The Right Complexity Measure in Locally Private Estimation: It is not the Fisher Information

John C. Duchi Feng Ruan
Stanford University

September 2020

Abstract

We identify fundamental tradeoffs between statistical utility and privacy under local models of privacy in which data is kept private even from the statistician, providing instance-specific bounds for private estimation and learning problems by developing the *local minimax risk*. In contrast to approaches based on worst-case (minimax) error, which are conservative, this allows us to evaluate the difficulty of individual problem instances and delineate the possibilities for adaptation in private estimation and inference. Our main results show that the local modulus of continuity of the estimand with respect to the variation distance—as opposed to the Hellinger distance central to classical statistics—characterizes rates of convergence under locally private estimation for many notions of privacy, including differential privacy and its relaxations. As consequences of these results, we identify an alternative to the Fisher information for private estimation, giving a more nuanced understanding of the challenges of adaptivity and optimality.

1 Introduction

The increasing collection of data at large scale—medical records, location information from cell phones, internet browsing history—points to the importance of a deeper understanding of the tradeoffs inherent between privacy and the utility of using the data collected. Classical mechanisms for preserving privacy, such as permutation, small noise addition, releasing only mean information, or basic anonymization are insufficient, and notable privacy compromises with genomic data [37] and movie rating information [44] have caused the NIH to temporarily stop releasing genetic information and Netflix to cancel a proposed competition for predicting movie ratings. Balancing the tension between utility and the risk of disclosure of sensitive information is thus essential.

In response to these challenges, researchers in the statistics, databases, and computer science communities have studied *differential privacy* [55, 33, 29, 28, 34, 25, 21] as a formalization of disclosure risk limitation. This literature discusses two notions of privacy: *local* privacy, in which data is privatized before it is even shared with a data collector, and *central* privacy, where a centralized curator maintains the sample and guarantees that any information it releases is appropriately private. The local model is stronger and entails some necessary loss of statistical efficiency, yet its strong privacy protections encourage its adoption. Whether for ease of regulatory compliance, for example with European Union privacy rules [32]; for transparency and belief in the importance of privacy; or to avoid risks proximate to holding sensitive data, like hacking or subpoena risk; major technology companies have adopted local differential privacy protections in their data collection and machine learning tools. Apple provides local differential privacy in many of its iPhone systems [3], and Google has built systems supplying central and local differential privacy [30, 1]. The broad impact of privacy protections in billions of devices suggest we should carefully understand the fundamental limitations and possibilities of learning with local notions of privacy.

To address this challenge, we borrow from Cai and Low [10] to study the *local minimax complexity* of estimation and learning under local privacy. Worst-case notions of complexity may be too stringent for statistical practice, and we wish to understand how difficult the *actual* problem we have is and whether we can adapt to this problem difficulty, so that our procedures more efficiently

solve easy problems—as opposed to being tuned to worst-case scenarios. Our adoption of local minimax complexity is thus driven by three desiderata, which Cai and Low [10] identify: we seek fundamental limits on estimation and learning that (i) are instance specific, applying to the particular problem at hand, (ii) are (uniformly) attainable, in that there exist procedures to achieve the instance-specific difficulty, and (iii) have super-efficiency limitations, so that if a procedure achieves *better* behavior than the lower bounds suggest is possible, there should be problem instances in which the procedure must have substantially worse behavior. We provide characterize the local minimax complexity of locally private estimation of one-dimensional quantities, showing that this benchmark (nearly) always satisfies desiderata (i) and (iii). Via a series of examples—some specific, others general—we show that there are procedures whose risk is of the order of the local minimax risk for all underlying (unknown) populations. As an essential part of this program is that the complexity is (ii) attainable—which, to our knowledge, remains open even in the non-private case—we view this paper as an initial foray into understanding problem-specific optimality in locally private estimation.

1.1 Contributions, outline, and related work

Our development of instance-specific complexity notions under privacy constraints allows us to quantify the statistical price of privacy. Identifying the tension here is of course of substantial interest, and Duchi et al. [25, 24] develop a set of statistical and information-theoretic tools for understanding the *minimax* risk in locally differentially private settings, providing the point of departure for our work. To understand their and our coming approach, we formalize our setting.

We have i.i.d. data $X_1, \ldots, X_n$ drawn according to a distribution P on a space $\mathcal{X}$. Instead of observing the original sample $\{X_i\}$, however, the statistician or learner sees only *privatized* data $\{Z_i\}$, where Z_i is drawn from a Markov kernel $Q(\cdot \mid X_i)$ conditional on X_i (following information-theory, we call Q the privacy *channel* [15]). We allow the channel to be *sequentially interactive* [25], meaning that Z_i may depend on the previous (private) observations $Z_1, \ldots, Z_{i-1}$, i.e.

$$Z_i \mid X_i = x, Z_1, \ldots, Z_{i-1} \sim Q(\cdot \mid x, Z_{1:i-1}). \tag{1}$$

This notion of interactivity is important for procedures, such as stochastic gradient methods [25] or the one-step-corrected estimators we develop in the sequel, which modify the mechanism after some number of observations to more accurately perform inference.

The statistical problems we consider are, abstractly, as follows. Let $\mathcal{P}$ be a family of distributions, and let $\theta : \mathcal{P} \to \Theta \subset \mathbb{R}^d$ be a parameter we wish to estimate and belonging to Θ, where $\theta(P)$ denotes the target parameter. Let $L : \mathbb{R}^d \to \mathbb{R}_+$ be a symmetric quasiconvex loss, where we assume that $L(\mathbf{0}) = 0$. A typical example is the mean $\theta(P) = \mathbb{E}_P[X]$ with squared error $L(\theta - \theta(P)) = (\theta - \mathbb{E}_P[X])^2$. Let $\mathcal{Q}$ be a collection of private channels, for example, ε-differentially private channels (which we define in the sequel). The *private minimax risk* [25] is

$$\mathfrak{M}_n(L, \mathcal{P}, \mathcal{Q}) := \inf_{\widehat{\theta}, Q \in \mathcal{Q}} \sup_{P \in \mathcal{P}} \mathbb{E}_{Q \circ P} \left[L(\widehat{\theta}(Z_1, \ldots, Z_n) - \theta(P)) \right] \tag{2}$$

where $Q \circ P$ denotes the marginal $X_i \sim P$ and Z_i drawn conditionally (1). Duchi et al. [25] provide upper and lower bounds on this quantity when $\mathcal{Q}$ is the collection of ε-locally differentially private channels, developing strong data processing inequalities to quantify the costs of privacy.

The worst-case nature of the formulation (2) gives lower bounds that may be too pessimistic for practice, and it prohibits a characterization of problem-specific difficulty. Accordingly, we adopt

a *local minimax approach*, which builds out of the classical statistical literature on hardest one-dimensional alternatives that begins with Stein [48, 6, 18, 19, 20, 10, 13]. To that end, we define the *local minimax risk* at the distribution P_0 for the set of channels $\mathcal{Q}$ as

$$\mathfrak{M}_n^{\loc}(P_0, L, \mathcal{P}, \mathcal{Q}) := \sup_{P_1 \in \mathcal{P}} \inf_{\widehat{\theta}, Q \in \mathcal{Q}} \max_{P \in \{P_0, P_1\}} \mathbb{E}_{Q \circ P} \left[L(\widehat{\theta}(Z_1, \ldots, Z_n) - \theta(P)) \right]. \tag{3}$$

The quantity (3) measures the difficulty of the loss minimization problem for a *particular distribution* P_0 under the privacy constraints $\mathcal{Q}$ characterizes, and at this distinguished distribution, we look for the hardest alternative distribution $P_1 \in \mathcal{P}$. As we shall see, the definition (3) indeed becomes local, if P_1 is far from P_0, then it is easy to develop an estimator $\widehat{\theta}$ distinguishing P_0 and P_1, so that (for large n) the supremum is essentially constrained to a neighborhood of P_0.

