Submitted to the Annals of Statistics

SCORE ATTACK: A LOWER BOUND TECHNIQUE FOR OPTIMAL DIFFERENTIALLY PRIVATE LEARNING

BY T. TONY CAI[1,a], YICHEN WANG[2,b] AND LINJUN ZHANG[3,c]

[1]*Department of Statistics and Data Science, The Wharton School, University of Pennsylvania ,* [a]*tcai@wharton.upenn.edu*

[2]*Independent researcher,* [b]*wangyichen2012@gmail.com*

[3]*Department of Statistics, Rutgers University,* [c]*linjun.zhang@rutgers.edu*

Achieving optimal statistical performance while ensuring the privacy of personal data is a challenging yet crucial objective in modern data analysis. However, characterizing the optimality, particularly the minimax lower bound, under privacy constraints is technically difficult.

To address this issue, we propose a novel approach called the score attack, which provides a lower bound on the differential-privacy-constrained minimax risk of parameter estimation. The score attack method is based on the tracing attack concept in differential privacy and can be applied to any statistical model with a well-defined score statistic. It can optimally lower bound the minimax risk of estimating unknown model parameters, up to a logarithmic factor, while ensuring differential privacy for a range of statistical problems. We demonstrate the effectiveness and optimality of this general method in various examples, such as the generalized linear model in both classical and high-dimensional sparse settings, the Bradley-Terry-Luce model for pairwise comparisons, and nonparametric regression over the Sobolev class.

1. Introduction. With the vast amount of data being generated by individuals, businesses, and governments, statistical and machine learning algorithms are widely employed to facilitate informed decision-making in domains such as healthcare, finance, public policy, transportation, education, and scientific discoveries. The extensive use of algorithms underscores the importance of safeguarding data privacy. As a result, the differential privacy framework [18, 19] for privacy-preserving data processing has garnered substantial attention. Notably, the US Census Bureau utilized differentially private methods for the first time in the 2020 US Census [27] to publish demographic data.

In essence, a differentially private algorithm protects data privacy by ensuring that an observer of the algorithm's output cannot ascertain the presence or absence of any individual record in the input dataset. The design and analysis of differentially private algorithms is a rapidly evolving research field, with many differentially private solutions available in the literature for essential statistical and machine learning problems. These include mean estimation [9, 31, 32, 14], top-k selection [7, 54], linear regression [59, 14], multiple testing [23], causal inference [36, 37], and deep learning [1, 45]. Achieving optimal statistical performance while preserving privacy is a challenging yet crucial objective in modern data analysis.

While desirable for many reasons, differential privacy imposes a constraint on algorithms and may compromise their accuracy in statistical inference. In the decision-theoretical framework, the accuracy of parameter estimation is often measured by the minimax risk, which is

The research of T. Tony Cai was supported in part by NSF Grant DMS-2015259 and NIH grant R01-GM129781. The research of Yichen Wang was conducted prior to and outside of his current employment at Amazon.com Services LLC. The research of Linjun Zhang was supported in part by NSF Grant DMS-2015378.

MSC2020 subject classifications: Primary 62F30; secondary 62J12, 62G05.

Keywords and phrases: differential privacy, parameter estimation, minimax optimality.

defined as the best possible worst-case performance among *all* procedures. When the class of procedures considered is limited to differentially private ones, we arrive at the *privacy-constrained* minimax risk, which represents the optimal statistical performance among all differentially private methods in the worst-case scenario.

The difference between the unconstrained minimax risk and the privacy-constrained minimax risk quantifies the cost of differential privacy, or the amount of accuracy that is inevitably lost due to differential privacy, regardless of how well the differentially private algorithm is designed. Characterizing optimality, particularly the minimax lower bound, under privacy constraints is technically difficult. There have been active efforts to quantify the cost of differential privacy, in such problems as mean estimation [9, 31, 32, 14], top-k selection [7, 54], linear regression [14], and so on.

A key step in establishing minimax theory, whether constrained or unconstrained, is the derivation of minimax lower bounds. In the classical unconstrained setting, several effective lower bound techniques have been developed in the literature, including Le Cam's two-point argument, Assouad's Lemma, and Fano's Lemma. (See [35, 57] for more detailed discussions on minimax lower bound arguments.) However, these methods are not directly applicable to the privacy-constrained setting, and new technical tools are needed. Another line of work [9, 33, 2, 3] lower bounds the privacy-constrained minimax risk using differentially private analogs of traditional Le Cam's, Fano's, and Assouad's inequalities. Similar to the original versions, these differentially private analogs can, in principle, be applied to general estimation and testing problems and lead to tight lower bound results in discrete distribution estimation [2, 3], but their effectiveness in other statistical problems has yet to be fully explored.

In this paper, we introduce a general technique named the "score attack" to establish lower bounds on the privacy-constrained minimax risk. The method is applicable to any statistical model with a well-defined score statistic, which is simply the gradient of the log-likelihood function with respect to the model parameters. After presenting the technique in general terms in Section 2, we use it to derive precise privacy-constrained minimax lower bounds across four statistical models: the low-dimensional generalized linear models (GLMs), the Bradley-Terry-Luce model for pairwise comparisons, the high-dimensional sparse GLMs, and nonparametric regression.

1.1. *Main Results and Our Contribution.*

The score attack technique. The score attack technique generalizes the "tracing adversary" argument, which was first developed by [13, 22]. It has been further applied to various statistical problems, including sharp lower bounds for classical Gaussian mean estimation and linear regression [31, 14], as well as lower bounds for high-dimensional sparse mean estimation and linear regression [54, 14]. In these previous works, the design of tracing attacks is largely ad hoc and specific to statistical models such as Gaussian or Beta-Binomial; a general principle for designing attacks has not been observed. Although some promising proposals have been made in this direction [49, 42], it is unclear whether the suggested attacks in these works actually imply any lower bound results.

The proposed score attack technique is a general method for lower bounding the privacy-constrained minimax risk in statistical models that have a well-defined score statistic, which is the gradient of the likelihood function with respect to the model parameters. As explained in Section 2, the score attack method reduces lower bounding the privacy-constrained minimax risk to computing the score statistic and choosing an appropriate prior distribution over the parameter space. This approach is reminiscent of the classical method of lower bounding the minimax risk by the Bayes risk.

Optimal differentially private algorithms. In this paper, we establish the minimax optimal rate of convergence, up to a logarithmic factor, under the differential privacy constraint for

four statistical estimation problems, namely parameter estimation in low-dimensional generalized linear models (GLMs), the Bradley-Terry-Luce (BTL) model, the high-dimensional sparse GLMs, and nonparametric regression over the Sobolev class. We design optimal algorithms that ensure differential privacy by leveraging established techniques in differential privacy, such as the Laplace and Gaussian mechanisms [19], the K-norm mechanism [26], and differentially private optimization methods [11, 10, 17, 34]. In each of the four problems, we use the score attack technique to establish minimax lower bounds, demonstrating the sharpness of these bounds and the versatility of the score attack method. The main results are summarized as follows.

- Low-dimensional GLMs: Theorem 3.1 presents a minimax lower bound for estimating the parameters and Theorem 3.2 shows that this lower bound is achieved, up to a logarithmic factor, by a noisy gradient descent algorithm.
- BTL model for pairwise comparisons: Similarly, Theorem 4.1 establishes a minimax lower bound for parameter estimation and Theorem 4.2 shows that this lower bound can be attained up to a logarithmic factor by an objective perturbation algorithm.
- High-dimensional sparse GLMs: Theorem 5.1 proves a minimax lower bound which scales only logarithmically with the total dimension and linearly with the sparsity, and Theorem 5.2 shows that this minimax lower bound can be achieved up to a logarithmic factor by an iterative hard-thresholding algorithm.
- Nonparametric regression over the Sobolev class: unlike the previous problems, where the number of parameters is finite, this problem deals with estimating an entire function with a differential privacy guarantee. Here, we establish a matching lower bound in Theorem 6.1 and an upper bound in Theorem 6.2 for the minimax mean integrated squared risk.

1.2. *Related Work.* There is a relatively large body of literature on differentially private GLMs, particularly the logistic regression model [16, 17, 62, 50, 51, 5, 6]; in particular, [62] studies sparse logistic regression with differential privacy via the perspective of graphical models. Our paper, while inspired by previous work, is distinct from previous research in its emphasis on parameter estimation accuracy rather than the excess risk of the solution. For problems of ranking based on pairwise comparisons, several studies have investigated differentially private rank aggregation, including [48, 28, 50, 39, 61]. However, to the best of our knowledge, no prior research had explored optimal differentially private parameter estimation in the BTL model. For nonparametric function estimation under differential privacy, [60] and [38] studied the convergence rate of noisy histogram estimators, but did not investigate optimality or explore the lower bound. On the other hand, [25] introduced general mechanisms for releasing differentially private functional data, while [9] proposed a minimax optimal differentially private histogram estimator for Lipschitz functions.

1.3. *Organization of the Paper.* The rest of the paper is organized as follows. We finish this section with notational conventions in this paper. Section 2 describes differential privacy and the privacy-constrained minimax risk in technical terms, and formulates the score attack method for general parametric distribution families. The general formulation is then specialized to four examples: the low-dimensional GLMs in Section 3, the Bradley-Terry-Luce model in Section 4, the high-dimensional sparse GLMs in Section 5, and finally nonparametric regression in Section 6. We discuss possible extensions in Section 7, present the proof of one main result in Section 8, and defer the rest of the proofs to the supplement [15] due to the space limit.

1.4. *Notation.* For real-valued sequences $\{a_n\}, \{b_n\}$, we write $a_n \lesssim b_n$ if $a_n \le cb_n$ for some universal constant $c \in (0, \infty)$, and $a_n \gtrsim b_n$ if $a_n \ge c'b_n$ for some universal constant $c' \in (0, \infty)$. We say $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. $c, C, c_0, c_1, c_2, \cdots$, and so on refer to universal constants in the paper, with their specific values possibly varying from place to place.

For a vector $\boldsymbol{v} \in \mathbb{R}^d$ and a subset $S \subseteq [d]$, $\boldsymbol{v}_S$ denotes the restriction of vector $\boldsymbol{v}$ to the index set S. Define $\mathrm{supp}(\boldsymbol{v}) := \{j \in [d] : v_j \neq 0\}$. $\|\boldsymbol{v}\|_p$ denotes the vector ℓ_p norm for $1 \le p \le \infty$, with an additional convention that $\|\boldsymbol{v}\|_0$ denotes the number of non-zero coordinates of $\boldsymbol{v}$. For a square matrix $\boldsymbol{A}$, $\lambda_j(\boldsymbol{A})$ refers to its jth smallest eigenvalue, and $\lambda_{\max}(\boldsymbol{A}), \lambda_{\min}(\boldsymbol{A})$ refer to its largest and smallest eigenvalues respectively. For a function $f : \mathbb{R} \to \mathbb{R}$, $\|f\|_\infty$ denotes the the essential supremum of $|f|$. For $t \in \mathbb{R}$ and $R > 0$, let $\Pi_R(t)$ denote the projection of t onto the closed interval $[-R, R]$.

2. The Score Attack.

This section presents the general framework for the score attack to ensure that the high-level concept is not obscured when we examine specific models later in the paper. We commence by defining the privacy-constrained minimax risk in Section 2.1 and then introduce the score attack in Section 2.2.

2.1. *Differential Privacy and the Minimax Risk.* The notion of differential privacy formalizes an intuitive idea: an algorithm M compromises the privacy of input data set $\boldsymbol{X}$ if an observer of the output $M(\boldsymbol{X})$ only can infer better than randomly guessing whether an individual datum $\boldsymbol{x}$ belongs to the input $\boldsymbol{X}$ or not. A differentially algorithm M therefore guarantees that, for every pair of data sets $\boldsymbol{X}$ and $\boldsymbol{X}'$ that differ by a single datum ("adjacent data sets"), the probability distributions of $M(\boldsymbol{X})$ and of $M(\boldsymbol{X}')$ are close to each other.

DEFINITION 1 (Differential Privacy [19]). A randomized algorithm $M : \mathcal{X}^n \to \mathcal{R}$ is (ε, δ)-differentially private if for every pair of adjacent data sets $\boldsymbol{X}, \boldsymbol{X}' \in \mathcal{X}^n$ that differ by one individual datum and every measurable $S \subseteq \mathcal{R}$,

$$\mathbb{P}\left(M(\boldsymbol{X}) \in S\right) \le e^{\varepsilon} \cdot \mathbb{P}\left(M(\boldsymbol{X}') \in S\right) + \delta,$$

where the probability measure $\mathbb{P}$ is induced by the randomness of M only.

If an algorithm is (ε, δ)-differentially private for small values of $\varepsilon, \delta \ge 0$, the distributions of $M(\boldsymbol{X})$ and $M(\boldsymbol{X}')$ are almost indistinguishable. The popularity of differential privacy in applications partially lies in the ease of constructing differentially private algorithms. For example, adding random noise often suffices to achieve differential privacy for many non-private algorithms.

EXAMPLE 2.1 (The Laplace and Gaussian Mechanisms [19, 20]). Let $M : \mathcal{X}^n \to \mathbb{R}^d$ be an algorithm that is not necessarily differentially private.

- Suppose $\sup_{\boldsymbol{X}, \boldsymbol{X}' \text{adjacent}} \|M(\boldsymbol{X}) - M(\boldsymbol{X}')\|_1 < B < \infty$. For $\boldsymbol{w} \in \mathbb{R}^d$ with its coordinates $w_1, w_2, \cdots, w_d \overset{\text{i.i.d.}}{\sim} \text{Laplace}(B/\varepsilon)$, $M(\boldsymbol{X}) + \boldsymbol{w}$ is $(\varepsilon, 0)$-differentially private.
- If instead we have $\sup_{\boldsymbol{X}, \boldsymbol{X}' \text{adjacent}} \|M(\boldsymbol{X}) - M(\boldsymbol{X}')\|_2 < B < \infty$, for $\boldsymbol{w} \sim N_d(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ with $\sigma^2 = 2B^2 \log(2/\delta)/\varepsilon$, $M(\boldsymbol{X}) + \boldsymbol{w}$ is (ε, δ)-differentially private.

That is, if a non-private algorithm's output is not too sensitive to changing any single datum in the input data set, perturbing the algorithm with Laplace or Gaussian noises produces a differentially private algorithm.

Differential privacy is a desirable property, but it is also a constraint that may come at the expense of statistical accuracy. It is important to understand the effect, or cost, of the differential privacy constraint to statistical accuracy that is naturally measured by the privacy-constrained minimax risk. The formal definition of minimax risk consists of the following elements.

- $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is a family of statistical models supported over $\mathcal{X}$.
- $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ is an i.i.d. sample drawn from $f_{\boldsymbol{\theta}^*}$ for some unknown $\boldsymbol{\theta}^* \in \Theta$, and $M : \mathcal{X}^n \to \Theta$ is an estimator of $\boldsymbol{\theta}^*$.
- $\ell : \Theta \times \Theta \to \mathbb{R}_+$ is a metric on Θ and $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing function.

