Statistically Optimal Mechanism Design for Differential Privacy

Anish Jayant
Department of Computer Science
USC
jayant@usc.edu

Spencer Cockerell
Department of Computer Science
USC
scockere@usc.edu

Abstract

In this work, we survey the theoretical analysis of the *inverse sensitivity mechanism* and *instance optimality* proposed by [Asi and Duchi, 2020] and applied in [Hopkins et al., 2022]. We present an information-theoretic and instance-dependent lower bound on notions of loss for private estimators and show that inverse sensitivity mechanisms are indeed optimal. We'll also study a particularly interesting application of this mechanism through statistical estimation of Gaussians in high-dimension.

1 Introduction

As machine learning algorithms become ubiquitous and gain access to increasingly sensitive data, it seems natural to introduce the desiderata that these algorithms be *private*. There are several intuitive interpretations of this requirement, one of the most common being that, while the algorithm learns something (potentially) useful about the population, it must learn very little about any given individual. More precisely, the administrator must know no more about any individual at the end of the analysis than she knew at the beginning. The formalization of these requirements compares the output of the algorithm on two possible datasets with Hamming distance 1 (i.e. differing on exactly one member) and stipulates that the outcomes must be similar.

Definition 1 (Differential Privacy, [Dwork et al., 2006])**.** *Consider an algorithm $A : \mathcal{X}^n \to \mathcal{T}$, which operates over datasets of size n of members of $\mathcal{X}$ and outputs a result in $\mathcal{T}$ [1]. If, for every $x, y \in \mathcal{X}^n$ such that $d(x, y) = 1$, and every subset $\mathcal{O} \subseteq \mathcal{T}$,*

$$\Pr[A(x) \in \mathcal{O}] \leq e^{\varepsilon} \Pr[A(y) \in \mathcal{O}] + \delta$$

for some $\varepsilon \geq 0$, $\delta \in [0, 1]$, we say A is (ε, δ)-differentially private.

Algorithms that satisfy this worst-case definition of differential privacy admit many strong properties, shown in post-processing and composability lemmas, that we'd expect a truly *private* method to satisfy.

Lemma 2 (Post-Processing, [Dwork et al., 2006])**.** *Let $A : \mathcal{X} \to \mathcal{T}$ be a (ε, δ)-d.p. algorithm, and $f : \mathcal{T} \to \mathcal{T}'$ be any randomized mapping. Then, $f \circ A : \mathcal{X} \to \mathcal{T}'$ is (ε, δ)-d.p.*

Lemma 3 (Composability, [Dwork et al., 2006])**.** *Suppose $A_1 : \mathcal{X} \to \mathcal{T}_1$ and $A_2 : \mathcal{X} \to \mathcal{T}_2$ are $(\varepsilon_1, 0)$ and $(\varepsilon_2, 0)$ d.p. respectively. Then, $A_3 : \mathcal{X} \to \mathcal{T}_1 \times \mathcal{T}_2$ given by $A_3(x) = (A_1(x), A_2(x))$ is $(\varepsilon_1 + \varepsilon_2, 0)$-d.p.*

[1] For the purposes of this review, we'll assume $(\mathcal{T}, \|\cdot\|)$ is a well-defined metric space. Note that $(\mathcal{X}^n, d)$ is also a metric space over n-length strings.

Course project for CSCI699.

1.1 Folklore Differential Privacy Constructions

Over the years, researchers have found increasingly clever ways to construct differential privacy in various situations, of which we'll present just three. The first two, *privacy through Laplacian noise* and *exponential mechanism* are simple in statement and proof, but lay the groundwork for many more modern ideas. The third, *smooth sensitivity*, is a canonical example of 'optimization', which strictly improves on an earlier construction, but exhibits a more complicated construction.

Definition 4 (Privacy by Noise, [Dwork et al., 2006])**.** *Given any query function $f : \mathcal{X} \to \mathbb{R}^k$, calculate its global sensitivity:*

$$GF_f \triangleq \sup_{x,x':d(x,x')=1} \|f(x') - f(x)\|. \tag{1}$$

The Laplacian mechanism, defined as

$$M_{Lap}(x, f(\cdot), \varepsilon) = f(x) + \frac{GF_f}{\varepsilon} Lap(1).^{2}$$

is $(\varepsilon, 0)$-d.p.