To situate our contributions, let us first consider the non-private local minimax complexity, when $\mathcal{Q} = \{\mathrm{id}\}$ (the identity mapping). Throughout, we will use the shorthand

$$\mathfrak{M}_n^{\loc}(P_0, L, \mathcal{P}) := \mathfrak{M}_n^{\loc}(P_0, L, \mathcal{P}, \{\mathrm{id}\})$$

for the non-private local minimax risk. We wish to estimate a linear function $v^T\theta$ of the parameter θ with (projected) square loss $L_{\mathrm{sq},v}(t) = (v^T t)^2$. In the classical setting of a parametric family $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ with Fisher information matrix I_θ, then (as we describe more formally in Section 2.2) the Fisher information bound for the parameter θ_0 is

$$\mathfrak{M}_n^{\loc}(P_{\theta_0}, L_{\mathrm{sq},v}, \mathcal{P}, \{\mathrm{id}\}) \asymp \frac{1}{n} \mathbb{E}\left[\left(v^T Z\right)^2 \right] \quad \text{for} \quad Z \sim \mathsf{N}\left(0, I_{\theta_0}^{-1}\right), \tag{4}$$

where $\asymp$ denotes equality to within numerical constants. More generally, if we wish to estimate a functional $\theta(P) \in \mathbb{R}$ of P, Donoho and Liu [18, 19, 20] show how the *modulus of continuity* takes the place of the classical information bound. Again considering the squared error $L_{\mathrm{sq}}(t) = t^2$, define the Hellinger modulus of continuity of $\theta(\cdot)$ at $P_0 \in \mathcal{P}$ by

$$\omega_{\mathrm{hel}}(\delta; P_0, \mathcal{P}) := \sup_{P_1 \in \mathcal{P}} \left\{ |\theta(P_0) - \theta(P_1)| \text{ s.t. } P_1 \in \mathcal{P}, d_{\mathrm{hel}}(P_0, P_1) \le \delta \right\} \tag{5}$$

where $d_{\mathrm{hel}}^2(P_0, P_1) = \frac{1}{2} \int (\sqrt{dP_0} - \sqrt{dP_1})^2$. In the local minimax case, characterizations via a local modulus are available in some problems [10, 13], where $\mathfrak{M}_n^{\loc}(P_0, L_{\mathrm{sq}}, \mathcal{P}) \asymp \omega_{\mathrm{hel}}^2(n^{-1/2}; P_0, \mathcal{P})$, while under mild regularity conditions, the *global modulus* $\sup_{P \in \mathcal{P}} \omega_{\mathrm{hel}}(\delta; P, \mathcal{P})$ governs non-private global minimax risk: (often) one has $\mathfrak{M}_n(L_{\mathrm{sq}}, \mathcal{P}) \asymp \sup_{P_0 \in \mathcal{P}} \omega_{\mathrm{hel}}(n^{-1/2}; P_0, \mathcal{P})$ [6, 18, 19, 20].

In contrast, the work of Duchi et al. [25, 24] suggests that for ε-locally differentially private estimation, we should replace the Hellinger distance by *variation distance*. In the case of higher-dimensional problems, there are additional dimension-dependent penalties in estimation that local differential privacy makes unavoidable, at least in a minimax sense [25]. In work independent of and contemporaneous to our own, Rohde and Steinberger [47] build off of [25] to show that (non-local) minimax rates of convergence under ε-local differential privacy are frequently governed by a global modulus of continuity, except that the variation distance $\|P_0 - P_1\|_{\mathrm{TV}} = \sup_A |P_0(A) - P_1(A)|$ replaces the Hellinger distance d_{hel}. They also exhibit a mechanism that is minimax optimal for "nearly" linear functionals based on randomized response [55, 47, Sec. 4]. Thus, locally differentially private procedures give rise to a different geometry than classical statistical problems.

We are now in a position for a high-level description of our results, which apply in a variety of locally private estimation settings consisting of weakenings of ε-differential privacy, whose definitions we formalize in Section 2.1. We provide a precise characterization of the local minimax complexity (3) in these settings. If we define the local modulus of continuity at P_0 by

$$\omega_{\mathrm{TV}}(\delta; P_0, \mathcal{P}) := \sup_{P \in \mathcal{P}} \left\{ |\theta(P_0) - \theta(P)| \text{ s.t. } \|P - P_0\|_{\mathrm{TV}} \le \delta \right\},$$

then a consequence of Theorem 1 is that for the squared loss and ε-locally private channels $\mathcal{Q}_\varepsilon$,

$$\mathfrak{M}_n^{\rm loc}(P_0, L_{\rm sq}, \mathcal{P}, \mathcal{Q}_\varepsilon) \asymp \omega_{\rm TV}^2\left((n\varepsilon^2)^{-1/2}; P_0, \mathcal{P}\right).$$

We provide this characterization in more detail and for general losses in Section 3. Moreover, we show a super-efficiency result that any procedure that achieves risk better than the local minimax complexity at a distribution P_0 *must* suffer higher risk at another distribution P_1, so that this characterization does indeed satisfy our desiderata of an instance-specific complexity measure.

The departure of these risk bounds from the typical Hellinger modulus (5) has consequences for locally private estimation and adaptivity of estimators, which we address for parametric problems and examples in Section 4 and for general estimation in Section 5. Instead of the Fisher information, an alternative we term the L^1*-information* characterizes the complexity of locally private estimation. A challenging consequence of these results is that, for some parametric models (including Bernoulli estimation and binomial logistic regression), the local complexity (3) is *independent* of the underlying parameter: nominally easy problems (in the Fisher information sense) are not so easy under local privacy constraints. Our proofs rely on novel Markov contraction inequalities for divergence measures, which strengthen classical strong data processing inequalities [14, 16, 25].

Developing procedures achieving the local minimax risk (3) is challenging, but we show that locally uniform convergence is asymptotically possible in a number of cases in Sections 4 and 5, including well- and mis-specified exponential family models, using stochastic gradient methods or one-step corrected estimators. An important point of our results (Sec. 5.3) is that the local private minimax risk—sometimes in distinction from the non-private case—depends strongly on the assumed family $\mathcal{P}$, making the development of private adaptive estimators challenging. We use a protein expression-prediction problem in Section 6 to compare our locally optimal procedures with minimax optimal procedures [25]; the experimental results suggests that the locally optimal procedures outperform global minimax procedures, though costs of privacy still exist.

Notation: We use a precise big-O notation throughout the paper, where for functions $f, g : \mathcal{X} \to \mathbb{R}_+$, $g(x) = O(f(x))$ means that there exists a numerical (universal) constant $C < \infty$ such that $g(x) \le Cf(x)$; we use $g(x) \lesssim f(x)$ to mean the same. We write $O_t(\cdot)$ when the constant C may depend on an auxiliary parameter t. We write $g(x) \asymp f(x)$ if both $g(x) \lesssim f(x)$ and $f(x) \lesssim g(x)$. If $g(x) = o(f(x))$ as $x \to x_0$, we mean that $\limsup_{x\to x_0} g(x)/f(x) = 0$. We let $L^p(P)$ be the collection of $g : \mathcal{X} \to \mathbb{R}^d$ with $\int \|g(x)\|^p \, dP(x) < \infty$, where $p = \infty$ is the set of essentially bounded g; the dimension d is tacit. For a sequence of distributions P_n, we write convergence in distribution $X_n \xrightarrow{d}_{P_n} X$ to mean that for any bounded continuous f, $\mathbb{E}_{P_n}[f(X_n)] \to \mathbb{E}[f(X)]$.

2 Preliminaries

We begin in Section 2.1 with definitions and some brief discussion of the definitions of privacy we consider. To help situate our approach, we discuss local minimax complexity without privacy in Section 2.2. There are several plausible notions of attainment of the local minimax risk—all related to desideratum (ii) in the introduction that the risk be achievable—so we conclude in Section 2.3 by giving several related results, including an asymptotic and locally uniform convergence guarantee that will be what we typically demonstrate for our procedures. In spite of the (sometimes) asymptotic focus, which builds out of Le Cam's quadratic mean differentiability theory and various notions of efficiency in semiparametric models [38, 39, 45, 52, 5], we will typically achieve optimality only to within numerical constants—getting sharp constants appears challenging when we allow arbitrary privatization schemes and sequential interactivity (1).