Then, the (statistical) risk of M is given by $\mathbb{E}\rho(\ell(M(\boldsymbol{X}), \boldsymbol{\theta}^*))$, where the expectation is taken over the data distribution $f_{\boldsymbol{\theta}^*}$ and the randomness of estimator M. Because the risk $\mathbb{E}\rho(\ell(M(\boldsymbol{X}), \boldsymbol{\theta}^*))$ depends on the unknown $\boldsymbol{\theta}^*$ and can be minimized by choosing $M(\boldsymbol{X}) \equiv \boldsymbol{\theta}^*$, a more sensible measure of performance is the maximum risk over the entire class of distributions $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, $\sup_{\boldsymbol{\theta} \in \Theta} \mathbb{E}\rho(\ell(M(\boldsymbol{X}), \boldsymbol{\theta}))$.

The minimax risk of estimating $\boldsymbol{\theta} \in \Theta$ is then given by

$$\inf_{M} \sup_{\boldsymbol{\theta} \in \Theta} \mathbb{E}\rho(\ell(M(\boldsymbol{X}), \boldsymbol{\theta})). \tag{2.1}$$

By definition, this quantity characterizes the best possible worst-case performance that an estimator can hope to achieve over the class of models $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$.

In this paper, we study a *privacy-constrained* minimax risk: let $\mathcal{M}_{\varepsilon,\delta}$ be the collection of all (ε, δ)-differentially private algorithms mapping from $\mathcal{X}^n$ to Θ, we consider

$$\inf_{M \in \mathcal{M}_{\varepsilon,\delta}} \sup_{\boldsymbol{\theta} \in \Theta} \mathbb{E}\rho(\ell(M(\boldsymbol{X}), \boldsymbol{\theta})). \tag{2.2}$$

As $\mathcal{M}_{\varepsilon,\delta}$ is a proper subset of all possible estimators, the privacy-constrained minimax risk as defined above will be at least as large as the unconstrained minimax risk, with the difference between these two minimax risks, (2.1) and (2.2) being the "cost of privacy".

Either the unconstrained minimax risk (2.1) or the constrained (2.2) is often characterized from two opposing directions. While analyzing the risk of *any* concrete algorithm for every $\boldsymbol{\theta} \in \Theta$ leads to an upper bound of the minimax risk, lower bounding the minimax risk requires reasoning abstractly about *all* estimators and understanding their fundamental limits at estimating the parameter $\boldsymbol{\theta}$. The score attack provides a general and effective method for lower bounding the privacy-constrained minimax risk.

2.2. *The Score Attack.* The score attack is a type of tracing attack [13, 22, 21]. A tracing attack is an algorithm which takes a single "candidate" datum as input and attempts to infer whether this candidate belongs to a given data set or not, by comparing the candidate with some summary statistics computed from the data set. Statisticians may envision a tracing attack as a hypothesis test which rejects the null hypothesis that the candidate is out of the data set when some test statistic takes a large value. This hypothesis testing formulation motivates some desirable properties for a tracing attack.

- Soundness (type I error control): if the candidate does not belong to the data set, the tracing attack is likely to takes small values.
- Completeness (type II error control): if the candidate does belong, the tracing attack is likely to take large values.

For example, [22, 31, 14] show that, if the random sample $\boldsymbol{X}$ and the candidate $\boldsymbol{z}$ are drawn from a Gaussian distribution with mean $\boldsymbol{\mu}$, tracing attacks of the form $\langle M(\boldsymbol{X}) - \boldsymbol{\mu}, \boldsymbol{z} - \boldsymbol{\mu} \rangle$ is sound and complete provided that $M(\boldsymbol{X})$ is an accurate estimator of $\boldsymbol{\mu}$.

It is this accuracy requirement that connects tracing attacks with risk lower bounds for differentially private algorithms: if an estimator $M(\boldsymbol{X})$ is differentially private, it cannot possibly be too close to the estimand, or the existence of tracing attacks leads to a contradiction with the guarantees of differential privacy. Designing sound and complete tracing attacks, therefore, is crucial to the sharpness of privacy-constrained minimax lower bounds. Besides the Gaussian mean tracing attack mentioned above, there are some successful tracing attacks proposed for specific problems, such as top-k selection [54] or linear regression [14], but a general recipe for the design and analysis of tracing attacks has not been available.

The score attack is a form of tracing attack applicable to general parametric families of distributions. Given a parametric family of distributions $\{f_{\boldsymbol{\theta}}(\boldsymbol{x}) : \boldsymbol{\theta} \in \Theta\}$ with $\Theta \subseteq \mathbb{R}^d$, the score statistics, or simply the score, is given by $S_{\boldsymbol{\theta}}(\boldsymbol{x}) := \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\boldsymbol{x})$. If $\boldsymbol{x} \sim f_{\boldsymbol{\theta}}$, we have $\mathbb{E}S_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathbf{0}$ and $\mathrm{Var}S_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathcal{I}(\boldsymbol{\theta})$, where $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher information matrix of $f_{\boldsymbol{\theta}}$. Based on the score statistic, the score attack is defined as

$$\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{z}, M(\boldsymbol{X})) := \langle M(\boldsymbol{X}) - \boldsymbol{\theta}, S_{\boldsymbol{\theta}}(\boldsymbol{z}) \rangle. \tag{2.3}$$

The score attack conjectures that $\boldsymbol{z}$ belongs to $\boldsymbol{X}$ for large values of $\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{z}, M(\boldsymbol{X}))$. In particular, if $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ is the density of $N(\boldsymbol{\theta}, \boldsymbol{I})$, the score attack coincides with the tracing attacks for Gaussian means studied in [22, 31, 14].

As argued earlier, a tracing attack should ideally be "sound" (low type I error probability) and "complete" (low Type II error probability). This is indeed the case for our score attack (2.3).

THEOREM 2.1. *Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ be an i.i.d. sample drawn from $f_{\boldsymbol{\theta}}$. For each $i \in [n]$, let $\boldsymbol{X}_i'$ denote an adjacent data set of $\boldsymbol{X}$ obtained by replacing $\boldsymbol{x}_i$ with an independent copy $\boldsymbol{x}_i' \sim f_{\boldsymbol{\theta}}$.*

1. *Soundness: for each $i \in [n]$,*

$$\mathbb{E}\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X}_i')) = 0;\ \mathbb{E}|\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X}_i'))| \le \sqrt{\mathbb{E}\|M(\boldsymbol{X}) - \boldsymbol{\theta}\|_2^2}\sqrt{\lambda_{\max}(\mathcal{I}(\boldsymbol{\theta}))}. \tag{2.4}$$

2. *Completeness: if for every $j \in [d]$, $\log f_{\boldsymbol{\theta}}(\boldsymbol{X})$ is continuously differentiable with respect to θ_j and $|\frac{\partial}{\partial \theta_j} \log f_{\boldsymbol{\theta}}(\boldsymbol{X})| < g_j(X)$ such that $\mathbb{E}|g_j(\boldsymbol{X})M(\boldsymbol{X})_j| < \infty$, we have*

$$\sum_{i \in [n]} \mathbb{E}\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X})) = \sum_{j \in [d]} \frac{\partial}{\partial \theta_j} \mathbb{E}M(\boldsymbol{X})_j. \tag{2.5}$$

Theorem 2.1 is proved in Section 8.1. The special form of "completeness" for Gaussian and Beta-Binomial families has been discovered as "fingerprinting lemma" in the literature [56, 13, 54, 31]. It may not be clear yet how the soundness and completeness properties would imply lower bounds for $\mathbb{E}\|M(\boldsymbol{X}) - \boldsymbol{\theta}\|_2^2$. For the specific attacks designed for Gaussian mean estimation [31] and top-k selection [54], it has been observed that, if M is an (ε, δ)-differentially private algorithm, one can prove inequalities of the form $\mathbb{E}\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X})) \le \mathbb{E}\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X}_i')) + O(\varepsilon)\mathbb{E}|\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X}_i'))|$. Suppose such relations hold for the score attack as well, the soundness property (2.4) would then imply

$$\sum_{i \in [n]} \mathbb{E}\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X})) \le \sqrt{\mathbb{E}\|M(\boldsymbol{X}) - \boldsymbol{\theta}\|_2^2} \cdot n\sqrt{\lambda_{\max}(\mathcal{I}(\boldsymbol{\theta}))}O(\varepsilon).$$

We give a precise statement of such an inequality in Section 2.2.1.

On the other hand, if we can also bound $\sum_{i \in [n]} \mathbb{E}\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X}))$ from below by some positive quantity, a lower bound for $\mathbb{E}\|M(\boldsymbol{X}) - \boldsymbol{\theta}\|_2^2$ is immediately implied. Completeness

may help us in this regard: when $\mathbb{E}M(\boldsymbol{X})_j$ is close to θ_j, it is reasonable to expect that $\frac{\partial}{\partial\theta_j}\mathbb{E}M(\boldsymbol{X})_j$ is bounded away from zero. Indeed several versions of this argument, often termed "strong distribution", exist in the literature [22, 55] and have led to lower bounds for Gaussian mean estimation and top-k selection. In Section 2.2.2, we suggest a systematic approach to lower bounding $\frac{\partial}{\partial\theta_j}\mathbb{E}M(\boldsymbol{X})_j$ via Stein's Lemma [52, 53]. The results in Sections 2.2.1 and 2.2.2 combined with Theorem 2.1 would enable us to later prove concrete minimax lower bounds for a variety of statistical problems.

2.2.1. *Score Attack and Differential Privacy.* In Theorem 2.1, we have found that, when the data set $\boldsymbol{X}_i'$ does not include $\boldsymbol{x}_i$, the score attack is unlikely to take large values:

$$\mathbb{E}\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X}_i')) = 0;\ \mathbb{E}|\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X}_i'))| \le \sqrt{\mathbb{E}\|M(\boldsymbol{X}) - \boldsymbol{\theta}\|_2^2}\sqrt{\lambda_{\max}(\mathcal{I}(\boldsymbol{\theta}))}.$$

If M is differentially private, the distribution of $M(\boldsymbol{X}_i')$ is close to that of $M(\boldsymbol{X})$; as a result, the inequalities above can be related to the case where the data set $\boldsymbol{X}$ does include the candidate $\boldsymbol{x}_i$.

PROPOSITION 2.1. *If M is an (ε, δ)-differentially private algorithm with $0 < \varepsilon < 1$ and $\delta \ge 0$, then for every $T > 0$,*

(2.6)
$$\mathbb{E}\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X})) \le 2\varepsilon\sqrt{\mathbb{E}\|M(\boldsymbol{X}) - \boldsymbol{\theta}\|_2^2}\sqrt{\lambda_{\max}(\mathcal{I}(\boldsymbol{\theta}))} + 2\delta T + \int_T^\infty \mathbb{P}\left(|\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X}))| > t\right).$$

Proposition 2.1 is proved in Section 8.1.1. The quantity on the right side of (2.6) is determined by the statistical model $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ and the choice of T.

2.2.2. *Score Attack and Stein's Lemma.* Let us denote $\mathbb{E}_{X|\boldsymbol{\theta}}M(\boldsymbol{X})$ by $g(\boldsymbol{\theta})$, then g is a map from Θ to Θ, and we are interested in bounding $\frac{\partial}{\partial\theta_j}g_j(\boldsymbol{\theta})$ from below. Stein's Lemma [52, 53], is helpful.

LEMMA 2.1 (Stein's Lemma). *Let Z be distributed according to some density $p(z)$ which is supported on $[a, b]$ for some $-\infty \le a < b \le \infty$ and continuously differentiable over (a, b). Suppose a function $h : [a, b] \to \mathbb{R}$ is differentiable and satisfies $\mathbb{E}|h'(Z)| < \infty$, $\mathbb{E}|h'(Z)p'(Z)/p(Z)| < \infty$, then*

$$\mathbb{E}h'(Z) = \mathbb{E}\left[\frac{-h(Z)p'(Z)}{p(Z)}\right] + h(b-)p(b-) - h(a+)p(a+), \tag{2.7}$$

where $h(b-), p(b-)$ are the left limits of h and p at b and $h(a+), p(a+)$ are the right limits of h and p at a. In particular, if $p(z) = (2\pi)^{-1/2}e^{-z^2/2}$, we have $\mathbb{E}h'(Z) = \mathbb{E}Zh(Z)$.

Stein's Lemma implies that, by imposing appropriate prior distributions on $\boldsymbol{\theta}$, one can obtain a lower bound for $\frac{\partial}{\partial\theta_j}g_j(\boldsymbol{\theta})$ on average over the prior distribution of $\boldsymbol{\theta}$, as follows.

PROPOSITION 2.2. *Let $\boldsymbol{\theta}$ be distributed according to a density $\boldsymbol{\pi}$ with marginal densities $\{\pi_j\}_{j\in[d]}$. If for every $j \in [d]$, π_j, g_j satisfy the regularity conditions in Lemma 2.1 and additionally each π_j converges to 0 at the endpoints of its support, we have*

(2.8)
$$\mathbb{E}_{\boldsymbol{\pi}}\left(\sum_{j\in[d]}\frac{\partial}{\partial\theta_j}g_j(\boldsymbol{\theta})\right) \ge \mathbb{E}_{\boldsymbol{\pi}}\left(\sum_{j\in[d]}\frac{-\theta_j\pi_j'(\theta_j)}{\pi_j(\theta_j)}\right) - \sqrt{\mathbb{E}_{\boldsymbol{\pi}}\|g(\boldsymbol{\theta}) - \boldsymbol{\theta}\|_2^2\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{j\in[d]}\left(\frac{\pi_j'(\theta_j)}{\pi_j(\theta_j)}\right)^2\right]}.$$

Proposition 2.2 is proved in Section 8.1.2. Despite the cumbersome expression, the right side is in fact convenient: often we may assume that $\mathbb{E}_{\boldsymbol{\pi}}\|g(\boldsymbol{\theta})-\boldsymbol{\theta}\|_2^2 \le \mathbb{E}_{\boldsymbol{\pi}}\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\theta}}\|M(\boldsymbol{X})-\boldsymbol{\theta}\|_2^2 < C$ for some constant C when the sample size n is sufficiently large; the right side is then completely determined by the choice of π, for example:

EXAMPLE 2.2. Let $\boldsymbol{\pi}$ be the density of $N(\mathbf{0},\boldsymbol{I})$, then (2.8) reduces to

$$\mathbb{E}_{\boldsymbol{\pi}}\left(\sum_{j\in[d]}\frac{\partial}{\partial\theta_j}g_j(\boldsymbol{\theta})\right) \ge \sum_{j\in[d]}\mathbb{E}_{\pi_j}\theta_j^2 - \sqrt{C}\sqrt{\sum_{j\in[d]}\mathbb{E}_{\pi_j}\theta_j^2} = d-\sqrt{Cd} \gtrsim d.$$

In view of the completeness property (2.5), Proposition 2.2 suggests an *average* lower bound for $\sum_{i\in[n]}\mathbb{E}\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X}))$ over some prior distribution $\boldsymbol{\pi}(\boldsymbol{\theta})$, with the specific form of this average lower bound entirely determined by the choice of $\boldsymbol{\pi}$. This connection between lower bound and choosing a prior over the parameter space may be reminiscent of the familiar fact that the the Bayes risk always lower bounds the minimax risk, which is the exact reasoning we rely on to finish our minimax lower bound argument.

In addition to the standard regularity conditions of Stein's Lemma, Proposition 2.2 assumes that the marginal priors all converge to zero at the boundary of their supports, in order to simplify the right side of (2.8) and highlight the main idea. For those prior distributions not satisfying the vanishing assumption, Proposition 2.2 can be readily extended by adding the last two terms on the right side of Stein's Lemma, equation (2.7), to the right side of equation (2.8). This extension is carried out in Section 5.1 for truncated normal priors and 6.1 for uniform priors.