Intuitively, adding noise to the output of $f(x)$ obscures peculiarities that an adversary could attack. Additionally, adding noise proportionally to the *global sensitivity* makes sense if we think about sensitivity as a measure of discrimination abilities (how far apart can we send two similar points x, x'). However, there is a very clear degradation of information associated with adding noise, which causes stronger privacy guarantees to be put at odds with accuracy. The exponential mechanism avoids adding noise:

Definition 5 (Exponential Mechanism, [McSherry and Talwar, 2007])**.** *Let $h : \mathcal{X}^n \times \mathcal{T} \to \mathbb{R}_+$ be 1-Lipschitz with respect to Hamming distance on $\mathcal{X}$ (that is, $|h(x,t) - h(x',t)| \leq 1$ for adjacent x, x'). Let μ be some measure over $\mathcal{T}$. Then, the mechanism L_{exp} which samples $X \sim \pi$ constructed as*

$$d\pi(t) = \frac{\exp(-h(x,t)\varepsilon/2) \cdot d\mu(t)}{\int_{s\in\mathcal{T}} \exp(-h(x,s)\varepsilon/2) \cdot d\mu(s)}$$

satisfies ε-d.p.

Although instances of the exponential mechanism tend to be computationally intractable (think about L_{exp} constructing a distribution π and then randomly sampling from it), they are useful in 'private optimization'. It can often be shown that we select r.v.'s from π that minimize h with high probability while remaining differentially private. We'll see a particularly strong application of this in Section 2.6.

But can we do better? Perhaps we feel that the noise added in Definition 4 is too extreme, and should be tailored to our dataset x. A notion for this is *local sensitivity*, calculated as $\text{LS}_f(x) \triangleq \sup_{x':d(x,x')=1} \|f(x') - f(x)\|$. However, just calculating the local sensitivity of a given sample x cannot be done privately (intuitively, we reveal information about x revealing how different it is from its neighbors x'). So, we might seek a 'smoothened' upper bound $S(x) \geq \text{LS}_f(x)$ that is β-d.p. It is shown in [Nissim et al., 2007] that the smallest such S is defined as follows.

Definition 6 (Smooth Sensitivity, [Nissim et al., 2007])**.** *For a function f, the β-smooth sensitivity is*

$$S^*_{f,\beta}(x) \triangleq \max_{y\in D^n} LS_f(x) \exp(-\beta d(x,y))$$

As a sanity check, recall that the constant function $S(x) = \text{GF}_f = \sup_x \text{LS}_f(x)$ satisfies the conditions as well, so we have $S^*_{f,\beta} \leq \text{GF}_f$ as a (conservative) upper-bound, just by construction. It turns out that a mechanism $M_{smooth,\beta}(x, f(\cdot), \varepsilon)$ that adds noise according to

$$M_{smooth,\beta} = f(x) + \frac{S^*_{f,\beta}}{\varepsilon/2}\text{Lap}(1)$$

can be shown to be ε-d.p.! Additionally, as we add less noise than the Laplace mechanism, we should expect less corruption and thus a more accurate result for most $x \in \mathcal{X}$.

Is there a limit to how far can we go? Through the rest of this survey, we'll look into the horizon of differentially private construction, and touch on the following pertinent research questions:

- Is there an optimal (accuracy) lower-bound that all DP mechanisms obey?
- If there is a limit, what method(s) achieve it? What does it mean to be optimal?