2.1 Definitions of Local Privacy

With the notion (1) of sequentially interactive channels, where the ith private observation is drawn conditionally on the past as $Z_i \mid X_i = x, Z_1, \ldots, Z_{i-1} \sim Q(\cdot \mid x, Z_{1:i-1})$, we consider several privacy definitions. First is *local differential privacy*, which Warner [55] proposes (implicitly) in his 1965 work on survey sampling, and which Evfimievski et al. [33] and Dwork et al. [29] make explicit.

Definition 1. The channel Q is ε-*locally differentially private* if for all $i \in \mathbb{N}$, $x, x' \in \mathcal{X}$, and $z_{1:i-1} \in \mathcal{Z}^{i-1}$,

$$\sup_{A \in \sigma(\mathcal{Z})} \frac{Q(A \mid x, z_{1:i-1})}{Q(A \mid x', z_{1:i-1})} \le e^{\varepsilon}.$$

The channel Q is *non-interactive* if for all $z_{1:i-1} \in \mathcal{Z}^{i-1}$ and $A \in \sigma(\mathcal{Z})$,

$$Q(A \mid x, z_{1:i-1}) = Q(A \mid x)$$

Duchi et al. [25] consider this notion of privacy, developing its consequences for minimax optimal estimation. An equivalent view [56] is that an adversary knowing the data is either x or x' cannot accurately test, even conditional on the output Z, whether the generating data was x or x' (the sum of Type I and II errors is at least $\frac{1}{1+e^{\varepsilon}}$). To mitigate the consequent difficulties for estimation and learning with differentially private procedures, researchers have proposed weakenings of Definition 1, which we also consider.[1] These repose on α-Rényi-divergences, defined for $\alpha \ge 1$ by

$$D_\alpha \left(P \| Q\right) := \frac{1}{\alpha - 1} \log \int \left(\frac{dP}{dQ}\right)^{\alpha} dQ.$$

For $\alpha = 1$ one takes the limit $\alpha \downarrow 1$, yielding $D_\alpha \left(P \| Q\right) = D_{\rm kl} \left(P \| Q\right)$, and for $\alpha = \infty$ one has $D_\alpha \left(P \| Q\right) = \operatorname{ess\,sup} \log \frac{dP}{dQ}$. Mironov [43] then proposes the following definition:

Definition 2. The channel Q is (α, ε)-*Rényi locally differentially private (RDP)* if for all $x, x' \in \mathcal{X}$, and $z_{1:i-1} \in \mathcal{Z}$, we have

$$D_\alpha \left(Q(\cdot \mid x, z_{1:i-1}) \| Q(\cdot \mid x', z_{1:i-1})\right) \le \varepsilon.$$

This definition simplifies *concentrated differential privacy* [27, 9] by requiring that it hold only for a single fixed α, and it has allowed effective private methods for large scale machine learning [1].

The choice $\alpha = 2$ in Definition 2 is salient and important in our analysis. Consider a prior on points x, x', represented by $\pi(x) \in [0, 1]$ and $\pi(x') = 1 - \pi(x)$, and the posterior $\pi(x \mid Z)$ and $\pi(x' \mid Z)$ after observing the private quantity $Z \sim Q(\cdot \mid x)$. Then $(2, \varepsilon)$-Rényi privacy is equivalent [43, Sec. VII] to the the prior and posterior odds of x against x' being close in expectation:

$$\mathbb{E}\left[\frac{\pi(x \mid Z)/\pi(x' \mid Z)}{\pi(x)/\pi(x')} \mid x\right] \le e^{\varepsilon}$$

for all two-point priors π, where the expectation is taken over $Z \mid x$. (For ε-differential privacy, the inequality holds for *all* Z without expectation). As Rényi divergences are monotonic in α (cf. [53, Thm. 3]), any (α, ε)-Rényi private channel is (α', ε)-Rényi private for $\alpha' \le \alpha$. Thus, any lower bound we prove on estimation for $(\alpha = 2, \varepsilon)$-local RDP implies an identical lower bound for $\alpha' \ge 2$.

The definitions provide varying levels of privacy. It is immediate that if a channel is ε-differentially private, then it is (α, ε)-Rényi locally private for any α. More sophisticated bounds

[1] We ignore (ε, δ)-approximate differential privacy, as for locally private estimation, it is essentially equivalent to ε-differential privacy [e.g. 21, Appendix D.1].

are possible. Most importantly, ε-differential privacy (Definition 1) implies $(\alpha, 2\alpha\varepsilon^2)$-Rényi differential privacy (Definition 2) for all $\alpha \geq 1$. For $\alpha = 2$, we can tighten this to $(2, \min\{\frac{3}{2}\varepsilon^2, 2\varepsilon\})$-RDP. We therefore write our lower bounds to apply for $(2, \varepsilon^2)$-Rényi differentially private channels; this implies lower bounds for all (α, ε^2)-RDP channels, and (as differential privacy is stronger than Rényi privacy) implies lower bounds for any ε-locally differentially private channels.

2.2 A primer on local minimax complexity

We briefly review local minimax complexity to give intuition for and motivate our approach. The starting point is Stein [48], who considers estimating a nonparametric functional $\theta(P)$, proposing that the "information" about θ at P_0 should be the least Fisher information over all one-dimensional subfamilies of distributions that include P_0, leading to the local minimax risk (3) with $\mathcal{Q} = \{\mathrm{id}\}$. Specializing to the squared error L_{sq}, in the non-private case, one then defines

$$\mathfrak{M}_n^{\mathrm{loc}}(P_0, L_{\mathrm{sq}}, \mathcal{P}) := \sup_{P_1 \in \mathcal{P}} \inf_{\widehat{\theta}} \max_{P \in \{P_0, P_1\}} \mathbb{E}_P \left[(\widehat{\theta} - \theta(P))^2 \right]. \tag{6}$$

Then the Hellinger modulus (5) typically characterizes the local minimax risk (6) to numerical constants [19, 10], as the next proposition shows (we include a proof for completeness in Appendix A.1).

Proposition 1. *For each $n \in \mathbb{N}$ and any $P_0 \in \mathcal{P}$,*

$$\frac{\sqrt{2}-1}{8\sqrt{2}} \omega_{\mathrm{hel}}^2(n^{-1/2}/2; P_0, \mathcal{P}) \leq \mathfrak{M}_n^{\mathrm{loc}}(P_0, L_{\mathrm{sq}}, \mathcal{P}) \leq \sup_{r \geq 0} \left\{ \omega_{\mathrm{hel}}^2(r; P_0, \mathcal{P}) \exp(-nr^2) \right\}.$$

Whenever the modulus of continuity behaves nicely, the upper bound shows that the lower is tight to within constant factors. For example, under a polynomial growth assumption that there exist $B, \beta < \infty$ such that $\omega_{\mathrm{hel}}(c\delta; P_0, \mathcal{P}) \leq B c^\beta \omega_{\mathrm{hel}}(\delta; P_0, \mathcal{P})$ for all $c > 1$, then

$$\mathfrak{M}_n^{\mathrm{loc}}(P_0, \mathcal{P}) \leq (B\beta^{\beta/2} e^{-\beta/2}) \cdot \omega_{\mathrm{hel}}^2 \left(n^{-1/2}/2; P_0, \mathcal{P} \right) \tag{7}$$

(cf. Appendix A.1). The global modulus of continuity of the parameter $\theta(P)$ with respect to Hellinger distance also characterizes global minimax error for estimation of linear functionals on convex spaces of distributions [6, 18, 19] and gives lower bounds generically.