2.2.3. *From Score Attack to Lower Bounds.* Theorem 2.1 combined with Propositions 2.1 and 2.2 reveals the connection between the score attack and privacy-constrained minimax lower bounds.

Let $\boldsymbol{\pi}$ be a prior distribution supported over the parameter space Θ with marginal densities $\{\pi_j\}_{j\in[d]}$, and assume without the loss of generality that $\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\theta}}\|M(\boldsymbol{X})-\boldsymbol{\theta}\|_2^2 < C$ for every $\boldsymbol{\theta}\in\Theta$. The completeness part of Theorem 2.1 and Lemma 2.2 imply that

$$\sum_{i\in[n]}\mathbb{E}_{\boldsymbol{\pi}}\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\theta}}\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X})) \ge \mathbb{E}_{\boldsymbol{\pi}}\left(\sum_{j\in[d]}\frac{-\theta_j\pi_j'(\theta_j)}{\pi_j(\theta_j)}\right) - \sqrt{C}\sqrt{\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{j\in[d]}\left(\frac{\pi_j'(\theta_j)}{\pi_j(\theta_j)}\right)^2\right]}$$

Since Lemma 2.1 holds for every $\boldsymbol{\theta}$, it follows from the Lemma that

$$\sum_{i\in[n]}\mathbb{E}_{\boldsymbol{\pi}}\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\theta}}\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X}))$$

$$\le 2n\varepsilon\sqrt{\mathbb{E}_{\boldsymbol{\pi}}\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\theta}}\|M(\boldsymbol{X})-\boldsymbol{\theta}\|_2^2}\sqrt{\lambda_{\max}(\mathcal{I}(\boldsymbol{\theta}))} + 2n\delta T + \sum_{i\in[n]}\int_T^\infty \mathbb{P}\left(|\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X}))| > t\right).$$

These two inequalities are true for every (ε,δ)-differentially private M, and they therefore suggest a lower bound for $\inf_{M\in\mathcal{M}_{\varepsilon,\delta}}\mathbb{E}_{\boldsymbol{\pi}}\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\theta}}\|M(\boldsymbol{X})-\boldsymbol{\theta}\|_2^2$, which in turn lower bounds $\inf_{M\in\mathcal{M}_{\varepsilon,\delta}}\sup_{\boldsymbol{\theta}\in\Theta}\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\theta}}\|M(\boldsymbol{X})-\boldsymbol{\theta}\|_2^2$, since the maximum risk is greater than the average risk over any prior distribution.

2.3. *The Utility of Score Attack.* The analysis in Section 2.2 amounts to a reduction from lower bounding the privacy-constrained minimax risk (2.2) to analyzing the expectation of score attack,

$$\sum_{i\in[n]} \mathbb{E}_{\boldsymbol{X}|\boldsymbol{\theta}} \mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X})).$$

Specifically, the analysis of score attack consists of upper bounding the expectation via differential privacy, and lower bounding the expectation "on average" by choosing a prior over the parameter space Θ.

The proposed score attack method is only as valuable as the concrete minimax lower bound results it implies. In the coming sections, we specialize the general method to a variety of problems.

- Parameter estimation in classical models: the generalized linear model (Section 3), and the Bradley-Terry-Luce model (Section 4).
- High-dimensional sparse parameter estimation (Section 5).
- Non-parametric function estimation (Section 6).

In each example, we shall analyze the score attack following the recipe outlined in Section 2.2 and prove the implied minimax risk lower bound; the sharpness of the lower bound is then demonstrated by a concrete differentially private algorithm with matching risk upper bound. These examples will collectively make a strong case for the utility of score attack as a general lower bound technique. While some of them require no more than a straightforward application of the aforementioned method, a few examples involve non-trivial modifications of the general score attack approach which will be highlighted as appropriate.

3. The Generalized Linear Model. As a first example, we consider the privacy-constrained minimax risk of estimating parameters $\boldsymbol{\beta} \in \mathbb{R}^d$ in the generalized linear model

$$f_{\boldsymbol{\beta}}(y|\boldsymbol{x}) = h(y,\sigma)\exp\left(\frac{y\boldsymbol{x}^\top\boldsymbol{\beta} - \psi(\boldsymbol{x}^\top\boldsymbol{\beta})}{c(\sigma)}\right); \boldsymbol{x} \sim f_{\boldsymbol{x}} \tag{3.1}$$

using an i.i.d. sample $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i\in[n]} = \{(y_i, \boldsymbol{x}_i)\}_{i\in[n]}$ drawn from the model (3.1). The functional form of the model, including the partition function ψ and the normalizing factor h, is assumed to be fixed and known; the sole parameter of interest is the vector $\boldsymbol{\beta}$.

In Section 3.1, we prove a minimax risk lower bound applying the score attack method to the generalized linear model. The lower bound is then shown to be sharp up to a logarithmic factor, via a noisy gradient descent algorithm for estimating $\boldsymbol{\beta}$ in Section 3.2.

3.1. *The Privacy-Constrained Minimax Lower Bound.* For the generalized linear model (3.1) and a candidate datum $(\tilde{y}, \tilde{\boldsymbol{x}})$, the score attack (2.3) takes the form

$$\mathcal{A}_{\boldsymbol{\beta}}((\tilde{y},\tilde{\boldsymbol{x}}), M(\boldsymbol{y},\boldsymbol{X})) = \frac{1}{c(\sigma)}\big\langle M(\boldsymbol{y},\boldsymbol{X}) - \boldsymbol{\beta}, [\tilde{y} - \psi'(\tilde{\boldsymbol{x}}^\top\boldsymbol{\beta})]\tilde{\boldsymbol{x}}\big\rangle. \tag{3.2}$$

As outlined in Section 2.2, we establish a privacy-constrained minimax lower bound for estimating $\boldsymbol{\beta}$ by analyzing the sum of expectations $\sum_{i\in[n]} \mathbb{E}\mathcal{A}_{\boldsymbol{\beta}}((y_i,\boldsymbol{x}_i), M(\boldsymbol{y},\boldsymbol{X}))$. When the reference to data $(\boldsymbol{y},\boldsymbol{X})$ and estimator M is clear, we abbreviate $\mathcal{A}_{\boldsymbol{\beta}}((y_i,\boldsymbol{x}_i), M(\boldsymbol{y},\boldsymbol{X}))$ as A_i.

We begin with upper bounding the $\sum_{i\in[n]} \mathbb{E}A_i$, which amounts to specializing the soundness part of Theorem 2.1 and Proposition 2.1 to the GLM score attack (3.2).

PROPOSITION 3.1. *Consider i.i.d. observations* $(y_1, \boldsymbol{x}_1), \cdots, (y_n, \boldsymbol{x}_n)$ *drawn from* (3.1). *Suppose* $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top)$ *is diagonal and* $\lambda_{\max}(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top)) < C < \infty$, $\|\boldsymbol{x}\|_2 \lesssim \sqrt{d}$ *almost surely, and* $\|\psi^{''}\|_\infty < c_2 < \infty$. *If the estimator* M *is* (ε, δ)*-differentially private with* $0 < \varepsilon < 1$ *and satisfies* $\|M(\boldsymbol{y}, \boldsymbol{X}) - \boldsymbol{\beta}\|_2^2 \lesssim d$, *then*

(3.3)
$$\sum_{i\in[n]} \mathbb{E}_{\boldsymbol{y},\boldsymbol{X}|\boldsymbol{\beta}} A_i \leq 2n\varepsilon\sqrt{\mathbb{E}\|M(\boldsymbol{y}, \boldsymbol{X}) - \boldsymbol{\beta}\|_2^2}\sqrt{Cc_2/c(\sigma)} + 4\sqrt{2}\delta d\sqrt{c_2\log(1/\delta)/c(\sigma)}.$$

Based on the general results, Theorem 2.1 and Proposition 2.1, proving Proposition 3.1 essentially entails computing the Fisher information matrix and choosing an appropriate T in equation (2.6). We defer the details to Section A.1 and move on to deriving an average lower bound of $\sum_{i\in[n]} \mathbb{E}A_i$.

PROPOSITION 3.2. *Let the coordinates of* $\boldsymbol{\beta} \in \mathbb{R}^d$ *be drawn i.i.d. from the Beta*$(3, 3)$ *distribution. For every* M *satisfying* $\mathbb{E}_{\boldsymbol{y},\boldsymbol{X}|\boldsymbol{\beta}}\|M(\boldsymbol{y}, \boldsymbol{X}) - \boldsymbol{\beta}\|_2^2 \lesssim 1$ *at every* $\boldsymbol{\beta}$, *we have*

$$\sum_{i\in[n]} \mathbb{E}_{\boldsymbol{\pi}}\mathbb{E}_{\boldsymbol{y},\boldsymbol{X}|\boldsymbol{\beta}} A_i \gtrsim d, \tag{3.4}$$

where $\boldsymbol{\pi}$ *refers to the i.i.d. Beta prior for* $\boldsymbol{\beta}$.

The proof of Proposition 3.2, which involves plugging the appropriate $\boldsymbol{\pi}$ into the general Proposition 2.2, is in Section A.2. We are now ready to establish the minimax risk lower bound for estimating $\boldsymbol{\beta}$, by combining the bounds for $\sum_{i\in[n]} \mathbb{E}A_i$ in both directions. The result is presented in the next theorem.

THEOREM 3.1. *Consider i.i.d. observations* $(y_1, \boldsymbol{x}_1), \cdots, (y_n, \boldsymbol{x}_n)$ *drawn from* (3.1). *Suppose* $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top)$ *is diagonal and* $\lambda_{\max}(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top)) < C < \infty$, $\|\boldsymbol{x}\|_2 \lesssim \sqrt{d}$ *almost surely, and* $\|\psi^{''}\|_\infty < c_2 < \infty$. *If* $d \lesssim n\varepsilon$, $0 < \varepsilon < 1$ *and* $\delta \lesssim n^{-(1+\gamma)}$ *for some* $\gamma > 0$, *then*

$$\inf_{M\in\mathcal{M}_{\varepsilon,\delta}} \sup_{\boldsymbol{\beta}\in\mathbb{R}^d} \mathbb{E}\|M(\boldsymbol{y}, \boldsymbol{X}) - \boldsymbol{\beta}\|_2^2 \gtrsim c(\sigma)\left(\frac{d}{n} + \frac{d^2}{n^2\varepsilon^2}\right). \tag{3.5}$$

The first term in (3.5) is the non-private minimax risk lower bound, and the second term is the "cost of differential privacy". We show in the next section that the lower bound is attainable, up to a logarithmic term, by a noisy gradient descent algorithm.

3.2. *Optimality of the Private GLM Lower Bound.* We consider minimizing the negative GLM log-likelihood

$$\mathcal{L}_n(\boldsymbol{\beta}; \boldsymbol{Z}) = \frac{1}{n}\sum_{i=1}^n \Big(\psi(\boldsymbol{x}_i^\top\boldsymbol{\beta}) - y_i\boldsymbol{x}_i^\top\boldsymbol{\beta}\Big)$$

by noisy gradient descent algorithm, first proposed by [11] in its generic form for arbitrary convex functions. The following algorithm specializes the generic algorithm to GLMs.

Algorithm 1: Differentially Private Generalized Linear Regression

Input : $\mathcal{L}_n(\boldsymbol{\beta}, \boldsymbol{Z})$, data set $\boldsymbol{Z}$, step size η^0, privacy parameters ε, δ, noise scale B, number of iterations T, truncation parameter R, initial value $\boldsymbol{\beta}^0 \in \mathbb{R}^d$.

for t *in* 0 **to** $T-1$ **do**

  Generate $\boldsymbol{w}_t \in \mathbb{R}^d$ with $w_{t1}, w_{t2}, \cdots, w_{td} \overset{\text{i.i.d.}}{\sim} N\left(0, (\eta^0)^2 2B^2 \frac{d\log(2T/\delta)}{n^2(\varepsilon/T)^2}\right)$;

  Compute $\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - (\eta_0/n)\sum_{i=1}^n (\psi'(\boldsymbol{x}_i^\top \boldsymbol{\beta}^t) - \Pi_R(y_i))\boldsymbol{x}_i + \boldsymbol{w}_t$;

end

Output: $\boldsymbol{\beta}^T$.

For analyzing the privacy guarantee and rate of convergence of Algorithm 1, we collect here some useful assumptions.

(D1) Bounded design: there is a constant $\sigma_{\boldsymbol{x}} < \infty$ such that $\|\boldsymbol{x}\|_2 < \sigma_{\boldsymbol{x}}\sqrt{d}$ almost surely.
(D2) Bounded moments of design: $\mathbb{E}\boldsymbol{x} = \boldsymbol{0}$, and the covariance matrix $\Sigma_{\boldsymbol{x}} = \mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top$ satisfies $0 < 1/C < \lambda_{\min}(\Sigma_{\boldsymbol{x}}) \le \lambda_{\max}(\Sigma_{\boldsymbol{x}}) < C$ for some constant $0 < C < \infty$.
(G1) The function ψ in the GLM (3.1) satisfies $\|\psi'\|_\infty < c_1$ for some constant $c_1 < \infty$.
(G2) The function ψ satisfies $\|\psi''\|_\infty < c_2$ for some constant $c_2 < \infty$.

These assumptions are comparable to those required for the theoretical analysis of GLMs in the non-private setting; for examples, see [44, 40, 58] and the references therein.

Because the algorithm is a composition of T individual steps, if each step is $(\varepsilon/T, \delta/T)$-differentially private, the overall algorithm would be (ε, δ)-differentially private by the composition property of differential privacy. This is indeed the case under appropriate assumptions.

PROPOSITION 3.3. *If assumptions (D1) and (G1) hold, then choosing $B = 4(R + c_1)\sigma_{\boldsymbol{x}}$ guarantees that Algorithm 1 is (ε, δ)-differentially private.*

Proposition 3.3 is proved in Section A.4. Although the privacy guarantee holds for any number of iterations T, choosing T properly is crucial for the accuracy of Algorithm 1, as a larger value of T introduces a greater amount noise into Algorithm 1 to achieve privacy.

Existing results on noisy gradient descent typically require $O(n)$ [10] or $O(n^2)$ [11] iterations for minimizing generic convex functions. For the GLM problem, it turns out that $O(\log n)$ iterations suffice, thanks to the restricted strong convexity and restricted smoothness of generalized linear models (see, for example, [40], Proposition 1).

These weaker versions of strong convexity and smoothness are sufficient for Algoirthm 1 to attain linear convergence, which is the same rate for minimizing strongly convex and smooth functions. Therefore, $O(\log n)$ iterations would allow the algorithm to converge to an accuracy of $O(n^{-1})$ within $\hat{\boldsymbol{\beta}}$, the true minimizer of $\mathcal{L}_n$, in terms of squared ℓ_2 norm; as the squared ℓ_2 risk of $\hat{\boldsymbol{\beta}}$, $\mathbb{E}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$, is of order d/n, there is little reason from a statistical perspective to run the algorithm further than $O(\log n)$ iterations.

