem, thus, disallowing

the desired level of recall. The optimal choice is $\beta \propto \alpha^{-2} \cdot (k/d)^{1/p}$, as long as this quantity is ≥ 1. This choice yields

$$\mathrm{Tr}_\kappa(\mathcal{P}_{k,p}; n, \alpha) \geq \mathbb{E}_{S_n \sim \mathcal{D}^{\otimes n}, \hat{\theta} \sim \mathcal{A}_n(S_n)} \left[\frac{1}{n} \sum_{i=1}^{n} \phi(\hat{\theta}, Z_i) \right] \gtrsim \frac{d^{1-1/p}}{k^{1/2-1/p} n\alpha},$$

where $\kappa \in \Theta(1)$, and, for some universal constant $c > 0$, we let

$$\phi(\theta, Z) := \frac{c d^{1/p}}{\sqrt{k}} \left\langle \theta, \left(Z - \frac{d}{k}\mu \right) \right\rangle_{\mathrm{supp}(Z)}.$$

Note that the $d^{1/p}/\sqrt{k}$ scaling ensures ϕ induces a 1-subgaussian process. Finally, it remains to choose a suitable value for k, for each pair (p, α). Recall the definition of $\mathcal{P}_{k,p}$ from Equation (7).

Theorem 5.1. *Let $\mathcal{P}_{k,p}$ be the family of problems described in Equation (7). There exist universal constants $c_1, c_2 > 0$ such that, for all $\alpha \in (0, 1/6]$ and $d \in \mathbb{N}$, the following subgaussian trace value lower bounds hold for all $p \in [1, \infty)$ and $\kappa \leq c_1\sqrt{d}$:*

(i) For $p \leq 2$ and $k = d$, we have $\mathrm{Tr}_\kappa(\mathcal{P}_{k,p}; n, \alpha) \geq c_2 \frac{\sqrt{d}}{n\alpha}$.

(ii) For $p \geq 2$ and $k = (6\alpha)^p d \vee 1$, we have $\mathrm{Tr}_\kappa(\mathcal{P}_{k,p}; n, \alpha) \geq c_2 \left[\frac{\sqrt{d}}{n(6\alpha)^{p/2}} \wedge \frac{d^{1-1/p}}{n\alpha} \right]$.

Using the reduction Theorem 3.5, the above establishes Theorems 2.5 and 2.7.

Refinement for $p = 1$. While the above construction also yields a traceability result for $p = 1$, it is suboptimal for the following simple reason: for $k = d$, the problem in Equation (7) only requires $\Theta(1/\alpha^2)$ samples to learn, thus, it is impossible to trace out $\Omega(\log(d)/\alpha^2)$ samples. On the other hand, the problem in Equation (6) requires $\Theta(\log(d)/\alpha^2)$ samples to learn but is not traceable. The intuition we follow here is to modify the construction in Equation (7) to make Θ "look" more like an ℓ_1-ball to drive up the sample complexity while still avoiding the counterexample with an ERM learner from the beginning of the section. In particular, we consider the following ℓ_1-problem,

$$\Theta = \mathcal{B}_1(1) \cap \mathcal{B}_\infty(1/s), \quad \mathcal{Z} = \{\pm 1\}^d, \quad f(\theta, z) = -\langle z, \theta \rangle, \tag{8}$$

for a suitably chosen $s \in [d]$. Note that, if we choose $s \gg 1$, Θ above is a polytope with much more vertices ($2^s\binom{d}{s}$) than an ℓ_1 ball ($2d$), which would intuitively force a learner like an ERM to reveal more information about the training sample. On a technical level, selecting large s improves the subgaussian constant of a tracer; however, selecting s that is too large shrinks the diameter of the set, and thus, the problem becomes easier to learn. We must trade off these two aspects, and carefully set the value of s. As it turns out, the optimal choice is $s \propto d^{1-c}$ for any small $c > 0$ in order to establish Theorem 2.6. The remainder of the proof is rather technical and hence is deferred to Appendix F.2.

6 Limitations

We conclude by stating an intriguing open problem. We conjecture Theorem 2.8 is tight, and the dichotomy between traceability and SCO also holds for $p > 2$. In particular, we conjecture that the optimal DP-SCO excess risk for ℓ_p with $p > 2$ scales as

$$\min\left\{ \frac{d^{1/2-1/p}}{\sqrt{n}}, \left(\frac{1}{n}\right)^{1/p} \right\} + \min\left\{ \frac{d^{1-1/p}}{n\varepsilon}, \left(\frac{d}{\varepsilon^2 n^2}\right)^{1/p} \right\},$$

ignoring $\log(1/\delta)$ factors. If the conjecture is true, we have a complete understanding of traceability in SCO. If it is false, it reveals that there is something fundamentally different about settings with $p > 2$, which would also significantly enrich our understanding of DP-SCO.

Acknowledgments

The authors would like to thank Mufan Li and Ziyi Liu for their comments on the drafts of this work.

Funding

This work was completed while Sasha Voitovych was a student at the University of Toronto and supported by an Undergraduate Student Research Award from the Natural Sciences and Engineering Research Council of Canada. Mahdi Haghifam is supported by a Khoury College of Computer Sciences Distinguished Postdoctoral Fellowship. Idan Attias is supported by the National Science Foundation under Grant ECCS-2217023, through the Institute for Data, Econometrics, Algorithms, and Learning (IDEAL). Roi Livni is supported by a Google fellowship, a Vatat grant and the research has been funded, in parts, by an ERC grant (FoG - 101116258). Daniel M. Roy is supported by the funding through NSERC Discovery Grant and Canada CIFAR AI Chair at the Vector Institute.

References

[AWBR09] A. Agarwal, M. J. Wainwright, P. Bartlett, and P. Ravikumar. "Information-theoretic lower bounds on the oracle complexity of convex optimization". *Advances in Neural Information Processing Systems* 22 (2009).

[ALMM19] N. Alon, R. Livni, M. Malliaris, and S. Moran. "Private PAC learning implies finite Littlestone dimension". In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 2019, pp. 852–860.

[AKL21] I. Amir, T. Koren, and R. Livni. "SGD generalizes better than GD (and regularization doesn't help)". In: *Conference on Learning Theory*. PMLR. 2021, pp. 63–92.

[ABGMU22] R. Arora, R. Bassily, C. Guzmán, M. Menart, and E. Ullah. "Differentially private generalized linear models revisited". *Advances in neural information processing systems* 35 (2022), pp. 22505–22517.

[AFKT21] H. Asi, V. Feldman, T. Koren, and K. Talwar. "Private stochastic convex optimization: Optimal rates in l1 geometry". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 393–403.

[ADHLR24] I. Attias, G. K. Dziugaite, M. Haghifam, R. Livni, and D. M. Roy. "Information Complexity of Stochastic Convex Optimization: Applications to Generalization and Memorization". *arXiv preprint arXiv:2402.09327* (2024).

[BFTG19] R. Bassily, V. Feldman, K. Talwar, and A. Guha Thakurta. "Private stochastic convex optimization with optimal rates". *Advances in neural information processing systems* 32 (2019).

[BGN21] R. Bassily, C. Guzmán, and A. Nandi. "Non-Euclidean differentially private stochastic convex optimization". In: *Conference on Learning Theory*. PMLR. 2021, pp. 474–499.

[BMNSY18] R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff. "Learners that Use Little Information". In: *Algorithmic Learning Theory*. 2018, pp. 25–55.

[BST14] R. Bassily, A. Smith, and A. Thakurta. "Private empirical risk minimization: Efficient algorithms and tight error bounds". In: *2014 IEEE 55th annual symposium on foundations of computer science*. IEEE. 2014, pp. 464–473.

[Bat08] N. Batir. "Inequalities for the gamma function". *Archiv der Mathematik* 91.6 (2008), pp. 554–563.

[BHMZ20] O. Bousquet, S. Hanneke, S. Moran, and N. Zhivotovskiy. "Proper learning, Helly number, and an optimal SVM bound". In: *Conference on Learning Theory*. PMLR. 2020, pp. 582–609.

[BBFST21] G. Brown, M. Bun, V. Feldman, A. Smith, and K. Talwar. "When is memorization of irrelevant training data necessary for high-accuracy learning?" In: *Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing*. 2021, pp. 123–132.

[BLM20] M. Bun, R. Livni, and S. Moran. "An equivalence between private classification and online prediction". In: *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2020, pp. 389–402.

[BNSV15] M. Bun, K. Nissim, U. Stemmer, and S. Vadhan. "Differentially private release and learning of threshold functions". In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE. 2015, pp. 634–649.

[BUV14] M. Bun, J. Ullman, and S. Vadhan. "Fingerprinting codes and the price of approximate differential privacy". In: *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. 2014, pp. 1–10.

[CWZ23] T. T. Cai, Y. Wang, and L. Zhang. "Score attack: A lower bound technique for optimal differentially private learning". *arXiv preprint arXiv:2303.07152* (2023).

[CCNSTT22] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer. "Membership inference attacks from first principles". In: *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2022, pp. 1897–1914.

[CMS11] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. "Differentially private empirical risk minimization." *Journal of Machine Learning Research* 12.3 (2011).

[CDK22] C. Cheng, J. Duchi, and R. Kuditipudi. "Memorize to generalize: on the necessity of interpolation in high dimensional linear regression". In: *Conference on Learning Theory*. PMLR. 2022, pp. 5528–5560.

[DMNS06] C. Dwork, F. McSherry, K. Nissim, and A. Smith. "Calibrating noise to sensitivity in private data analysis". In: *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer. 2006, pp. 265–284.

[DSSUV15] C. Dwork, A. Smith, T. Steinke, J. Ullman, and S. Vadhan. "Robust traceability from trace amounts". In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE. 2015, pp. 650–669.

[Fel16] V. Feldman. "Generalization of ERM in Stochastic Convex Optimization: The Dimension Strikes Back". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc., 2016.

[Fel20] V. Feldman. "Does learning require memorization? a short tale about a long tail". In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. 2020, pp. 954–959.

[FKT20] V. Feldman, T. Koren, and K. Talwar. "Private stochastic convex optimization: optimal rates in linear time". In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. 2020, pp. 439–449.

[FS17] V. Feldman and T. Steinke. "Generalization for adaptively-chosen estimators via stable median". In: *Conference on learning theory*. PMLR. 2017, pp. 728–757.

[GLLST23] S. Gopi, Y. T. Lee, D. Liu, R. Shen, and K. Tian. "Private convex optimization in general norms". In: *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2023, pp. 5068–5089.

[HDMR21] M. Haghifam, G. K. Dziugaite, S. Moran, and D. Roy. "Towards a unified information-theoretic framework for generalization". *Advances in Neural Information Processing Systems* 34 (2021), pp. 26370–26381.

[HNKRD20] M. Haghifam, J. Negrea, A. Khisti, D. M. Roy, and G. K. Dziugaite. "Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms". *Advances in Neural Information Processing Systems* 33 (2020), pp. 9925–9935.

[HRTSMK23] M. Haghifam, B. Rodríguez-Gálvez, R. Thobaben, M. Skoglund, D. M. Roy, and G. Karolina Dziugaite. "Limitations of Information-Theoretic Generalization Bounds for Gradient Descent Methods in Stochastic Convex Optimization". In: *Proceedings of The 34th International Conference on Algorithmic Learning Theory*. Vol. 201. Proceedings of Machine Learning Research. PMLR, 2023, pp. 663–706.

[HRS16] M. Hardt, B. Recht, and Y. Singer. "Train faster, generalize better: Stability of stochastic gradient descent". In: *International conference on machine learning*. PMLR. 2016, pp. 1225–1234.

[HSRDTMPSNC08] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays". *PLoS genetics* 4.8 (2008), e1000167.

[KOV17] P. Kairouz, S. Oh, and P. Viswanath. "The Composition Theorem for Differential Privacy". *IEEE Transactions on Information Theory* 63.6 (2017), pp. 4037–4049.

[KLSU18] G. Kamath, J. Li, V. Singhal, and J. Ullman. "Privately Learning High-Dimensional Distributions". *arXiv preprint arXiv:1805.00216* (2018).

[KLSU19] G. Kamath, J. Li, V. Singhal, and J. Ullman. "Privately learning high-dimensional distributions". In: *Conference on Learning Theory*. PMLR. 2019, pp. 1853–1902.

[KMS22] G. Kamath, A. Mouzakis, and V. Singhal. "New lower bounds for private estimation and a generalized fingerprinting lemma". *Advances in neural information processing systems* 35 (2022), pp. 24405–24418.

[LLL24] Y. T. Lee, D. Liu, and Z. Lu. *The Power of Sampling: Dimension-free Risk Bounds in Private ERM*. 2024. arXiv: 2105.13637 [cs.LG].

[LW86] N. Littlestone and M. Warmuth. "Relating data compression and learnability" (1986).

[Liv24] R. Livni. "Information theoretic lower bounds for information theoretic upper bounds". *Advances in Neural Information Processing Systems* 36 (2024).

[LT24] X. Lyu and K. Talwar. "Fingerprinting Codes Meet Geometry: Improved Lower Bounds for Private Query Release and Adaptive Data Analysis". *arXiv preprint arXiv:2412.14396* (2024).

[MY16] S. Moran and A. Yehudayoff. "Sample compression schemes for VC classes". *Journal of the ACM (JACM)* 63.3 (2016), pp. 1–10.

[NY83] A. S. Nemirovskij and D. B. Yudin. "Problem complexity and method efficiency in optimization" (1983).

[SDSOJ19] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jégou. "White-box vs black-box: Bayes optimal strategies for membership inference". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 5558–5567.

[SSSS09] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. "Stochastic Convex Optimization." In: *COLT*. Vol. 2. 4. 2009, p. 5.

[SSSS17] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. "Membership Inference Attacks against Machine Learning Models". In: *IEEE Symposium on Security and Privacy*. 2017.

[ST10] K. Sridharan and A. Tewari. "Convex Games in Banach Spaces". In: *COLT*. 2010, pp. 1–13.

[SU17] T. Steinke and J. Ullman. "Tight lower bounds for differentially private selection". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 552–563.

[SZ20] T. Steinke and L. Zakynthinou. "Reasoning about generalization via conditional mutual information". In: *Conference on Learning Theory*. PMLR. 2020, pp. 3437–3452.

[Ste16] T. A. Steinke. "Upper and lower bounds for privacy and adaptivity in algorithmic data analysis". PhD thesis. 2016.

[Ver18] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

[Wai19] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press, 2019.

[XR17] A. Xu and M. Raginsky. "Information-theoretic analysis of generalization capability of learning algorithms". In: *Advances in Neural Information Processing Systems*. 2017, pp. 2524–2533.

Appendix Contents

A Additional preliminaries

A.1 Background on SCO

The next proposition summarizes the known minimax rates for learning SCO problems in general geometries. A proof can be found in [NY83; AWBR09; ST10].

Theorem A.1. *Fix $p \in [1, \infty]$, $d \in \mathbb{N}$, and $n \in \mathbb{N}$. Let $\alpha_{\mathsf{stat}}(\mathcal{L}_p^d, n)$ be the minimax excess risk rate of learning ℓ_p-Lipschitz-bounded problems, as defined in Equation (1). Then,*

1. *For $p = 1$, we have*
$$\alpha_{\mathsf{stat}}(\mathcal{L}_p^d, n) \in \Theta\left(\sqrt{\frac{\log(d)}{n}}\right).$$
2. *For $1 < p \le 2$, we have*
$$\alpha_{\mathsf{stat}}(\mathcal{L}_p^d, n) \in \Theta\left(\sqrt{\frac{\log(d)}{n}} \wedge \frac{1}{(p-1)\sqrt{n}}\right)$$
3. *For $2 \le p < \infty$, we have*
$$\alpha_{\mathsf{stat}}(\mathcal{L}_p^d, n) \in \Theta\left(\frac{d^{1/2-1/p}}{\sqrt{n}} \wedge \frac{1}{n^{1/p}}\right)$$
4. *For $p = \infty$, we have*
$$\alpha_{\mathsf{stat}}(\mathcal{L}_p^d, n) \in \Theta\left(\sqrt{\frac{d}{n}}\right).$$

Remark A.2. Notice that, in the overparameterized regime ($d \ge n$), the minimax excess risk for $p \ge 2$ is $\Theta\big(\big(\frac{1}{n}\big)^{1/p}\big)$ which is dimension-independent. This shows that for $d \ge n$, in all geometries except $p = \{1, \infty\}$, the minimax excess risk is dimension-free. ◁

This proposition implies the following corollary on the minimum number of samples required for α-learners.

Corollary A.3. *Fix $p \in [1, \infty]$, $d \in \mathbb{N}$, and $\alpha \in (0, 1]$. Let $N_{\mathsf{stat}}(\mathcal{L}_p^d, n)$ be the sample complexity of learning problems $\mathcal{L}_p^d$ up to excess risk α, i.e.,*
$$N_{\mathsf{stat}}(\mathcal{L}_p^d, n) = \min\left\{n \colon \alpha_{\mathsf{stat}}(\mathcal{L}_p^d, n) \le \alpha\right\}.$$
Then,

1. *For $p = 1$, we have*
$$N_{\mathsf{stat}}(\mathcal{L}_p^d, n) \in \Theta\left(\frac{\log(d)}{\alpha^2}\right).$$
2. *For $1 < p \le 2$, we have*
$$N_{\mathsf{stat}}(\mathcal{L}_p^d, n) \in \Theta\left(\frac{\log(d)}{\alpha^2} \wedge \frac{1}{((p-1)\alpha)^2}\right).$$
3. *For $2 \le p < \infty$, we have*
$$N_{\mathsf{stat}}(\mathcal{L}_p^d, n) \in \Theta\left(\frac{d^{1-2/p}}{\alpha^2} \wedge \frac{1}{\alpha^p}\right).$$
4. *For $p = 1$, we have*
$$N_{\mathsf{stat}}(\mathcal{L}_p^d, n) \in \Theta\left(\frac{d}{\alpha^2}\right).$$

A.2 Differential Privacy

Definition A.4. Let $\varepsilon > 0$ and $\delta \in [0, 1)$. A randomized mechanism $\mathcal{A}_n : \mathcal{Z}^n \to \mathcal{M}_1(\Theta)$ is (ε, δ)-DP, iff, for every two neighboring datasets $S_n \in \mathcal{Z}^n$ and $S'_n \in \mathcal{Z}^n$ (that is, S_n, S'_n differ in one element), and for every measurable subset $M \subseteq \Theta$, it holds

$$\Pr_{\hat{\theta} \sim \mathcal{A}_n(S_n)} \left(\hat{\theta} \in M \right) \leq e^{\varepsilon} \cdot \Pr_{\hat{\theta} \sim \mathcal{A}_n(S'_n)} \left(\hat{\theta} \in M \right) + \delta.$$

Algorithms that satisfy DP are not traceable in the sense of Definition 2.4 [KOV17]. The following simple proposition formalizes this observation.