These calculations are abstract, so it is instructive to specialize to more familiar families, where we recover the information bound (4). Consider a parametric family of distributions $\mathcal{P} := \{P_\theta\}_{\theta \in \Theta}$, $\Theta \subset \mathbb{R}^d$, with dominating measure μ. We assume $\mathcal{P}$ is *quadratic mean differentiable* (QMD) at θ [52, Ch. 7.1], meaning there exists a score $\dot{\ell}_\theta : \mathcal{X} \to \mathbb{R}^d$ such that

$$\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h^T \dot{\ell}_\theta \sqrt{p_\theta} \right)^2 d\mu = o(\|h\|^2) \tag{8}$$

as $h \to 0$. Most classical families of distributions (e.g. exponential families) are QMD with the familiar score $\dot{\ell}_\theta(x) = \nabla_\theta \log p_\theta(x)$ (cf. [40, 52]). The Fisher information $I_\theta = \int \dot{\ell}_\theta \dot{\ell}_\theta^T p_\theta d\mu \in \mathbb{R}^{d \times d}$ then exists, and we have the asymptotic expansion

$$d_{\mathrm{hel}}^2(P_{\theta+h}, P_\theta) = \frac{1}{8} h^T I_\theta h + o(\|h\|^2). \tag{9}$$

When the parameter θ is identifiable, the local minimax risk (6) coincides with the standard Fisher information bounds to within numerical constants. Indeed, consider the following identifiability

Assumption A1. *For $\delta > 0$, there exists $\gamma > 0$ such that $\|\theta - \theta_0\| > \delta$ implies $d_{\rm hel}^2(P_\theta, P_{\theta_0}) > \gamma$.*

We can then make the approximation (4) for estimating $v^T\theta_0$ rigorous (see Appendix A.2):

Claim 2.1. *Let $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$ be quadratic mean differentiable at θ_0 with positive definite Fisher information I_{θ_0}, assume that Θ is bounded, and that θ_0 is identifiable (A1). Then for large $n \in \mathbb{N}$,*

$$\frac{1}{21} \cdot \frac{1}{n} v^T I_{\theta_0}^{-1} v \le \mathfrak{M}_n^{\rm loc}(P_{\theta_0}, L_{{\rm sq},v}, \mathcal{P}) \le \frac{9}{e} \cdot \frac{1}{n} v^T I_{\theta_0}^{-1} v.$$

We cannot expect to achieve the correct numerical constants with the two-point lower bounds in the local minimax risk [12], but Claim 2.1 recovers the correct scaling in problem parameters.

2.3 Measuring attainment of the local minimax risk

As we note in the introduction, we would like a procedure that uniformly achieves the local minimax benchmark, that is, for a given loss L, returning to the more general notation (3), we would like

$$\sup_{Q\circ P_0 \in \mathcal{P}} \frac{\mathbb{E}_{P_0}[L(\widehat{\theta}_n - \theta(P_0))]}{\mathfrak{M}_n^{\rm loc}(P_0, L, \mathcal{P}, \mathcal{Q})} \lesssim 1.$$

Achieving this generally is challenging (and for many families $\mathcal{P}$, impossible [5]); indeed, a major contribution of Cai and Low [10] is to show that it *is* possible to achieve this uniform benchmark for the squared error and various functionals in convexity-constrained nonparametric regression.

As a consequence, we often consider a weakening to achieve the local minimax risk (to within numerical constants). We describe this precisely at the end of this section, first reviewing some of the necessary parametric and semi-parametric theory [52, 5]. In parametric cases, $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$, we consider sequences of scaled losses, taking the form $L_n(\widehat{\theta}_n - \theta(P_0)) = L(\sqrt{n}(\widehat{\theta}_n - \theta(P_0)))$. An estimator $\widehat{\theta}_n$ is *asymptotically local minimax rate optimal* if

$$\sup_c \limsup_{n\to\infty} \sup_{\|h\|\le c/\sqrt{n}} \frac{\mathbb{E}_{P_{\theta_0+h}}[L(\sqrt{n}(\widehat{\theta}_n - (\theta_0 + h)))]}{\mathfrak{M}_n^{\rm loc}(P_{\theta_0}, L_n, \mathcal{P}, \{\rm id\})} \lesssim 1 \tag{10}$$

for all $\theta_0 \in \operatorname{int}\Theta$. Le Cam's local asymptotic normality theory of course allows much more, even achieving correct constants [38, 39, 45, 52]. We emphasize that while many of our ideas build out of semiparametric efficiency, typically we only achieve optimality to within numerical constants.

We will generally demonstrate procedures that achieve the (private) local minimax risk in some locally uniform sense, and with this in mind, we review a few necessary concepts in semi-parametric estimation on regularity, sub-models, and tangent spaces [cf. 52, Chapters 8.5 & 25.3] that will be important for developing our asymptotics. Let $\mathcal{P}$ be a collection of distributions, and for some $P_0 \in \mathcal{P}$ let $\mathcal{P}_{{\rm sub},0} := \{P_h\}_{h\in\mathbb{R}^d} \subset \mathcal{P}$ be a sub-model within $\mathcal{P}$ indexed by $h \in \mathbb{R}^d$, where we assume that $\mathcal{P}_{{\rm sub},0}$ is quadratic mean differentiable (QMD) (8) at P_0 for a score function $g : \mathcal{X} \to \mathbb{R}^d$ (usually this score will simply be $g(x) = \nabla_h \log dP_h(x)|_{h=0}$) [52, Ch. 25.3], that is,

$$\int \left| dP_h^{1/2} - dP_0^{1/2} - \frac{1}{2} h^T g dP_0^{1/2} \right|^2 = o(\|h\|^2) \tag{11}$$

as $h \to 0$. Considering different QMD sub-models $h \mapsto P_h$ around P_0 yields the tangent set $\dot{\mathcal{P}}_0$, which is a collection of mean-zero score functions $g : \mathcal{X} \to \mathbb{R}^d$ with $g \in L^2(P_0)$. Then a parameter

$\theta : \mathcal{P} \to \mathbb{R}^k$ is *differentiable relative to* $\dot{\mathcal{P}}_0$ if there exists a mean-zero *influence function* $\dot{\theta}_0 : \mathcal{X} \to \mathbb{R}^k$, where for each submodel $\mathcal{P}_{\sub,0} = \{P_h\}_{h \in \mathbb{R}^d}$ and associated score $g : \mathcal{X} \to \mathbb{R}^d$,[2]

$$\theta(P_h) = \theta(P_0) + \int \dot{\theta}_0(x) \langle g(x), h \rangle dP_0(x) + o(\|h\|). \tag{12}$$

We turn now away from properties of the parameter θ to properties of estimators that will be useful. An estimator $\widehat{\theta}_n$ is *regular for* θ at P_0 if for all h and sequences $h_n \to h \in \mathbb{R}^d$,

$$\sqrt{n}(\widehat{\theta}_n - \theta(P_{h_n/\sqrt{n}})) \xrightarrow[P^n_{h_n/\sqrt{n}}]{d} Z$$

for a random variable Z (which is identical for each h); such estimators are classically central [52]. In our constructions, the (private) estimators $\widehat{\theta}_n$ depend both on the variables X_i and, as we construct $Z_i \sim Q(\cdot \mid X_i, Z_{1:i-1})$, we can assume w.l.o.g. that there is an independent sequence of auxiliary random variables $\xi_i \stackrel{\iid}{\sim} P_{\aux}$ such that $\widehat{\theta}_n = \widehat{\theta}_n(X_{1:n}, \xi_{1:n})$. Then under the sampling distribution $P_0 \times P_{\aux}$, we shall often establish the asymptotic linearity of $\widehat{\theta}_n$ at $P_0 \times P_{\aux}$, meaning

$$\sqrt{n}(\widehat{\theta}_n - \theta(P_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\theta}_0(X_i) + \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_{\aux}(\xi_i) + o_{P_0}(1), \tag{13}$$

where $\mathbb{E}[\dot{\theta}_0(X)] = \mathbb{E}[\phi_{\aux}(\xi)] = 0$, and $\Cov(\dot{\theta}_0) = \Sigma_0$ and $\Cov(\phi_{\aux}) = \Sigma_{\aux}$. Such expansions, with $\phi_{\aux} \equiv 0$, frequently occur in classical parametric, semi-parametric, and nonparametric statistics [cf. 52, Chs. 8 & 25]. For example, in parametric cases with $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$, standard score $\dot{\ell}_\theta = \nabla_\theta \log p_\theta$, and Fisher information $I_\theta = \mathbb{E}_\theta[\dot{\ell}_\theta \dot{\ell}_\theta^T]$, if $\widehat{\theta}_n$ is the MLE (without privacy), then $\phi_{\aux} \equiv 0$ and $\dot{\theta}_0(x) = -I_{\theta_0}^{-1} \dot{\ell}_{\theta_0}(x)$. We have the following regularity result, which essentially appears as [52, Lemmas 8.14 & 25.23], though we include a proof in Appendix A.3.