THEOREM 3.2. *Let $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n]}$ be an i.i.d. sample from the GLM (3.1), and let the true regression coefficients be denoted by $\hat{\boldsymbol{\beta}}^* \in \mathbb{R}^d$. Suppose assumptions (D1), (D2), (G1) and (G2) are true. There exist data-agnostic choices of tuning parameters $\eta^0 = O(1)$, $R = O(\sqrt{\log n})$, $B = O(\sqrt{\log n})$, $T = O(\log n)$, and initial value $\boldsymbol{\beta}^0 \in \mathbb{R}^d$ such that, if $n \gtrsim c(\sigma)\left(d\sqrt{\log(1/\delta)}\log^2 n/\varepsilon\right)$ for a sufficiently large constant K, the output of Algorithm*

1 satisfies

$$\|\boldsymbol{\beta}^T - \boldsymbol{\beta}^*\|_2^2 \lesssim c(\sigma)\left(\frac{d}{n} + \frac{d^2\log(1/\delta)\log^4 n}{n^2\varepsilon^2}\right) \tag{3.6}$$

with probability at least $1 - c_3\exp(-c_4\log n)$ *for some absolute constants* $c_3, c_4 > 0$.

Theorem 3.2 is proved in Section A.5. The requisite scaling of n versus d, ε and δ is reasonable, as our lower bound result, Theorem 3.1, implies that no estimator can achieve low ℓ_2-error unless the assumed scaling holds. More importantly, comparing the rate of convergence (3.6) and the lower bound Theorem 3.1 reveals that the latter is tight up to at most a logarithmic factor in n, under the usual setting of $\delta \asymp n^{-\alpha}$ with $\alpha > 1$.

4. The Bradley-Terry-Luce Model. Rank aggregation based on pairwise comparisons is a common problem in a range of applications, including recommendation systems [8], sports tournaments [41], and education [29]. The Bradley-Terry-Luce (BTL) model is widely recognized as the most popular model for analyzing pairwise comparisons. In this section, we investigate parameter estimation with differential privacy in the BTL model, where each of the n items is associated with an unobserved parameter that represents its "strength" or "quality". The probability of one item winning a comparison over another is determined by their latent parameters. The statistical problem is to estimate these parameters using the observed random comparison outcomes while preserving data privacy through differential privacy techniques. Accurate parameter estimation allows for the ranking of the items.

Suppose there are n items indexed by $[n] = \{1, 2, \cdots, n\}$. We observe comparisons between pairs of items as follows.

- A pair of items indexed by $1 \le i < j \le n$ is compared with probability $0 < p < 1$ and independent of any other pair. The n items form a "comparison graph" where an edge (i, j) is present if and only if items i and j are compared. Let $\mathcal{G}$ denote the edge set of this comparison graph.
- Each item i is associated with a latent parameter $\theta_i \in [-1, 1]$. Given $\mathcal{G}$, the outcome of a comparison between items i and j is encoded by a Bernoulli random variable Y_{ij} which takes the value 1 if i wins. The distribution of Y_{ij} is independent of any other pair and determined by the latent parameters:

$$\mathbb{P}(Y_{ij} = 1) = \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}}.$$

The goal is to estimate the latent parameters $\boldsymbol{\theta} = \{\theta_i\}_{i\in[n]}$ based on the observed comparison outcomes $\{Y_{ij}\}_{(i,j)\in\mathcal{G}}$ with a differentially private algorithm. Here, we would like to protect the privacy of the comparison results for each individual given the algorithmic output. More specifically, two data sets are considered adjacent if and only if there exists one individual whose comparison outcomes in one data set differ from the same individual's outcomes in the other data set; the underlying comparison graph is assumed to be identical between adjacent data sets.

Let the parameter space be denoted by $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^n : \|\boldsymbol{\theta}\|_\infty \le 1\}$. The quantity of interest is the privacy-constrained minimax risk $\inf_{M\in\mathcal{M}_{\varepsilon,\delta}} \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\|M(\boldsymbol{Y}) - \boldsymbol{\theta}\|_2^2$. A privacy-constrained minimax lower bound for this problem is established via the score attack technique in Section 4.1. We then propose a differentially private estimator via maximizing a randomly perturbed and ℓ_2-penalized version of the likelihood function in Section 4.2. The minimax lower bound is shown to be optimal by analyzing the performance of this differentially private estimator.

4.1. *The Privacy-constrained Minimax Lower Bound.* To lower bound the privacy-constrained minimax risk, we consider the score attack that traces if the comparison results of item i are in the training data set for the pairwise comparison model. Let $\{\boldsymbol{e}_k\}_{k\in[n]}$ denote the standard basis of $\mathbb{R}^n$; for each item i with $1 \le i \le n$ and any estimator $M(\boldsymbol{Y})$ of $\boldsymbol{\theta} \in \Theta$, we have the score attack

$$\mathcal{A}(M(\boldsymbol{Y}), i) = \sum_{j=1}^{n} \mathbb{1}((i,j) \in \mathcal{G}) \left\langle M(\boldsymbol{Y}) - \boldsymbol{\theta}, \left(Y_{ij} - \frac{1}{1+\exp(-(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{\theta})} \right) (\boldsymbol{e}_i - \boldsymbol{e}_j) \right\rangle.$$

When the reference to M and $\boldsymbol{Y}$ is unambiguous, it is convenient to notate $A_i := \mathcal{A}(M(\boldsymbol{Y}), i)$. The strategy for establishing a lower bound, as usual, is to analyze $\sum_{i=1}^n \mathbb{E}A_i$, the expected value of score attacks summed over an entire data set.

When M is a differentially private estimator, the soundness of score attack, 2.1 and Proposition 2.1 yield an upper bound of $\sum_{i=1}^n \mathbb{E}A_i$. Unlike the GLM example in Section 3, the upper bound is not obtained by directly plugging in the Fisher information matrix on the right side, but requires some analysis tailored to the random comparison graph and the BTL model. The detailed proof is deferred to Section B.1.

PROPOSITION 4.1. *If M is an (ε, δ)-differentially private algorithm with $0 < \varepsilon < 1$ and $p > 1/2n$, then for sufficiently large n and every $\boldsymbol{\theta} \in \Theta$, it holds that*

$$\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}} A_i \le 16np\varepsilon \cdot \sqrt{\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}} \|M(\boldsymbol{Y}) - \boldsymbol{\theta}\|_2^2} + 16n^2\delta. \tag{4.1}$$

After upper bounding $\sum_{i=1}^n \mathbb{E}_{Y|\boldsymbol{\theta}} A_i$ at every $\boldsymbol{\theta} \in \Theta$, we show that $\sum_{i=1}^n \mathbb{E}_{Y|\boldsymbol{\theta}} A_i$ is bounded away from zero in an "average" sense: there exists a prior distribution $\boldsymbol{\pi}$ over Θ such that $\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta}} \mathbb{E}_{Y|\boldsymbol{\theta}} A_i$ is lower bounded. Specifically, let the density of each coordinate of $\boldsymbol{\theta}$ be $\pi(t) = \mathbb{1}(|t| < 1)(15/16)(1-t^2)^2$, and we have the following result.

PROPOSITION 4.2. *Suppose M is an estimator of $\boldsymbol{\theta}$ such that $\sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\|M(\boldsymbol{Y}) - \boldsymbol{\theta}\|_2^2 \le c_0 n$ for a sufficiently small constant c_0. If each coordinate of $\boldsymbol{\theta}$ has density $\pi(t) = \mathbb{1}(|t| < 1)(15/16)(1-t^2)^2$, then there is some constant $C > 0$ such that*

$$\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}} \mathbb{E}_{Y|\boldsymbol{\theta}} A_i > Cn. \tag{4.2}$$

We are now ready to state the privacy-constrained minimax lower bound for estimating $\boldsymbol{\theta}$, by combining the bounds on $\sum_{i=1}^n \mathbb{E}A_i$ in Propositions 4.1 and 4.2.

THEOREM 4.1. *If $\sqrt{n}p\varepsilon > 1$, $0 < \varepsilon < 1$ and $\delta < cn^{-1}$ for a sufficiently small constant $c > 0$, it holds that*

$$\inf_{M \in \mathcal{M}_{\varepsilon,\delta}} \sup_{\theta \in \Theta} \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}} \|M(\boldsymbol{Y}) - \boldsymbol{\theta}\|_2^2 \gtrsim \frac{1}{p} + \frac{1}{p^2\varepsilon^2}. \tag{4.3}$$

The proof is in Section B.3. The privacy-constrained minimax risk lower bound, similar to its GLM counterpart, consists of the "statistical" term which holds regardless of privacy [43, 47], and a term attributable to the differential privacy constraint. The next step is to show the lower bound (4.3) is optimal, by constructing a differentially private algorithm with matching rate of convergence.

4.2. *Optimality of the Private BTL Minimax Lower Bound.* For constructing an (ε, δ)-differentially private estimator of $\boldsymbol{\theta}$, our approach is to maximize a randomly perturbed and ℓ_2-penalized version of the likelihood function. The negative log-likelihood function is given by

$$\mathcal{L}(\boldsymbol{\theta}; y) = \sum_{(i,j)\in\mathcal{G}} -y_{ij}(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{\theta} + \log(1 + \exp((\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{\theta})).$$

As the model is invariant to translations of $\boldsymbol{\theta}$, we further assume that the true parameter $\boldsymbol{\theta}$ is centered: $\mathbf{1}^\top \boldsymbol{\theta} = 0$. Define the feasible set $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^n : \|\boldsymbol{\theta}\|_\infty \le 1, \mathbf{1}^\top \boldsymbol{\theta} = 0\}$ and consider an estimator

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta}; y) + \frac{\gamma}{2}\|\boldsymbol{\theta}\|_2^2 + \boldsymbol{w}^\top \boldsymbol{\theta}, \quad \boldsymbol{w} = (w_1, w_2, \cdots, w_n) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2), \tag{4.4}$$

The choices of γ and σ to ensure differential privacy and estimation accuracy of $\hat{\boldsymbol{\theta}}$ are to be specified next.

PROPOSITION 4.3. *If $\sigma \ge \frac{\sqrt{n}\sqrt{8\log(2/\delta)+4\varepsilon}}{\varepsilon}$ and $\gamma \ge 1/\varepsilon$, $\hat{\boldsymbol{\theta}}$ is (ε, δ)-differentially private.*

Intuitively, the noise term added to the objective function in (4.4) is equivalent to perturbing the stationary condition of the original problem, and the ℓ_2-regularization coefficient γ ensures that the objective function is strongly convex, so that perturbing the gradient maps to sufficient perturbation to the solution. This perturbation method is an instance of the general "objective perturbation" method in differentially private optimization.

While larger values of hyper-parameters σ, γ lead to stronger privacy guarantees, they also lead to slower convergence of the estimator. The next proposition quantifies this effect in terms of σ.

PROPOSITION 4.4. *If $\gamma = c_0\sqrt{np}$ for some absolute constant c_0, $p \ge c_1 \log n/n$ for some sufficiently large constant c_1, then*

$$\mathbb{E}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \lesssim \frac{1}{p} + \frac{\sigma^2}{np^2}.$$

Proposition 4.4 is proved in Section B.5. Comparing the privacy guarantee, Proposition 4.3 with the rate of convergence, Proposition 4.4, tells us the best choice of γ and σ, which leads to the optimal risk upper bound for the estimator $\hat{\boldsymbol{\theta}}$.

THEOREM 4.2. *If $\varepsilon \lesssim \log(1/\delta)$, $\varepsilon > c_0(np)^{-1/2}$ for some absolute constant $c_0 > 0$, $p \ge c_1 \log n/n$ for some absolute constant $c_1 > 0$ and $\lambda = \varepsilon/16$, then the estimator $\hat{\boldsymbol{\theta}}$ defined in* (4.4) *is (ε, δ)-differentially private and satisfies*

$$\mathbb{E}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \lesssim \frac{1}{p} + \frac{\log(1/\delta)}{p^2\varepsilon^2}. \tag{4.5}$$

The condition $\varepsilon > c_0(np)^{-1/2}$ ensures that the choice of $\gamma \asymp \sqrt{np}$ in Proposition 4.4 satisfies the requirement $\gamma > 1/\varepsilon$ in Proposition 4.3. The other regularity conditions are inherited from the two propositions. The bound (4.5) is obtained by plugging $\sigma = 16\sqrt{n}\log(1/\delta)/\varepsilon$ into Proposition 4.4.

Theorem 4.2 implies that the privacy-constrained minimax lower bound in Theorem 4.1 is rate-optimal up to logarithm factors.

5. The High-dimensional Sparse GLMs. High-dimensional generalized linear models (GLMs) are widely used in contemporary data-driven scientific research and have a vast range of applications in various fields, such as genetics, metabolomics, finance, and econometrics. In this section, we consider privacy-preserving parameter estimation under the generalized linear model

$$f_{\boldsymbol{\beta}}(y|\boldsymbol{x}) = h(y,\sigma)\exp\left(\frac{y\boldsymbol{x}^\top\boldsymbol{\beta} - \psi(\boldsymbol{x}^\top\boldsymbol{\beta})}{c(\sigma)}\right); \boldsymbol{x}\sim f_{\boldsymbol{x}} \tag{5.1}$$

in a high-dimensional setting where d, the dimension of $\boldsymbol{\beta}$, dominates the sample size n, but the vector of regression coefficients $\boldsymbol{\beta}$ is assumed to be s^*-sparse: $\|\boldsymbol{\beta}\|_0 \le s^*$. Under the sparsity assumption, the privacy-constrained minimax risk will scale linearly with the sparsity, or the "intrinsic dimension" of $\boldsymbol{\beta}$, and only logarithmically with the "ambient dimension" d. This much different setting from the non-sparse GLM considered in Section 3 also calls for new methods: we study a sparse score attack in Section 5.1 to establish the minimax risk lower bound, and propose a iterative hard thresholding algorithm in Section 5.2 with matching risk upper bound.

5.1. *The Sparse Score Attack for Minimax Lower Bound .* For the high-dimensional sparse GLM, we consider a modification of the classical GLM score attack (3.2), the sparse GLM score attack:

$$\mathcal{A}_{\boldsymbol{\beta},s^*}((\widetilde{y},\widetilde{\boldsymbol{x}}), M(\boldsymbol{y},\boldsymbol{X})) = \frac{1}{c(\sigma)}\big\langle (M(\boldsymbol{y},\boldsymbol{X})-\boldsymbol{\beta})_{\mathrm{supp}(\boldsymbol{\beta})}, [\widetilde{y}-\psi'(\widetilde{\boldsymbol{x}}^\top\boldsymbol{\beta})]\widetilde{\boldsymbol{x}}\big\rangle. \tag{5.2}$$

It is called a sparse score attack because we are restricting the inner product to the non-zero coordinates of $\boldsymbol{\beta}$, which is a small fraction of all d coordinates. For each $i\in[n]$, we denote $\mathcal{A}_{\boldsymbol{\beta},s^*}((y_i,\boldsymbol{x}_i), M(\boldsymbol{y},\boldsymbol{X}))$ by A_i and try to bound the sum of expectations $\sum_{i\in[n]}\mathbb{E}A_i$. As usual, upper bounding $\sum_{i\in[n]}\mathbb{E}A_i$ relies on the soundness of score attack, Theorem 2.1, and the differential privacy of estimator M.