Proposition A.5. *Fix $n \in \mathbb{N}$ and $\varepsilon, \delta > 0$. Let $\mathcal{A}_n$ be an (ε, δ)-DP algorithm. Then, if $\mathcal{A}_n$ is (ξ, m)-traceable, it holds that*

$$m \leq n \exp(\varepsilon)\xi + n\delta.$$

A.3 Concentration inequalities

First, we collect lemmata on the subgaussian norm, introduced in Section 3, which we use to derive concentration inequalities. The following is Equation (2.14) in [Ver18], and shows that a bound on subgaussian norm immediately leads to concentration inequalities.

Lemma A.6 (Subgaussian concentration)**.** *There exists a universal constant C such that the following holds for every random variable X with $\|X\|_{\psi_2} < \infty$: for every $t \geq 0$,*

$$\Pr\left[|X| \geq t\right] \leq 2 \exp\left(- \frac{ct^2}{\|X\|_{\psi_2}^2} \right)$$

The subgaussian norm behaves nicely under the summation of independent random variables. The following is Proposition 2.6.1 in [Ver18].

Lemma A.7 (Sum of subgaussian variables)**.** *Let $C > 0$ be a universal constant. Let $X_1, \ldots, X_n$ be a collection of arbitrary independent real random variables. Then,*

$$\left\| \sum_{i=1}^{n} X_i \right\|_{\psi_2}^2 \leq C \sum_{i=1}^{n} \|X_i\|_{\psi_2}^2 .$$

Subgaussian norm also behaves nicely under mixtures. In particular, we have the following proposition.

Proposition A.8 (Subgaussian mixtures)**.** *Let $\{X_\alpha\}_{\alpha \in A}$ be σ-subgaussian random variables, and let π be a distribution over the index set A. Then, a mixture of $\{X_\alpha\}_{\alpha \in A}$ under $\alpha \sim \pi$ is also σ-subgaussian.*

Proof. Let Y be such mixture. Then, for every $t > 0$, we have

$$\mathbb{E}[\exp(Y^2/t^2)] = \mathbb{E}_{\alpha \sim \pi} \mathbb{E}[\exp(X_\alpha^2/t^2)].$$

Plugging in $t = \sigma$ into above, and using that $\|X_\alpha\|_{\psi_2} \leq \sigma$ for all α, we have

$$\mathbb{E}[\exp(Y^2/\sigma^2)] = \mathbb{E}_{\alpha \sim \pi} \mathbb{E}[\exp(X_\alpha^2/t^2)] \leq 2,$$

i.e., $\|Y\|_{\psi_2} \leq \sigma$, as desired. □

It is well-known that bounded random variables are subgaussian (Equation (2.17) of [Ver18]).

Proposition A.9. *Suppose X is a random variable such that $X \in [-b, b]$ almost surely. Then,*

$$\|X\|_{\psi_2} \leq Cb,$$

for some universal constant $C > 0$.

We will heavily use the following result for the supremum of subgaussian processes (which follows from [Ver18, Theorem 8.1.6]). Let $\mathcal{N}(\Theta, \|\cdot\|, \varepsilon)$ denote the covering number of Θ in norm $\|\cdot\|$ at scale $\varepsilon > 0$.

Proposition A.10. *Let $\{X_\theta\}_{\theta\in\Theta}$ be a σ-subgaussian process w.r.t. a metric space $(\Theta, \|\cdot\|)$ as per Definition 3.1, and further assume that Θ is contained in the unit ball of $\|\cdot\|$. Let $t \geq 0$ be arbitrary. Then, with probability at least $1 - 4\exp(-t^2)$*

$$\sup_\theta X_\theta \leq C\sigma \left[\int_0^1 \sqrt{\log \mathcal{N}(\Theta, \|\cdot\|, \varepsilon)} d\varepsilon + t \right],$$

for some universal constant $C > 0$.

Proof. Fix an arbitrary $\theta_0 \in \Theta$. Using Theorem 8.1.6 [Ver18], we obtain the following bound for the *increment* of the subgaussian process $\{X_\theta\}$,

$$\Pr\left[\sup_{\theta\in\Theta} |X_\theta - X_{\theta_0}| \leq C\sigma \left(\int_0^\infty \sqrt{\log \mathcal{N}(\Theta, \|\cdot\|, \varepsilon)} d\varepsilon + t \right)\right] \geq 1 - 2\exp(-t^2).$$

First, note that for $\varepsilon \geq 1$, $\mathcal{N}(\Theta, \|\cdot\|, \varepsilon) = 1$, since Θ lies in the unit ball of $\|\cdot\|$. Thus,

$$\Pr\left[\sup_{\theta\in\Theta} |X_\theta - X_{\theta_0}| \leq C\sigma \left(\int_0^1 \sqrt{\log \mathcal{N}(\Theta, \|\cdot\|, \varepsilon)} d\varepsilon + t \right)\right] \geq 1 - 2\exp(-t^2). \tag{9}$$

Note that, by triangle inequality, we have

$$\sup_{\theta\in\Theta} |X_\theta - X_{\theta_0}| \geq \sup_{\theta\in\Theta} |X_\theta| - X_{\theta_0}. \tag{10}$$

Since $\{X_\theta\}_{\theta\in\Theta}$ satisfies Definition 3.1, we have

$$\|X_{\theta_0}\|_{\psi_2} \leq 2\sigma.$$

From Lemma A.6, we then have

$$\Pr\left[|X_{\theta_0}| \leq c\sigma t\right] \geq 1 - 2\exp(-t^2),$$

for some constant $c > 0$. Combining this with Equation (10) and taking a union bound with Equation (9), we get

$$\Pr\left[\sup_{\theta\in\Theta} |X_\theta| \leq C'\sigma \left(\int_0^1 \sqrt{\log \mathcal{N}(\Theta, \|\cdot\|, \varepsilon)} d\varepsilon + t \right)\right] \geq 1 - 4\exp(-t^2),$$

for some absolute constant $C' > 0$. □

The following lemma is an anti-concentration inequality based on Paley–Zygmund inequality. It shows that if the sum of variables is large, one can conclude that many of them are large given an appropriate control over their sum of squares. It is given as Lemma A.4 in [ADHLR24], and it is also similar to Lemma 25 in [DSSUV15].

Lemma A.11. *Fix $n \in \mathbb{N}$ and $(a_1, \ldots, a_n) \in \mathbb{R}^n$. Let $A_1 := \sum_{i\in[n]} a_i$ and $A_2 := \sum_{i\in[n]} (a_i)^2$. Then, for every $\beta \in \mathbb{R}$, $\left|\{i \in [n] \ : \ a_i \geq \beta/n\}\right| \geq \frac{(\max\{A_1 - \beta, 0\})^2}{A_2}$.*

A.4 Beta distributions

Next definitions are the versions of beta distributions that we use in this paper. Recall that, classically, beta distribution is supported on $[0, 1]$. However, in our results, it is convenient to consider the rescaled and centered variants.

Definition A.12. Fix $\beta > 0$. A (symmetric) beta distribution denoted by $\mathsf{s\text{-}beta}\,(\beta, \beta)$ is a continuous distribution, such that, if $X \sim \mathsf{s\text{-}beta}\,(\beta, \beta)$, then, for every $a \in [-1, 1]$, we have

$$\Pr(X \leq a) = \int_{-1}^a \frac{(1 - x^2)^{\beta-1}}{B(\beta)} dx,$$

where $B(\beta) = 2^{2\beta-1}\Gamma(\beta)^2/\Gamma(2\beta)$.

Definition A.13. Fix $\beta > 0$ and $\gamma \in (0, 1]$. We define rescaled (symmetric) beta distribution, denoted by $\mathsf{s\text{-}beta}_{[-\gamma,\gamma]}(\beta, \beta)$, where for $a \in [-\gamma, \gamma]$, its distribution is given by

$$\Pr(X \le a) = \frac{1}{\gamma B(\beta)} \int_{-\gamma}^{a} \left(1 - \left(\frac{x}{\gamma}\right)^2\right)^{\beta-1} dx,$$

where $B(\beta) = 2^{2\beta-1}\Gamma(\beta)^2/\Gamma(2\beta)$.

We have the following result on the first moment of the beta distribution.

Lemma A.14. *Fix $\beta > 0$. Let $X \sim \mathsf{s\text{-}beta}(\beta, \beta)$ where $\beta \ge 1$. Then,*

$$\mathbb{E}|X| \ge \frac{1}{3\sqrt{\beta}}.$$

Proof. Let $B(\beta) = 2^{2\beta-1}\Gamma(\beta)^2/\Gamma(2\beta)$ be the normalization constant. We have

$$\begin{aligned}
\mathbb{E}|X| &= \frac{1}{B(\beta)} \int_{-1}^{1} |x|(1-x^2)^{\beta-1} dx \\
&= \frac{1}{B(\beta)} \int_{0}^{1} 2x(1-x^2)^{\beta-1} dx \\
&= \frac{1}{\beta \cdot B(\beta)}.
\end{aligned}$$

It remains to upper bound $B(\beta)$. It follows from Theorem 1.5 of [Bat08] that, for every $x \ge 1$, we have

$$a\left(\frac{x-1/2}{e}\right)^{x-1/2} \le \Gamma(x) \le b\left(\frac{x-1/2}{e}\right)^{x-1/2},$$

where $a = \sqrt{2e}$ and $b = \sqrt{2\pi}$ are absolute constants. Thus,

$$\begin{aligned}
B(\beta) = \frac{2^{2\beta-1}\Gamma(\beta)^2}{\Gamma(2\beta)} &\le \frac{2^{2\beta-1} b^2 \left(\frac{\beta-1/2}{e}\right)^{2\beta-1}}{a\left(\frac{2\beta-1/2}{e}\right)^{2\beta-1/2}} \\
&= \frac{b^2\sqrt{e}}{a}(2\beta-1/2)^{-1/2}\left(\frac{2\beta-1}{2\beta-1/2}\right)^{2\beta-1} \\
&\le \frac{b^2\sqrt{e}}{a}(2\beta-1/2)^{-1/2} \\
&= \frac{2\pi\sqrt{e}}{\sqrt{2e}}(2\beta-1/2)^{-1/2} \\
&= \pi\,(\beta-1/4)^{-1/2} \\
&\le \pi\left(\frac{3}{4\beta}\right)^{1/2},
\end{aligned}$$

where in the last line we used $\beta - 1/4 \ge \frac{3}{4}\beta$ which holds as $\beta \ge 1$. Thus,

$$\mathbb{E}|X| \ge \frac{1}{\pi(3/4)^{1/2}} \frac{1}{\sqrt{\beta}} \ge \frac{1}{3\sqrt{\beta}},$$

as desired. □

Since the density of the rescaled beta distribution is homogeneous w.r.t. γ, we have the following result.

Corollary A.15. *Fix $\beta \ge 1$ and $\gamma \in (0, 1]$. Let $X \sim \mathsf{s\text{-}beta}_{[-\gamma,\gamma]}(\beta, \beta)$. Then,*

$$\mathbb{E}|X| \ge \frac{\gamma}{3\sqrt{\beta}}.$$

B Additional Related Work

Necessity of memorization in learning. A parallel line of work investigated memorization using the notion of *label memorization* in supervised setups. As per this definition, a learner is said to memorize its training samples if it "overfits" at these points. Feldman [Fel20] showed that, in some classification tasks, if the underlying distribution is *long-tailed*, then a learner is forced to memorize many training labels. Cheng, Duchi, and Kuditipudi [CDK22] showed this phenomenon also occurs in the setting of linear regression. While this framework is suitable to study memorization in supervised tasks, the notion of "labels" in SCO in not well-defined and thus calls for alternative definitions.

Another line of work studied memorization through the lens of information theoretic measures. Brown, Bun, Feldman, Smith, and Talwar [BBFST21] used *input-output mutual information* (IOMI) as a memorization metric and showed that IOMI can scale linearly with the training sample's entropy, indicating that a constant fraction of bits is memorized. In the context of SCO in ℓ_2 geometry, lower bounds on IOMI have been studied in [HRTSMK23; Liv24]. Specifically [Liv24] demonstrated that, for every accurate algorithm, its IOMI must scale with dimension d. Our approach to the study of memorization is conceptually different since we focus on the number of *samples* memorized as opposed to the number of *bits*. Nevertheless, it can be shown using Lemma H.3 and [HNKRD20, Thm. 2.1] that the recall *lower bounds* IOMI of an algorithm (provided that soundness parameter ξ is small enough, e.g., $\xi = 1/n^2$). However, because of the Lipschitzness of loss functions in ℓ_p SCO, we can use discretization of Θ and design algorithms with IOMI that is significantly smaller that the entropy of the training set, thus, memorization in the sense of [BBFST21] does not arise here.

Membership inference. Membership inference is an important practical problem [HSRDTMPSNC08; SSSS17; CCNSTT22]. In these works, the focus is on devising strategies for the tracer in modern machine learning settings, particularly neural networks. Our work takes a more fundamental perspective, aiming to determine whether membership inference is inherently unavoidable or simply a byproduct of specific training algorithms. An interesting aspect of our results is that, for $1 < p \leq 2$, the optimal strategy for tracing depends *only* on the loss function, which is in line with empirical studies [SDSOJ19].

Private Stochastic Convex Optimization. DP-SCO has been extensively studied in ℓ_2 geometry (see, for instance, [CMS11; BST14; BFTG19; FKT20]). For ℓ_p with $p \in [1, 2)$, the optimal DP excess risk was established in [AFKT21; BGN21]. The best known upper bounds for DP-SCO in ℓ_p geometry for $p > 2$ are due to [BGN21; GLLST23]. In this setting, there is a long-standing gap between upper and lower bounds, and the best known lower bounds are due to [ABGMU22; LLL24], which our paper improves on.

B.1 Detailed comparison with [DSSUV15; BST14].

One might hope that existing traceability results (such as [DSSUV15]) and a clever reduction to mean estimation (such as [BST14, Section 5.1]) might yield optimal results for SCO. Here, we will demonstrate rigorously that merely combining results and techniques of [DSSUV15; BST14] yields suboptimal results for the setup of SCO, even in the simple setting of ℓ_2 geometry. [BST14] considers the following ℓ_2 problem:

$$\Theta = \mathcal{B}_2(1), \quad \mathcal{Z} = \left\{ \pm \frac{1}{\sqrt{d}} \right\}^d, \quad f(\theta, Z) = -\langle \theta, Z \rangle.$$

To apply fingerprinting to establish traceability, we first need to posit a prior distribution over the unknown distribution. [DSSUV15] does so by considering product distributions over $\mathcal{Z}$, and placing a uniform prior over the mean $\mu \in [-1/\sqrt{d}, 1/\sqrt{d}]^d$. We now show that this (Bayesian) problem requires only $O(1/\alpha)$ samples to learn, and thus, tracing $\Omega(1/\alpha^2)$ samples is clearly impossible. Consider the ERM learner $\hat{\theta}$. It is easy to see that $\hat{\theta}$ can be written as:

$$\hat{\theta} = \frac{\hat{\mu}}{\|\hat{\mu}\|_2},$$

where $\hat{\mu}$ is the empirical mean of the dataset, that is, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Similarly, the population risk minimizer $\theta^\star$ is $\mu/\left\|\mu\right\|_2$. The expected excess risk of $\hat{\theta}$ is then:

$$\begin{aligned}
\mathbb{E}\left\langle \frac{\mu}{\left\|\mu\right\|_2}, \mu \right\rangle - \left\langle \frac{\hat{\mu}}{\left\|\hat{\mu}\right\|_2}, \mu \right\rangle &= \mathbb{E}\frac{\left\|\mu\right\|_2 \left\|\hat{\mu}\right\|_2 - \langle \hat{\mu}, \mu\rangle}{\left\|\hat{\mu}\right\|_2} \\
&\leq^{(a)} \mathbb{E}\left[\frac{\frac{1}{2}\left\|\mu\right\|_2^2 + \frac{1}{2}\left\|\hat{\mu}\right\|_2^2 - \langle \hat{\mu}, \mu\rangle}{\left\|\hat{\mu}\right\|_2} \wedge 2\right] \\
&= \frac{1}{2}\mathbb{E}\left[\frac{\left\|\mu - \hat{\mu}\right\|_2^2}{\left\|\hat{\mu}\right\|_2} \wedge 4\right],
\end{aligned}$$

where in (a) we used the AM-GM inequality, and the fact that the expression on the preceding line is always bounded by 2. The intuition behind the rest of the argument is that, due to the uniform prior on μ, we have $\left\|\mu\right\|, \left\|\hat{\mu}\right\| \in \Omega(1)$ with high probability. At the same time

$$\mathbb{E}\left[\left\|\mu - \hat{\mu}\right\|^2\right] = \mathbb{E}\frac{1}{dn}\sum_{i=1}^{n} 2\mu_i(1-\mu_i) \leq \frac{1}{2n},$$

thus, expected risk will be on the order of $O(1/n)$. To formalize this, note that $\mathbb{E}\left\|\mu\right\|_2^2 = 1/3$, and $\left\|\mu\right\|^2$ is a sum of d independent random variables bounded by $1/\sqrt{d}$ in absolute value. Hoeffding's inequality then yields that we have $\left\|\mu\right\|_2^2 \geq 1/6$ with very high probability. Similarly, we can obtain $\left\|\mu - \hat{\mu}\right\|_2^2 \leq 1/(2n) + 1/36 \leq 1/12$ for large enough n, with high probability. Then, with high probability, event

$$\mathcal{E} := \left\{\left\|\hat{\mu}\right\| \geq \frac{1}{\sqrt{6}} - \frac{1}{\sqrt{12}} \geq 0.1\right\}$$

holds Then, the excess risk is upper bounded by

$$\begin{aligned}
\mathbb{E}\left\langle \frac{\mu}{\left\|\mu\right\|_2}, \mu \right\rangle - \left\langle \frac{\hat{\mu}}{\left\|\hat{\mu}\right\|_2}, \mu \right\rangle &\leq \frac{1}{2}\mathbb{E}\left[\frac{\left\|\mu - \hat{\mu}\right\|_2^2}{\left\|\hat{\mu}\right\|_2} \wedge 4\right] \\
&= \frac{1}{2}\mathbb{E}\left[\mathbb{1}(\mathcal{E})\left(\frac{\left\|\mu - \hat{\mu}\right\|_2^2}{\left\|\hat{\mu}\right\|_2} \wedge 4\right)\right] + \frac{1}{2}\mathbb{E}\left[\mathbb{1}(\mathcal{E}^c)\left(\frac{\left\|\mu - \hat{\mu}\right\|_2^2}{\left\|\hat{\mu}\right\|_2} \wedge 4\right)\right] \\
&\leq \frac{1}{2}\mathbb{E}\left[10\left\|\mu - \hat{\mu}\right\|_2^2\right] + 2\Pr(\mathcal{E}^c) \\
&\in O(1/n),
\end{aligned}$$

as desired.