Lemma 1. *Let $\mathcal{P}_{\sub,0} = \{P_h\}_{h \in \mathbb{R}^d} \subset \mathcal{P}$ be a QMD (8) sub-model at P_0 with score g, and assume that $\theta : \mathcal{P} \to \mathbb{R}^k$ is differentiable (12) relative to $\dot{\mathcal{P}}_0$ at P_0. Let $\widehat{\theta}_n$ be asymptotically linear (13) at $P_0 \times P_{\aux}$. Then for any sequence $h_n \to h \in \mathbb{R}^d$,*

$$\sqrt{n}\left(\widehat{\theta}_n - \theta(P_{h_n/\sqrt{n}})\right) \xrightarrow[P_{h_n/\sqrt{n}} \times P_{\aux}]{d} \mathsf{N}\left(0, \Sigma_0 + \Sigma_{\aux}\right).$$

Additionally, for any bounded continuous $L : \mathbb{R}^k \to \mathbb{R}_+$ and any $c < \infty$,

$$\lim_{n \to \infty} \sup_{\|h\| \le c} \mathbb{E}_{P_{h/\sqrt{n}}}\left[L(\sqrt{n}(\widehat{\theta}_n - \theta(P_{h/\sqrt{n}})))\right] = \mathbb{E}[L(Z)] \quad \text{where} \quad Z \sim \mathsf{N}(0, \Sigma_0 + \Sigma_{\aux}).$$

We use Lemma 1 to describe the local uniform convergence we seek. Define the rescaled losses $L_n(t) = L(\sqrt{n} \cdot t)$. We say an estimator $\widehat{\theta}_n$ and channel $Q \in \mathcal{Q}$ are *local minimax rate optimal* if for all $P_0 \in \mathcal{P}$ with QMD submodel $\mathcal{P}_{\sub,0} = \{P_h\} \subset \mathcal{P}$ passing through P_0 with score function g,

$$\sup_{c < \infty} \limsup_{n \to \infty} \sup_{\|h\| \le c/\sqrt{n}} \frac{\mathbb{E}_{Q \circ P_h}[L(\sqrt{n}(\widehat{\theta}_n(Z_1, \ldots, Z_n) - \theta(P_h)))]}{\mathfrak{M}_n^{\loc}(P_0, L_n, \mathcal{P}, \mathcal{Q})} \le C, \tag{14}$$

where the constant C is a numerical constant independent of L and P_0. Our general recipe is now apparent: demonstrate an asymptotically linear (14) locally private estimator $\widehat{\theta}_n$ with covariance $\Sigma_0 + \Sigma_{\aux}$. Then for *any* collection of losses $\{L\}$ for which we can lower bound $\mathfrak{M}_n^{\loc}(P_0, L_n, \mathcal{P}, \mathcal{Q}) \gtrsim \mathbb{E}[L(Z)]$ when $Z \sim \mathsf{N}(0, \Sigma_0 + \Sigma_{\aux})$, we obtain the convergence (14).

[2] Recalling [52, Ch. 25.3] and the Riesz representation theorem, the existence of this influence function is equivalent to the exists of a continuous linear map $\varphi : L^2(P_0) \to \mathbb{R}^k$ such that $\theta(P_h) - \theta(P_0) = \varphi(h^T g) + o(\|h\|)$.

3 Local minimax complexity and private estimation

We turn to our main goal of establishing localized minimax complexities for locally private estimation. We focus first on the squared error for simplicity in Section 3.1, giving consequences of our results. Instead of the Hellinger modulus (5), we show upper and lower bounds on the local minimax minimax complexity for private estimation using a local total variation modulus. We then give several example calculations, and provide a super-efficiency result. In Sections 3.2 and 3.3, we generalize to show how a total variation modulus characterizes local minimax complexity for nearly arbitrary losses, making our initial results on squared error corollaries.

3.1 Local minimax squared error and the variation distance modulus

We begin with a somewhat simplified setting, where we wish to estimate a parameter $\theta(P) \in \mathbb{R}$ of a distribution $P \in \mathcal{P}$, a collection of possible distributions, and we measure performance of an estimand θ via the squared error $L_{\rm sq}(\theta, P) = (\theta - \theta(P))^2$. For a family of distributions $\mathcal{P}$, the *modulus of continuity* with respect to the variation distance at distribution P_0 is

$$\omega_{\rm TV}(\delta; P_0, \mathcal{P}) := \sup_{P \in \mathcal{P}} \left\{|\theta(P) - \theta(P_0)| \text{ s.t. } \|P - P_0\|_{\rm TV} \le \delta\right\}. \tag{15}$$

As we shall see, this modulus of continuity fairly precisely characterizes the difficulty of locally private estimation of functionals. The key is that the modulus is with respect to *variation distance*. This is in contrast to the classical results we review in the introduction and Section 2.2 on optimal estimation, where the more familiar modulus of continuity with respect to Hellinger distance characterizes problem difficulty. As we illustrate, the difference between the Hellinger (5) and variation (15) moduli leads to different behavior for private and non-private estimation problems.

With this, we come to a corollary of our Theorem 1, to come in Section 3.2:

Corollary 1. *Let $\mathcal{Q}_\varepsilon$ be the collection of $(2, \varepsilon^2)$-locally Rényi private channels (Definition 2). Then*

$$\mathfrak{M}_n^{\rm loc}(P_0, L_{\rm sq}, \mathcal{P}, \mathcal{Q}_\varepsilon) \ge \frac{1}{16}\omega_{\rm TV}^2\left(\frac{1}{2\sqrt{2n\varepsilon^2}}; P_0, \mathcal{P}\right).$$

An identical bound (to within numerical constants) holds for ε-locally differentially private channels, as (recall Section 2.1) any ε-differentially private channel is $(2, O(1)\varepsilon^2)$-Rényi private. In (nearly) simultaneous independent work to the original version of this paper on the `arXiv`, Rohde and Steinberger [47] provide a global (2) minimax lower bound via a global modulus of continuity with respect to variation distance, extending [18, 19, 20] to the private case. The main difference is our focus: while they, similar to [25], demonstrate that private minimax rates depart from non-private ones, our focus is on instance-specific bounds. Consequently, Rohde and Steinberger study linear functionals $\theta(P)$, designing estimators to achieve the global minimax risk, while we allow nonlinear functionals and develop estimators that must achieve the refined local minimax complexity, with the hope that we may calculate practically useful quantities akin to classical information bounds [52, 39]. (As an aside, we also provide lower bounds for weaker forms of privacy.)