PROPOSITION 5.1. *Consider i.i.d. observations $(y_1,\boldsymbol{x}_1),\cdots,(y_n,\boldsymbol{x}_n)$ drawn from* (5.1) *with $\|\beta\|_0\le s^*$. Suppose $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top)$ is diagonal and $\lambda_{\max}(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top)) < C < \infty$, $\|\boldsymbol{x}\|_\infty < c < \infty$ almost surely, and $\|\psi^{''}\|_\infty < c_2 < \infty$. If the estimator M is (ε,δ)-differentially private with $0<\varepsilon<1$ and satisfies $\|M(\boldsymbol{y},\boldsymbol{X})-\boldsymbol{\beta}\|_2^2 \lesssim s^*$, then*

$$\sum_{i\in[n]}\mathbb{E}A_i \le 2n\varepsilon\sqrt{\mathbb{E}\|M(\boldsymbol{y},\boldsymbol{X})-\boldsymbol{\beta}\|_2^2}\sqrt{Cc_2/c(\sigma)} + 4\sqrt{2}\delta s^*\sqrt{c_2\log(1/\delta)/c(\sigma)}. \tag{5.3}$$

The proposition is proved in Section C.1.

For lower bounding $\sum_{i\in[n]}\mathbb{E}_{\boldsymbol{y},\boldsymbol{X}|\boldsymbol{\beta}}A_i$ on average over some prior distribution of $\boldsymbol{\beta}$, a major difference from its counterpart in the non-sparse GLM case is that we have to choose a prior distribution over the set of s^*-sparse vectors, $\{\boldsymbol{\beta}:\boldsymbol{\beta}\in\mathbb{R}^d, \|\boldsymbol{\beta}\|_0\le s^*\}$. Specifically, we consider $\boldsymbol{\beta}$ generated as follows: let $\widetilde{\beta}_1,\widetilde{\beta}_2,\cdots,\widetilde{\beta}_d$ be an i.i.d. sample from the truncated normal $N(0,\gamma^2)$ distribution with truncation at -1 and 1, let I_{s^*} be be the index set of $\widetilde{\boldsymbol{\beta}}$ with top s^* greatest absolute values so that $|I_{s^*}| = s^*$ by definition, and define $\beta_j = \widetilde{\beta}_j\mathbb{1}(j\in I_{s^*})$.

Then, by the Stein's Lemma argument in Section 2.2.2, we obtain a lower bound of $\sum_{i\in[n]}\mathbb{E}_{\boldsymbol{pi}}\mathbb{E}_{\boldsymbol{y},\boldsymbol{X}|\boldsymbol{\beta}}A_i$, where $\boldsymbol{\pi}$ refers to the sparse truncated normal prior described above.

PROPOSITION 5.2. *Suppose $s^* = o\left(d^{1-\gamma}\right)$ for some $\gamma>0$. For every M satisfying $\mathbb{E}_{\boldsymbol{y},\boldsymbol{X}|\boldsymbol{\beta}}\|M(\boldsymbol{y},\boldsymbol{X})-\boldsymbol{\beta}\|_2^2 \lesssim 1$ at every $\boldsymbol{\beta}$, we have*

$$\sum_{i\in[n]}\mathbb{E}_{\boldsymbol{\pi}}\mathbb{E}_{\boldsymbol{y},\boldsymbol{X}|\boldsymbol{\beta}}A_i \gtrsim s^*\log(d/s^*), \tag{5.4}$$

where $\boldsymbol{\pi}$ refers to the sparse truncated normal prior for $\boldsymbol{\beta}$.

Proposition 5.2 is proved in Section C.2. It is noteworthy that, as a result of the sparse prior, the right side $s^* \log(d/s)$ is different from its non-sparse counterpart in Proposition 3.2. We are now ready to combine the two propositions to obtain a minimax risk lower bound for sparse GLMs.

THEOREM 5.1. *Consider i.i.d. observations $(y_1, \boldsymbol{x}_1), \cdots, (y_n, \boldsymbol{x}_n)$ drawn from (5.1) with $\|\boldsymbol{\beta}\|_0 \leq s^*$, and $s^* = o(d^{1-\gamma})$ for some $\gamma > 0$. Suppose $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top)$ is diagonal and $\lambda_{\max}(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top)) < C < \infty$, $\|\boldsymbol{x}\|_2 \infty < c < \infty$, and $\|\psi''\|_\infty < c_2 < \infty$. If $s^* \log(d/s^*) \lesssim n\varepsilon$, $0 < \varepsilon < 1$ and $\delta \lesssim n^{-(1+\gamma)}$ for some $\gamma > 0$, then*

(5.5)
$$\inf_{M \in \mathcal{M}_{\varepsilon,\delta}} \sup_{\boldsymbol{\beta} \in \mathbb{R}^d, \|\boldsymbol{\beta}\|_0 \leq s^*} \mathbb{E}\|M(\boldsymbol{y}, \boldsymbol{X}) - \boldsymbol{\beta}\|_2^2 \gtrsim c(\sigma) \left(\frac{s^* \log(d/s^*)}{n} + \frac{(s^* \log(d/s^*))^2}{n^2 \varepsilon^2} \right).$$

Theorem 5.1 is proved in Section C.3. To show that the lower bound is tight, we propose in the next section an algorithm for estimating the sparse $\boldsymbol{\beta}$ with differential privacy. From the desired rate of convergence (5.5), it is already apparent that the noisy gradient descent algorithm considered in Section 3 is unlikely to succeed, for its requisite noise scales with the full dimension d. Our iterative hard thresholding algorithm manages to add noise which scales with sparsity and shows the lower bound (5.5) is achievable up to a logarithmic factor in n.

5.2. *Optimality of the Private Sparse GLM Lower Bound.* In this section, we construct a differentially private algorithm for estimating GLM parameters when the dimension d dominates the sample size n. Even without privacy requirements, directly minimizing the negative log-likelihood function $\mathcal{L}_n(\boldsymbol{\beta})$ no longer achieves any meaningful statistical accuracy, because the objective function $\mathcal{L}_n$ can have infinitely many minimizers due to a rank-deficient Hessian matrix $\nabla^2 \mathcal{L}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \psi''(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \boldsymbol{x}_i \boldsymbol{x}_i^\top$.

The problem is nevertheless solvable when the true parameter vector $\boldsymbol{\beta}^*$ is s^*-sparse with $s^* = o(d)$, that is when at most s^* out of d coordinates of $\boldsymbol{\beta}^*$ are non-zero. For estimating a sparse $\boldsymbol{\beta}^*$, the primary challenge lies in (approximately) solving the non-convex optimization problem $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}: \|\boldsymbol{\beta}\|_0 \leq s^*} \mathcal{L}_n(\boldsymbol{\beta}; \boldsymbol{Z})$. Some popular non-private approaches include convex relaxation via ℓ_1 regularization of $\mathcal{L}_n$ [44, 4], or projected gradient descent onto the non-convex feasible set $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_0 \leq s^*\}$, also known as iterative hard thresholding [12, 30]:

Algorithm 2: Iterative Hard Thresholding (IHT)

Input : Objective function $f(\boldsymbol{\theta})$, sparsity s, step size η, number of iterations T.

Initialize $\boldsymbol{\theta}^0$ with $\|\boldsymbol{\theta}^0\|_0 \leq s$, set $t = 0$;

for t *in* 0 **to** $T - 1$ **do**

  $\boldsymbol{\theta}^{t+1} = P_s\left(\boldsymbol{\theta}^t - \eta \nabla f(\boldsymbol{\theta}^t)\right)$, where $P_s(\boldsymbol{v}) = \arg\min_{\boldsymbol{z}: \|\boldsymbol{z}\|_0 = s} \|\boldsymbol{v} - \boldsymbol{z}\|_2^2$;

end

Output: $\boldsymbol{\theta}^T$.

In each iteration, the algorithm updates the solution via gradient descent, keeps its largest s coordinates in magnitude, and sets the other coordinates to 0.

For privately fitting high-dimensional sparse GLMs, we shall construct a noisy version of Algorithm 2, and show in Section 5.2.2 that it enjoys a linear rate of convergence similar to the noisy gradient descent, Algorithm 1. As a first step, we consider in Section 5.2.1 a noisy, differentially private version of the projection operator P_s, as well as a noisy iterative hard thresholding algorithm applicable to any objective function that satisfies restricted strong convexity and restricted smoothness.

5.2.1. *The Noisy Iterative Hard Thresholding Algorithm.* At the core of our algoirthm is a noisy, differentially private algorithm that identifies the top-s largest coordinates of a given vector with good accuracy. The following "Peeling" algorithm [23] serves this purpose, with fresh Laplace noises added to the underlying vector and one coordinate "peeled" from the vector in each iteration.

Algorithm 3: Noisy Hard Thresholding (NoisyHT)

Input : vector-valued function $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{Z}) \in \mathbb{R}^d$, data $\boldsymbol{Z}$, sparsity s, privacy parameters ε, δ, noise scale λ.

Initialize $S = \emptyset$;

for i *in* 1 **to** s **do**

  Generate $\boldsymbol{w}_i \in \mathbb{R}^d$ with $w_{i1}, w_{i2}, \cdots, w_{id} \overset{\text{i.i.d.}}{\sim} \text{Laplace}\left(\lambda \cdot \frac{2\sqrt{3s\log(1/\delta)}}{\varepsilon}\right)$;

  Append $j^* = \arg\max_{j \in [d] \setminus S} |v_j| + w_{ij}$ to S;

end

Set $\widetilde{P}_s(\boldsymbol{v}) = \boldsymbol{v}_S$;

Generate $\widetilde{\boldsymbol{w}}$ with $\widetilde{w}_1, \cdots, \widetilde{w}_d \overset{\text{i.i.d.}}{\sim} \text{Laplace}\left(\lambda \cdot \frac{2\sqrt{3s\log(1/\delta)}}{\varepsilon}\right)$;

Output: $\widetilde{P}_s(\boldsymbol{v}) + \widetilde{\boldsymbol{w}}_S$.

The algorithm is guaranteed to be (ε, δ)-differentially private when the vector-valued function $\boldsymbol{v}(\boldsymbol{Z})$ is not sensitive to replacing any single datum.

LEMMA 5.1 ([23**?**]). *If for every pair of adjacent data sets $\boldsymbol{Z}, \boldsymbol{Z}'$ we have $\|\boldsymbol{v}(\boldsymbol{Z}) - \boldsymbol{v}(\boldsymbol{Z}')\|_\infty < \lambda$, then NoisyHT is an (ε, δ)-differentially private algorithm.*

The accuracy of Algorithm 3 is quantified by the next lemma.

LEMMA 5.2. *Let $\widetilde{P}_s$ be defined as in Algorithm* 3*. For any index set I, any $\boldsymbol{v} \in \mathbb{R}^I$ and $\hat{\boldsymbol{v}}$ such that $\|\hat{\boldsymbol{v}}\|_0 \le \hat{s} \le s$, we have that for every $c > 0$,*

$$\|\widetilde{P}_s(\boldsymbol{v}) - \boldsymbol{v}\|_2^2 \le (1 + 1/c)\frac{|I| - s}{|I| - \hat{s}}\|\hat{\boldsymbol{v}} - \boldsymbol{v}\|_2^2 + 4(1 + c)\sum_{i \in [s]} \|\boldsymbol{w}_i\|_\infty^2.$$

Lemma 5.2 is proved in Section C.4. In comparison, the exact, non-private projection operator P_s satisfies ([30], Lemma 1) $\|P_s(\boldsymbol{v}) - \boldsymbol{v}\|_2^2 \le \frac{|I|-s}{|I|-\hat{s}}\|\hat{\boldsymbol{v}} - \boldsymbol{v}\|_2^2$. Algorithm 3, therefore, is as accurate as its non-private counterpart up to a constant multiplicative factor and some additive noise. Taking the private top-s projection algorithm, we have the following noisy iterative hard thresholding algorithm.

Compared to the non-private Algorithm 2, we simply replaced the exact projection P_s with the noisy projection given by Algorithm 3. The privacy guarantee of Algorithm 4 is then inherited from that of Algorithm 3.

Algorithm 4: Noisy Iterative Hard Thresholding (NoisyIHT)

Input : Objective function $\mathcal{L}_n(\boldsymbol{\theta}, \boldsymbol{Z}) = n^{-1}\sum_{i=1}^n l(\boldsymbol{\theta}, \boldsymbol{z}_i)$, data set $\boldsymbol{Z}$, sparsity level s, step size η^0, privacy parameters ε, δ, noise scale B, number of iterations T.

Initialize $\boldsymbol{\theta}^0$ with $\|\boldsymbol{\theta}^0\|_0 \leq s$, set $t = 0$;

for t *in* 0 **to** $T-1$ **do**

  $\boldsymbol{\theta}^{t+1} = \texttt{NoisyHT}\left(\boldsymbol{\theta}^t - \eta^0 \nabla \mathcal{L}_n(\boldsymbol{\theta}^t; \boldsymbol{Z}), \boldsymbol{Z}, s, \varepsilon/T, \delta/T, (\eta^0/n)B\right)$;

end

Output: $\boldsymbol{\theta}^T$.

LEMMA 5.3. *If for every pair of adjacent data $\boldsymbol{z}, \boldsymbol{z}'$ and every $\boldsymbol{\theta} \in \Theta$ we have $\|\nabla l(\boldsymbol{\theta}; \boldsymbol{z}) - \nabla l(\boldsymbol{\theta}; \boldsymbol{z}')\|_\infty < B$, then NoisyIHT is an (ε, δ)-differentially private algorithm.*

The lemma is proved in Section C.5. Similar to the noisy gradient descent (Algorithm 1), the privacy guarantee of Algorithm 4 is valid for any choice of T, however a fast rate of convergence would allow us to select a small T and thereby introducing less noise into the algorithm. To our delight, restricted strong convexity and restricted smoothness again lead to a linear rate of convergence even in the high-dimensional sparse setting.

PROPOSITION 5.3. *Let $\hat{\boldsymbol{\theta}} = \arg\min_{\|\boldsymbol{\theta}\|_0 \leq s^*} \mathcal{L}_n(\boldsymbol{\theta}; \boldsymbol{Z})$. For iteration number $t \geq 0$, suppose*

$$\langle \nabla \mathcal{L}_n(\boldsymbol{\theta}^t) - \nabla \mathcal{L}_n(\hat{\boldsymbol{\theta}}), \boldsymbol{\theta}^t - \hat{\boldsymbol{\theta}} \rangle \geq \alpha \|\boldsymbol{\theta}^t - \hat{\boldsymbol{\theta}}\|_2^2 \tag{5.6}$$

$$\langle \nabla \mathcal{L}_n(\boldsymbol{\theta}^{t+1}) - \nabla \mathcal{L}_n(\hat{\boldsymbol{\theta}}), \boldsymbol{\theta}^{t+1} - \hat{\boldsymbol{\theta}} \rangle \leq \gamma \|\boldsymbol{\theta}^{t+1} - \hat{\boldsymbol{\theta}}\|_2^2. \tag{5.7}$$

for constants $0 < \alpha < \gamma$. Let $\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_s$ be the noise vectors added to $\boldsymbol{\theta}^t - \eta^0 \nabla \mathcal{L}_n(\boldsymbol{\theta}^t; \boldsymbol{Z})$ when the support of $\boldsymbol{\theta}^{t+1}$ is iteratively selected, S^{t+1} be the support of $\boldsymbol{\theta}^{t+1}$, and $\widetilde{\boldsymbol{w}}$ be the noise vector added to the selected s-sparse vector. Then, for $\eta_0 = 2/3\gamma$, there exists an absolute constant c_0 so that, choosing $s \geq c_0(\gamma/\alpha)^2 s^*$ guarantees

$$\mathcal{L}_n(\boldsymbol{\theta}^{t+1}) - \mathcal{L}_n(\hat{\boldsymbol{\theta}}) \leq \left(1 - \rho \cdot \frac{\alpha}{\gamma} - \frac{2s^*}{s + s^*}\right)\left(\mathcal{L}_n(\boldsymbol{\theta}^t) - \mathcal{L}_n(\hat{\boldsymbol{\theta}})\right) + C_\gamma \left(\sum_{i \in [s]} \|\boldsymbol{w}_i\|_\infty^2 + \|\widetilde{\boldsymbol{w}}_{S^{t+1}}\|_2^2\right),$$

where $0 < \rho < 1$ is an absolute constant, and $C_\gamma > 0$ is a constant depending on γ.