2 Survey of Past Works

2.1 Problem Setup

We'll begin by introducing technical definitions and lemmas to understand statistical optimality in the context of [Asi and Duchi, 2020]. For a learning problem with ground truth $f : \mathcal{X} \to \mathcal{T}$, we define a loss function $L : \mathcal{T} \times \mathcal{T} \to \mathbb{R}_+$ to judge the responses of some (randomized) learner M. In class, we extensively studied the *(expected) loss* of our learner through $\mathbb{E}\left[L(M(x), f(x))\right]$. For the sake of simplicity, almost all results will concern this definition of loss, but we'll also mention the idea of *local minimax risk*

$$\mathcal{R}(x, L, \mathcal{M}) \triangleq \sup_{x' \in \mathcal{X}^n} \inf_{M \in \mathcal{M}} \max_{\tilde{x} \in \{x, x'\}} \mathbb{E}\left[L(M(\tilde{x}), f(\tilde{x}))\right]. \tag{2}$$

This construction will allow us to consider a particular class of mechanisms $\mathcal{M}$, such as differentially private and unbiased estimators (Definition 1, 7). Most importantly, it offers us a worst-case perspective while still being data-dependent (unlike, say GF_f). Some particularly difficult theorems involving expected loss in [Asi and Duchi, 2020] are demonstrated using minimax risk.

As is in classical statistics, we'll see that optimal mechanisms are often *unbiased*.

Definition 7 (Unbiased, Definition 1.4, folklore)**.** *A randomized algorithm $M : \mathcal{X}^n \to \mathcal{T}$ is L-unbiased if, for any $x \in \mathcal{X}^n$ and $t \in \mathcal{T}$, $\mathbb{E}\left[L(M(x), f(x))\right] \leq \mathbb{E}\left[L(M(x), t)\right]$. That is, the algorithm outperforms any constant predictor in the expected loss paradigm.*

Intuitively, Definition 7 demands that M outperform any constant predictor in the output space for all datasets. If we take $L(s, t) = (s - t)^2$, for example, we recover the $\mathbb{E}[M(x)] = f(x)$, which may be familiar from classical notions.

With these definitions in place, we can set up two notions of optimality:

Definition 8 (Local minimax optimal, Definition 1.3, [Asi and Duchi, 2020])**.** *A mechanism $M : \mathcal{X}^n \to \mathcal{T}$ is local minimax optimal w.r.t. $\mathcal{M}$ if there exists $C < \infty$ such that*

$$\mathbb{E}\left[L(M(x), f(x))\right] \leq C \cdot \mathcal{R}(x, L, \mathcal{M}).$$

Definition 9 (Local unbiased-private optimal, Definition 1.5, [Asi and Duchi, 2020])**.** *Let $C \geq 1$. A randomized mechanism M is C-optimal against L-unbiased mechanisms if M is $C\varepsilon$-d.p., and for any ε-d.p., L-unbiased mechanism $M_{unb} : \mathcal{X}^n \to \mathcal{T}$, and dataset $x \in \mathcal{X}^n$*

$$\mathbb{E}\left[L(M(x), f(x))\right] \leq \mathbb{E}\left[L(M_{unb}(x), f(x))\right]$$

So, we might expect a competitive mechanism to be, say 'local minimax optimal', or at least competitive with unbiased opponents by being C-optimal against L-unbiased mechanisms.

2.1.1 Cramér-Rao Lower Bound

To motivate the discussion that follows, we review a parallel information lower-bound from classical statistics that demonstrates optimality of a parameter estimate. Consider an estimator $\hat{\theta}$ which aims to estimate quantity θ_0 given access to a finite sample set. The mean-squared error can be decomposed,

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{var}(\hat{\theta}) + (\mathbb{E}(\hat{\theta} - \theta_0))^2,$$

where the first term is known as variance and the second as bias. If we only consider unbiased estimators, the *efficiency* of $\hat{\theta}$ is determined entirely through its variance – the smaller this quantity the better $\hat{\theta}$ does in terms of error.

Lemma 10 (Cramér-Rao)**.** *Let $X_1, ..., X_n$ be i.i.d. with density function $f(x|\theta_0)$, and $\hat{\theta}$ be an unbiased estimate of θ_0. Then,*

$$\mathit{var}(\hat{\theta}) \geq \frac{1}{nI(\theta_0)}$$

where I is the Fisher information.