C Proofs from Section 3

C.1 Proof of Lemma 3.4

We first prove a slightly more general concentration statement to bound the supremum in Lemma 3.4, which will be useful to reuse in other proofs. Let $\mathcal{N}(\Theta, \left\|\cdot\right\|, \varepsilon)$ denote the size of the minimal cover of Θ in norm $\left\|\cdot\right\|$ at scale $\varepsilon > 0$. Then, the more general statement is given below.

Lemma C.1. *Fix $n, d \in \mathbb{N}$. Suppose $\Theta \subset \mathbb{R}^d$ is a subset of a unit ball in some norm $\left\|\cdot\right\|$. Let $\phi\colon \Theta \times \mathcal{Z} \to \mathbb{R}$ and $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ be such that, as $Z \sim \mathcal{D}$, $\{\phi(\theta, Z)\}$ is a σ-subgaussian process w.r.t. $(\Theta, \left\|\cdot\right\|)$ and for every $\theta \in \Theta$, $\mathbb{E}[\phi(\theta, Z)] = 0$. Let $(Z_1, \ldots, Z_n) \sim \mathcal{D}^{\otimes n}$. Then, there exist a universal constant $C > 0$, such that for every $t \geq 0$,*

$$\Pr\left[\sup_{\theta\in\Theta}\sqrt{\sum_{i=1}^{n}\left[\phi(\theta, Z_i)\right]^2} \leq C\sigma\left(\sqrt{n} + \int_0^1 \sqrt{\log \mathcal{N}\left(\Theta; \left\|\cdot\right\|, \varepsilon\right)}d\varepsilon + t\right)\right] \geq 1 - 4\exp(-t^2).$$

Proof. Let Φ_θ denote the following random vector

$$\Phi_\theta = \begin{bmatrix} \phi(\theta, Z_1) \\ \vdots \\ \phi(\theta, Z_n) \end{bmatrix}.$$

Then, observe that, the desired quantity is equal to

$$\sup_{\theta\in\Theta} \sqrt{\sum_{i=1}^{n} [\phi(\theta, Z_i)]^2} = \sup_{\theta\in\Theta} \|\Phi_\theta\|_2 = \sup_{\theta\in\Theta, x\in\mathbb{S}^{n-1}} \langle x, \Phi_\theta\rangle .$$

Then, $\langle x, \Phi_\theta\rangle$ can be seen to be a random process parameterized by a pair (x, θ). We will show that it is, in fact, a subgaussian process. Indeed, note that, by triangle inequality,

$$\|\langle x, \Phi_\theta\rangle - \langle x', \Phi_{\theta'}\rangle\|_{\psi_2} \le \|\langle x - x', \Phi_\theta\rangle\|_{\psi_2} + \|\langle x', \Phi_\theta - \Phi_{\theta'}\rangle\|_{\psi_2} . \tag{11}$$

Since Φ_θ^i is σ-subgaussian for each i, we have by Lemma A.7,

$$\|\langle x - x', \Phi_\theta\rangle\|_{\psi_2} \le C\sigma \|x - x'\|_2 ,$$

for some universal constant $C > 0$. Now, for every i, $(\Phi_\theta - \Phi_{\theta'})^i$ is $\sigma\|\theta - \theta'\|$-subgaussian. Therefore, by Lemma A.7, we have

$$\|\langle x', \Phi_\theta - \Phi_{\theta'}\rangle\|_{\psi_2} = \left\| \sum_{i=1}^{n} (x')^i (\Phi_\theta^i - \Phi_{\theta'}^i) \right\|_{\psi_2} \le C\sigma \|\theta - \theta'\| .$$

Combining the two inequalities, we get

$$\begin{aligned} \|\langle x, \Phi_\theta\rangle - \langle x', \Phi_{\theta'}\rangle\|_{\psi_2} &\le C\sigma \|\theta - \theta'\| + C\sigma \|x - x'\|_2 \\ &= 2C\sigma \cdot \frac{1}{2} \left[\|\theta - \theta'\| + \|x - x'\|_2\right] . \end{aligned}$$

Thus, $\langle x, \Phi_\theta\rangle$ is $(2C\sigma)$-subgaussian process w.r.t the norm γ, defined as

$$\gamma((x, \theta)) := \frac{1}{2} \left[\|x\|_2 + \|\theta\|\right] .$$

Moreover, we can see that $\Theta \times \mathbb{S}^{n-1}$ is a subset of a unit ball in γ. By definition of γ, we have

$$\mathcal{N}\left(\mathbb{S}^{n-1} \times \Theta; \gamma, \varepsilon\right) \le \mathcal{N}\left(\mathbb{S}^{n-1}; \|\cdot\|_2, \varepsilon\right) \cdot \mathcal{N}\left(\Theta; \|\cdot\|, \varepsilon\right) . \tag{12}$$

Then, using Proposition A.10, we have, for some constant $K > 0$, that with probability $1 - 4\exp(-t^2)$

$$\begin{aligned} \sup_\theta \|\Phi_\theta\|_2 &\le K\sigma \left[\int_0^1 \sqrt{\log \mathcal{N}\left(\Theta \times \mathbb{S}^{n-1}; \gamma, \varepsilon\right)} d\varepsilon + t\right] \\ &\le K\sigma \left[\int_0^1 \sqrt{\log \mathcal{N}\left(\mathbb{S}^{n-1}; \|\cdot\|_2, \varepsilon\right) + \log \mathcal{N}\left(\Theta; \|\cdot\|, \varepsilon\right)} d\varepsilon + t\right] \\ &\le^{(a)} K\sigma \left[\int_0^1 \sqrt{n \log\left(1 + \frac{4}{\varepsilon}\right)} d\varepsilon + \int_0^1 \sqrt{\log \mathcal{N}\left(\Theta; \|\cdot\|, \varepsilon\right)} d\varepsilon + t\right] \\ &\le K'\sigma \left[\sqrt{n} + \int_0^1 \sqrt{\log \mathcal{N}\left(\Theta; \|\cdot\|, \varepsilon\right)} d\varepsilon + t\right] , \end{aligned}$$

as desired, where in (a) we used Example 5.8 from [Wai19], and $K' > 0$ is some other universal constant. □

Using Example 5.8 from [Wai19] once again to upper bound $\sqrt{\log \mathcal{N}\left(\Theta; \|\cdot\|, \varepsilon\right)}$, we have the proof of Lemma 3.4.

Lemma 3.4. *Fix $n, d \in \mathbb{N}$. Suppose $\Theta \subset \mathbb{R}^d$ is a subset of a unit ball in some norm $\|\cdot\|$. Let $\phi\colon \Theta \times \mathcal{Z} \to \mathbb{R}$ and $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ be such that, as $Z \sim \mathcal{D}$, $\{\phi(\theta, Z)\}$ is a σ-subgaussian process w.r.t. $(\Theta, \|\cdot\|)$. Let $(Z_1, \dots, Z_n) \sim \mathcal{D}^{\otimes n}$. Then, there is a constant $C > 0$, such that*

$$\Pr\left[\sup_{\theta\in\Theta} \sqrt{\sum_{i=1}^n [\phi(\theta, Z_i)]^2} \le C\sigma\left(\sqrt{n} + \sqrt{d} + t\right)\right] \ge 1 - 4\exp(-t^2), \quad \forall t \ge 0.$$

Proof. From Example 5.8 in [Wai19], we have

$$\log \mathcal{N}(\Theta; \|\cdot\|, \varepsilon) \le d \log\left(1 + \frac{2}{\varepsilon}\right).$$

Plugging this into the result of Lemma C.1, with probability at least $1 - 4\exp(-t^2)$, we have

$$\begin{aligned}
\sup_{\theta\in\Theta} \sqrt{\sum_{i=1}^n [\phi(\theta, Z_i)]^2} &\le C\sigma\left(\sqrt{n} + \int_0^1 \sqrt{\log \mathcal{N}(\Theta; \|\cdot\|, \varepsilon)}d\varepsilon + t\right) \\
&\le C\sigma\left(\sqrt{n} + \int_0^1 \sqrt{d \log\left(1 + \frac{2}{\varepsilon}\right)}d\varepsilon + t\right) \\
&\le C'\sigma\left(\sqrt{n} + \sqrt{d} + t\right),
\end{aligned}$$

for some other universal constant $C' > 0$. □

C.2 Proof of Theorem 3.5

Theorem 3.5. *Fix $n \in \mathbb{N}$, $d \in \mathbb{N}$, $\kappa > 0$ and $\alpha \in [0, 1]$. Consider an arbitrary SCO problem $\mathcal{P} = (\Theta, \mathcal{Z}, f)$. Let $T = \mathrm{Tr}_\kappa(\mathcal{P}; n, \alpha)$ be the subgaussian trace value of $\mathcal{P}$. Then, for some constant $c > 0$, every α-learner $\mathcal{A}_n$ is (ξ, m)-traceable with*

$$\xi = \exp(-cT^2), \quad m = c\left[\frac{n^2T^2}{n+d} - \frac{16\kappa^2 n}{\exp(n+d)}\right].$$

Proof. We set

$$\lambda := \frac{T}{2}$$

First, we show that the soundness condition holds. Since Z and $\hat{\theta}$ are independent, and using the subgaussian nature of $\phi(\hat{\theta}, Z)$, we have by Lemma A.6

$$\Pr_{Z\sim\mathcal{D}}\left[\phi(\hat{\theta}, Z) \ge \lambda\right] \le \exp\left(-c\lambda^2\right) \le \exp\left(-cT^2/4\right),$$

where $c > 0$ is some constant. For recall, let's define the set $\mathcal{I}$ as follows

$$\mathcal{I} = \{i \in [n]\colon \phi(\hat{\theta}, Z_i) \ge \lambda\}.$$

Using Lemma A.11, we have

$$\begin{aligned}
\mathbb{E}\left[|\mathcal{I}|\right] &= \mathbb{E}\left|\{i \in [n]\colon \phi(\hat{\theta}, Z_i) \ge \lambda\}\right| \\
&\ge \mathbb{E}\left[\frac{\left(\sum_{i=1}^n \phi(\hat{\theta}, Z_i) - n\lambda\right)_+^2}{\sum_{i=1}^n \phi(\hat{\theta}, Z_i)^2}\right] \\
&\ge \mathbb{E}\left[\frac{\left(\sum_{i=1}^n \phi(\hat{\theta}, Z_i) - n\lambda\right)_+^2}{\sup_\theta \|\{\phi(\theta, Z_i)\}_{i=1}^n\|_2^2}\right],
\end{aligned}$$

where for every $x \in \mathbb{R}$, we define $(x)_+ = \max\{x, 0\}$. Then, Lemma 3.4 tells us that, for $t := \sqrt{n+d}$, we have with probability $1 - 4\exp(-t^2)$,

$$\sup_{\theta \in \Theta} \left\| \left[\phi(\hat{\theta}, Z_1), \ldots, \phi(\hat{\theta}, Z_n) \right]^\top \right\|_2^2 \leq C\,(n+d)\,,$$

for some constant $C > 0$. Thus,

$$\Pr\left[\mathcal{E} := \left\{ \sup_{\theta \in \Theta} \left\| \left[\phi(\hat{\theta}, Z_1), \ldots, \phi(\hat{\theta}, Z_n) \right]^\top \right\|_2^2 \leq C\,(n+d) \right\} \right] \geq 1 - 4\exp(-t^2).$$

This implies,

$$\begin{aligned}
\mathbb{E}|\mathcal{I}| &\geq \mathbb{E}\left[\frac{\left(\sum_{i=1}^n \phi(\hat{\theta}, Z_i) - n\lambda \right)_+^2}{\sup_\theta \|\{\phi(\theta, Z_i)\}_{i=1}^n\|_2^2} \right] \\
&\geq \mathbb{E}\left[\frac{\left(\sum_{i=1}^n \phi(\hat{\theta}, Z_i) - n\lambda \right)_+^2}{\sup_\theta \|\{\phi(\theta, Z_i)\}_{i=1}^n\|_2^2} \mathbb{1}(\mathcal{E}) \right] \\
&= \mathbb{E}\left[\frac{\left(\sum_{i=1}^n \phi(\hat{\theta}, Z_i) - n\lambda \right)_+^2}{C\,(n+d)} \mathbb{1}(\mathcal{E}) \right] \\
&\geq \mathbb{E}\left[\frac{\left(\sum_{i=1}^n \phi(\hat{\theta}, Z_i) - n\lambda \right)_+^2}{C\,(n+d)} \right] - \mathbb{E}\left[\frac{\left(\sum_{i=1}^n \phi(\hat{\theta}, Z_i) - n\lambda \right)_+^2}{C\,(n+d)} \mathbb{1}(\mathcal{E}^c) \right].
\end{aligned}$$

We know that, almost surely,

$$\phi(\hat{\theta}, Z)^2 \leq \kappa^2.$$

Thus, almost surely,

$$\left(\sum_{i=1}^n \phi(\hat{\theta}, Z_i) - n\lambda \right)_+^2 \leq \left(\sum_{i=1}^n \left(\phi(\hat{\theta}, Z_i) - \lambda \right)_+ \right)^2 \leq n \sum_{i=1}^n \phi(\hat{\theta}, Z_i)^2 \leq \kappa^2 n^2.$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\frac{\left(\sum_{i=1}^n \phi(\hat{\theta}, Z_i) - n\lambda \right)_+^2}{C\,(n+d)} \mathbb{1}(\mathcal{E}^c) \right] &\leq \Pr[\mathcal{E}^c] \cdot \frac{\kappa^2 n^2}{C(n+d)} \\
&\leq \frac{4\kappa^2 n}{C \exp(n+d)}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[|\mathcal{I}|\right] &\geq \mathbb{E}\left[\frac{\left(\sum_{i=1}^{n}\phi(\hat{\theta},Z_i)-n\lambda\right)_+^2}{C\left(n+d\right)}\right]-\mathbb{E}\left[\frac{\left(\sum_{i=1}^{n}\phi(\hat{\theta},Z_i)-n\lambda\right)_+^2}{C\left(n+d\right)}\mathbb{1}(\mathcal{E}^c)\right]\\
&\geq \mathbb{E}\left[\frac{\left(\sum_{i=1}^{n}\phi(\hat{\theta},Z_i)-n\lambda\right)_+^2}{C\left(n+d\right)}\right]-\frac{4\kappa^2 n}{C\exp(n+d)}\\
&\geq^{(a)} \frac{\left(\sum_{i=1}^{n}\mathbb{E}\phi(\hat{\theta},Z_i)-n\lambda\right)_+^2}{C\left(n+d\right)}-\frac{4\kappa^2 n}{C\exp(n+d)}\\
&\geq \frac{n^2T^2/4}{C\left(n+d\right)}-\frac{4\kappa^2 n}{C\exp(n+d)}\\
&= c\left[\frac{n^2T^2}{n+d}-\frac{16\kappa^2 n}{\exp(n+d)}\right],
\end{aligned}$$

where (a) follows by Jensen's inequality and $c=1/4C$. □

C.3 Proof of Theorem 3.6

Theorem 3.6. *There exists a universal constant $c>0$, such that the following holds. Fix $p\in[1,\infty)$, $n\in\mathbb{N}$, $d\in\mathbb{N}$, $\alpha\in[0,1]$, $\kappa>0$ $\varepsilon>0$, and $\delta\in[0,1]$. Consider an arbitrary SCO problem $\mathcal{P}=(\Theta,\mathcal{Z},f)$ in $\mathbb{R}^d$. Let $T=\mathrm{Tr}_\kappa(\mathcal{P};n,\alpha)$ be the subgaussian trace value of problem $\mathcal{P}$. Then, for every (ε,δ)-DP α-learner $\mathcal{A}_n$, we have* $\exp(\varepsilon)-1\geq c\left(T-2\delta\kappa\right)$.