Proposition 5.3 is proved in Section C.6. While conditions (5.6) and (5.7) are similar to the ordinary strong convexity and smoothness conditions in appearance, they are in fact much weaker because $\hat{\boldsymbol{\theta}}$, $\boldsymbol{\theta}^t$ are both s-sparse. In the next section, we apply the iterative hard thresholding algorithm to the GLM likelihood function and obtain its rate of convergence to the truth $\boldsymbol{\beta}^*$.

5.2.2. *Noisy Iterative Hard Thresholding for the Sparse GLM.* Assuming that the true GLM parameter vector $\boldsymbol{\beta}^*$ satisfies $\|\boldsymbol{\beta}^*\|_0 \leq s^*$, we now specialize the results of Section 5.2.1 to the GLM negative log-likelihood function

$$\mathcal{L}_n(\boldsymbol{\beta}; \boldsymbol{Z}) = \frac{1}{n}\sum_{i=1}^n \left(\psi(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right).$$

Algorithm 5: Differentially Private Sparse Generalized Linear Regression

Input : $\mathcal{L}_n(\boldsymbol{\beta}, \boldsymbol{Z})$, data set $\boldsymbol{Z}$, sparsity level s, step size η^0, privacy parameters ε, δ, noise scale B, number of iterations T, truncation parameter R.

Initialize $\boldsymbol{\beta}^0$ with $\|\boldsymbol{\beta}^0\|_0 \le s$, set $t = 0$;

for t *in* 0 **to** $T-1$ **do**

  Compute $\boldsymbol{\beta}^{t+0.5} = \boldsymbol{\beta}^t - (\eta_0/n)\sum_{i=1}^n (\psi'(\boldsymbol{x}_i^\top \boldsymbol{\beta}^t) - \Pi_R(y_i))\boldsymbol{x}_i$;

  $\boldsymbol{\beta}^{t+1} = \texttt{NoisyHT}\left(\boldsymbol{\beta}^{t+0.5}, \boldsymbol{Z}, s, \varepsilon/T, \delta/T, \eta^0 B/n\right)$;

end

Output: $\boldsymbol{\beta}^T$.

Some assumptions about the data set $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n]}$ and its distribution will be helpful for analyzing the accuracy and privacy guarantees of Algorithm 5. The necessary assumptions for the high-dimensional sparse case are identical to those for the low-dimensional case, except with (D1) replaced by (D1'), as follows.

(D1') Bounded design: there is a constant $\sigma_{\boldsymbol{x}} < \infty$ such that $\|\boldsymbol{x}\|_\infty < \sigma_{\boldsymbol{x}}$ almost surely.

Because Algorithm 5 is a special case of the general Algorithm 4, the privacy guarantee of Algorithm 5 reduces to specializing Lemma 5.3 to GLMs, as follows.

LEMMA 5.4. *If assumptions (D1') and (G1) are true, then choosing* $B = 4(R + c_1)\sigma_{\boldsymbol{x}}$ *guarantees that Algorithm 5 is* (ε, δ)*-differentially private.*

The lemma is proved in Section C.7. For the rate of convergence of Algorithm 5, the restricted strong convexity and restricted smoothness of the GLM likelihood (see, for example, [40], Proposition 1) combined with the sparsity of $\hat{\boldsymbol{\beta}}$, $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^t$ for every t are sufficient for conditions (5.6) and (5.7) in Proposition 5.3 to hold. Applying Proposition 5.3 in a proof by induction leads to an upper bound for $\|\boldsymbol{\beta}^T - \boldsymbol{\beta}^*\|_2^2$. Below we state the main result; the detailed proof is in Section C.8.

THEOREM 5.2. *Let* $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n]}$ *be an i.i.d. sample from the model* (5.1) *where the true parameter vector* $\boldsymbol{\beta}^*$ *satisfies* $\|\boldsymbol{\beta}^*\|_0 \le s^*$*. Suppose assumptions (D1'), (D2), (G1) and (G2) are true. There exist data-agnostic choices of tuning parameters* $s \asymp s^*, \eta^0 = O(1)$*,* $R = O(\sqrt{\log n})$*,* $B = O(\sqrt{\log n})$*,* $T = O(\log n)$*, and initial value* $\boldsymbol{\beta}^0 \in \mathbb{R}^d$ *such that, if* $n \gtrsim c(\sigma)\left(s^* \log d \sqrt{\log(1/\delta)} \log^{3/2} n/\varepsilon\right)$*, the output of Algorithm 5 satisfies*

$$\|\boldsymbol{\beta}^T - \boldsymbol{\beta}^*\|_2^2 \lesssim c(\sigma)\left(\frac{s^* \log d}{n} + \frac{(s^* \log d)^2 \log(1/\delta) \log^3 n}{n^2 \varepsilon^2}\right). \tag{5.8}$$

with probability at least $1 - c_3 \exp(-c_4 \log(d/s^* \log n)) - c_3 \exp(-c_4 \log n)$ *for some absolute constants* $c_3, c_4 > 0$.

The assumed scaling of n versus d, s^*, ε and δ in Theorem 5.2 is reasonable, as the minimax lower bound, Theorem 5.1, shows that no estimator can achieve low ℓ_2-error unless the assumed scaling holds. The rate of convergence of Algorithm 5 implies that the minimax lower bound (5.5) established via score attack is optimal except possibly for factors of $\log n$, when δ is set at the usual level $\delta \asymp n^{-\alpha}$ for some $\alpha > 1$.

6. Nonparametric Function Estimation. Although the score statistics is a fundamentally parametric concept, the score attack method can still lead to optimal minimax lower bounds in nonparametric problems, as this example demonstrates.

Consider n pairs of random variables $\{(Y_i, X_i)\}_{i\in[n]}$ drawn i.i.d. from the model

$$Y_i = f(X_i) + \xi_i, X_i \sim U[0,1],$$

where the noise term ξ_i is independent of X_i and follows the $N(0,\sigma^2)$ distribution. We would like to estimate the unknown mean function $f : [0,1] \to \mathbb{R}$ with (ε,δ) differential privacy. For an estimator $\hat{f}$ of the true f, a reasonable metric for its performance is the mean integrated squared risk (MISE),

$$R(\hat{f}, f) = \mathbb{E}\left[\int_0^1 (\hat{f}(x) - f(x))^2 \mathrm{d}x\right],$$

where the expectation is taken over the joint distribution of $\{(Y_i, X_i)\}_{i\in[n]}$. As the true f is unknown, we cannot hope to know $R(\hat{f}, f)$ in general and assume instead that f belongs to some pre-specified class of functions $\mathcal{F}$. We may then circumvent the dependence on unknown f by considering the maximum MISE of $\hat{f}$ over the entire class $\mathcal{F}$,

$$\sup_{f\in\mathcal{F}} R(\hat{f}, f) = \sup_{f\in\mathcal{F}} \mathbb{E}\left[\int_0^1 (\hat{f}(x) - f(x))^2 \mathrm{d}x\right].$$

That is, $R(\hat{f}, \mathcal{F})$ measures the worst-case performance of $\hat{f}$ over the function class $\mathcal{F}$. In this example, we take $\mathcal{F}$ to be the periodic Sobolev class $\tilde{W}(\alpha, C)$ over $[0,1]$: for $\alpha \in \mathbb{N}$ and $C > 0$,

$$\tilde{W}(\alpha, C) = \left\{ f : [0,1] \to \mathbb{R} \middle| \int_0^1 (f^{(\alpha)}(x))^2 \mathrm{d}x \le C^2, f^{(j)}(0) = f^{(j)}(1) \text{ for } j \in [\alpha - 1] \right\}.$$

As usual, let the collection of all (ε,δ)-differentially private estimators be denoted by $\mathcal{M}_{\varepsilon,\delta}$. The privacy-constrained minimax risk of estimating f is therefore

$$\inf_{\hat{f}\in\mathcal{M}_{\varepsilon,\delta}} \sup_{f\in\tilde{W}(\alpha,C)} \mathbb{E}\left[\int_0^1 (\hat{f}(x) - f(x))^2 \mathrm{d}x\right].$$

We shall characterize the privacy-constrained minimax risk by first deriving a lower bound via the score attack method in Section 6.1, and then exhibit an estimator with matching risk upper bound in Section 6.2.

6.1. *The Nonparametric Minimax Lower Bound.* Lower bounding the onparametric privacy-constrained minimax risk is made easier by a sequence of reductions to parametric lower bound problems. The first step is to consider the orthogonal series expansion of $f \in \tilde{W}(\alpha, C)$ with respect to the Fourier basis

$$\varphi_1(t) = 1; \varphi_{2k}(t) = \sqrt{2}\cos(2\pi k t), \varphi_{2k}(t) = \sqrt{2}\sin(2\pi k t), k = 1, 2, 3 \cdots.$$

We have $f = \sum_{j=1}^{\infty} \theta_j \varphi_j(x)$, where the Fourier coefficients are given by $\theta_j = \int_0^1 f(x)\varphi_j(x)\mathrm{d}x, j = 1, 2, 3, \cdots$. The Fourier coefficients allow a convenient representation of the periodic Sobolev class $\tilde{W}(\alpha, C)$: a function f belongs to $\tilde{W}(\alpha, C)$ if and only if its Fourier coefficients belong to the "Sobolev ellipsoid",

$$\Theta(\alpha, C) = \left\{ \theta \in \mathbb{R}^{\mathbb{Z}^+} : \sum_{j=1}^{\infty} \tau_j^2 \theta_j^2 < C^2/\pi^{2\alpha} \right\}, \tag{6.1}$$

where $\tau_j = j^\alpha$ for even j and $\tau_j = (j-1)^\alpha$ for odd j. We can therefore define $\tilde{W}(\alpha, C)$ equivalently as

$$\tilde{W}(\alpha, C) = \left\{ f = \sum_{j=1}^{\infty} \theta_j \varphi_j : \theta \in \Theta(\alpha, C) \right\}.$$

This alternative definition of $\tilde{W}(\alpha, C)$ motivates a reduction from the original lower bound problem over an infinite-dimensional space, $\tilde{W}(\alpha, C)$, to a finite-dimensional lower bound problem. Specifically, for $k \in \mathbb{N}$, consider the k-dimensional subspace

$$\tilde{W}_k(\alpha, C) = \left\{ f = \sum_{j=1}^{\infty} \theta_j \varphi_j : \theta \in \Theta(\alpha, C), \theta_j = 0 \text{ for every } j > k \right\}.$$

It follows that $\tilde{W}_k(\alpha, C) \subseteq \tilde{W}(\alpha, C)$ for every k; in other words, for every k we have

(6.2)
$$\inf_{\hat{f} \in \mathcal{M}_{\varepsilon,\delta}} \sup_{f \in \tilde{W}(\alpha,C)} \mathbb{E}\left[\int_0^1 (\hat{f}(x) - f(x))^2 \mathrm{d}x \right] \geq \inf_{\hat{f} \in \mathcal{M}_{\varepsilon,\delta}} \sup_{f \in \tilde{W}_k(\alpha,C)} \mathbb{E}\left[\int_0^1 (\hat{f}(x) - f(x))^2 \mathrm{d}x \right].$$

The next step is to find a minimax lower bound over each k-dimensional subspace, and optimize over k to solve the original problem.

6.1.1. *Finite-dimensional Minimax Lower Bounds via Score Attack.* Once we focus on the k-dimensional subspace, the problem can be further simplified. For an estimator $\hat{f}$ and some $f \in \tilde{W}_k(\alpha, C)$, let $\{\hat{\theta}_j\}_{j \in \mathbb{N}}$ and $\{\theta_j\}_{j \in \mathbb{N}}$ be their respective Fourier coefficients. By the orthonormality of the Fourier basis, we have

$$\mathbb{E}\left[\int_0^1 (\hat{f}(x) - f(x))^2 \mathrm{d}x \right] \geq \mathbb{E} \sum_{j=1}^{k} (\hat{\theta}_j - \theta_j)^2, \tag{6.3}$$

reducing the original problem into lower bounding the minimax mean squared risk of estimating a finite-dimensional parameter. Let $\Theta_k(\alpha, C)$ denote a finite-dimensional restriction of the Sobolev ellipsoid,

$$\Theta_k(\alpha, C) = \left\{ \theta \in \mathbb{R}^k : \sum_{j=1}^{k} \tau_j^2 \theta_j^2 < C^2/\pi^{2\alpha} \right\},$$

and suppose $M(\boldsymbol{X}, \boldsymbol{Y})$ is a differentially private estimator of $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k) \in \Theta_k(\alpha, C)$. For $i \in [n]$, consider the score attack given by

$$\mathcal{A}(M(\boldsymbol{X}, \boldsymbol{Y}), (X_i, Y_i)) = \left\langle M(\boldsymbol{X}, \boldsymbol{Y}) - \boldsymbol{\theta}, \sigma^{-2} \left(Y_i - \sum_{j=1}^{k} \theta_j \varphi_j(X_i) \right) \boldsymbol{\varphi}(X_i) \right\rangle,$$

where $\boldsymbol{\varphi}$ denotes the vector valued function $\boldsymbol{\varphi} : \mathbb{R} \to \mathbb{R}^k$, $\boldsymbol{\varphi}(x) = (\varphi_1(x), \varphi_2(x), \cdots, \varphi_k(x))$.

When the reference to M and $(\boldsymbol{X}, \boldsymbol{Y})$ is clear, we notate $A_i := \mathcal{A}(M(\boldsymbol{X}, \boldsymbol{Y}), (X_i, Y_i))$. To establish a lower bound of $\sup_{\boldsymbol{\theta} \in \Theta_k(\alpha,C)} \mathbb{E}\|M(\boldsymbol{X}, \boldsymbol{Y}) - \boldsymbol{\theta}\|_2^2$, we shall analyze $\sum_{i \in [n]} \mathbb{E} A_i$, the expected value of score attacks summed over an entire data set.

PROPOSITION 6.1. *If M is an (ε, δ)-differentially private algorithm with $0 < \varepsilon < 1$, then for sufficiently large n and every $\boldsymbol{\theta} \in \Theta_k(\alpha, C)$, it holds that*

$$\sum_{i \in [n]} \mathbb{E}_{\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{\theta}} A_i \leq \sigma^{-1} \left(2n\varepsilon \sqrt{\mathbb{E}_{\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{\theta}} \|M(\boldsymbol{X}, \boldsymbol{Y}) - \boldsymbol{\theta}\|_2^2} + 8Cn\sqrt{k \log(1/\delta)}\delta \right). \tag{6.4}$$

The proof of Proposition 6.1 is deferred to Section D.1.