So, from a statistical sense, it makes sense to call $\hat{\theta}$ information-theoretically optimal if it tightly satisfies this lower-bound; $\mathrm{var}(\hat{\theta}) = 1/nI(\theta)$.

2.2 Loss Lower Bounds in Cramér-Rao fashion

We'll introduce this section with some instance-dependent quantities. One that the reader might already be familiar with is *inverse local sensitivity*, which essentially reverses the phrasing of local sensitivity.

Definition 11 (Inverse (local) sensitivity, folklore)**.** *Consider $f : \mathcal{X}^n \to \mathcal{T}$ and a sampled dataset x. The inverse local sensitivity of x with respect to label $t \in \mathcal{T}$ is*

$$len_f(x;t) = \inf_{x'} \{d(x,x') | f(x') = t\}, \tag{3}$$

First, it is easy to see that inverse sensitivity is instance (or data) dependent due to its construction. Informally, $\text{len}_f(x;t)$ measures the corruption required of a true dataset x such that a query f returns value t. So, values of t that characterize $f(x)$ better will tend to have smaller $\text{len}_f(x;t)$ than more obscure values. A slightly deeper analysis notes that, if x is in a stable space (i.e. datasets close in Hamming distance share the same outcome under f), then we have $\text{len}_f(x;t)$ is large, so this metric detects how robust (or inversely sensitive) f is to changes around x.

With these observations in mind, we are ready to show our first instance-dependent risk lower-bound for private/unbiased estimators. We start with the reduced learning setting where the ground truth $f : \mathcal{X}^n \to \mathcal{T}$ has a finite output $|\mathcal{T}| < \infty$ and we are judged by 0-1 loss $\ell_{0-1}(s,t) = \mathbf{1}(s \neq t)$.

Theorem 12 (Lower bounds for 0-1 loss[3], [Asi and Duchi, 2020])**.** *If M is ε-d.p., then*

$$\inf_{x \in \mathcal{X}^n} \mathbb{P}(M(x) = f(x)) \leq \inf_{x \in \mathcal{X}^n} \frac{1}{\sum_{t \in \mathcal{T}} e^{-len_f(x;t)\varepsilon}}.$$

Furthermore, if M is also ℓ_{0-1}-unbiased, then

$$\mathbb{P}(M(x) = f(x)) \leq \frac{1}{\sum_{t \in \mathcal{T}} e^{-2len_f(x;t)\varepsilon}}$$

Even in this toy case, considering just a finite output space and 0-1 loss, we can see why global methods might be loose. As Theorem 12 presents, we can upper-bound agreements between M and f (thus lower-bound loss) entirely using the instance-dependent expression $\text{len}_f(x;t)$. Unsurprisingly, the losses of unbiased, differentially private mechanisms built using less sensitive locations $x \in \mathcal{X}^n$ have tighter lower-bounds, which would not have been caught by global measures.

The instance-dependent bounds of Theorem 12 can be expanded to a more general loss setting where $L(s,t) = \ell(\|s - t\|)$, for some non-decreasing $\ell : \mathbb{R}_+ \to \mathbb{R}_+$, but will require the use of a more generic instance-dependent quantity, dubbed *local modulus of continuity*.

Definition 13 (Local Modulus of Continuity, [Asi and Duchi, 2020])**.** *The local modulus of continuity of $f : \mathcal{X}^n \to \mathcal{T}$ at $x \in \mathcal{X}^n$ is*

$$\omega_f(x;k) = \sup_{x' \in \mathcal{X}^n} \{\|f(x) - f(x')\| : d(x,x') \leq k\}$$

Note that ω_f is instance-dependent (by construction) and is always upper-bounded by the global sensitivity, $\text{GF}_f = \sup_{x \in \mathcal{X}} \omega_f(x;1)$. It captures a similar idea as our analysis of inverse local sensitivity (Definition 11) in that it expresses some judgement of stability around x; if x is near (in the metric space $(\mathcal{X}^n, d)$) points of different labels, $\omega_f(x)$ will tend to be larger and vice versa. Having introduced this quantity, we are ready to generalize Theorem 12 to a general loss for our main result.