Proof. Consider an arbitrary distribution $\mathcal{D}$ and a function ϕ s.t. $\{\phi(\theta,Z)\}_{\theta\in\Theta}$ is a 1-subgaussian process w.r.t. $(\Theta,\|\cdot\|_\Theta)$ and $|\phi|\leq\kappa$ almost surely. Consider a sample $S_n=(Z_1,\ldots,Z_n)$ and let Z_0 be a freshly sampled point; let $S_n^{(i)}$ be a sample with Z_i substituted by Z_0. Let $\hat{\theta}$ be a learner trained on S_n and $\hat{\theta}^{(i)}$ be a learner trained on $S_n^{(i)}$. Then, since $\hat{\theta}$ is (ε,δ)-DP and noting that $\phi(\theta,Z)$ is supported on $[-\kappa,\kappa]$, we may apply Lemma A.1 of [FS17] and get

$$\left|\mathbb{E}\phi(\hat{\theta},Z_i)-\mathbb{E}\phi(\hat{\theta}^{(i)},Z_i)\right|\leq\mathbb{E}\left|\phi(\hat{\theta}^{(i)},Z_i)\right|\left(\exp(\varepsilon)-1\right)+2\delta\kappa.$$

By independence of $\hat{\theta}$ and Z_i, we conclude that $\phi(\hat{\theta},Z_i)$ is 1-subgaussian random variable. It is well-known that $\mathbb{E}|X|\leq C\sigma$ if X is σ-subgaussian for some constant C (see part (ii) of Proposition 2.5.2 of [Ver18] for $p=1$), thus the above gives

$$\left|\mathbb{E}\phi(\hat{\theta},Z_i)\right|\leq C(\exp(\varepsilon)-1)+2\delta\kappa.$$

Then, for every $\mathcal{D}$ and ϕ we get that,

$$\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\phi(\hat{\theta},Z_i)\leq C(\exp(\varepsilon)-1)+2\delta\kappa.$$

Thus,

$$T\leq C(\exp(\varepsilon)-1)+2\delta\kappa,$$

which, after rearranging, implies the desired result. □

D Proofs of fingerprinting lemmas (Section 4)

D.1 Proof of Lemma 4.2

Lemma 4.2 (Sparse fingerprinting)**.** *Fix $d,n\in\mathbb{N}$ and let $k\in[d]$. For each $\mu\in[-k/d,k/d]^d$, let $\mathcal{Z}_k$ and $\mathcal{D}_{\mu,k}$ be as in Definition 4.1 . Let $\pi=\mathsf{s\text{-}beta}_{[-k/d,k/d]}\left(\beta,\beta\right)^{\otimes d}$ be a prior and set*

$$\phi_\mu(\theta,Z):=\left\langle\theta,\left(Z-\frac{d}{k}\mu\right)\right\rangle_{\mathrm{supp}(Z)}.$$

Then, for every learning algorithm $\mathcal{A}_n : \mathcal{Z}^n \to \mathcal{M}_1(\mathbb{R}^d)$ *with sample* $S_n = (Z_1, \ldots, Z_n)$,

$$\mathbb{E}_{\mu\sim\pi}\mathbb{E}_{S_n\sim\mathcal{D}_{\mu,k}^{\otimes n},\hat{\theta}\sim\mathcal{A}_n(S_n)}\left[\sum_{i=1}^{n}\phi_\mu(\hat{\theta}, Z_i)\right] = \frac{2\beta d}{k}\mathbb{E}_{\mu\sim\pi}\left\langle \mu, \mathbb{E}_{S_n\sim\mathcal{D}_{\mu,k}^{\otimes n},\hat{\theta}\sim\mathcal{A}_n(S_n)}[\hat{\theta}]\right\rangle.$$

Proof. For each $j \in [d]$, let $\mathcal{I}_j := \{i \in [n] : Z_i^j \neq 0\}$ as the index of the training points such that their j-th coordinate is non-zero. Then, we have

$$\mathbb{E}\left[\sum_{i=1}^{n}\phi_\mu(\hat{\theta}, Z_i)\right] = \mathbb{E}\left[\sum_{j=1}^{d}\sum_{i\in\mathcal{I}_j}\left((\hat{\theta})^j\left(Z_i^j - \frac{d}{k}\mu^j\right)\right)\right] \tag{13}$$

$$= \sum_{j=1}^{d}\mathbb{E}\left[\sum_{i\in\mathcal{I}_j}(\hat{\theta})^j\left(Z_i^j - \frac{d}{k}\mu^j\right)\right]. \tag{14}$$

Then, define the following function

$$g^j(\mu^j) := \mathbb{E}\left[(\hat{\theta})^j \middle| \{\mathcal{I}_r\}_{r\in[d]}, \{Z_i^m\}_{m\neq j, i\in[n]}\right].$$

We claim

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}_j}(\hat{\theta})^j \cdot \left(Z_i^j - \frac{d}{k}\mu^j\right) \middle| \{\mathcal{I}_r\}_{r\in[d]}, \{Z_i^m\}_{m\neq j, i\in[n]}\right] = \frac{k}{d}\left(1 - \left(\frac{d}{k}\mu^j\right)^2\right)\frac{d}{d\mu^j}g^j(\mu^j). \tag{15}$$

The proof is based on the following two observations: 1) conditioned on $\{Z_i^m\}_{m\neq j, i\in[n]}$, $\hat{\theta}$ is a function of $\{Z_i^j\}_{i\in[n]}$, 2) conditioned on $\{\mathcal{I}_r\}_{r\in[d]}$ the non-zero elements in $\{Z_i^j\}_{i\in[n]}$ are sampled i.i.d from $\{\pm 1\}$ with mean $\frac{d}{k}\mu^j$. Then, based on these observations Equation (15) follows as an straightforward application of [Ste16, Lemma 4.3.7].

Recall the definition of π and notice that π is a product measure. Let π^j be the distribution on the j-th coordinate. By the definition of the prior distribution, we can write

$$\begin{aligned}
&\mathbb{E}_{\mu^j\sim\pi^j}\left[\frac{k}{d}\left(1 - \left(\frac{d}{k}\mu^j\right)^2\right)\frac{d}{d\mu^j}g^j(\mu^j)\right] \\
&= \frac{1}{C}\int_{-\frac{k}{d}}^{+\frac{k}{d}}\frac{k}{d}\left(1 - \left(\frac{d}{k}v\right)^2\right)\frac{d}{dv}g^j(v)\left(1 - \left(\frac{d}{k}v\right)^2\right)^{\beta-1}dv \\
&= \frac{1}{C}\int_{-\frac{k}{d}}^{+\frac{k}{d}}\frac{k}{d}\frac{d}{d\mu^j}g^j(v)\left(1 - \left(\frac{d}{k}v\right)^2\right)^{\beta}dv \\
&= \frac{2}{C}\int_{-\frac{k}{d}}^{+\frac{k}{d}}\frac{d}{k}\beta v\left(1 - \left(\frac{d}{k}v\right)^2\right)^{\beta-1}g^j(v)dv \\
&= 2\beta\frac{d}{k}\mathbb{E}_{\mu^j\sim\pi^j}\left[g^j(\mu^j)\mu^j\right].
\end{aligned} \tag{16}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}_{\mu\sim\pi}\mathbb{E}_{S_n\sim\mathcal{D}_{\mu,k}^{\otimes n},\hat{\theta}\sim\mathcal{A}_n(S_n)}\left[\sum_{i=1}^{n}\phi_\mu(\hat{\theta},Z_i)\right]\\
&=\sum_{j=1}^{d}\mathbb{E}\left[\sum_{i\in\mathcal{I}_j}(\hat{\theta})^j\left(Z_i^j-\frac{d}{k}\mu^j\right)\right]\\
&=\sum_{j=1}^{d}\mathbb{E}\left[\mathbb{E}\left[\sum_{i\in\mathcal{I}_j}(\hat{\theta})^j\left(Z_i^j-\frac{d}{k}\mu^j\right)\middle|\{\mathcal{I}_r\}_{r\in[d]},\{Z_i^m\}_{m\neq j,i\in[n]}\right]\right]\\
&=2\beta\frac{d}{k}\sum_{j=1}^{d}\mathbb{E}\left[g^j(\mu^j)\cdot\mu^j\right],
\end{aligned}$$

where the last step follow from Equations (15) and (16). Then, notice that

$$\begin{aligned}
\mathbb{E}\left[g^j(\mu^j)\cdot\mu^j\right]&=\mathbb{E}\left[\mathbb{E}\left[(\hat{\theta})^j\cdot\mu^j\middle|\{\mathcal{I}_r\}_{r\in[d]},\{Z_i^m\}_{m\neq j,i\in[n]}\right]\right]\\
&=\mathbb{E}\left[(\hat{\theta})^j\cdot\mu^j\right].
\end{aligned}$$

Therefore, by the definition of inner product in $\mathbb{R}^d$, we have

$$2\beta\frac{d}{k}\sum_{j=1}^{d}\mathbb{E}\left[g^j(\mu^j)\cdot\mu^j\right]=2\beta\frac{d}{k}\mathbb{E}\left[\left\langle\hat{\theta},\mu\right\rangle\right],$$

as was to be shown. □

D.2 Fingerprinting for ℓ_1 setup.

Additionally, to prove Theorem 2.6, we will need the following fingerprinting lemma. It can be seen as a generalization of beta-fingerprinting lemma in [SU17] using the scaling matrix technique of [KLSU18].

Lemma D.1 (Fingerprinting lemma with a scaling matrix)**.** *Fix $d\in\mathbb{N}$. Let $\mathcal{Z}=\{\pm1\}^d$ and let $\beta>0$ be arbitrary. Consider arbitrary $0<\gamma\leq1$. For every $\mu\in[-\gamma,\gamma]^d$, let $\mathcal{D}_\mu$ be the product distribution on $\mathcal{Z}$ with mean μ, i.e., for every $z\in\mathcal{Z}$, we have $\mathcal{D}_\mu=\prod_{k=1}^{d}\left(\frac{1+z^k\mu^k}{2}\right)$ let Λ_μ be a diagonal matrix of size d where the i-th diagonal element is given by $\Lambda_\mu^{ii}=\frac{1-(\mu^i/\gamma)^2}{1-(\mu^i)^2}$, and let $\phi_\mu(\theta,z)=\langle\theta,\Lambda_\mu(z-\mu)\rangle$. Let $\pi=\mathsf{s\text{-}beta}_{[-\gamma,\gamma]}(\beta,\beta)^{\otimes d}$ be a prior. Then, for every algorithm $\mathcal{A}_n:\mathcal{Z}^n\to\mathcal{M}_1(\mathbb{R}^d)$, we have*

$$\mathbb{E}_{\mu\sim\pi}\mathbb{E}_{S_n\sim\mathcal{D}_\mu^{\otimes n},\hat{\theta}\sim\mathcal{A}_n(S_n)}\sum_{Z\in S_n}\phi_\mu(\hat{\theta},Z)=\frac{2\beta}{\gamma^2}\mathbb{E}_{\mu\sim\pi}\left\langle\mu,\mathbb{E}_{S_n\sim\mathcal{D}_\mu^{\otimes n},\hat{\theta}\sim\mathcal{A}_n(S_n)}[\hat{\theta}]\right\rangle.$$

This fingerprinting lemma is handy for the following reason. To ensure the problem is hard to learn, entries of μ typically need to inversely scale with α. To achieve this, one can select small γ in the above to shrink the beta-prior to a smaller scale, while simultaneously having the freedom to set β to any value. In particular, this allows us to choose $\beta\in\Theta(\log(d))$ in the proof of Theorem 2.6 to leverage the anti-concentration result of [SU17, Prop. 5].

Before we proceed with the proof, we state the necessary lemmata. Throughout this section, for a real number $p\in[-1,1]$, we will write $X\sim p$ to denote the fact that X is a random variable on $\{\pm1\}$ with mean p. The following is a classical fingerprinting result.

Lemma D.2 (Lemma 5 of [DSSUV15])**.** *Let $f\colon\{\pm1\}^n\to\mathbb{R}$ be arbitrary. Define $g\colon[-1,1]\to\mathbb{R}$ by*

$$g(p)=\mathbb{E}_{X\sim p^{\otimes n}}[f(X)].$$

Then,

$$\mathbb{E}_{X\sim p^{\otimes n}}\left[f(X)\sum_{i\in[n]}(X_i-p)\right]=(1-p^2)g'(p).$$

Armed with the above result, we proceed to the proof of Lemma D.1. We will first prove a *per-coordinate version* of Lemma D.1. We make a note that the proofs combine techniques for beta-fingerprinting results of [SU17] and the scaling matrix technique of [KLSU19; ADHLR24].

Lemma D.3 (Per-coordinate version of Lemma D.1)**.** *Let $f\colon\{\pm1\}^n\to\mathbb{R}$ be arbitrary. Let $\pi=\mathsf{s\text{-}beta}_{[-\gamma,\gamma]}(\beta,\beta)$ be a prior distribution. Then,*

$$\mathbb{E}_{p\sim\pi}\mathbb{E}_{X\sim p^{\otimes n}}\left[\frac{1-(p/\gamma)^2}{1-p^2}f(X)\sum_{i=1}^{n}(X_i-p)\right]=\frac{2\beta}{\gamma^2}\mathbb{E}_{p\sim\pi}\left[p\cdot\mathbb{E}_{X\sim p^{\otimes n}}f(X)\right].$$

Proof. Let

$$g(p)=\mathbb{E}_{X\sim p^{\otimes n}}[f(X)].$$

Then, by Lemma D.2, we have for every $p\in[-1,1]$,

$$\mathbb{E}_{X\sim p^{\otimes n}}\left[\frac{1-(p/\gamma)^2}{1-p^2}f(X)\sum_{i\in[n]}(X_i-p)\right]=\left(1-\left(\frac{p}{\gamma}\right)^2\right)g'(p).$$

Recalling the definition of scaled symmetric beta distribution from Definition A.13, we have

$$\begin{aligned}
&\mathbb{E}_{p\sim\pi}\mathbb{E}_{X\sim p^{\otimes n}}\left[\frac{1-(p/\gamma)^2}{1-p^2}f(X)\sum_{i\in[n]}(X_i-p)\right]\\
&=\mathbb{E}_{p\sim\pi}\left[\left(1-\left(\frac{p}{\gamma}\right)^2\right)g'(p)\right]\\
&=\frac{1}{\gamma B(\beta)}\int_{-\gamma}^{\gamma}\left(1-\left(\frac{p}{\gamma}\right)^2\right)^{\beta-1}\cdot\left(1-\left(\frac{p}{\gamma}\right)^2\right)g'(p)dp\\
&=\frac{1}{\gamma B(\beta)}\int_{-\gamma}^{\gamma}\left(1-\left(\frac{p}{\gamma}\right)^2\right)^{\beta}g'(p)dp\\
&=^{(a)}\frac{1}{\gamma B(\beta)}\left[\left(1-\left(\frac{p}{\gamma}\right)^2\right)^{\beta}g(p)\Bigg|_{-\gamma}^{\gamma}-\int_{-\gamma}^{\gamma}\left(\left(1-\left(\frac{p}{\gamma}\right)^2\right)^{\beta}\right)'g(p)dp\right]\\
&=\frac{1}{\gamma B(\beta)}\left[\int_{-\gamma}^{\gamma}\left(1-\left(\frac{p}{\gamma}\right)^2\right)^{\beta-1}\cdot\frac{2\beta p}{\gamma^2}g(p)dp\right]\\
&=\frac{2\beta}{\gamma^2}\mathbb{E}_{p\sim\pi}\left[p\cdot g(p)\right],
\end{aligned}$$

where in (a) we used integration by parts. This concludes the proof. □

Applying the above results to each coordinate and summing the equalities gives Lemma D.1.

Lemma D.1 (Fingerprinting lemma with a scaling matrix)**.** *Fix $d\in\mathbb{N}$. Let $\mathcal{Z}=\{\pm1\}^d$ and let $\beta>0$ be arbitrary. Consider arbitrary $0<\gamma\le1$. For every $\mu\in[-\gamma,\gamma]^d$, let $\mathcal{D}_\mu$ be the product distribution on $\mathcal{Z}$ with mean μ, i.e., for every $z\in\mathcal{Z}$, we have $\mathcal{D}_\mu=\prod_{k=1}^{d}\left(\frac{1+z^k\mu^k}{2}\right)$ let Λ_μ be a diagonal matrix of size d where the i-th diagonal element*

is given by $\Lambda_\mu^{ii} = \frac{1-(\mu^i/\gamma)^2}{1-(\mu^i)^2}$, *and let* $\phi_\mu(\theta, z) = \langle \theta, \Lambda_\mu(z-\mu)\rangle$. *Let* $\pi = \mathsf{s\text{-}beta}_{[-\gamma,\gamma]}(\beta,\beta)^{\otimes d}$ *be a prior. Then, for every algorithm* $\mathcal{A}_n : \mathcal{Z}^n \to \mathcal{M}_1(\mathbb{R}^d)$, *we have*

$$\mathbb{E}_{\mu\sim\pi}\mathbb{E}_{S_n\sim\mathcal{D}_\mu^{\otimes n},\hat\theta\sim\mathcal{A}_n(S_n)} \sum_{Z\in S_n} \phi_\mu(\hat\theta, Z) = \frac{2\beta}{\gamma^2}\mathbb{E}_{\mu\sim\pi}\left\langle \mu, \mathbb{E}_{S_n\sim\mathcal{D}_\mu^{\otimes n},\hat\theta\sim\mathcal{A}_n(S_n)}[\hat\theta]\right\rangle.$$

Proof. For a sample $S_n = (Z_1,\dots,Z_n)$ and $j\in[j]$, we will use $S_n^j \in \mathbb{R}^n$ to denote a vector $(Z_1^j,\dots,Z_n^j)$ of j^{th} coordinates. For each coordinate $j\in[d]$, let $f_j\colon \{\pm1\}^n \to \mathbb{R}$ the function such that

$$f_j(S_n^j) = \mathbb{E}_{\mu\sim\pi}\mathbb{E}_{S_n\sim\mathcal{D}_\mu^{\otimes n},\hat\theta\sim\mathcal{A}_n(S_n)}\left[\hat\theta^i \;\middle|\; S_n^j\right]$$

In other words, $f_j(X)$ is the expected value of θ^j, given that j^{th} coordinates of samples in S_n are given by X. Applying the result of Lemma D.1 to f_j, we have

$$\begin{aligned}
&\mathbb{E}_{\mu^j\sim\pi^j}\mathbb{E}_{S_n^j\sim(\mu^j)^{\otimes n}}\left[\Lambda_\mu^{jj}\cdot f_j(S_n^j)\sum_{i=1}^n (Z_i^j-\mu^j)\right]\\
&\quad= \mathbb{E}_{\mu^j\sim\pi^j}\mathbb{E}_{S_n^j\sim(\mu^j)^{\otimes n}}\left[\frac{1-(\mu^j/\gamma)^2}{1-(\mu^j)^2} f_j(S_n^j)\sum_{i=1}^n (Z_i^j-\mu^j)\right]\\
&\quad= \frac{2\beta}{\gamma^2}\mathbb{E}_{\mu^j\sim\pi^j}\left[\mu^j\cdot \mathbb{E}_{S_n^j\sim(\mu^j)^{\otimes n}} f_j(S_n^j)\right].
\end{aligned}$$

By the law of total expectation, we get

$$\mathbb{E}_{\mu\sim\pi}\mathbb{E}_{S_n\sim\mathcal{D}_\mu^{\otimes n},\hat\theta\sim\mathcal{A}_n(S_n)}\left[\Lambda_\mu^{jj}\cdot\hat\theta^j\sum_{i=1}^n(Z_i^j-\mu^j)\right] = \frac{2\beta}{\gamma^2}\mathbb{E}_{\mu\sim\pi}\left[\mu^j\cdot\mathbb{E}_{S_n\sim\mathcal{D}_\mu^{\otimes n},\hat\theta\sim\mathcal{A}_n(S_n)}\hat\theta^j\right].$$

Finally, summing the above over all coordinates $j\in[d]$, we obtain

$$\mathbb{E}_{\mu\sim\pi}\mathbb{E}_{S_n\sim\mathcal{D}_\mu^{\otimes n},\hat\theta\sim\mathcal{A}_n(S_n)}\left[\left\langle\Lambda_\mu\hat\theta, \sum_{i=1}^n(Z_i-\mu)\right\rangle\right] = \frac{2\beta}{\gamma^2}\mathbb{E}_{\mu\sim\pi}\left[\left\langle\mu, \mathbb{E}_{S_n\sim\mathcal{D}_\mu^{\otimes n},\hat\theta\sim\mathcal{A}_n(S_n)}\hat\theta\right\rangle\right],$$

as desired. □

E Hard problem constructions and proofs of subgaussian trace value lower bounds

E.1 Proofs for ℓ_p-geometries (Theorem 5.1)

First, recall here the construction of the hard problems $\mathcal{P}_{k,p}$ in Equation (7), parameterized by $k\in[d]$

$$\Theta = \mathcal{B}_\infty(d^{-1/p}),\quad \mathcal{Z} = \{z\in\{0,\pm1\}^d\colon \|z\|_0 = k\},\quad f(\theta,z) = -k^{-1/q}\langle\theta,z\rangle. \tag{$\mathcal{P}_{k,p}$}$$

First, we show in the simple proposition below that α-learners for linear problems must agree with the distribution mean.