After upper bounding $\sum_{i\in[n]} \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta}} A_i$ at every $\boldsymbol{\theta} \in \Theta_k(\alpha, C)$, we show that $\sum_{i\in[n]} \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta}} A_i$ is bounded away from zero in an "average" sense: there is a prior distribution $\boldsymbol{\pi}$ over $\boldsymbol{\theta} \in \Theta_k(\alpha, C)$ such that $\sum_{i\in[n]} \mathbb{E}_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta}} A_i$ is lower bounded. Specifically, let each θ_j follow the uniform distribution between $-B$ and B, where $B^2 = \frac{C^2}{2\pi^{2\alpha}} \left(\int_1^{k+1} t^{2\alpha} \mathrm{d}t\right)^{-1} \asymp k^{-(2\alpha+1)}$, so that

$$\sum_{j=1}^{k} \tau_j^2 \theta_j^2 \le B^2 \sum_{j=1}^{k} j^{2\alpha} \le \frac{C^2}{2\pi^{2\alpha}}$$

ensures the chosen prior distribution is supported within $\Theta_k(\alpha, C)$.

PROPOSITION 6.2. *Let* $B^2 = \frac{C^2}{2\pi^{2\alpha}} \left(\int_1^{k+1} t^{2\alpha} \mathrm{d}t\right)^{-1}$. *Suppose* M *is an estimator of* $\boldsymbol{\theta}$ *satisfying*

$$\sup_{\theta\in\Theta_k(\alpha,C)} \mathbb{E}\|M(\boldsymbol{X}, \boldsymbol{Y}) - \boldsymbol{\theta}\|_2^2 \le \frac{kB^2}{20}.$$

If each coordinate of $\boldsymbol{\theta}$ *follows the uniform distribution between* $-B$ *and* B, *then there is some constant* $c > 0$ *such that*

$$\sum_{i\in[n]} \mathbb{E}_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta}} A_i > ck. \tag{6.5}$$

The proposition is proved in Section D.2. Like in every parametric example we have considered so far, the bounds of the score attack's expectations, Propositions 6.1 and 6.2, imply a finite-dimensional minimax lower bound.

PROPOSITION 6.3. *If* $0 < \varepsilon < 1$ *and* $0 < \delta < cn^{-2}$ *for a sufficiently small constant* $c > 0$, *it holds that*

$$\inf_{M\in\mathcal{M}_{\varepsilon,\delta}} \sup_{\boldsymbol{\theta}\in\Theta_k(\alpha,C)} \mathbb{E}\|M(\boldsymbol{X}, \boldsymbol{Y}) - \boldsymbol{\theta}\|_2^2 \gtrsim \min\left(k^{-2\alpha}, \frac{k^2}{n^2\varepsilon^2}\right). \tag{6.6}$$

The finite-dimensional lower bound is proved in Section D.3. We are now ready to recover the nonparametric lower bound by optimizing over k.

6.1.2. *Optimizing the Finite-dimensional Lower Bounds.* By the reductions (6.2) and (6.3), it suffices to optimize the finite-dimensional lower bound (6.6) with respect to k to obtain the desired lower bound over $\tilde{W}(\alpha, C)$, by setting $k \asymp (n\varepsilon)^{\frac{1}{\alpha+1}}$.

THEOREM 6.1. *If* $0 < \varepsilon < 1$, $0 < \delta < cn^{-2}$ *for a sufficiently small constant* $c > 0$ *and* $n\varepsilon \gtrsim 1$, *it holds that*

$$\inf_{\hat{f}\in\mathcal{M}_{\varepsilon,\delta}} \sup_{f\in\tilde{W}(\alpha,C)} \mathbb{E}\left[\int_0^1 (\hat{f}(x) - f(x))^2 \mathrm{d}x\right] \gtrsim n^{-\frac{2\alpha}{2\alpha+1}} + (n\varepsilon)^{-\frac{2\alpha}{\alpha+1}}. \tag{6.7}$$

The first term can be recognized as the optimal MISE of function estimation in the periodic Sobolev class of order α, and the second term is the cost of differential privacy. The next section shows the optimality of this nonparametric privacy-constrained lower bound, by exhibiting an estimator with matching MISE up to a logarithmic factor in n.

6.2. *Optimality of the Nonparametric Lower Bound.* Absent the differential privacy constraint, the jth Fourier coefficient of the mean function f can be estimated by its empirical version, $\hat{\theta}_j = n^{-1}\sum_{i=1}^n Y_i\varphi_j(X_i)$, and the function f is then estimated by $\hat{f}(x) = \sum_{j=1}^K \hat{\theta}_j\varphi_j(x)$ for some appropriately chosen K.

We construct an estimator of f also by estimating the Fourier coefficients with differential privacy, then the estimator of f would be differentially private as well by post-processing. The sample mean $\hat{\theta}_j = n^{-1}\sum_{i=1}^n Y_i\varphi_j(X_i)$ lends itself naturally to the noise addition mechanisms, except that the Gaussian-distributed Y_i are unbounded. Truncating the Y_i's before computing the empirical coefficient enables bounding their sensitivity over adjacent data sets and informing our choice of random noise distribution.

We fix the number of terms in the estimator at K, and let $\boldsymbol{\varphi}$ denote the vector valued function $\boldsymbol{\varphi}:\mathbb{R}\to\mathbb{R}^K$, $\boldsymbol{\varphi}(x) = (\varphi_1(x), \varphi_2(x), \cdots, \varphi_K(x))$. With the aforementioned truncation, the empirical Fourier coefficients with truncation are given by

$$\frac{1}{n}\sum_{i=1}^n Y_i\mathbb{1}(|Y_i|\le T)\cdot\boldsymbol{\varphi}(X_i).$$

Over two adjacent data sets D, D' with symmetric difference $\{(Y_i, X_i), (Y_i', X_i')\}$, their empirical coefficients differ by

$$\boldsymbol{\Delta}_{D,D'} = \frac{1}{n}\left(Y_i\mathbb{1}(|Y_i|\le T)\cdot\boldsymbol{\varphi}(X_i) - Y_i'\mathbb{1}(|Y_i'|\le T)\cdot\boldsymbol{\varphi}(X_i')\right)\in\mathbb{R}^K.$$

Although the truncation of Y and the boundedness of $\boldsymbol{\varphi}$ imply straightforward ℓ_p-norms bounds of $\boldsymbol{\Delta}_{D,D'}$ which scales with the dimension K, [24] observes that noise addition according to the K-norm mechanism [26] (the "K" in "K-norm" is unrelated to the dimension K of the estimator) can achieve much improved accuracy compared to the usual Laplace or Gaussian mechanisms based on ℓ_1 or ℓ_2 sensitivities.

Specifically, observe that$\boldsymbol{\Delta}_{D,D'}$ belongs to a scaled version of the set

$$\mathcal{S} = \mathrm{conv}\{\pm\boldsymbol{\varphi}(x), x\in[0,1]\}\subseteq\mathbb{R}^K,$$

where $\mathrm{conv}\{\cdot\}$ refers to the convex hull. The set $\mathcal{S}$, known as the Universal Caratheodory orbitope [24, 46], is convex, compact, centro-symmetric and has an non-empty interior, and therefore induces a norm on $\mathbb{R}^k$: $\|\boldsymbol{v}\|_{\mathcal{S}} = \inf\{r>0 : x\in r\cdot\mathcal{S}\}$. It then follows that $\|\boldsymbol{\Delta}_{D,D'}\|_{\mathcal{S}}\le 2T/n$ for any adjacent D, D', and the K-norm mechanism [26] implies that $(\varepsilon, 0)$-differential privacy is achieved by

$$\tilde{\boldsymbol{\theta}}_{K,T} = \frac{1}{n}\sum_{i=1}^n Y_i\mathbb{1}(|Y_i|\le T)\cdot\boldsymbol{\varphi}(X_i) + \boldsymbol{w},$$

where $\boldsymbol{w}$ is drawn from the density $g_{\boldsymbol{w}}(\boldsymbol{t})\propto\exp\left(-\frac{2n\varepsilon}{T}\|\boldsymbol{t}\|_{\mathcal{S}}\right)$. While sampling from this unconventional distribution is highly non-trivial, Section 4.4.4. of [24] proposes an efficient sampling algorithm, and we focus on the statistical accuracy of $\tilde{\boldsymbol{\theta}}_{K,T}$ and the associated function estimator

$$\tilde{f}_{K,T} = \sum_{j=1}^K\left(\tilde{\boldsymbol{\theta}}_{K,T}\right)_j\varphi_j(x). \tag{6.8}$$

THEOREM 6.2. *If $T = 4\sigma\sqrt{\log n}$ and $\sigma^2\le c_0$ for some absolute constant c_0, and $K = c_1\min(n^{-\frac{1}{2\alpha+1}}, (n\varepsilon)^{-\frac{1}{\alpha+1}})$ for some absolute constant $c_1>0$, then*

$$\sup_{f\in\tilde{W}(\alpha,C)}\mathbb{E}\left[\int_0^1(\tilde{f}_{K,T}(x) - f(x))^2\mathrm{d}x\right]\lesssim n^{-\frac{2\alpha}{2\alpha+1}} + (n\varepsilon)^{-\frac{2\alpha}{\alpha+1}}\cdot\log n. \tag{6.9}$$

Theorem 6.2 is proved in Section D.4. The risk upper bound (6.9) matches the privacy-constrained minimax lower bound (6.7), up to a logarithmic factor in n.

7. Discussion. In the present paper, we introduced a new technique, the score attack, for lower bounding the privacy-constrained minimax risk in differentially private learning. We demonstrated the effectiveness of this novel technique through a range of examples covering classical statistical problems, a ranking problem, a high-dimensional sparse problem, and a nonparametric problem. In each example, we were able to obtain an optimal minimax lower bound (up to at most a logarithmic factor) by defining an appropriate form of score attack and carrying out the analysis introduced in Section 2.2. These results suggest that the score attack technique could be useful for characterizing the necessary cost of differential privacy in other statistical problems.

The logarithmic gaps between upper and lower bounds in this paper may warrant further investigation. Some of them appear to be artifacts of truncating unbounded data or compositing iterative steps, and can potentially be eliminated by constructing more efficient algorithms. Some other gaps related to the privacy parameter δ may suggest interesting questions about the inherent difficulty of parameter estimation with differential privacy, for example, whether, or when, the "approximate", (ε, δ)-differential privacy is less costly in statistical inference than the "pure", $(\varepsilon, 0)$-differential privacy in all statistical problems.

At present, the score attack method has only been applied to the ℓ_2-loss, but it would be useful to extend it to other loss functions for statistical problems, such as model selection, where the ℓ_2-distance may not be the most appropriate metric. Additionally, it would be interesting to explore whether the score attack method can be generalized to interval estimation and testing problems, as many lower bound methods in non-private statistical theory are unified across point estimation, confidence intervals, and hypothesis testing.

8. Proofs. We prove Theorem 2.1 in this section. For reasons of space, the proofs of other results and technical lemmas are given in the supplement [15].

8.1. *Proof of Theorem 2.1.*

PROOF. For soundness, we note that $\boldsymbol{x}_i$ and $M(\boldsymbol{X}_i')$ are independent, and therefore

$$\mathbb{E}\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X}_i')) = \mathbb{E}\langle M(\boldsymbol{X}_i') - \boldsymbol{\theta}, S_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\rangle = \langle \mathbb{E}M(\boldsymbol{X}_i') - \boldsymbol{\theta}, \mathbb{E}S_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\rangle = \mathbf{0}.$$

The last equality is true by the property of the score that $\mathbb{E}S_{\boldsymbol{\theta}}(\boldsymbol{z}) = \mathbf{0}$ for any $\boldsymbol{z} \sim f_{\boldsymbol{\theta}}$. As to the first absolute moment, we apply Jensen's inequality,

$$\mathbb{E}|\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X}_i'))| \le \sqrt{\mathbb{E}\langle M(\boldsymbol{X}_i') - \boldsymbol{\theta}, S_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\rangle^2}$$

$$\le \sqrt{\mathbb{E}(M(\boldsymbol{X}_i') - \boldsymbol{\theta})^\top (\mathrm{Var} S_{\boldsymbol{\theta}}(\boldsymbol{x}_i))(M(\boldsymbol{X}_i') - \boldsymbol{\theta})} \le \sqrt{\mathbb{E}\|M(\boldsymbol{X}) - \boldsymbol{\theta}\|_2^2}\sqrt{\lambda_{\max}(\mathcal{I}(\boldsymbol{\theta}))}.$$

For completeness, we first simplify

$$\sum_{i\in[n]} \mathbb{E}\mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X})) = \mathbb{E}\Big\langle M(\boldsymbol{X}) - \boldsymbol{\theta}, \sum_{i\in[n]} S_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\Big\rangle = \mathbb{E}\Big\langle M(\boldsymbol{X}), \sum_{i\in[n]} S_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\Big\rangle.$$

By the definition of score and that $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ are i.i.d., $\sum_{i\in[n]} S_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = S_{\boldsymbol{\theta}}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) = S_{\boldsymbol{\theta}}(\boldsymbol{X})$. It follows that

$$\mathbb{E}\Big\langle M(\boldsymbol{X}), \sum_{i\in[n]} S_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\Big\rangle = \mathbb{E}\Big\langle M(\boldsymbol{X}), S_{\boldsymbol{\theta}}(\boldsymbol{X})\Big\rangle = \sum_{j\in[d]} \mathbb{E}\left[M(\boldsymbol{X})_j \frac{\partial}{\partial\theta_j}\log f_{\boldsymbol{\theta}}(\boldsymbol{X})\right].$$

For each term in the right-side summation, one may exchange differentiation and integration thanks to the regularity conditions on $f_{\boldsymbol{\theta}}$, and therefore

$$\begin{aligned}\mathbb{E}\left[M(\boldsymbol{X})_j \frac{\partial}{\partial\theta_j}\log f_{\boldsymbol{\theta}}(\boldsymbol{X})\right] &= \mathbb{E}\left[M(\boldsymbol{X})_j (f_{\boldsymbol{\theta}}(\boldsymbol{X}))^{-1}\frac{\partial}{\partial\theta_j} f_{\boldsymbol{\theta}}(\boldsymbol{X})\right] \\ &= \frac{\partial}{\partial\theta_j}\mathbb{E}\left[M(\boldsymbol{X})_j (f_{\boldsymbol{\theta}}(\boldsymbol{X}))^{-1} f_{\boldsymbol{\theta}}(\boldsymbol{X})\right] = \frac{\partial}{\partial\theta_j}\mathbb{E}M(\boldsymbol{X})_j.\end{aligned}$$

□

8.1.1. *Proof of Lemma 2.1.*

PROOF. Let $A_i := \mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X}))$, $A_i' := \mathcal{A}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, M(\boldsymbol{X}_i'))$, and let $Z^+ = \max(Z, 0)$ and $Z^- = -\min(Z, 0)$ denote the positive and negative parts of a random variables Z respectively. We have

$$\mathbb{E}A_i = \mathbb{E}A_i^+ - \mathbb{E}A_i^- = \int_0^\infty \mathbb{P}(A_i^+ > t)\mathrm{d}t - \int_0^\infty \mathbb{P}(A_i^- > t)\mathrm{d}t.$$