We can provide a converse to Corollary 1 that (nearly) characterizes the local minimax error by the modulus of continuity. Indeed, Proposition 2 to come implies that for $\varepsilon \le \frac{3}{2}$, we have

Corollary 2. *Let $\mathcal{Q}_\varepsilon$ be all non-interactive ε-differentially private channels (Def. 1). Then*

$$\mathfrak{M}_n^{\rm loc}(P_0, L_{\rm sq}, \mathcal{P}, \mathcal{Q}_\varepsilon) \le 2\sup_{\tau \ge 0} \omega_{\rm TV}^2\left(\frac{5\sqrt{2}\tau}{\sqrt{n\varepsilon^2}}; P_0, \mathcal{P}\right) e^{-\tau^2}.$$

Exactly as in inequality (7), whenever the modulus $\omega_{\rm TV}$ grows at most polynomially—so that there exist $B, \beta < \infty$ such that $\omega_{\rm TV}(c\delta; P_0, \mathcal{P}) \le Bc^\beta \omega_{\rm TV}(\delta; P_0, \mathcal{P})$ for $c > 1$, we have

$$\mathfrak{M}_n^{\rm loc}(P_0, L_{\rm sq}, \mathcal{P}, \mathcal{Q}) \le C_{B,\beta} \omega_{\rm TV}^2 \left(\frac{C_1}{\sqrt{n\varepsilon^2}}; P_0, \mathcal{P} \right)$$

where C_1 is a numerical constant and $C_{B,\beta}$ depends on B, β only. We note that we have thus far characterized the local minimax benchmark but have provided no estimator uniformly achieving it.

3.1.1 Example moduli of continuity

It is instructive to give examples of the local modulus and connect them to estimation rates. We give three mean estimation examples—a fully nonparametric setting, a collection of distributions $\mathcal{P}$ with bounded variance, and a Bernoulli estimation problem—where we see that the variation modulus (15) is essentially independent of the distribution P_0, in distinction with the Hellinger modulus (5). After these, additional examples will highlight that this is not always the case.

Example 1 (Bounded mean estimation)**:** Let $\mathcal{X} \subset \mathbb{R}$ be a bounded set and $\mathcal{P} := \{P : \operatorname{supp} P \subset \mathcal{X}\}$ be the collection of distributions supported on $\mathcal{X}$. Using the shorthand $\theta_0 = \theta(P_0) = \mathbb{E}_{P_0}[X]$, we claim the following upper and lower bounds:

$$\delta \cdot \sup_{x \in \mathcal{X}} |x - \theta_0| \le \omega_{\rm TV}(\delta; P_0, \mathcal{P}) \le 2\delta \cdot \sup_{x \in \mathcal{X}} |x - \theta_0|, \tag{16}$$

so that the local modulus is nearly independent of P_0. To see the lower bound (16), for any $x \in \mathcal{X}$, define $P_x = (1 - \delta)P_0 + \delta \cdot \mathbf{1}_x$, where $\mathbf{1}_x$ denotes a point mass at x. Then $\|P_x - P_0\|_{\rm TV} \le \delta$, so $\omega_{\rm TV}(\delta) \ge \sup_{x \in \mathcal{X}} |\theta_0 - \theta(P_x)| = \delta \cdot \sup_{x \in \mathcal{X}} |x - \theta_0|$. The upper bound (16) is straightforward:

$$|\theta(P) - \theta_0| = \left| \int (x - \theta_0)(dP(x) - dP_0(x)) \right| \le 2 \sup_{x \in \mathcal{X}} |x - \theta_0| \, \|P - P_0\|_{\rm TV}$$

for all $P \in \mathcal{P}$, by the triangle inequality, which is our desired result.

On the other hand, the Hellinger modulus (5) (asymptotically) smaller. Let $\mathcal{P}$ be any collection of distributions with uniformly bounded fourth moment. We claim (see Appendix A.4 for proof) that there exist numerical constants $0 < c_0 \le c_1 < \infty$ such that for all small enough $\delta > 0$,

$$c_0 \sqrt{\operatorname{Var}_{P_0}(X)} \cdot \delta \le \omega_{\rm hel}(\delta; P_0, \mathcal{P}) \le c_1 \sqrt{\operatorname{Var}_{P_0}(X)} \cdot \delta \quad \text{and} \quad \lim_{\delta \downarrow 0} \frac{\omega_{\rm hel}(\delta; P_0, \mathcal{P})}{\sqrt{8 \operatorname{Var}_{P_0}(X) \delta}} = 1. \tag{17}$$

The variance $\operatorname{Var}_{P_0}(X)$ of the distribution P_0 thus determines the local Hellinger modulus (5). $\diamond$

Example 2 (Means with bounded variance)**:** We specialize Example 1 by considering distributions on $\mathcal{X}$ with a variance bound σ^2, defining $\mathcal{P} := \{P : \operatorname{supp} P \subset \mathcal{X}, \operatorname{Var}_P(X) \le \sigma^2\}$. We consider the case that $\operatorname{Var}_{P_0}(X) < \sigma^2$; we claim that the bounds (16) again hold for small $\delta > 0$. The upper bound is immediate. The lower bound follows by noting that if $P_x = (1 - \delta)P_0 + \delta \cdot \mathbf{1}_x$, then $\operatorname{Var}_{P_x}(X) = \operatorname{Var}_{P_x}(X - \theta_0) = (1 - \delta)\operatorname{Var}_{P_0}(X) + \delta(1 - \delta)(x - \theta_0)^2$, so that for small enough δ we have $\operatorname{Var}_{P_x}(X) \le \sigma^2$ and the identical lower bound (16) holds. $\diamond$

Example 3 (Modulus of continuity for Bernoulli parameters)**:** We further restrict to binary random variables, so that the problem is parametric. Let $\mathsf{Bern}(\theta)$ be the Bernoulli distribution with mean θ, and $\mathcal{P} = \{\mathsf{Bern}(\theta)\}_{\theta \in [0,1]}$. We have $\|P_{\theta_0} - P_\theta\|_{\rm TV} = |\theta - \theta_0|$ and for $\delta \le \frac{1}{2}$, $\omega_{\rm TV}(\delta; P_{\theta_0}, \mathcal{P}) = \delta$. On the other hand, Eq. (17) shows that $\omega_{\rm hel}^2(\delta; P_{\theta_0}, \mathcal{P}) = 8 \frac{\delta^2}{\theta_0(1 - \theta_0)}(1 + o(1))$. The Hellinger modulus is local to θ_0, while the local variation modulus is global. $\diamond$

Summarizing examples 1–3, in each case the local TV-modulus (15) of distributions supported on $\mathcal{X}$ *must* scale as the diameter of $\mathcal{X}$—essentially identical to a global modulus of continuity over the full set $\mathcal{P} = \{P : \operatorname{supp} P \subset \mathcal{X}\}$—while the Hellinger modulus (5) scales linearly in $\sqrt{\operatorname{Var}_{P_0}(X)}$. This lack of locality in the local modulus for variation distance has consequences for estimation, which we can detail by applying Corollary 1:

Corollary 3 (Locally private mean estimation)**.** *Let $\mathcal{X}$ be bounded, let $\mathcal{P}$ be any of the collections of distributions in Examples 1–3, and let $\mathcal{Q}_\varepsilon$ be the collection of $(2, \varepsilon^2)$-Rényi locally private channels. There exists a numerical constant $c > 0$ such that for any $P_0 \in \mathcal{P}$ (where in the case of Example 2 we require $\operatorname{Var}_{P_0}(X) < \sigma^2$), for all large enough n*

$$c\frac{\operatorname{diam}(\mathcal{X})^2}{n\varepsilon^2} + c\frac{\operatorname{Var}_0(X)}{n} \le \mathfrak{M}_n(P_0, L_{\mathrm{sq}}, \mathcal{P}, \mathcal{Q}) \le \frac{\operatorname{diam}(\mathcal{X})^2}{2n\varepsilon^2} + \frac{\operatorname{Var}(X)}{n}.$$

Standard mechanisms [25] achieve the upper bound in Corollary 3: letting $Z_i = X_i + \frac{\operatorname{diam}(\mathcal{X})}{\varepsilon} W_i$ for $W_i \stackrel{\mathrm{iid}}{\sim} \mathsf{Lap}(1)$ gives an ε-differentially private view of X_i; define the estimator $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n Z_i$. This highlights the difference with the non-private case, where the matching upper and lower bounds are $\operatorname{Var}_{P_0(X)}/n$, while in the private case the diameter of $\mathcal{X}$ is central.