Proposition E.1. *Let* $\mathcal{A}_n$ *be an* α*-learner for* $\mathcal{P}_{k,p}$. *Let* $\mathcal{D}\in\mathcal{M}_1(\mathcal{Z})$ *be a distribution with mean* $\mu = \mathbb{E}_{Z\sim\mathcal{D}}[Z]$. *Then, we have*

$$\mathbb{E}_{S_n\sim\mathcal{D}^{\otimes n},\hat\theta\sim\mathcal{A}_n(S_n)}\left[\left\langle\mu,\hat\theta\right\rangle\right] \ge d^{-1/p}\|\mu\|_1 - k^{1/q}\alpha.$$

Proof. Since $\mathcal{A}_n$ is an α-learner, we have

$$\begin{aligned}
\alpha &\ge \mathbb{E}\left[F_\mathcal{D}(\hat\theta)\right] - \inf_{\theta\in\Theta}F_\mathcal{D}(\theta)\\
&= \mathbb{E}_{S_n\sim\mathcal{D}^{\otimes n},\hat\theta\sim\mathcal{A}_n(S_n)}\mathbb{E}_{Z\sim\mathcal{D}}\left[f(\hat\theta,Z)\right] - \inf_{\theta\in\Theta}\mathbb{E}_{Z\sim\mathcal{D}}f(\theta,Z)\\
&= k^{-1/q}\left[\sup_{\theta\in\Theta}\mathbb{E}_{Z\sim\mathcal{D}}\langle\theta,Z\rangle - \mathbb{E}_{S_n\sim\mathcal{D}^{\otimes n},\hat\theta\sim\mathcal{A}_n(S_n)}\mathbb{E}_{Z\sim\mathcal{D}}\left\langle\hat\theta,Z\right\rangle\right]\\
&= k^{-1/q}\left[\sup_{\theta\in\Theta}\langle\theta,\mu\rangle - \mathbb{E}_{S_n\sim\mathcal{D}^{\otimes n},\hat\theta\sim\mathcal{A}_n(S_n)}\left\langle\hat\theta,\mu\right\rangle\right],
\end{aligned}$$

which, after rearranging, becomes

$$\begin{aligned}\mathbb{E}_{S_n\sim\mathcal{D}^{\otimes n},\hat{\theta}\sim\mathcal{A}_n(S_n)}\left[\left\langle\mu,\hat{\theta}\right\rangle\right] &\geq \sup_{\theta\in\Theta}\langle\mu,\theta\rangle - k^{1/q}\cdot\alpha \\ &= d^{-1/p}\left\|\mu\right\|_1 - k^{1/q}\alpha,\end{aligned}$$

where in the last transition we used duality of ℓ_∞ and ℓ_1 norms and the fact that $\Theta = \mathcal{B}_\infty(d^{-1/p})$. This concludes the proof. □

In the next lemma, we show that every α-learner for $\mathcal{P}_{k,p}$ needs to have a large correlation with the training samples in order to achieve small excess risk. The proof is an application of Lemma 4.2 combined with Proposition E.1.

Lemma E.2. *Let $\alpha \leq 1/6$, and suppose $k \in [d]$ is such that $k \geq (6\alpha)^p d$. Then, for every α-learner $\mathcal{A}_n$ for $\mathcal{P}_{k,p}$, there exists $\mu \in [-k/d, k/d]^d$ and distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z}_k)$ with mean μ such that the following holds: let*

$$\phi(\theta, Z) := \frac{d^{1/p}}{\sqrt{k}}\left\langle\theta, \left(Z - \frac{d}{k}\mu\right)\right\rangle_{\mathrm{supp}(Z)},$$

then,

$$\mathbb{E}_{S_n\sim\mathcal{D}^{\otimes n},\hat{\theta}\sim\mathcal{A}_n(S_n)}\left[\sum_{i=1}^n \phi(\hat{\theta}, Z_i)\right] \geq \frac{d^{1-1/p}}{18k^{1/2-1/p}\alpha}$$

Proof. Let

$$\beta = \left(\frac{k^{1/p}}{6d^{1/p}\alpha}\right)^2 \geq 1,$$

and $\pi = \mathsf{s\text{-}beta}_{[-k/d,k/d]}(\beta,\beta)$. Then, using Corollary A.15, we have

$$\begin{aligned}\mathbb{E}_{\mu\sim\pi}[\left\|\mu\right\|_1] &= d\mathbb{E}_{\mu\sim\pi}|\mu^1| \\ &\geq d\cdot\frac{k/d}{3\sqrt{\beta}} \\ &= \frac{k}{3(k^{1/p}/6d^{1/p}\alpha)} \\ &= 2\alpha d^{1/p}k^{1-1/p} \\ &= 2\alpha d^{1/p}k^{1/q}.\end{aligned}$$

Then, using Lemma 4.2 we have

$$\begin{aligned}\mathbb{E}_{\mu\sim\pi}\mathbb{E}_{S_n\sim\mathcal{D}_{\mu,k}^{\otimes n}}\sum_{i=1}^n\left\langle\hat{\theta}, \left(Z_i - \frac{d}{k}\mu\right)_{\mathrm{supp}(Z_i)}\right\rangle &= \frac{2d\beta}{k}\mathbb{E}_{\mu\sim\pi}\mathbb{E}_{S_n\sim\mathcal{D}_{\mu,k}^{\otimes n},\hat{\theta}\sim\mathcal{A}_n(S_n)}\left[\left\langle\mu,\hat{\theta}\right\rangle\right] \\ &\geq^{(a)} \frac{2d\beta}{k}\mathbb{E}_{\mu\sim\pi}\left[d^{-1/p}\left\|\mu\right\|_1 - k^{1/q}\alpha\right] \\ &\geq^{(b)} \frac{2d\beta}{k}\left[d^{-1/p}\cdot 2\alpha d^{1/p}k^{1/q} - k^{1/q}\alpha\right] \\ &= \frac{2d}{k}\cdot\left(\frac{k^{1/p}}{6d^{1/p}\alpha}\right)^2\cdot k^{1/q}\alpha \\ &= \frac{2d}{k}\cdot\left(\frac{k^{1/p}}{6d^{1/p}\alpha}\right)^2\cdot k^{1/q}\alpha \\ &= \frac{d^{1-2/p}k^{1/p}}{18\alpha}.\end{aligned}$$

Since the above holds in expectation over draws of μ, there exists at least one value of μ for which the above holds; let $\mathcal{D} = \mathcal{D}_{k,\mu}$. Then, letting

$$\phi(\theta, Z) := \frac{d^{1/p}}{\sqrt{k}}\left\langle\theta, \left(Z - \frac{d}{k}\mu\right)\right\rangle_{\mathrm{supp}(Z)},$$

we obtain

$$\mathbb{E}_{S_n\sim\mathcal{D}^{\otimes n},\hat{\theta}\sim\mathcal{A}_n(S_n)}\left[\sum_{i=1}^n \phi(\hat{\theta},Z_i)\right] = \mathbb{E}_{\mu\sim\pi}\mathbb{E}_{S_n\sim\mathcal{D}_{\mu,k}^{\otimes n}}\sum_{i=1}^n \frac{d^{1/p}}{\sqrt{k}}\left\langle \hat{\theta}, \left(Z_i - \frac{d}{k}\mu\right)_{\mathrm{supp}(Z_i)}\right\rangle$$
$$\geq \frac{d^{1-1/p}}{18k^{1/2-1/p}\alpha},$$

as desired. □

We now argue that the pair $(\phi,\mathcal{D})$ from the lemma above (with ϕ scaled by some constant) constitutes a valid subgaussian tracer. In particular, the lemma below shows that $\{\phi(\theta,Z)\}_{\theta\in\Theta}$ induces a $O(1)$-subgaussian process w.r.t. $(\Theta,\|\cdot\|_\Theta)$ norm.

Lemma E.3. *Fix $d\in\mathbb{N}$. Let $\mu\in[-k/d,k/d]^d$ be arbitrary. Let $\phi:\Theta\times\mathcal{Z}_k\to\mathbb{R}$ be as in Lemma E.2. Let $\mathcal{D}_{\mu,k}$ be the data distribution from Definition 4.1 for some μ, and consider $Z\sim\mathcal{D}_{\mu,k}$. Then, $\{\phi(\theta,Z)\}_{\theta\in\Theta}$ is a C-subg process w.r.t. to $(\Theta,\|\cdot\|_\Theta)$ for some universal constant $C>0$.*

Proof. Let $J\in\binom{[d]}{k}$ be an arbitrary coordinate subset of size k, and, recalling Definition 4.1, let Z_J be a random variable with PMF given by $P_{\mu,k,J}$. Then, Z is a uniform mixture of $\{Z_J\}_{J\in\binom{[d]}{k}}$.

Fix $J\in\binom{[d]}{k}$, and let $\theta_1,\theta_2\in\Theta$ be two arbitrary points. First, we upper bound a subgaussian norm of $\phi(\theta_1,Z_J)-\phi(\theta_2,Z_J)$. We have

$$\begin{aligned}
\|\phi(\theta_1,Z_J)-\phi(\theta_2,Z_J)\|_{\psi_2} &= \frac{d^{1/p}}{\sqrt{k}}\|\langle\theta_1-\theta_2,Z_J\rangle_J\|_{\psi_2}\\
&= \frac{d^{1/p}}{\sqrt{k}}\left\|\sum_{j\in J}(\theta_1^j-\theta_2^j)Z_J^j\right\|_{\psi_2}\\
&\leq^{(a)} \frac{d^{1/p}}{\sqrt{k}}\sqrt{C_1\sum_{j\in J}\left\|(\theta_1^j-\theta_2^j)Z_J^j\right\|_{\psi_2}^2}\\
&\leq^{(b)} \frac{d^{1/p}}{\sqrt{k}}\sqrt{C_1\sum_{j\in J}C_2|\theta_1^j-\theta_2^j|^2}\\
&= \frac{d^{1/p}}{\sqrt{k}}\cdot\sqrt{C_1C_2}\sqrt{k}\,\|\theta_1-\theta_2\|_\infty\\
&=^{(c)} \sqrt{C_1C_2}\,\|\theta_1-\theta_2\|_\Theta,
\end{aligned}$$

where $C_{1,2}>0$ are universal constants, in (a) we apply Lemma A.7, in (b) we apply Proposition A.9, and in (c) we use that, since $\Theta=\mathcal{B}_\infty(d^{-1/p})$, we have $\|\cdot\|_\Theta=d^{1/p}\|\cdot\|_\infty$. Thus, letting $C=\sqrt{C_1C_2}$, we have

$$\|\phi(\theta_1,Z_J)-\phi(\theta_2,Z_J)\|_{\psi_2}\leq C\,\|\theta_1-\theta_2\|_\Theta.$$

Now, note that $\phi(\theta_1,Z)-\phi(\theta_2,Z)$ has the same distribution as a uniform mixture of $\{\phi(\theta_1,Z_J)-\phi(\theta_2,Z_J)\}_{J\in\binom{[d]}{k}}$. Then, by Proposition A.8, we also have

$$\|\phi(\theta_1,Z)-\phi(\theta_2,Z)\|_{\psi_2}\leq C\,\|\theta_1-\theta_2\|_\Theta,$$

which satisfies the first condition in Definition 3.1. Finally, by plugging $\theta_2=0$ into the above, we have

$$\|\phi(\theta_1,Z)\|_{\psi_2}\leq C\,\|\theta_1\|_\Theta\leq C,$$

which satisfies the second condition in Definition 3.1. Thus, $\{\phi(\theta,Z)\}_{\theta\in\Theta}$ is a C-subgaussian process w.r.t. $(\Theta,\|\cdot\|_\Theta)$, as desired. □

Finally, we lower bound the subgaussian trace value of $\mathcal{P}_{k,p}$.

Lemma E.4. *Let $\alpha \le 1/6$ and $d \in \mathbb{N}$ be arbitrary, and let $k \in [d]$ be such that $k \le (6\alpha)^p d$. Let $\mathcal{P}_{k,p}$ be as in Equation (7). Then, the following subgaussian trace value lower bounds hold for every $p \in [1,\infty)$ and some $\kappa \le c_1\sqrt{d}$,*

$$\mathrm{Tr}_\kappa(\mathcal{P}_{k,p}; n, \alpha) \ge c_2 \frac{d^{1-1/p}}{k^{1/2-1/p} n\alpha}$$

where $c_{1,2} > 0$ are universal constants.

Proof. Let $C > 0$ be the constant from Lemma E.3, and ϕ be as in Lemma E.2. Then, $\{\phi(\theta, Z)/C\}_{\theta\in\Theta}$ is a 1-subgaussian process w.r.t. $(\Theta, \|\cdot\|_\Theta)$. Moreover, $\phi(\theta, Z)/C \le \sqrt{k}/C \le \sqrt{d}/C$. Then, letting $\kappa = \sqrt{d}/C$ and using Lemma E.2, the subgaussian trace value of $\mathcal{P}_{k,p}$ can be lower bounded by

$$\mathrm{Tr}_\kappa(\mathcal{P}_{k,p}; n, \alpha) \ge \frac{1}{C}\mathbb{E}_{S_n\sim\mathcal{D}^{\otimes n}, \hat\theta\sim\mathcal{A}_n(S_n)} \frac{1}{n}\left[\sum_{i=1}^n \phi(\hat\theta, Z_i)\right] \ge \frac{d^{1-1/p}}{18C\cdot k^{1/2-1/p} n\alpha},$$

as desired. □

Now we are ready to prove Theorem 5.1.

Theorem 5.1. *Let $\mathcal{P}_{k,p}$ be the family of problems described in Equation (7). There exist universal constants $c_1, c_2 > 0$ such that, for all $\alpha \in (0, 1/6]$ and $d \in \mathbb{N}$, the following subgaussian trace value lower bounds hold for all $p \in [1,\infty)$ and $\kappa \le c_1\sqrt{d}$:*

(i) For $p \le 2$ and $k = d$, we have $\mathrm{Tr}_\kappa(\mathcal{P}_{k,p}; n, \alpha) \ge c_2 \frac{\sqrt{d}}{n\alpha}$.

(ii) For $p \ge 2$ and $k = (6\alpha)^p d \vee 1$, we have $\mathrm{Tr}_\kappa(\mathcal{P}_{k,p}; n, \alpha) \ge c_2 \left[\frac{\sqrt{d}}{n(6\alpha)^{p/2}} \wedge \frac{d^{1-1/p}}{n\alpha}\right]$.

Proof. The theorem is a direct consequence of Lemma E.4. For $p \le 2$, plug in $k = d$ into the statement of Lemma E.4. We obtain

$$\mathrm{Tr}_\kappa(\mathcal{P}_{k,p}; n, \alpha) \ge c_2 \frac{d^{1-1/p}}{k^{1/2-1/p} n\alpha} = c_2 \frac{\sqrt{d}}{n\alpha}.$$

For $p \ge 2$, plug in $k = (6\alpha)^p \vee 1$. We obtain,

$$\begin{aligned}
\mathrm{Tr}_\kappa(\mathcal{P}_{k,p}; n, \alpha) &\ge c_2 \frac{d^{1-1/p}}{k^{1/2-1/p} n\alpha} \\
&= c_2 \left[\frac{d^{1-1/p}}{n\alpha} \wedge \frac{\sqrt{d}}{(6\alpha)^{p(1/2-1/p)} n\alpha}\right] \\
&= c_2 \left[\frac{d^{1-1/p}}{n\alpha} \wedge \frac{\sqrt{d}}{(6\alpha)^{p/2} n}\right],
\end{aligned}$$

as desired. □

E.2 Proofs for ℓ_1-geometry

E.2.1 Intuition

Refinement for $p = 1$. While the above construction also yields a traceability result for $p = 1$, it is suboptimal for the following simple reason: for $k = d$, the problem in Equation (7) only requires $\Theta(1/\alpha^2)$ samples to learn, thus, it is impossible to trace out $\Omega(\log(d)/\alpha^2)$ samples. On the other hand, the problem in Equation (6) requires $\Theta(\log(d)/\alpha^2)$ samples to learn but is not traceable. The intuition we follow here is to modify the construction in Equation (7) to make Θ "look" more like an ℓ_1-ball to drive up the sample complexity while still avoiding the counterexample with an ERM learner from the beginning of the section. In particular, we consider the following ℓ_1-problem,

$$\Theta = \mathcal{B}_1(1) \cap \mathcal{B}_\infty(1/s), \quad \mathcal{Z} = \{\pm 1\}^d, \quad f(\theta, z) = -\langle z, \theta\rangle, \tag{17}$$

for a suitably chosen $s \in [d]$. Note that, if we choose $s \gg 1$, Θ above is a polytope with much more vertices ($2^s\binom{d}{s}$) than an ℓ_1 ball ($2d$), which would intuitively force a learner like an ERM to reveal more information about the training sample. On a technical level, selecting large s improves the subgaussian constant of a tracer; however, selecting s that is too large shrinks the diameter of the set, and thus, the problem becomes easier to learn. We must trade off these two aspects, and carefully set the value of s. As it turns out, the optimal choice is $s \propto d^{1-c}$ for an arbitrary small $c > 0$ in order to establish Theorem 2.6. The remainder of the proof is rather technical and hence is deferred to Appendix F.2.