Theorem 14 (Lower bounds for general loss, [Asi and Duchi, 2020])**.** *If M is ε-d.p., then for any $k \geq 1$*

$$\sup_{x \in \mathcal{X}^n} \mathbb{E}[L(M(x), f(x))] \geq \sup_{x \in \mathcal{X}^n} \frac{\ell(\omega_f(x;k)/2}{e^{k\varepsilon} + 1}.$$

Furthermore, if M is also L-unbiased, then

$$\mathbb{E}[L(M(x), f(x))] \geq \frac{\ell(\omega_f(x;k)/2)}{e^{2k\varepsilon} + 1}.$$

[3]Note that these are indeed expected loss lower bounds because $\mathbb{E}[L(M(x), f(x))] = 1 - \mathbb{P}(M(x) = f(x))$ in the 0-1 loss setting

Informally, Theorem 14 demonstrates an instance-dependent limit for the performance of an unbiased estimator satisfying notions of privacy. In the realm of expected loss analysis, we can not expect to (consistently) perform under a certain threshold. As the authors note, this is reminiscent of the Cramer-Rao bound in classical statistics, which lower-bound the variance of an unbiased estimator using Fisher-information (10). The following result shows that this is tight.

Theorem 15 (Theorem 14 is optimal, $\mathcal{T} = \mathbb{R}$ case)**.** *Consider the case of $f : \mathcal{X}^n \to \mathbb{R}$ and let $L(s,t) = \ell(|s-t|)$ where $\ell : \mathbb{R}_+ \to \mathbb{R}_+$ is non-decreasing. Suppose a mechanism M is ε-d.p. and, at some point $x \in \mathcal{X}^n$ achieves*

$$\mathbb{E}[\ell(|M(x) - f(x)|)] \le \gamma \ell(\omega_f(x, 1/\varepsilon)/2)$$

where $\gamma \le \dfrac{1}{2\varepsilon}$, then there exists a sample $x' \in \mathcal{X}^n$ with $d(x,x') \le \dfrac{\log(1/2\gamma)}{2\varepsilon}$ with

$$\mathbb{E}[\ell(|M(x') - f(x')|))] \ge \frac{1}{4}\ell\left(\frac{1}{4}\omega_f\left(x'; \frac{\log(1/2\gamma)}{2\varepsilon}\right)\right)$$

Though the tools needed to prove Theorem 15 involve minimax risk, which we have yet to discuss, it amounts to a stronger guarantee of Theorem 15. Considering the toy case where our output space $\mathcal{T}$ is the real numbers, and our loss function corresponds to $|\cdot|$, we cannot consistently beat the threshold from Theorem 14 up to constants on an arbitrary point x, since we will pay dearly in our prediction for a neighboring point x'.

For the final statistical result, we'll show that minimax risk itself can be strongly bounded by instance dependent quantities.

Theorem 16 (Local-minimax bounds)**.** *If $\mathcal{M}_\varepsilon$ is the collection of ε-d.p. private mechanisms, then*

$$\frac{1}{4}\max_{k \le n}\{\ell(\omega_f(x;k)\exp(-k\varepsilon)\} \le \mathcal{R}(x, L, \mathcal{M}_\varepsilon) \le \max_{k \le n}\left\{\frac{1}{1+\exp(k\varepsilon/2)}\ell(\omega_f(x;k))\right\}.$$

It follows $\mathcal{R}(x, L, \mathcal{M}_\varepsilon) \asymp \ell\left(\omega_f\left(x; 1/\varepsilon\right)\right)$.

So, for the class of ε-dp mechanisms, minimax risk is described up to constants just by the local modulus of continuity. When this result is paired with previous theorems about expected loss, it becomes easy to show C-optimality through relation to $\ell(\omega_f(x; 1/\varepsilon))$, which is precisely our goal.