Yet the local total-variation (private) modulus can depend strongly on the distribution P_0 and set $\mathcal{P}$ of potential alternatives, a point to which we will return later. Two simple examples illustrate.
Example 4 (Modulus of continuity for a normal mean)**:** Let $\mathcal{P} = \{\mathsf{N}(\theta, \sigma^2)\}_{\theta \in \mathbb{R}}$ for a known variance σ^2. Letting ϕ and Φ be the standard normal p.d.f. and c.d.f., respectively, for any pair $\theta_0, \theta_1 \in \mathbb{R}$ with $\Delta = |\theta_0 - \theta_1|$, we then have $\|\mathsf{N}(\theta_0, \sigma^2) - \mathsf{N}(\theta_1, \sigma^2)\|_{\mathrm{TV}} = \Phi(\Delta/2\sigma) - \Phi(-\Delta/2\sigma)$. Solving for the modulus gives that for any $P_0 \in \mathcal{P}$,

$$\omega_{\mathrm{TV}}(\delta; P_0, \mathcal{P}) = \frac{\sigma\delta}{\phi(0)}(1 + O_\sigma(\delta))$$

as $\delta \to 0$. It is possible but tedious to extend this to cases with an unknown variance, so that $\mathcal{P} = \{\mathsf{N}(\theta, \sigma^2)\}_{\theta \in \mathbb{R}, \sigma^2 < \infty}$, in which case we obtain $\omega_{\mathrm{TV}}(\delta; P_0, \mathcal{P}) \asymp \sqrt{\operatorname{Var}_{P_0}(X)}\delta$ as $\delta \to 0$. $\diamond$

Other parametric families also have stronger dependence on the local distribution P.
Example 5 (Exponential distributions)**:** Let $p_\theta(x) = \frac{1}{\theta}\exp(-\frac{x}{\theta})\mathbf{1}\{x \ge 0\}$ be the density of an exponential distribution with scale θ, and $\mathcal{P}$ be the collection of such distributions. Let $\tau, \theta > 0$, and set $x_\star = \frac{\theta\tau}{\theta - \tau}\log\frac{\theta}{\tau}$. The variation distance between two exponential distributions is then $\|P_\theta - P_\tau\|_{\mathrm{TV}} = |e^{-x_\star/\theta} - e^{-x_\star/\tau}|$. For $\theta = \tau + \delta$ (or $\tau = \theta - \delta$), we thus obtain that

$$\|P_\theta - P_{\theta-\delta}\|_{\mathrm{TV}} = \exp\left(-\frac{\theta}{\delta}\log\frac{1}{1-\delta/\theta}\right)\left|\frac{1}{1-\delta/\theta} - 1\right| = e^{-1}\left[\frac{|\delta|}{\theta} + O_\theta(\delta)\right],$$

and $\|P_\theta - P_{\theta-\delta}\|_{\mathrm{TV}}$ is monotonic in $|\delta|$. Eliding details, we thus find that

$$\omega_{\mathrm{TV}}(\delta; P_\theta, \mathcal{P}) = \theta\delta \cdot (e + O_\theta(\delta)),$$

which evidently is local to θ. $\diamond$

3.1.2 Super-efficiency for squared error

To demonstrate that the local modulus of continuity is the "correct" lower bound on estimation, we consider the third of the desiderata for a strong lower bound that we idenfity in the introduction: a super-efficiency result [8, 10, 49] showing that any estimator substantially outperforming the local minimax benchmark at a given distribution P_0 necessarily suffers higher expected error for some other distribution P_1. As a corollary of Proposition 3 to come, we establish the following result.

Corollary 4. *Let Q be a sequentially interactive $(2, \varepsilon^2)$-Rényi-private channel (Def. 2). If for some $\eta \in [0, 1]$ the estimator $\widehat{\theta}$ satisfies*

$$\mathbb{E}_{Q \circ P_0}[(\widehat{\theta}(Z_{1:n}) - \theta_0)^2] \le \eta \omega_{\rm TV}^2 \left(\frac{1}{\sqrt{4n\varepsilon^2}}; P_0, \mathcal{P} \right),$$

then for all $t \in [0, 1]$ there exists a distribution $P_1 \in \mathcal{P}$ such that

$$\mathbb{E}_{Q \circ P_1}[(\widehat{\theta}(Z_{1:n}) - \theta(P_1))^2] \ge \frac{1}{8} \left[1 - \eta^{\frac{(1-t)}{2}} \right]_+^2 \omega_{\rm TV}^2 \left(\frac{1}{4} \sqrt{\frac{t \log \frac{1}{\eta}}{n\varepsilon^2}}; P_1, \mathcal{P} \right).$$

Unpacking the corollary by ignoring constants (e.g., set $t = \frac{1}{2}$), we see (roughly) the following result: if an estimator achieves expected squared error less (by a factor $\eta < 1$) than the squared modulus of continuity at P_0, it must have squared error scaling with the modulus for a radius $\sqrt{\log \frac{1}{\eta}}$-times larger. For example, considering the sample mean examples 1–3, we see that in any of the settings, there exists a numerical constant $c > 0$ such that if $\widehat{\theta}_n$ is locally private and

$$\mathbb{E}_{P_0} \left[(\widehat{\theta}_n - \theta(P_0))^2 \right] \le \eta \frac{\operatorname{diam}(\mathcal{X})^2}{n\varepsilon^2}$$

for some $0 < \eta < 1$, then there exists $P_1 \in \mathcal{P}$ such that for all large enough n,

$$\mathbb{E}_{P_1} \left[(\widehat{\theta}_n - \theta(P_1))^2 \right] \ge c \frac{\operatorname{diam}(\mathcal{X})^2}{n\varepsilon^2} \cdot \log \frac{1}{\eta}.$$

3.2 Local private minimax risk for general losses

We return to prove our local minimax upper and lower bounds for general losses, along the way proving the claimed corollaries. Recall that we use any symmetric quasiconvex loss $L : \mathbb{R}^d \to \mathbb{R}_+$ satisfying $L(\mathbf{0}) = 0$. Then for a family of distributions $\mathcal{P}$, the *modulus of continuity* associated with the loss L at the distribution P_0 is

$$\omega_{L,\rm TV}(\delta; P_0, \mathcal{P}) := \sup_{P \in \mathcal{P}} \left\{ L \left(\frac{\theta(P_0) - \theta(P)}{2} \right) \text{ s.t. } \|P - P_0\|_{\rm TV} \le \delta \right\}, \tag{18}$$

where the normalization by $\frac{1}{2}$ is convenient for our proofs. We then have our first main theorem, which lower bounds the local minimax risk using the modulus (18) in analogy to Proposition 1. We defer the proof to Section 7.2, where we also present a number of new strong data-processing inequalities to prove it.

Theorem 1. *Let $\mathcal{Q}$ be the collection of $(2, \varepsilon^2)$-locally Rényi differentially private channels (Definition 2). Let $c_{\rm conv} = 1$ if L is convex and 2 otherwise. Then for any distribution P_0, we have*

$$\mathfrak{M}_n^{\rm loc}(P_0, L, \mathcal{P}, \mathcal{Q}) \ge \frac{1}{4c_{\rm conv}} \omega_{L,\rm TV} \left(\frac{1}{2\sqrt{2n\varepsilon^2}}; P_0, \mathcal{P} \right).$$

Corollary 1 is then immediate: for the squared error, $L(\frac{1}{2}(\theta(P_0) - \theta(P_1))) = \frac{1}{4}(\theta(P_0) - \theta(P_1))^2$. Note also, as in the discussion after Corollary 1, that this implies the lower bound $\omega_{L,\rm TV}(O(1)/\sqrt{n\varepsilon^2}; P_0, \mathcal{P})$ on any ε-locally differentially private procedure.