E.2.2 Formal Proof

For technical reasons, we will need the following refinement of Lemma 3.4 for the special case when the function ϕ is convex and Θ is a polytope. In the proof, we use Lemma C.1.

Lemma E.5. *Fix $n, d \in \mathbb{N}$. Suppose $\Theta \subset \mathbb{R}^d$ is (i) a subset of a unit ball in some norm $\|\cdot\|$, and (ii) Θ is a polytope with N vertices. Let $\phi\colon \Theta \times \mathcal{Z} \to \mathbb{R}$ be a measurable function that is convex in its first argument. Let $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ be such that $\phi(\theta, Z)$ is a σ-subgaussian process w.r.t. $(\Theta, \|\cdot\|)$.Let $(Z_1, \dots, Z_n) \sim \mathcal{D}^{\otimes n}$. Then, for every $t \geq 0$,*

$$\Pr\left(\sup_{\theta\in\Theta}\sqrt{\sum_{i=1}^{n}\left[\phi(\theta, Z_i)\right]^2} \leq C\sigma\left[\sqrt{n} + \sqrt{\log(N)} + t\right]\right) \geq 1 - 2\exp(-t^2)$$

where $C > 0$ is some universal constant.

Proof. Similarly to the proof of Lemma C.1, let Φ_θ denote the following random vector

$$\Phi_\theta = \begin{bmatrix}\phi(\theta, Z_1)\\ \vdots \\ \phi(\theta, Z_n)\end{bmatrix}.$$

Then, observe that, the desired quantity is equal to

$$\sup_{\theta\in\Theta}\sqrt{\sum_{i=1}^{n}\left[\phi(\theta, Z_i)\right]^2} = \sup_{\theta\in\Theta}\|\Phi_\theta\|_2 = \sup_{\theta\in\Theta, x\in\mathbb{S}^{n-1}}\langle x, \Phi_\theta\rangle.$$

Let V be the set of vertices of Θ with $|V| \leq N$. Since ϕ is convex in its first argument, the supremum above is attained in one of the vertices of Θ. Thus,

$$\sup_{\theta\in\Theta}\sqrt{\sum_{i=1}^{n}\left[\phi(\theta, Z_i)\right]^2} = \sup_{\theta\in V, x\in\mathbb{S}^{n-1}}\langle x, \Phi_\theta\rangle = \sup_{\theta\in V}\sqrt{\sum_{i=1}^{n}\left[\phi(\theta, Z_i)\right]^2}.$$

Thus, we may apply Lemma C.1 to V instead of Θ. Trivially, we have

$$\mathcal{N}(V, \|\cdot\|, \varepsilon) \leq |V| = N.$$

Then, we have with probability $1 - 2\exp(-t^2)$,

$$\begin{aligned}\sup_{\theta\in\Theta}\sqrt{\sum_{i=1}^{n}\left[\phi(\theta, Z_i)\right]^2} &\leq C\sigma\left(\sqrt{n} + \int_0^1\sqrt{\log\mathcal{N}(\Theta; \|\cdot\|, \varepsilon)}d\varepsilon + t\right)\\ &\leq C\sigma\left(\sqrt{n} + \sqrt{\log(N)} + t\right),\end{aligned}$$

as desired. □

Intuitively, in the special case when Θ is a polytope, the log-number of vertices becomes "effective dimension" instead of d, due to the fact that ϕ satisfies the convexity requirement. In some cases, we can have $d \gg \log(N)$, in which the above gives a tighter concentration. In particular, this is a case in our construction for ℓ_1 geometry in Equation (17). With the above result, we can also establish the following refinement of Theorem 3.5.

Theorem E.6. *Fix $n \in \mathbb{N}$, $d \in \mathbb{N}$, $\kappa > 0$ and $\alpha \in [0,1]$. Consider an arbitrary SCO problem $\mathcal{P} = (\Theta, \mathcal{Z}, f)$, and suppose Θ is a polytope with $N > 0$ vertices. Let T be defined as,*

$$T := \mathrm{Tr}_\kappa(\mathcal{P}; n, \alpha).$$

Then, for some constant $c > 0$, every α-learner $\mathcal{A}_n$ is (ξ, m)-traceable with

$$\xi = \exp(-cT^2), \quad m = c\left[\frac{n^2T^2}{n + \log(N)} - \frac{16\kappa^2 n}{\exp(n + \log(N))}\right].$$

Proof. The proof is identical to Theorem 3.5, but using Lemma E.5 instead of Lemma 3.4, and thus, replacing d with $\log(N)$ everywhere. We omit the details. □

Now, recall the construction from Equation (17),

$$\mathcal{Z} = \{\pm 1\}^d, \quad \Theta = \mathcal{B}_1(1) \cap \mathcal{B}_\infty(1/s), \quad f(\theta, Z) = -\langle \theta, Z \rangle . \tag{18}$$

It is easy to see that $f(\cdot, Z)$ above is 1-Lipschitz w.r.t. ℓ_1 as $\mathcal{Z} \subset \mathcal{B}_\infty(1)$. We have the following claim.

Lemma E.7. *Let $\mathcal{P}_s$ be as in (18), $n \in \mathbb{N}$ and $1/8 > \alpha > 0$. Then,*

$$\mathrm{Tr}_\kappa(\mathcal{P}_s; n, \alpha) \geq \frac{c\sqrt{s}\log\left(\frac{d}{16(s \vee 14)}\right)}{n\alpha},$$

where $\kappa \leq c'\sqrt{s}$ for some constants $c, c' > 0$

Proof. We aim to use Lemma D.1 to characterize the subgaussian trace value. Consider the construction of the prior in Lemma D.1 with the following parameters: $\gamma = 8\alpha \leq 1$ and $\beta = 1 + \frac{1}{2}\log\left(\frac{d}{16(s \vee 14)}\right)$. Then, by combining Lemma D.1 with and Proposition E.1, there exist a prior π and a family $\{\Lambda_\mu\}$ of diagonal matrices with non-negative diagonal entries bounded by 1 from above, such that

$$\begin{aligned}
\mathbb{E}_{\mu\sim\pi}\mathbb{E}_{Z\sim\mu^{\otimes n}}\left\langle \hat{\theta}, \sum_{i=1}^n \Lambda_\mu(Z_i - \mu)\right\rangle &= \frac{2\beta}{\gamma^2}\mathbb{E}_{\mu\sim\pi}\left[\sup_\theta \langle \theta, \mu\rangle - \alpha\right] \\
&\geq \frac{2\beta}{\gamma^2}\mathbb{E}_{\mu\sim\pi}\left[\frac{1}{s}\sup_{\substack{I\subset[d]\\|I|=s}}\sum_{i\in I}|\mu_i| - \alpha\right] \\
&\geq^{(a)} \frac{\beta}{\gamma^2}\left(\frac{\gamma}{2} - 2\alpha\right) \\
&\geq \frac{\log\left(\frac{d}{16(s\vee 14)}\right)}{32\alpha^2}\cdot 2\alpha \\
&\geq \frac{\log\left(\frac{d}{16(s\vee 14)}\right)}{16\alpha},
\end{aligned}$$

where in (a) we used Proposition 5 of [SU17]. Since this holds in expectation over μ, it holds for at least one choice of μ. Let μ be that value. Now, let $\phi(\theta, Z) = C^{-1/2}\sqrt{s}\left\langle \hat{\theta}, \Lambda(Z - \mu)\right\rangle$, where C is the absolute constant from Lemma A.7. For all $\theta, \theta' \in \Theta$, we have

$$\begin{aligned}
\|\phi(\theta, Z) - \phi(\theta', Z)\|_{\psi_2} &\leq C^{-1/2}\sqrt{s}\,\|\langle \theta' - \theta, \Lambda(Z - \mu)\rangle\|_{\psi_2} \\
&\leq^{(a)} C^{-1/2}\sqrt{s}\,\|\theta' - \theta\|_2 \cdot C^{1/2}\max_i \left\|\Lambda^i\left(Z^i - \mu^i\right)\right\|_{\psi_2} \\
&\leq \sqrt{s}\,\|\theta' - \theta\|_2 \\
&\leq^{(b)} \|\theta' - \theta\|_\Theta ,
\end{aligned}$$

where in (a) we apply Lemma A.7, and in (b) we use that for every $\theta \in \Theta$, we have $\|\theta\|_2 \le 1/\sqrt{s}$, thus, $\sqrt{s}\,\|\cdot\|_2 \le \|\cdot\|_\Theta$. Plugging $\theta' = 0$ gives

$$\|\phi(\theta, Z)\|_{\psi_2} \le \|\theta\|_\Theta \le 1.$$

Thus, $\{\phi(\theta, Z)\}_{\theta\in\Theta}$ is a 1-subg process w.r.t. $(\Theta, \|\cdot\|_\Theta)$. Finally, we have

$$|\phi(\theta, Z)| \le C^{-1/2}\sqrt{s}$$

Therefore, setting $\kappa = C^{-1/2}\sqrt{s}$ and noting that ϕ is linear (and therefore convex) in its first argument, we have

$$\begin{aligned}\mathrm{Tr}_\kappa(\mathcal{P}_s; n, \alpha) &\ge \mathbb{E}_{\mu\sim\pi}\mathbb{E}_{Z\sim\mu^{\otimes n}} \sum_{i=1}^{n} \frac{1}{n}\phi(\theta, Z_i) \\ &\ge \frac{C^{-1/2}\sqrt{s}\log\left(\frac{d}{16(s\vee 14)}\right)}{n\alpha},\end{aligned}$$

as desired. □

F Proofs of the main results (Section 2.2)

F.1 Proof of Theorem 2.5

Theorem 2.5. *There exists a universal constant $c > 0$ such that, for all $p \in [1, 2)$, if d, n, $\xi \in (0, 1/e)$, and $\alpha > 0$ are such that*

$$\frac{c}{\sqrt{n}} \le \alpha \le \min\left\{c \cdot \sqrt{\frac{d}{n^2\log(1/\xi)}},\ \frac{1}{6}\right\}, \tag{3}$$

then there exist an ℓ_p SCO problem that every α-learner is (ξ, m)-traceable with $m \in \Omega\left(\alpha^{-2}\right)$.

Proof. We begin by noting that the interval for α in Equation (3) is non-empty only if

$$\frac{c}{\sqrt{n}} \le \min\left\{c \cdot \sqrt{\frac{d}{n^2\log(1/\xi)}},\ \frac{1}{6}\right\} \le c \cdot \sqrt{\frac{d}{n^2}},$$

where we used $\xi < 1/e$ in the last transition. Via straightforward algebra, the above implies $d \ge n$, thus, we may without loss of generality assume $d \ge n$ in the remainder of the proof.

Let $\mathcal{P}_{k,p}$ be as in Equation (7), and set k as in Theorem 5.1. Then, Theorem 5.1 gives the following lower bound on the subgaussian trace value in this case:

$$T := \mathrm{Tr}_\kappa(\mathcal{P}_{k,p}; n, \alpha) \ge c_2\frac{\sqrt{d}}{n\alpha} \ge \frac{c_2}{c}\sqrt{\log\left(\frac{1}{\xi}\right)},$$

for some $\kappa \le c_1\sqrt{d}$. Thus, provided c is small enough ($c \le c_2$), by Theorem 3.5, every α-learner is (ξ, m)-traceable with m satisfying, for some universal constant $c' > 0$,

$$\begin{aligned}m &\ge c'\left[\frac{n^2T^2}{n+d} - \frac{16\kappa^2 n}{\exp(n+d)}\right] \\ &\ge c'\left[\frac{c_2^2 d}{n+d}\cdot\frac{1}{\alpha^2} - \frac{16c_1^2 dn}{\exp(n+d)}\right] \\ &\ge c'\left[\frac{c_2^2}{2}\cdot\frac{1}{\alpha^2} - \frac{16c_1^2 d^2}{\exp(d)}\right],\end{aligned} \tag{19}$$

where we used $d \ge n$ in the last inequality. Then, we have

$$m \in \Omega\left(\frac{1}{\alpha^2}\right),$$

as desired.

□

F.2 Proof of Theorem 2.6

Then, the proof of Theorem 2.6 follows.

Theorem 2.6. *There exists a universal constant $c > 0$ such that, if d is large enough and n, $\xi \in (0, 1/e)$, and $\alpha > 0$ are such that*

$$c \cdot \sqrt{\frac{\log(d)}{n}} \leq \alpha \leq \min\left\{c \cdot \frac{d^{0.49}}{n\sqrt{\log(1/\xi)}}, \frac{1}{8}\right\}, \text{ then} \tag{4}$$

there exists a ℓ_1 SCO problem that every α-learner is (ξ, m)-traceable with $m \in \Omega\left(\log(d)/\alpha^2\right)$.

Proof. We begin by noting that the interval for α in Equation (4) is non-empty only if

$$c \cdot \sqrt{\frac{\log(d)}{n}} \leq c \cdot \frac{d^{0.49}}{n\sqrt{\log(1/\xi)}} \leq c \cdot \frac{d^{0.49}}{n}$$

where we used $\xi < 1/e$ and $c < 1/6$ in the last transition. Via straightforward algebra, the above implies $d^{0.98}/\log(d) \geq n$, thus, we may without loss of generality assume $d^{0.98}/\log(d) \geq n$ in the remainder of the proof.

Let $s = d^{0.98}$. Note that, letting V be the set of vertices of the polytope $\Theta = \mathcal{B}_1(1) \cap \mathcal{B}_\infty(1/s)$, we have

$$V = \left\{z \in \left\{0, \pm\frac{1}{s}\right\} : \|z\|_0 = s\right\}.$$

The proof of this fact is straightforward and it is based on showing that every point in Θ can be written as a convex combination of the points in V. Thus,

$$\log|V| = \log\left(2^s \binom{d}{s}\right) \leq s\log(de/s) + s = s\log(de^2/s).$$

By the choice of s and using Lemma E.7, we have, for some $\kappa \leq c_1\sqrt{s}$,

$$\begin{aligned}
T := \mathrm{Tr}_\kappa(\mathcal{P}_s; n, \alpha) &\geq c_2 \frac{d^{0.49}\log\left(\frac{d}{16(s\vee 14)}\right)}{n\alpha} \\
&\geq^{(a)} c_2 \frac{d^{0.49}\log\left(\frac{d}{16s}\right)}{n\alpha} \\
&= c_2 \frac{d^{0.49}\log\left(\frac{d^{0.02}}{16}\right)}{n\alpha} \\
&= c_2 \frac{d^{0.49}\left(0.02\cdot\log(d) - \log(16)\right)}{n\alpha} \\
&\geq^{(b)} c_2 \frac{d^{0.49}\cdot 0.01\cdot\log(d)}{n\alpha} \\
&\geq^{(c)} \sqrt{\log(1/\xi)},
\end{aligned}$$

where (a) and (b) hold provided d (and thus s) is large enough, and (c) holds provided $c > 0$ in Equation (4) is small enough. By Theorem E.6, every α-learner is (ξ, m)-traceable, where

$$\begin{aligned}
m &\geq c'\left[\frac{n^2T^2}{n + \log|V|} - \frac{16\kappa^2 n}{\exp(n + \log|V|)}\right] \\
&\geq c'\left[\frac{0.01^2 c_2^2 \cdot d^{0.98}\log(d)}{n + \log|V|}\cdot\frac{\log(d)}{\alpha^2} - \frac{16c_1^2\cdot sn}{\exp(n + \log|V|)}\right].
\end{aligned}$$

Recall that $d^{0.98} \geq n\log(d) \geq n$ (for $d \geq 3$). Moreover, $\log|V| \leq s\log(de^2/s) = d^{0.98}\log(e^2 d^{0.02}) \leq Cd^{0.98}\log(d)$, for some universal $C > 0$. Thus, for d large enough,

$$m \in \Omega\left(\frac{\log(d)}{\alpha^2}\right),$$

as desired. □

F.3 Proof of Theorem 2.7

Theorem 2.7. *There exists a universal constant $c > 0$ such that, for all $p \in [2, \infty)$, if d, n, $\xi \in (0, 1/e)$, and $\alpha > 0$ are such that*

$$\frac{1}{6} \cdot \min\left\{\frac{1}{n^{1/p}}, \frac{d^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{n}}\right\} \le \alpha \le \min\left\{c \cdot \left(\frac{d}{n^2 \log(1/\xi)}\right)^{1/p}, \frac{1}{6}\right\}, \text{ then} \tag{5}$$

there exist an ℓ_p SCO problem such that every α-learner is (ξ, m)-traceable with $m \in \Omega\left(1/(6\alpha)^p\right)$.

Proof. Throughout, we assume c is a sufficiently small constant. Assume $c < 1/6$. Then, we begin by noting that the interval for α in Equation (5) is non-empty only if

$$\frac{1}{6} \cdot \frac{1}{n^{1/p}} \le c \cdot \left(\frac{d}{n^2 \log(1/\xi)}\right)^{1/p} \le \frac{1}{6} \cdot \left(\frac{d}{n^2}\right)^{1/p},$$

where we used $\xi < 1/e$ and $c < 1/6$ in the last transition. Via straightforward algebra, the above implies $d \ge n$, thus, we may without loss of generality assume $d \ge n$ in the remainder of the proof.