2.3 Inverse-Sensitivity Mechanism is Instance-optimal

In this section, we'll introduce a particularly interesting instantiation of the exponential mechanism which involves inverse sensitivity (Definition 5, 11). Why should we expect anything promising from such an instantiation? Recall from the discussion surrounding Definition 5 that the distribution the exponential mechanism samples from strongly prefers small h values, and that inverse local sensitivity gives stable labels in $\mathcal{T}$ small values as well. Additionally, neither method violates privacy in calculation or result. So, we might expect a mechanism that combines these insights to be both differentially private *and* accurate, perhaps pushing tightness on bounds in Section 2.3.

Definition 17 (Inverse-sensitivity mechanism, folklore, [Asi and Duchi, 2020])**.** *Note that $len_f(x;t)$ is 1-Lipschitz w.r.t. Hamming distance on $\mathcal{X}$.* [4] *Thus, we may let $h(x,t) = len_f(x;t)$ as*

$$d\pi(t) = \frac{\exp(-f(x;t)\varepsilon/2) \cdot d\mu(t)}{\int_{s\in\mathcal{T}} \exp(-f(x;t)\varepsilon/2) \cdot d\mu(s)}$$

As we've done before, we begin by studying the reduced case of M_{disc} where $\mathcal{T}$ is discrete, and we use 0-1 loss. The following lemma follows by construction.

Lemma 18. *Consider an initialization M_{disc} of Definition 11 with discrete output $|\mathcal{T}| < \infty$ and loss $L(s,t) = \mathbf{1}(s \ne t)$. Then,*

$$\mathbb{P}(M_{disc}(x) = f(x)) = \frac{1}{\sum_{t\in\mathcal{T}} e^{-len_f(x;t)\varepsilon/2}}$$

[4] Use triangle inequality on adjacent x, x'

Recalling the statement of Theorem 12, we see that the exponential mechanism matches the upper-bound for accuracy (thus, the lower-bound for loss). So, it seems promising that the general case where we allow $\mathcal{T}$ to be unbounded and use a generic loss might meet the statistically optimal bounds of Theorem 14, and be optimal in the sense of Definitions 8, 9. Indeed this is the case:

Theorem 19 (Optimality, Corollary 3.3, [Asi and Duchi, 2020]). *A continuous form of the exponential mechanism is $C = O(\log n)$ optimal with respect to both minimax risk, and unbiased estimators.*

The fact that C is sample dependent rather than some constant may be an underwhelming result, however, it is shown that this can be tightened in many standard cases. In particular, [Asi and Duchi, 2020] defines and studies a class of *sample-monotone* query functions f where $f(x) \leq s \leq t$ or $t \leq s \leq f(x)$ implies $\text{len}_f(x; s) \leq \text{len}_f(x; t)$. Intuitively, this may be a resonable assumption to work under as it is 'easier' to reach s than t via swapping if t is further away (revisit 11). In this case, it is shown that inverse local sensitivity instantiation is usually $O(1)$-optimal and at worst $O(\log \log n)$-optimal, an improvement on Theorem 19.

2.4 Application to Parameter Estimation

In this section, we'll explore a particular instantiation of the inverse sensitivity mechanism explored extensively in [Hopkins et al., 2022] for *parameter estimation*.

Problem 20 ((Robust) Parameter Estimation, [Hopkins et al., 2022]). *Let $\mathcal{P}_\theta$ be a class of distributions parameterized by some parameter $\theta \in \Theta$. A parameter estimation algorithm $A : \mathcal{X}^n \to \Theta$ has access to n i.i.d. samples $x \in \mathcal{X}^n$ from $\mathcal{P}_{\theta^*}$ and aims to recover θ^*. A robust parameter estimation algorithm A_r is expected to perform the same task, but is instead given x', which is any dataset that satisfies $d(x', x) \leq \eta n$ (also called an η-(strong) corruption).*

Parameter estimation and its robust analogue are difficult and full research subjects in their own right, with known information-theoretic lower bounds. However, in the theme of this survey, we're most interested in enforcing privacy. It turns out that this is possible using tools we already understand in the inverse-sensitivity mechanism (17) and just rudimentary knowledge of a robust estimator, which we will now introduce.