For the positive part, if $0 < T < \infty$ and $0 < \varepsilon < 1$, we have

$$\begin{aligned}\int_0^\infty \mathbb{P}(A_i^+ > t)\mathrm{d}t &= \int_0^T \mathbb{P}(A_i^+ > t)\mathrm{d}t + \int_T^\infty \mathbb{P}(A_i^+ > t)\mathrm{d}t \\ &\le \int_0^T \left(e^{\varepsilon}\mathbb{P}(A_i^+ > t) + \delta\right)\mathrm{d}t + \int_T^\infty \mathbb{P}(A_i^+ > t)\mathrm{d}t \\ &\le \int_0^\infty \mathbb{P}(A_i'^+ > t)\mathrm{d}t + 2\varepsilon\int_0^\infty \mathbb{P}(A_i'^+ > t)\mathrm{d}t + \delta T + \int_T^\infty \mathbb{P}(|A_i| > t)\mathrm{d}t.\end{aligned}$$

Similarly for the negative part,

$$\begin{aligned}\int_0^\infty \mathbb{P}(A_i^- > t)\mathrm{d}t &= \int_0^T \mathbb{P}(A_i^- > t)\mathrm{d}t + \int_T^\infty \mathbb{P}(A_i^- > t)\mathrm{d}t \\ &\ge \int_0^T \left(e^{-\varepsilon}\mathbb{P}(A_i'^- > t) - \delta\right)\mathrm{d}t + \int_T^\infty \mathbb{P}(A_i^- > t)\mathrm{d}t \\ &\ge \int_0^T \mathbb{P}(A_i'^- > t)\mathrm{d}t - 2\varepsilon\int_0^T \mathbb{P}(A_i'^- > t) - \delta T + \int_T^\infty \mathbb{P}(A_i^- > t)\mathrm{d}t \\ &\ge \int_0^\infty \mathbb{P}(A_i'^- > t)\mathrm{d}t - 2\varepsilon\int_0^\infty \mathbb{P}(A_i'^- > t) - \delta T.\end{aligned}$$

It then follows that

$$\begin{aligned}\mathbb{E}A_i &\le \int_0^\infty \mathbb{P}(A_i'^+ > t)\mathrm{d}t - \int_0^\infty \mathbb{P}(A_i'^- > t)\mathrm{d}t + 2\varepsilon\int_0^\infty \mathbb{P}(|A_i'| > t)\mathrm{d}t + 2\delta T + \int_T^\infty \mathbb{P}(|A_i| > t)\mathrm{d}t \\ &= \mathbb{E}A_i' + 2\varepsilon\mathbb{E}|A_i| + 2\delta T + \int_T^\infty \mathbb{P}(|A_i| > t)\mathrm{d}t.\end{aligned}$$

The proof is now complete by soundness (2.4). □

8.1.2. *Proof of Lemma 2.2.*

PROOF. For each $j \in [d]$, by Lemma 2.1, we have

$$\mathbb{E}_{\pi_j}\left(\frac{\partial}{\partial\theta_j}g_j(\boldsymbol{\theta})\right) = \mathbb{E}_{\pi_j}\left(\frac{\partial}{\partial\theta_j}\mathbb{E}[g_j(\boldsymbol{\theta})|\theta_j]\right) = \mathbb{E}_{\pi_j}\left[\frac{-\mathbb{E}[g_j(\boldsymbol{\theta})|\theta_j]\pi_j'(\theta_j)}{\pi_j(\theta_j)}\right]$$

Because $|g_j(\boldsymbol{\theta}) - \theta_j| \le \|g(\boldsymbol{\theta}) - \boldsymbol{\theta}\|_2 \le \mathbb{E}_{\boldsymbol{X}|\boldsymbol{\theta}}\|M(\boldsymbol{X}) - \boldsymbol{\theta}\|_2$ for every $\boldsymbol{\theta} \in \Theta$, we have

$$\begin{aligned}\mathbb{E}_{\pi_j}\left[\frac{-\mathbb{E}[g(\boldsymbol{\theta})|\theta_j]\pi_j'(\theta_j)}{\pi_j(\theta_j)}\right] &\ge \mathbb{E}_{\pi_j}\left[\frac{-\theta_j\pi_j'(\theta_j)}{\pi_j(\theta_j)}\right] - \mathbb{E}_{\pi_j}\left[\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\theta}}\|M(\boldsymbol{X}) - \boldsymbol{\theta}\|_2 \cdot \left|\frac{\pi_j'(\theta_j)}{\pi_j(\theta_j)}\right|\right] \\ &\ge \mathbb{E}_{\pi_j}\left[\frac{-\theta_j\pi_j'(\theta_j)}{\pi_j(\theta_j)}\right] - \sqrt{\mathbb{E}_{\pi_j}\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\theta}}\|M(\boldsymbol{X}) - \boldsymbol{\theta}\|_2^2}\sqrt{\mathbb{E}_{\pi_j}\left[\left(\frac{\pi_j'(\theta_j)}{\pi_j(\theta_j)}\right)^2\right]}.\end{aligned}$$

So we have obtained

$$\mathbb{E}_{\pi_j}\left(\frac{\partial}{\partial\theta_j}g_j(\boldsymbol{\theta})\right) \ge \mathbb{E}_{\pi_j}\left[\frac{-\theta_j\pi_j'(\theta_j)}{\pi_j(\theta_j)}\right] - \sqrt{\mathbb{E}_{\pi_j}\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\theta}}\|M(\boldsymbol{X}) - \boldsymbol{\theta}\|_2^2}\sqrt{\mathbb{E}_{\pi_j}\left[\left(\frac{\pi_j'(\theta_j)}{\pi_j(\theta_j)}\right)^2\right]}.$$

Now we take expectation over $\boldsymbol{\pi}(\boldsymbol{\theta})/\pi_j(\theta_j)$ and sum over $j \in [d]$ to complete the proof. □

SUPPLEMENTARY MATERIAL

Supplement to "Score Attack: A Lower Bound Technique for Differentially Private Learning".

The supplement includes the proofs that were omitted from Sections 3 to 6. These proofs cover the minimax optimality results under differential privacy for classical generalized linear models, the Bradley-Terry-Luce ranking model, high-dimensional generalized linear models, and the nonparametric regression model. Additionally, the supplement provides proofs of the technical lemmas used in the main results.

REFERENCES

[1] ABADI, M., CHU, A., GOODFELLOW, I., MCMAHAN, H. B., MIRONOV, I., TALWAR, K. and ZHANG, L. (2016). Deep learning with differential privacy. In *ACM CCS 2016* 308–318. ACM.

[2] ACHARYA, J., SUN, Z. and ZHANG, H. (2018). Differentially private testing of identity and closeness of discrete distributions. In *Advances in Neural Information Processing Systems* 6878–6891.

[3] ACHARYA, J., SUN, Z. and ZHANG, H. (2021). Differentially private Assouad, Fano, and Le Cam. In *Algorithmic Learning Theory* 48–78. PMLR.

[4] AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2010). Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems* 37–45.

[5] AVELLA-MEDINA, M. (2021). Privacy-preserving parametric inference: a case for robust statistics. *Journal of the American Statistical Association* **116** 969–983.

[6] AVELLA-MEDINA, M., BRADSHAW, C. and LOH, P.-L. (2021). Differentially private inference via noisy optimization. *arXiv preprint arXiv:2103.11003*.

[7] BAFNA, M. and ULLMAN, J. (2017). The price of selection in differential privacy. *arXiv preprint arXiv:1702.02970*.

[8] BALAKRISHNAN, S. and CHOPRA, S. (2012). Two of a kind or the ratings game? Adaptive pairwise preferences and latent factor models. *Frontiers of Computer Science* **6** 197–208.

[9] BARBER, R. F. and DUCHI, J. C. (2014). Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*.

[10] BASSILY, R., FELDMAN, V., TALWAR, K. and THAKURTA, A. G. (2019). Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems* 11282–11291.

[11] BASSILY, R., SMITH, A. and THAKURTA, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *FOCS 2014* 464–473. IEEE.

[12] BLUMENSATH, T. and DAVIES, M. E. (2009). Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis* **27** 265–274.

[13] BUN, M., ULLMAN, J. and VADHAN, S. (2014). Fingerprinting codes and the price of approximate differential privacy. In *STOC 2014* 1–10. ACM.

[14] CAI, T. T., WANG, Y. and ZHANG, L. (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics* **49** 2825–2850.

[15] CAI, T. T., WANG, Y. and ZHANG, L. (2023). Supplement to "Score attack: a lower bound technique for differentially private learning ".

[16] CHAUDHURI, K. and MONTELEONI, C. (2009). Privacy-preserving logistic regression. In *Advances in neural information processing systems* 289–296.

[17] CHAUDHURI, K., MONTELEONI, C. and SARWATE, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research* **12** 1069–1109.

[18] DWORK, C., KENTHAPADI, K., MCSHERRY, F., MIRONOV, I. and NAOR, M. (2006). Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques* 486–503. Springer.

[19] DWORK, C., MCSHERRY, F., NISSIM, K. and SMITH, A. (2006). Calibrating noise to sensitivity in private data analysis. In *TCC 2006* 265–284. Springer.

[20] DWORK, C. and ROTH, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* **9** 211–407.

[21] DWORK, C., SMITH, A., STEINKE, T. and ULLMAN, J. (2017). Exposed! a survey of attacks on private data. *Annu. Rev. Stat. Appl.* **4** 61–84.

[22] DWORK, C., SMITH, A., STEINKE, T., ULLMAN, J. and VADHAN, S. (2015). Robust traceability from trace amounts. In *FOCS 2015* 650–669. IEEE.

[23] DWORK, C., SU, W. J. and ZHANG, L. (2018). Differentially Private False Discovery Rate Control. *arXiv preprint arXiv:1807.04209*.

[24] HALL, R. (2013). New Statistical Applications for Differential Privacy, PhD thesis, Carnegie Mellon University.

[25] HALL, R., RINALDO, A. and WASSERMAN, L. (2013). Differential privacy for functions and functional data. *The Journal of Machine Learning Research* **14** 703–727.

[26] HARDT, M. and TALWAR, K. (2010). On the geometry of differential privacy. In *Proceedings of the forty-second ACM symposium on Theory of computing* 705–714.

[27] HAWES, M. B. (2020). Implementing Differential Privacy: Seven Lessons From the 2020 United States Census. *Harvard Data Science Review* **2**. https://hdsr.mitpress.mit.edu/pub/dgg03vo6. https://doi.org/10.1162/99608f92.353c6f99

[28] HAY, M., ELAGINA, L. and MIKLAU, G. (2017). Differentially private rank aggregation. In *Proceedings of the 2017 SIAM International Conference on Data Mining* 669–677. SIAM.

[29] HELDSINGER, S. and HUMPHRY, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher* **37** 1–19.

[30] JAIN, P., TEWARI, A. and KAR, P. (2014). On iterative hard thresholding methods for high-dimensional m-estimation. In *NeurIPS 2014* 685–693.

[31] KAMATH, G., LI, J., SINGHAL, V. and ULLMAN, J. (2018). Privately learning high-dimensional distributions. *arXiv preprint arXiv:1805.00216*.

[32] KAMATH, G., SINGHAL, V. and ULLMAN, J. (2020). Private mean estimation of heavy-tailed distributions. *arXiv preprint arXiv:2002.09464*.

[33] KARWA, V. and VADHAN, S. (2017). Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*.

[34] KIFER, D., SMITH, A. and THAKURTA, A. (2012). Private convex empirical risk minimization and high-dimensional regression. In *COLT 2012* 25.1-25.40.

[35] LE CAM, L. (2012). *Asymptotic methods in statistical decision theory*. Springer Science & Business Media.

[36] LEE, S. K., GRESELE, L., PARK, M. and MUANDET, K. (2019). Private Causal Inference using Propensity Scores. *arXiv preprint arXiv:1905.12592*.

[37] LEE, S. K., GRESELE, L., PARK, M. and MUANDET, K. (2019). Privacy-Preserving Causal Inference via Inverse Probability Weighting. *arXiv preprint arXiv:1905.12592*.

[38] LEI, J. (2011). Differentially private m-estimators. In *NeurIPS 2011* 361–369.

[39] LI, Z., LIU, A., XIA, L., CAO, Y. and WANG, H. (2022). Differentially Private Condorcet Voting. *arXiv preprint arXiv:2206.13081*.

[40] LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research* **16** 559–616.

[41] MASAROTTO, G. and VARIN, C. (2012). The ranking lasso and its application to sport tournaments. *The Annals of Applied Statistics* **6** 1949–1970.

[42] MURAKONDA, S. K., SHOKRI, R. and THEODORAKOPOULOS, G. (2019). Ultimate Power of Inference Attacks: Privacy Risks of High-Dimensional Models. *arXiv preprint arXiv:1905.12774*.

[43] NEGAHBAN, S., OH, S. and SHAH, D. (2017). Rank centrality: Ranking from pairwise comparisons. *Operations Research* **65** 266–287.

[44] NEGAHBAN, S., YU, B., WAINWRIGHT, M. J. and RAVIKUMAR, P. K. (2009). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Advances in neural information processing systems* 1348–1356.

[45] PHAN, N., WANG, Y., WU, X. and DOU, D. (2016). Differential privacy preservation for deep auto-encoders: an application of human behavior prediction. In *Aaai* **16** 1309–1316.

[46] SANYAL, R., SOTTILE, F. and STURMFELS, B. (2011). Orbitopes. *Mathematika* **57** 275–314.

[47] SHAH, N., BALAKRISHNAN, S., BRADLEY, J., PAREKH, A., RAMCHANDRAN, K. and WAINWRIGHT, M. (2015). Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *Artificial Intelligence and Statistics* 856–865. PMLR.

[48] SHANG, S., WANG, T., CUFF, P. and KULKARNI, S. (2014). The application of differential privacy for rank aggregation: Privacy and accuracy. In *17th International Conference on Information Fusion (FUSION)* 1–7. IEEE.

[49] SHOKRI, R., STRONATI, M., SONG, C. and SHMATIKOV, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)* 3–18. IEEE.

[50] SONG, B., LAN, Q., LI, Y. and LI, G. (2022). Distributed Differentially Private Ranking Aggregation. *arXiv preprint arXiv:2202.03388*.

[51] SONG, S., STEINKE, T., THAKKAR, O. and THAKURTA, A. (2021). Evading the curse of dimensionality in unconstrained private glms. In *International Conference on Artificial Intelligence and Statistics* 2638–2646. PMLR.

[52] STEIN, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California.

[53] STEIN, C., DIACONIS, P., HOLMES, S. and REINERT, G. (2004). Use of exchangeable pairs in the analysis of simulations. In *Stein's Method* 1–25. Institute of Mathematical Statistics.

[54] STEINKE, T. and ULLMAN, J. (2017). Tight lower bounds for differentially private selection. In *FOCS 2017* 552–563. IEEE.

[55] STEINKE, T. and ULLMAN, J. (2017). Between Pure and Approximate Differential Privacy. *Journal of Privacy and Confidentiality* **7**.

[56] TARDOS, G. (2008). Optimal probabilistic fingerprint codes. *Journal of the ACM (JACM)* **55** 10.

[57] TSYBAKOV, A. B. (2009). *Introduction to nonparametric estimation*. Springer-Verlag.

[58] WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge University Press.

[59] WANG, Y.-X. (2018). Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1803.02596*.

[60] WASSERMAN, L. and ZHOU, S. (2010). A statistical framework for differential privacy. *J. Am. Stat. Assoc.* **105** 375–389.

[61] XU, S., SUN, W. W. and CHENG, G. (2023). Ranking Differential Privacy. *arXiv preprint arXiv:2301.00841*.

[62] ZHANG, H., KAMATH, G., KULKARNI, J. and WU, S. (2020). Privately learning Markov random fields. In *International Conference on Machine Learning* 11129–11140. PMLR.