Let $\mathcal{P}_{k,p}$ be as in Equation (7), and set k as in Theorem 5.1. Then, Theorem 5.1 gives the following lower bound on the subgaussian trace value

$$T := \mathrm{Tr}_\kappa(\mathcal{P}_{k,p}; n, \alpha) \ge c_2 \left[\frac{d^{1-1/p}}{n\alpha} \wedge \frac{\sqrt{d}}{(6\alpha)^{p/2} n}\right] \tag{20}$$

for some $\kappa \le c_1\sqrt{d}$. Note that, from (5), we have

$$\alpha \ge \frac{1}{6} \cdot \frac{1}{n^{1/p}} \ge \frac{1}{6} \cdot \frac{1}{d^{1/p}}$$

Now, note that, the minimum in Equation (20) is achieved in the second term iff $\alpha \ge d^{-1/p}/6$. Then, the lower bound on the subgaussian trace value becomes

$$T \ge c_2 \frac{\sqrt{d}}{(6\alpha)^{p/2} n} \ge \frac{c_2}{(6c)^{p/2}} \cdot \sqrt{\log(1/\xi)} \ge \sqrt{\log(1/\xi)}$$

where the second transition follows from Equation (5) and the third transition holds whenever $c > 0$ is small enough (e.g., when $c \le c_2^{2/p}/6$). Then, by Theorem 3.5 every α-learner is (ξ, m)-traceable with m satisfying, for some universal constant $c' > 0$,

$$\begin{aligned} m &\ge c' \left[\frac{n^2 T^2}{n+d} - \frac{16\kappa^2 n}{\exp(n+d)}\right] \\ &\ge c' \left[\frac{c_2^2 d}{n+d} \cdot \frac{1}{(6\alpha)^p} - \frac{16 c_1^2 dn}{\exp(n+d)}\right] \\ &\ge c' \left[\frac{c_2^2}{2} \cdot \frac{1}{(6\alpha)^p} - \frac{16 c_1^2 d^2}{\exp(d)}\right], \end{aligned} \tag{21}$$

where we used $d \ge n$ in the last transition. Thus,

$$m \in \Omega\left(\frac{1}{(6\alpha)^p}\right),$$

as desired. □

F.4 Proof of Theorem 2.8

Theorem 2.8. *Let $p \in [2, \infty)$. There exist a universal constant $c > 0$ and an ℓ_p SCO problem $\mathcal{P} = (\Theta, \mathcal{Z}, f)$ such that every (ε, δ)-DP learner of $\mathcal{P}$ with $\varepsilon \le 1$ and $\delta \le c/n$ satisfies,*

$$\alpha \ge c \cdot \min\left\{\left(\frac{d}{\varepsilon^2 n^2}\right)^{\frac{1}{p}}, \frac{d^{1-1/p}}{\varepsilon n}, 1\right\}.$$

Proof. By Theorem 5.1, for some problem $\mathcal{P}$, we have

$$T := \mathrm{Tr}_\kappa\left(\mathcal{P}; n, \alpha\right) \geq c' \left[\frac{d^{1-1/p}}{n\alpha} \wedge \frac{\sqrt{d}}{n\alpha^{p/2}}\right]$$

Then Theorem 3.6 implies

$$\exp(\varepsilon) - 1 \geq c'' \left[T - 2\delta\kappa\right],$$

Note that for all $\varepsilon \leq 1$, we have $2\varepsilon \geq \exp(\varepsilon) - 1$. Thus,

$$2\varepsilon \geq c' \left[T - 2\delta\kappa\right].$$

which implies,

$$2(\varepsilon/c'' + \delta\kappa) \geq T \geq c' \left[\frac{d^{1-1/p}}{n\alpha} \wedge \frac{\sqrt{d}}{n\alpha^{p/2}}\right].$$

Then, for some $C > 0$,

$$C(\varepsilon \vee \delta\kappa) \geq \frac{d^{1-1/p}}{n\alpha} \wedge \frac{\sqrt{d}}{n\alpha^{p/2}}$$

Rearranging gives

$$\alpha \geq \frac{d^{1-1/p}}{Cn(\varepsilon \vee \delta\kappa)} \wedge \left(\frac{\sqrt{d}}{Cn(\varepsilon \vee \delta\kappa)}\right)^{2/p}$$

Note that if $\varepsilon \geq \delta\kappa$, the desired bound is immediate. For $\delta\kappa \geq \varepsilon$, we have, since $\delta \leq c/n$ and $\kappa \leq c'\sqrt{d}$,

$$\alpha \geq \frac{d^{1-1/p}}{C'\sqrt{d}} \wedge \left(\frac{\sqrt{d}}{C'\sqrt{d}}\right)^{2/p} \geq (C')^{-2/p},$$

for some $C' > 0$, as desired. □

F.5 Proof of Corollary 2.9

Corollary 2.9. *Let $\mathcal{Z} = \{\pm 1\}^d$, and suppose an estimator is given such that, given access to i.i.d. samples $Z_1, \ldots, Z_n \in \mathcal{Z}$, outputs $\hat{\mu}$ with $\mathbb{E}\left\|\hat{\mu} - \mathbb{E}[Z_1]\right\|_\infty \leq \alpha/2$. Then, there exists a universal constant $c > 0$ such that, if d is large enough and n, $\xi \in (0, 1/e)$, and $\alpha > 0$ satisfy Equation (4), then the estimator $\hat{\mu}$ is (ξ, m)-traceable with $m \in \Omega\left(\log(d)/\alpha^2\right)$.*

Proof. We will first show that we can use the mean estimation algorithm to solve the corresponding hard problem for ℓ_1-SCO. Specifically, consider the SCO problem as in Equation (17), and define the following learning algorithm based on the mean estimation. Let $\hat{\mu}$ be the output of mean estimator based on the samples $Z_1, \ldots, Z_n$, and let

$$\hat{\theta} := \arg\max_{\theta \in \Theta} \langle \theta, \hat{\mu} \rangle.$$

Let $\theta^\star$ be the population risk minimizer, and let μ be the true mean, that is, $\mu = \mathbb{E}[Z_1]$. Then, the excess risk of $\hat{\theta}$ can be upper bounded as:

$$\begin{aligned}
\langle \theta^\star, \mu \rangle - \left\langle \hat{\theta}, \mu \right\rangle &= \left\langle \theta^\star - \hat{\theta}, \mu \right\rangle \\
&= \left\langle \theta^\star - \hat{\theta}, \mu - \hat{\mu} \right\rangle + \left\langle \theta^\star - \hat{\theta}, \hat{\mu} \right\rangle \\
&\leq \left\|\theta^\star - \hat{\theta}\right\|_1 \left\|\mu - \hat{\mu}\right\|_\infty + \left\langle \theta^\star - \hat{\theta}, \hat{\mu} \right\rangle \\
&\leq 2\left\|\mu - \hat{\mu}\right\|_\infty + \left\langle \theta^\star - \hat{\theta}, \hat{\mu} \right\rangle,
\end{aligned}$$

where the last transition follows since $\hat{\theta}, \theta^\star$ both lie inside $\Theta \subset \mathcal{B}_1$. Now, by the choice of $\hat{\theta}$, the second term is non-positive. Thus,

$$\langle \theta^\star, \mu \rangle - \left\langle \hat{\theta}, \mu \right\rangle \leq 2\left\|\mu - \hat{\mu}\right\|_\infty.$$

Taking expectations on both sides, we get

$$\mathbb{E}\langle \theta^\star, \mu \rangle - \left\langle \hat{\theta}, \mu \right\rangle \leq 2\mathbb{E}\left\|\mu - \hat{\mu}\right\|_\infty \leq \alpha.$$

Applying the result of Theorem 2.6 to $\hat{\theta}$, we conclude the proof. □

G Connection between subgaussian trace value and non-private sample complexity

From the main part of the paper, we observe that the subgaussian trace value is typically inversely proportional to α. We start by proving two innocuous results (Propositions G.1 and G.3) that establish an absolute upper bound on subgaussian trace value. It will then allow us to extract lower bounds on α by plugging in our lower bounds on subgaussian trace value (Theorems G.2 and G.4). We start with the $p \in (1, \infty)$ case, and then consider $p = 1$.

G.1 Lower bounds for $p \in (1, \infty)$

Proposition G.1. *Fix $n \in \mathbb{N}$, $d \in \mathbb{N}$, and $\alpha \in [0, 1]$. Consider an arbitrary SCO problem $\mathcal{P} = (\Theta, \mathcal{Z}, f)$ in $\mathbb{R}^d$. Let $\mathrm{Tr}_\kappa(\mathcal{P}; n, \alpha)$ be the subgaussian trace value of problem $\mathcal{P}$. Then, we have*

$$\mathrm{Tr}_\kappa(\mathcal{P}; n, \alpha) \cdot \sqrt{n} \le C \cdot \sqrt{d},$$

for some universal constant $C > 0$.

Proof. Let $(\phi, \mathcal{D})$ be an arbitrary subgaussian tracer. Consider the process $\{X_\theta\}_{\theta \in \Theta}$ defined as

$$X_\theta := \frac{1}{n} \sum_{i=1}^n \phi(\theta, Z_i),$$

where $S_n = (Z_1, \dots, Z_n) \sim \mathcal{D}^{\otimes n}$. We will argue that $\{X_\theta\}_{\theta \in \Theta}$ is $O(1/\sqrt{n})$-subgaussian process w.r.t. $(\Theta, \|\cdot\|_\Theta)$. First, consider arbitrary $\theta_1, \theta_2 \in \Theta$. We have

$$\begin{aligned}
\|X_{\theta_1} - X_{\theta_2}\|_{\psi_2} &= \frac{1}{n} \left\| \sum_{i=1}^n (\phi(Z_i, \theta_1) - \phi(Z_i, \theta_2)) \right\|_{\psi_2} \\
&\le^{(a)} \frac{C}{n} \sqrt{\sum_{i=1}^n \|\phi(Z_i, \theta_1) - \phi(Z_i, \theta_2)\|_{\psi_2}^2} \\
&\le^{(b)} \frac{C}{n} \sqrt{\sum_{i=1}^n \|\theta_1 - \theta_2\|_\Theta} \\
&= \frac{C}{\sqrt{n}} \|\theta_1 - \theta_2\|_\Theta ,
\end{aligned}$$

where in (a) we applied Lemma A.7, and in (b) we used the fact that $\{\phi(\theta, Z_i)\}_{\theta \in \Theta}$ is 1-subgaussian process w.r.t. $(\Theta, \|\cdot\|_\Theta)$ for every $i \in [n]$. Moreover, for every $\theta \in \Theta$, we similarly have

$$\begin{aligned}
\|X_\theta\|_{\psi_2} &= \frac{1}{n} \left\| \sum_{i=1}^n \phi(Z_i, \theta) \right\|_{\psi_2} \\
&\le^{(a)} \frac{C}{n} \sqrt{\sum_{i=1}^n \|\phi(Z_i, \theta)\|_{\psi_2}^2} \\
&\le^{(b)} \frac{C}{n} \sqrt{\sum_{i=1}^n \|\theta\|_\Theta} \\
&= \frac{C}{\sqrt{n}} \|\theta\|_\Theta ,
\end{aligned}$$

where in (a) we applied Lemma A.7, and in (b) we used the fact that $\{\phi(\theta, Z_i)\}_{\theta \in \Theta}$ is 1-subgaussian process w.r.t. $(\Theta, \|\cdot\|_\Theta)$ for every $i \in [n]$. Thus, $\{X_\theta\}_{\theta \in \Theta}$ is $C/\sqrt{n}$-subgaussian

process w.r.t. $(\Theta, \|\cdot\|)$ as per Definition 3.1. Therefore, by Proposition A.10, we have, with probability at least $1 - 4\exp(-t^2)$

$$\begin{aligned}
\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(\theta, Z_i)\right| &\le \frac{Ct}{\sqrt{n}} + \frac{C}{\sqrt{n}}\int_0^1 \sqrt{\log \mathcal{N}\left(\Theta; \|\cdot\|_\Theta, \varepsilon\right)}d\varepsilon \\
&\le \frac{Ct}{\sqrt{n}} + \frac{C}{\sqrt{n}}\int_0^1 \sqrt{d\log\left(1+\frac{2}{\varepsilon}\right)}d\varepsilon \\
&\le \frac{Ct}{\sqrt{n}} + \frac{C\sqrt{d}}{\sqrt{n}} \\
&= \frac{C}{\sqrt{n}}\left[\sqrt{d} + t\right], \qquad (22)
\end{aligned}$$

where in second inequality we use [Wai19, Example 5.8]. Hence,

$$\begin{aligned}
\mathbb{E}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(\theta, Z_i)\right| &= \int_0^\infty \Pr\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(\theta, Z_i)\right| > u\right] du \\
&\le C\sqrt{\frac{d}{n}} + \int_0^\infty \Pr\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(\theta, Z_i)\right| > C\sqrt{\frac{d}{n}} + u\right] du \\
&\le^{(a)} C\sqrt{\frac{d}{n}} + \frac{C}{\sqrt{n}}\int_0^\infty \Pr\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(\theta, Z_i)\right| > C\sqrt{\frac{d}{n}} + \frac{Ct}{\sqrt{n}}\right] dt \\
&\le^{(b)} C\sqrt{\frac{d}{n}} + \frac{C}{\sqrt{n}}\int_0^\infty 4\exp(-t^2)dt \\
&\le C\sqrt{\frac{d}{n}} + \frac{C'}{\sqrt{n}} \\
&\le 2(C \vee C')\sqrt{\frac{d}{n}},
\end{aligned}$$

where $C' > 0$ is some universal constant, (a) follows by a change of variables $u = Ct/\sqrt{n}$, and (b) follows from Equation (22). By Definition 2.3, we can write

$$\begin{aligned}
\mathrm{Tr}_\kappa(\mathcal{P}; n, \alpha) &= \inf_{\alpha\text{-learner}\mathcal{A}_n}\sup_{\mathcal{T}\in\mathfrak{T}_\kappa}\mathbb{E}_{S_n=(Z_1,\ldots,Z_n)\sim\mathcal{D}^{\otimes n}, \hat{\theta}\sim\mathcal{A}_n(S_n)}\left[\frac{1}{n}\sum_{i\in[n]}\phi(\hat{\theta}, Z_i)\right] \\
&\le \sup_{\mathcal{T}\in\mathfrak{T}_\kappa}\mathbb{E}_{S_n=(Z_1,\ldots,Z_n)\sim\mathcal{D}^{\otimes n}}\left[\sup_{\theta\in\Theta}\frac{1}{n}\sum_{i\in[n]}\phi(\theta, Z_i)\right] \\
&\le 2(C \vee C')\sqrt{\frac{d}{n}}.
\end{aligned}$$

By rearranging the terms we obtain the desired result. □

We now show that Theorem 5.1 implies lower bounds on the sample complexity of learning ℓ_p-Lipshitz-bounded problems for every $p \in [1, \infty)$. In particular, we show that the problems considered in Theorem 5.1 require many samples to learn (equivalently, we show a lower bound on optimal excess risk α).

Theorem G.2. *Let $\alpha > 0$, $p \in [1, \infty)$ and $n \in \mathbb{N}$ be arbitrary. Let $\mathcal{P}_{k,p}$ be as in Equation (7), and set k as in Theorem 5.1. Suppose there exist an α-learner for $\mathcal{P}_{k,p}$. Then,*

(i) for $p \in [1, 2]$, we have

$$\alpha \ge \frac{c}{\sqrt{n}},$$

(ii) for $p \in (2, \infty)$, we have

$$\alpha \geq c \left[\frac{1}{n^{1/p}} \wedge \frac{d^{1/2-1/p}}{\sqrt{n}} \right],$$

for some universal constant $c > 0$.

Proof. First, consider an arbitrary $k \in [d]$. We apply Proposition G.1 to the result of Lemma E.4. We then have the following double inequality

$$c_2 \frac{d^{1-1/p}}{k^{1/2-1/p} n\alpha} \leq \mathrm{Tr}_\kappa(\mathcal{P}_{k,p}; n, \alpha) \leq C\sqrt{\frac{d}{n}}.$$

Solving for α in the above, we have

$$\alpha \geq \frac{c_2}{C} \cdot \frac{(d/k)^{1/2-1/p}}{\sqrt{n}}.$$

First, consider the case $p \in [1, 2]$. Then $k = d$, and we have

$$\alpha \geq \frac{c_2}{C} \cdot \frac{1}{\sqrt{n}},$$

as desired. Now, consider the case $p \in (2, \infty)$. Then $k = (6\alpha)^p d \vee 1$, and we have

$$\alpha \geq \frac{c_2}{C} \cdot \frac{(d/k)^{1/2-1/p}}{\sqrt{n}} = \frac{c_2}{C} \left[\frac{(1/6\alpha)^{p/2-1}}{\sqrt{n}} \wedge \frac{d^{1/2-1/p}}{\sqrt{n}} \right].$$

Solving for α, we have

$$\alpha \geq \left[\left(\frac{c_2}{C 6^{p/2-1}} \right)^{2/p} \cdot \frac{1}{\sqrt{n}} \right] \wedge \left[\frac{c_2}{C} \cdot \frac{d^{1/2-1/p}}{\sqrt{n}} \right].$$

Note that, since $2/p \leq 1$, we have

$$\left(\frac{c_2}{C 6^{p/2-1}} \right)^{2/p} = \frac{1}{6} \left(\frac{6c_2}{C} \right)^{2/p} \geq \frac{1}{6} \left[1 \wedge \frac{6c_2}{C} \right].$$

Thus,

$$\alpha \geq \left[\frac{1}{6} \wedge \frac{c_2}{C} \right] \cdot \left[\frac{1}{\sqrt{n}} \wedge \frac{d^{1/2-1/p}}{\sqrt{n}} \right],$$

as desired. □

G.2 Lower bounds for $p = 1$.

Now, consider the case $p = 1$. For $p = 1$, we consider the problem as in Equation (17). We will need the following refinement of Proposition G.1 in a special case when Θ is a polytope with few vertices. Intuitively, d in the statement Proposition G.1 can be replaced by $\log N$ where N is the number of vertices of Θ.

Proposition G.3. *Fix $n \in \mathbb{N}$, $d \in \mathbb{N}$, and $\alpha \in [0, 1]$. Consider an arbitrary SCO problem $\mathcal{P} = (\Theta, \mathcal{Z}, f)$ in $\mathbb{R}^d$, where Θ is a polytope with N vertices. Let $\mathrm{Tr}_\kappa(\mathcal{P}; n, \alpha)$ be the subgaussian trace value of problem $\mathcal{P}$. Then, we have*

$$\mathrm{Tr}_\kappa(\mathcal{P}; n, \alpha) \cdot \sqrt{n} \leq C \cdot \sqrt{\log N},$$

for some universal constant $C > 0$.