Definition 21 (Robust Estimator). *Let $A_r : \mathcal{X}^n \to \Theta$ be a robust estimator for distribution $\mathcal{P}_\theta$. When evaluated on input $x \in \mathcal{X}^n$, where x is an η-corruption of data from $\mathcal{P}_\theta$, we have $\|A_r(x) - \theta\| \leq \alpha(\eta)$ with high probability, where $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ is nondecreasing.*

Using A_r we can build a score function similar to inverse local sensitivity (11).

Definition 22 (Scoring, 11). *Consider an inverse local sensitivity construction as*

$$s_x(\tilde{\theta}) = \min_{x'} \left(d(x, x') : \|A_r(x') - \tilde{\theta}\| \leq \alpha(\eta_0) \right)$$

for some predetermined value $\alpha(\eta_0)$.

We take a moment to reflect on this construction, starting with the set S_0 of parameters $\tilde{\theta}$ of value $s_x(\tilde{\theta}) = 0$. These proposed labels must lie within a radius of $\alpha(\eta_0)$ of the robust estimate $A_r(x')$, since we need not change any samples x to meet the condition. If we condition on the robust estimator working, we have that the true parameter $\theta^* \in S_0$. Now, consider a candidate label $\tilde{\theta}$ point of score ηn, $\eta < 1$. For the x' chosen, we of course have $\|A_r(x') - \tilde{\theta}\| \leq \alpha(\eta_0)$, but we also have that $\|\theta^* - A_r(x')\| \leq \alpha(\eta)$ if we think of x' as an η-corruption of x. Putting these together, we have that $\|\theta^* - \tilde{\theta}\| \leq 2\alpha(\eta_0)$ when $\eta \leq \eta_0$, and $\|\theta^* - \tilde{\theta}\| \leq 2\alpha(\eta)$ otherwise, or, informally, lower scores are better. Following from our discussions around the Definition 5, an exponential sampling mechanism seems to be an ideal candidate, as

$$A_{dp}(x = \theta) = \frac{\exp(-s_x(\theta) \cdot \varepsilon/2)}{\int_{\theta' \in \Theta} \exp(-s_x(\theta') \cdot \varepsilon/2) d\theta'}. \tag{4}$$

It can be shown that the samples A_{dp} concentrate around a ball of radius $2\alpha(\eta_0)$ from θ^* (i.e., have score at most η_0), so we get the following strong lemma:

Lemma 23 (Lemma 2.1, Hopkins et al. [2022]). *Suppose a dataset $X_1, ..., X_n \sim p_{\theta^*}$, where the parameter vector $\theta^* \in \Theta \subset \mathbb{R}^d$. For any threshold $\eta_0 \in [0, 1]$, a random θ drawn using A_{dp} has*

$\|\theta - \theta^*\| \leq 2\alpha(\eta_0)$ *with probability at least* $1 - 2\beta$ *if*

$$n \geq \max_{\eta_0 \leq \eta \leq 1} \frac{d \log \frac{2\alpha(\eta)}{\alpha(\eta_0)} + \log(1/\beta) + O(\log \eta n)}{\eta \varepsilon}.$$

Additionally, the mechanism A_{dp} *is* ε*-d.p.*[5]

A particularly interesting special case is where p_{θ^*} is the class of Guassians of bounded mean and identity covariance matrix, with the ℓ_2 metric. In this setting, robust estimators have been studied extensively, and known to produce estimates within α given $\tilde{O}(d/\alpha^2)$ samples. Using the black-box reduction given by A_{dp}, we can add privacy to this estimate at some complexity cost:

Corollary 24 (Private Mean Estimation, Guassian, [Hopkins et al., 2022])**.** *Let* $0 < \alpha, \beta, \varepsilon < 1$ *and* $R > 0$*. Let* $\mu \in \mathbb{R}^d$ *where* $\|\mu\|_2 \leq R$ *is to be estimated. There is an* ε*-dp mechanism that takes* n *i.i.d. samples from* $N(\mu, I_d)$ *and returns* $\hat{\mu}$ *that satisfies* $\|\mu - \hat{\mu}\|_2 \leq \alpha$ *with high probability, given*

$$n = \tilde{O}\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon} + \frac{d \log R}{\varepsilon}\right)$$