Proof. Let $(\phi, \mathcal{D})$ be an arbitrary subgaussian tracer. Similarly to the proof of Proposition G.1, consider the process $\{X_\theta\}_{\theta \in \Theta}$ defined as

$$X_\theta := \frac{1}{n} \sum_{i=1}^n \phi(\theta, Z_i),$$

where $S_n = (Z_1, \ldots, Z_n) \sim \mathcal{D}^{\otimes n}$. As in the proof of Proposition G.1, $\{X_\theta\}_{\theta \in \Theta}$ is a $C/\sqrt{n}$-subgaussian process w.r.t. $(\Theta, \|\cdot\|_\Theta)$ for some universal constant $C > 0$.

Let V be the set of vertices of Θ; then $|V| = N$, as per the proposition statement. Since ϕ is convex in its first argument, the mapping $\theta \mapsto X_\theta$ is also convex (almost surely). Then,

$$\sup_{\theta \in \Theta} |X_\theta| = \sup_{\theta \in V} |X_\theta|.$$

Therefore, by Proposition A.10, we have, with probability at least $1 - 4\exp(-t^2)$

$$\begin{aligned}
\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \phi(\theta, Z_i) \right| &\leq \frac{Ct}{\sqrt{n}} + \frac{C}{\sqrt{n}} \int_0^1 \sqrt{\log \mathcal{N}\left(V; \|\cdot\|_\Theta, \varepsilon\right)} d\varepsilon \\
&\leq \frac{Ct}{\sqrt{n}} + \frac{C}{\sqrt{n}} \int_0^1 \sqrt{\log N} d\varepsilon \\
&\leq \frac{Ct}{\sqrt{n}} + \frac{C\sqrt{\log N}}{\sqrt{n}}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E} \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \phi(\theta, Z_i) \right| &= \int_0^\infty \Pr\left[\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \phi(\theta, Z_i) \right| > u \right] du \\
&\leq C\sqrt{\frac{\log N}{n}} + \int_0^\infty \Pr\left[\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \phi(\theta, Z_i) \right| > C\sqrt{\frac{\log N}{n}} + u \right] du \\
&\leq^{(a)} C\sqrt{\frac{\log N}{n}} + \frac{C}{\sqrt{n}} \int_0^\infty \Pr\left[\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \phi(\theta, Z_i) \right| > C\sqrt{\frac{\log N}{n}} + \frac{Ct}{\sqrt{n}} \right] dt \\
&\leq^{(b)} C\sqrt{\frac{\log N}{n}} + \frac{C}{\sqrt{n}} \int_0^\infty 4\exp(-t^2) dt \\
&\leq C\sqrt{\frac{\log N}{n}} + \frac{C'}{\sqrt{n}} \\
&\leq 2(C \vee C')\sqrt{\frac{\log N}{n}},
\end{aligned}$$

where $C' > 0$ is some universal constant, (a) follows by a change of variables $u = Ct/\sqrt{n}$, and (b) follows from Equation (22). By Definition 2.3, we can write

$$\begin{aligned}
\mathrm{Tr}_\kappa(\mathcal{P}; n, \alpha) &= \inf_{\alpha\text{-learner} \mathcal{A}_n} \sup_{\mathcal{T} \in \mathfrak{T}_\kappa} \mathbb{E}_{S_n = (Z_1, \ldots, Z_n) \sim \mathcal{D}^{\otimes n}, \hat{\theta} \sim \mathcal{A}_n(S_n)} \left[\frac{1}{n} \sum_{i \in [n]} \phi(\hat{\theta}, Z_i) \right] \\
&\leq \sup_{\mathcal{T} \in \mathfrak{T}_\kappa} \mathbb{E}_{S_n = (Z_1, \ldots, Z_n) \sim \mathcal{D}^{\otimes n}} \left[\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i \in [n]} \phi(\theta, Z_i) \right] \\
&\leq 2(C \vee C')\sqrt{\frac{\log N}{n}}.
\end{aligned}$$

By rearranging the terms we obtain the desired result. □

Then, sample complexity lower bounds for ℓ_1 geometry follow.

Theorem G.4. *Let $\alpha > 0$ and $n \in \mathbb{N}$ be arbitrary. Let $\mathcal{P}_s$ be as in Equation (17) and set $s = d^{0.99}$. Suppose there exists an α-learner for $\mathcal{P}_s$. Then, for large enough d, we have*

$$\alpha \geq c\sqrt{\frac{\log(d)}{n}},$$

for some universal constant $c > 0$.

Proof. Lemma E.7 gives the following lower bound on the subgaussian trace value of $\mathcal{P}_s$,

$$\mathrm{Tr}_\kappa(\mathcal{P}_s; n, \alpha) \geq \frac{c\sqrt{s}\log\left(\frac{d}{16(s\vee 14)}\right)}{n\alpha}.$$

At the same time, noting that Θ is a polytope with vertices given by

$$V = \left\{ z \in \left\{0, \pm\frac{1}{s}\right\}^d : \|z\|_0 = s \right\},$$

which has cardinality

$$|V| = \binom{d}{s} \leq \left(\frac{de}{s}\right)^s,$$

we have by Proposition G.3

$$\mathrm{Tr}_\kappa(\mathcal{P}_s; n, \alpha) \leq C\sqrt{\frac{s\log(de/s)}{n}},$$

for some universal $C > 0$. Combining this with the lower bound on subgaussian trace value, we have

$$\frac{c\sqrt{s}\log\left(\frac{d}{16(s\vee 14)}\right)}{n\alpha} \leq C\sqrt{\frac{s\log(de/s)}{n}},$$

which gives

$$\alpha \geq \frac{c}{C}\frac{\log\left(\frac{d}{16(s\vee 14)}\right)}{\sqrt{\log(de/s)n}}.$$

Recall that $s = d^{0.99}$. For large enough d, we have $s \vee 14 = s$. Also, note that $\log(d/s) \geq \log(d)/100$. Then, for for large enough d, we have

$$\alpha \geq c'\sqrt{\frac{\log(d)}{n}},$$

for some universal $c' > 0$, as desired. □

H Traceability of VC classes (Section 1.1.1)

First, we state the main result.

Theorem H.1. *Fix $n \in \mathbb{N}$ and $\xi < 0.1$. Let $\mathcal{H}$ be an arbitrary VC concept class with VC dimension d_{vc}. Then, there exists an optimal algorithm in terms of number of samples such that it is $(\xi/(n\log(n)), m)$-traceable with $m \leq O\left(d_{\mathsf{vc}}\log^2(n)\right)$. Moreover, when $\mathcal{H}$ is the class of thresholds, we have $m \in O(1)$.*

To prove it we use an information-theoretic notion that controls the difficulty of tracing from [SZ20].

Definition H.2. Fix $n \in \mathbb{N}$. Let $\mathcal{D}$ be a data distribution, and $\mathcal{A}_n$ be a learning algorithm. For every $n \in \mathbb{N}$, let $\boldsymbol{Z} = (Z_{j,i})_{j\in\{0,1\}, i\in[n]}$ be an array of i.i.d. samples drawn from $\mathcal{D}$, and $\boldsymbol{U} = (U_1, \ldots, U_n) \sim \mathrm{Ber}\left(\frac{1}{2}\right)^{\otimes n}$, where $\boldsymbol{U}$ and $\boldsymbol{Z}$ are independent. Define training set $S_n = (Z_{U_i,i})_{i\in[n]}$. Then, define

$$\mathrm{CMI}_{\mathcal{D}}\left(\mathcal{A}_n\right) := I\left(\mathcal{A}_n(S_n); \boldsymbol{U} \middle| \boldsymbol{Z}\right).$$

In the next theorem, we show that the existence of a tracer for a learning algorithm provides a lower bound on the CMI of the algorithm. A similar observation is made in [ADHLR24].

Lemma H.3. *Fix $n \in \mathbb{N}$ such that $n \geq 2$ and $\xi < 1/2$. Let $\mathcal{A}_n$ be an arbitrary learning algorithm that is $(\xi/(n\log(n)), m)$-traceable. Then, it holds $\sup_{\mathcal{D}} \mathrm{CMI}_{\mathcal{D}}\left(\mathcal{A}_n\right) \geq m - 3\xi$.*

The following two results from [SZ20] and [HDMR21] provide upper bounds on the CMI of sample compression schemes. We skip the formal definitions of sample compression schemes and refer the reader to [LW86; MY16; BHMZ20].

Lemma H.4 (Thm 4.2. [SZ20]). *Let $\mathcal{H}$ be an arbitrary concept class with VC dimension d_{vc}. Then, there exists an algorithm such that for every data distribution $\mathcal{D}$, $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq O(d_{vc}\log^2(n))$.*

Lemma H.5 (Thm 3.4. [HDMR21]). *Let $\mathcal{H}$ be the concept class of threshold in $\mathbb{R}$. Then, there exists an algorithm such that for every data distribution $\mathcal{D}$ and $n \geq 2$, $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \leq 2\log 2$.*

Proof of Theorem H.1. The proof is simply by combining Lemma H.4 with Lemma H.3. For the case of the class of thresholds, we use Lemma H.5. □

Proof of Lemma H.3. Assume there exists a tracer with recall of m and soundness parameter of $\xi/(n\log(n)$. Let $\mathcal{D}$ denote the distribution used by the tracer (see Definition 2.3). Define the following random set

$$\mathcal{V}^1 = \{i \in [n] : \exists j \in \{0,1\}\ \phi(\hat{\theta}, Z_{j,i}) \geq \lambda \text{ and } \phi(\hat{\theta}, Z_{1-j,i}) < \lambda\}.$$

Also, for every $i \in [n]$, define the following random variable

$$\mathcal{G}_i = \mathbb{1}\{\phi(\hat{\theta}, Z_{\bar{U}_i,i}) < \lambda\}.$$

By the definition of mutual information and $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{Z}$, we have

$$\begin{aligned}\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) &= \mathrm{H}(\boldsymbol{U}) - \mathrm{H}(\boldsymbol{U}|\boldsymbol{Z}, \hat{\theta}) &(23)\\ &= n - \mathrm{H}(\boldsymbol{U}|\boldsymbol{Z}, \hat{\theta}). &(24)\end{aligned}$$

Therefore, to lower bound $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$, we need to upper bound $\mathrm{H}(\boldsymbol{U}|\boldsymbol{Z}, \hat{\theta})$. By the sub-additivity of the entropy, we have

$$\mathrm{H}(\boldsymbol{U}|\boldsymbol{Z}, \hat{\theta}) \leq \sum_{i=1}^{n} \mathrm{H}(U_i|\boldsymbol{Z}, \hat{\theta}).$$

Then, by monotonicity of the entropy and the chain rule, we have

$$\begin{aligned}\mathrm{H}(U_i|\boldsymbol{Z}, \hat{\theta}) &\leq \mathrm{H}(U_i, \mathcal{G}_i|\boldsymbol{Z}, \hat{\theta})\\ &\leq \mathrm{H}(\mathcal{G}_i) + \mathrm{H}(U_i|\mathcal{G}_i, \boldsymbol{Z}, \hat{\theta}). &(25)\end{aligned}$$

In the next step, by the definition of conditional entropy,

$$\mathrm{H}(U_i|\boldsymbol{Z}, \hat{\theta}, \mathcal{G}_i) \leq \mathrm{H}(U_i|\boldsymbol{Z}, \hat{\theta}, \mathcal{G}_i = 1) + \Pr(\mathcal{G}_i = 0).$$

Define the random variable $Y_i = \mathbb{1}\left(i \in \mathcal{V}^1\right)$. Notice that Y_i is $(\hat{\theta}, \boldsymbol{Z})$-measurable random variable. Then, using the notations for the disintegrated conditional entropy from [HDMR21], we have

$$\begin{aligned}\mathrm{H}(U_i|\boldsymbol{Z}, \hat{\theta}, \mathcal{G}_i = 1) &= \mathbb{E}\left[\mathrm{H}^{\boldsymbol{Z}, \hat{\theta}, \mathcal{G}_i = 1}(U_i)\right]\\ &= \mathbb{E}\left[\mathrm{H}^{\boldsymbol{Z}, \hat{\theta}, \mathcal{G}_i = 1}(U_i)\,\mathbb{1}\{Y_i = 1\}\right] + \mathbb{E}\left[\mathrm{H}^{\boldsymbol{Z}, \hat{\theta}, \mathcal{G}_i = 1}(U_i)\,\mathbb{1}\{Y_i = 0\}\right]. &(26)\end{aligned}$$

The main observation is that under the events $Y_i = 1$ and $\mathcal{G}_i = 1$, U_i is deterministically known from $(\boldsymbol{Z}, \hat{\theta})$. It follows because

$$\{Y_i = 1\} \wedge \{\mathcal{G}_i = 1\} \Leftrightarrow \{\phi(\hat{\theta}, Z_{j,i}) \geq \lambda \text{ and } \phi(\hat{\theta}, Z_{1-j,i}) < \lambda\} \wedge \{\phi(\hat{\theta}, Z_{\bar{U}_i,i}) < \lambda\} \Rightarrow U_i = j.$$

Therefore, since $\mathrm{H}^{\boldsymbol{Z}, \hat{\theta}, \mathcal{G}_i = 1}(U_i) \leq 1$ with probability one, by Equation (26)

$$\mathrm{H}(U_i|\boldsymbol{Z}, \hat{\theta}, \mathcal{G}_i = 1) \leq \Pr(Y_i = 0).$$

Thus, from Equations (25) and (26), we obtain the following upper bound

$$\mathrm{H}(U_i|\boldsymbol{Z}, \hat{\theta}) \leq \Pr(Y_i = 0) + \mathrm{H}_b(\Pr(\mathcal{G}_i = 0)) + \Pr(\mathcal{G}_i = 0),$$

where $\mathrm{H}_b(\cdot) : [0,1] \to [0,1]$ is the binary entropy function defined as $\mathrm{H}_b(x) = -x\log(x) - (1-x)\log(1-x)$. We can lower bound $\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n)$ as follows

$$\begin{aligned}
\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) &\geq n - \sum_{i=1}^{n} \mathrm{H}(U_i | \boldsymbol{Z}, \hat{\theta}) \\
&\geq n - \sum_{i=1}^{n} \left[\Pr(Y_i = 0) + \mathrm{H}_b\left(\Pr(\mathcal{G}_i = 0)\right) + \Pr(\mathcal{G}_i = 0)\right] \\
&= \sum_{i=1}^{n} \Pr(Y_i = 1) - \sum_{i=1}^{n} \left(\mathrm{H}_b\left(\Pr(\mathcal{G}_i = 0)\right) + \Pr(\mathcal{G}_i = 0)\right),
\end{aligned}$$

where the last step follows because $n - \sum_{i=1}^{n} \Pr(Y_i = 0) = \sum_{i=1}^{n} \Pr(Y_i = 1)$ as Y_i is an indicator random variable. In the next step, by the definition of soundness from Definition 2.3 and the definition of CMI in Definition H.2, we have

$$\begin{aligned}
\Pr(\mathcal{G}_i = 0) &= \mathbb{E}\left[\Pr\left(\mathcal{G}_i = 0 \middle| \boldsymbol{U}\right)\right] \\
&= \Pr_{S_n \sim \mathcal{D}^{\otimes n}, Z \sim \mathcal{D}, \hat{\theta} \sim \mathcal{A}_n(S_n)}\left(\phi(\hat{\theta}, Z) \geq \lambda\right) \\
&\leq \xi/(n.\log(n)),
\end{aligned}$$

Therefore,

$$\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) \geq \sum_{i=1}^{n} \Pr(Y_i = 1) - \frac{\xi}{\log(n)} - n\mathrm{H}_b\left(\frac{\xi}{n\log(n)}\right)$$

By the recall condition from Definition 2.3 and the definition of CMI in Definition H.2, we have

$$\begin{aligned}
m &= \mathbb{E}_{S_n = (Z_1, \ldots, Z_n) \sim \mathcal{D}^{\otimes n}, \hat{\theta} \sim \mathcal{A}_n(S_n)}\left[\left|i \in [n] : \phi(\hat{\theta}, Z_i) \geq \lambda\right|\right] \\
&= \sum_{i=1}^{n} \Pr\left(\phi(\hat{\theta}, Z_i) \geq \lambda\right) \\
&\leq \sum_{i=1}^{n} \Pr\left(\phi(\hat{\theta}, Z_{U_i, i}) \geq \lambda \wedge \mathcal{G}_i\right) + \Pr(\mathcal{G}_i^c) \\
&= \sum_{i=1}^{n} \Pr(Y_i = 1) + \sum_{i=1}^{n} \Pr(\mathcal{G}_i^c) \\
&\leq \sum_{i=1}^{n} \Pr(Y_i = 1) + \frac{\xi}{\log(n)}.
\end{aligned}$$

We also use the following well-known inequality, $\mathrm{H}_b(x) \leq -x\log(x) + x$ for $x \in [0,1]$. As a result, we obtain

$$\begin{aligned}
\mathrm{CMI}_{\mathcal{D}}(\mathcal{A}_n) &\geq \sum_{i=1}^{n} \Pr(Y_i = 1) - \frac{\xi}{\log(n)} - n\mathrm{H}_b\left(\frac{\xi}{n\log(n)}\right) \\
&\geq m - \frac{2\xi}{\log(n)} - n\left(\frac{\xi}{n.\log(n)} - \frac{\xi}{(n.\log(n))}\log\left(\frac{\xi}{(n.\log(n))}\right)\right) \\
&\geq m - \xi\left(\frac{1}{e} + \frac{3}{\log(n)}\right),
\end{aligned}$$

where the last step follows because $-x\log(x) \leq 1/e$ for $x \in [0,1]$.

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and the introduction capture the precise theoretical claims made in the main part.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: It has been discussed when comparing with prior work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Every theoretical result is presented in a form of a theorem stating all assumptions made.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer:[NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not include any experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not include any experiments.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include any experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and found that the research in this paper conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: It is a theory paper, and we forsee no such impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This is a theory paper, and it poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer:[NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.