As a final connection to the research question, we find that adding privacy onto a robust estimator via black-box reduction adds complexity $\tilde{O}(d/\alpha\varepsilon + d \log R/\varepsilon)$, recalling that $\tilde{O}(d/\alpha^2)$ is required just by constuction of A_r. Since A_{dp} is an inverse sensitivity mechanism (and also defined over a sample-monotone space), we have by [Asi and Duchi, 2020] that this is an optimal cost of privacy. The dependence on R is worrying but can be shown to be necessary in the pure-dp setting; if we allow approximate privacy, it relaxes to $1/\delta$. We can also allow for non-identity covariance matrices as $I \preccurlyeq \Sigma \preccurlyeq K \cdot I$, which introduces a $\log K$ dependency.

3 Commentary on Future Work

Since our results are so overwhelmingly theoretical, most of our suggested explorations will be in this direction, divided into statistical and algorithmic categories.

3.1 Theory/Statistical Interests

A key statistical interest is characterizing what regimes correspond with which C-optimalities. We've discussed the unconstrained continuous case and the sample-monotone case, seeing and improvement from $C = O(\log n)$ to $C = O(\log \log n)$, but we'd personally really like to understand the assumptions necessary to have $C = O(1)$. This would imply that the optimality of inverse sensitivity is as tight as possible, having no dependence on sample complexity.

A possible idea in this direction would be to consider a special case of sample monotonicity where $\text{len}_f(x; s) \leq h(s, t)\text{len}_f(x; t)$ holds for s in between $f(x)$ and t. This approach is inspired by the change from global sensitivity to smoothened sensitivity, where we add some instance-dependent condition described by $h(s, t)$. (Maybe, we could have $h(s, t) = e^{|s-t|}$ in the case of $\mathcal{T} = \mathbb{R}$ to get a 'differentially private' idea).

Another interesting direction to explore would be involving privacy in real-world application, such as exposure to noise (taking inspiration from [Hopkins et al., 2022]). For example, it's not difficult to show that the inverse sensitivity mechanism of A_{dp} performs as well under noise (up to constant factors) as it does in the noiseless regime, as robustness is built in. But, how does a generic ε-dp mechanism ($\mathcal{M}_\varepsilon$) measure up against unbiased algorithms in this setting (again, with respect to C-optimality). It seems intuitive that privacy should offer built-in protections against noise, since our mechanisms are conditioned not too give any datapoint too much weight. A result in this area would go towards showing privacy implies robustness to some extent.

3.2 Algorithmic Interests

A main algorithmic exploration is creating a polytime implementation of the exponential mechanism. As it is written theoretically, we must request an infinite number of queries to construct the distribution

[5] Note that $|s_x(\tilde{\theta}) - s_{x'}(\tilde{\theta})| \leq 1$ for any $d(x, x') \leq 1$.

and then sample at random from this, repeatedly – this is clearly intractable. In the 'General Sampling Algorithm', [Hopkins et al., 2022] presents a partial solution for this in the A_{dp} case. From what we can understand, this involves a reduction from the Sum-of-Squares used to compute robust estimates, and cannot be generalized beyond the black-box reduction setting. Though this is likely impossible to be established generally, it is still interesting to study, perhaps for other robust algorithms.

A final remark would be to study algorithms beyond the exponential mechanism initialization. Since this mechanism still has interesting properties to be studied (as elaborated in other points) and is by far the easiest to analyze mathematically, it may be productive to explore other DP algorithms with a better track record of tractability. For example, even if the noise-based methods discussed earlier have a worse constant C, their gains in computation may be great enough to make them worthwhile.

References

Hilal Asi and John C. Duchi. Near instance-optimality in differential privacy, 2020.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, 2006. URL https://api.semanticscholar.org/CorpusID:2468323.

Samuel B. Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation, 2022.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103, 2007. doi: 10.1109/FOCS.2007.66.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, STOC '07, page 75–84, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595936318. doi: 10.1145/1250790.1250803. URL https://doi.org/10.1145/1250790.1250803.