An upper bound in the theorem is a somewhat more delicate argument, and for now we do not provide procedures achieving the lower bound. Instead, under reasonable conditions on the loss, we can show the (partial) converse that the modulus $\omega_{L,\rm TV}$ describes the local minimax complexity.

Condition C.1 (Growth inequality). *There exists $\gamma < \infty$ such that for all $t \in \mathbb{R}^d$,*

$$L(t) \le \gamma L(t/2).$$

For example, for the squared error we have $L_{\rm sq}(t/2) = t^2/4 = L_{\rm sq}(t)/4$, giving $\gamma = 4$. In Appendix B.1, we prove the following partial converse to Theorem 1.

Proposition 2. *Let Condition C.1 on the loss L hold. Let $\varepsilon \ge 0$ and $\delta_\varepsilon = \frac{e^\varepsilon}{e^\varepsilon+1} - \frac{1}{2}$, and let $\mathcal{Q}$ be the collection of non-interactive ε-differentially private channels (Definition 1). Then*

$$\mathfrak{M}_n^{\rm loc}(P_0, L, \mathcal{P}, \mathcal{Q}) \le 2\gamma \sup_{\tau \ge 0} \left\{ \omega_{L,\rm TV}\left(\frac{\sqrt{2}\tau}{\delta_\varepsilon \sqrt{n}}; P_0, \mathcal{P}\right) e^{-\tau^2} \right\}.$$

The proposition as written is a bit unwieldy, so we unpack it slightly. We have $\delta_\varepsilon \ge \min\{\frac{\varepsilon}{5}, 1/3\}$, so for each $P_1 \in \mathcal{P}$ there exists a non-interactive ε-DP channel Q and estimator $\widehat{\theta}$ such that

$$\max_{P \in \{P_0, P_1\}} \mathbb{E}_{P,Q}\left[L(\widehat{\theta}(Z_{1:n}), P)\right] \le 2\gamma \cdot \sup_{\tau \ge 0} \omega_{L,\rm TV}\left(\frac{3\sqrt{2}\tau}{\sqrt{n \min\{9\varepsilon^2/25, 1\}}}, P_0, \mathcal{P}\right) e^{-\tau^2}.$$

Typically, this supremum is achieved at $\tau = O(1)$, so that Proposition 2 shows that the modulus (18) at radius $\frac{O(1)}{\sqrt{n\varepsilon^2}}$ characterizes the local minimax risk to constants for $\varepsilon \lesssim 1$. Appropriate assumptions, including the following condition on the modulus of continuity, allow more precision.

Condition C.2 (Polynomial growth). *For each P_0, there exist $\alpha, \beta < \infty$ such that for all $c \ge 1$*

$$\omega_{L,\rm TV}(c\delta; P_0, \mathcal{P}) \le (\beta c)^\alpha \omega_{L,\rm TV}(\delta; P_0, \mathcal{P}).$$

Condition C.2 is similar to the typical Hölder-type continuity properties assumed on the modulus of continuity for estimation problems [18, 19]. It holds, for example, for nonparametric mean estimation problems (recall Example 1), and we make this more concrete after the following corollary.

Corollary 5. *In addition to the conditions of Proposition 2, let Condition C.2 hold. Then*

$$\mathfrak{M}^{\rm loc}(P_0, L, \mathcal{P}, \mathcal{Q}) \le \gamma \beta^\alpha e^{\frac{\alpha}{2}[\log \frac{\alpha}{2} - 1]} \omega_{L,\rm TV}\left(\frac{\sqrt{2}}{\delta_\varepsilon \sqrt{n}}; P_0, \mathcal{P}\right).$$

Proof. We apply Proposition 2. For $\tau \le 1$, it already gives the result; otherwise, we use the growth condition C.2 to obtain $\mathbb{E}_0[L(\widehat{\theta} - \theta(P_0))] + \mathbb{E}_1[L(\widehat{\theta} - \theta(P_1))] \le 2\gamma \omega_{L,\rm TV}(\frac{\sqrt{2}}{\delta_\varepsilon\sqrt{n}}; P_0, \mathcal{P}) \beta^\alpha \sup_{\tau \ge 1} \tau^\alpha e^{-\tau^2}$. Noting that $\sup_{\tau \ge 0} \tau^\alpha e^{-\tau^2} = (\alpha/2)^{\alpha/2} e^{-\alpha/2}$ gives the result. □

We generally expect Condition C.2 to hold, so that the modulus describes the risk. Indeed, for any loss $L : \mathbb{R}^d \to \mathbb{R}_+$ satisfying Conditition C.1, we immediately obtain condition (C.2) whenever $\omega_{\rm TV}(\cdot)$ satisfies the condition, which it does for each of Examples 1–3 and (locally) 4–5.

3.3 Super-efficiency

We provide our general super-efficiency result via a constrained risk inequality [8, 22]. Our result applies in the typical setting in which the loss is $L(t) = \Phi(\|\theta - \theta(P)\|_2)$ for some increasing function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$, and we use the shorthand $R(\widehat{\theta}, \theta, P) := \mathbb{E}_P[\Phi(\|\widehat{\theta}(Z) - \theta\|_2)]$ for the risk (expected loss) of the estimator $\widehat{\theta}$ under the distribution P. We build off the approach of Brown and Low [8, Thm. 1], who show that if $\widehat{\theta}$ has squared error for a parameter θ under a distribution P_0, then its risk under a distribution P_1 close to P_0 may be large (see also [49, Thm. 6]). The next proposition, whose proof we provide in Section 7.3, extends this to show that improvement over our modulus of continuity lower bound at a point P_0 implies worse performance elsewhere.

Proposition 3. *Let Q be a sequentially interactive $(2, \varepsilon^2)$-Rényi private channel (Def. 2) with associated marginal distributions $M_a^n(\cdot) = \int Q(\cdot \mid x_{1:n}) dP_a^n(x_{1:n})$. Let Condition C.1 hold with parameter γ. If for some $\eta \in [0,1]$ the estimator $\widehat{\theta}$ satisfies*

$$R(\widehat{\theta}, \theta_0, M_0^n) \le \eta \omega_{L,\mathrm{TV}}\left(\frac{1}{\sqrt{4n\varepsilon^2}}; P_0, \mathcal{P}\right),$$

then for all $t \in [0,1]$ there exists a distribution $P_1 \in \mathcal{P}$ such that

$$R(\widehat{\theta}, \theta(P_1), M_1^n) \ge \frac{1}{2\gamma}\left[1 - \eta^{\frac{(1-t)}{2}}\right]_+^2 \omega_{L,\mathrm{TV}}\left(\frac{1}{4}\sqrt{\frac{t \log \frac{1}{\eta}}{n\varepsilon^2}}; P_1, \mathcal{P}\right).$$

The proposition depends on a number of constants, but roughly, it shows (for small enough η, where we simplify by taking $t = 1/2$) that if an estimator $\widehat{\theta}$ is super-efficient at P_0, in that $R(\widehat{\theta}, \theta_0, M_0^n) \le \eta \cdot \omega_{L,\mathrm{TV}}(1/\sqrt{4n\varepsilon^2}; P_0)$, then there exists $c > 0$ such that for some P_1 we have $R(\widehat{\theta}, \theta_1 M_1^n) \ge c \cdot \omega_{L,\mathrm{TV}}(\sqrt{\log(1/\eta)}/\sqrt{32n\varepsilon^2}; P_1)$. In this sense, our bounds are sharp: any estimator achieving much better risk than the local modulus at a distribution P_0 must pay elsewhere.

4 The private information

The ansatz of finding a locally most difficult problem via the local variation modulus of continuity (15) gives an approach to lower bounds that leads to non-standard behavior for a number of classical and not-so-classical problems in locally private estimation. In this section, we investigate examples in several one-dimensional parametric problems, showing how local privacy leads to a different geometry of local complexities than classical cases. Our first step is to define the L^1 *information*, a private analogue of the Fisher Information that governs the complexity of estimation under local privacy. We illustrate the private L^1 information for several examples, including of the mean of Bernoulli random variable, the scale of an exponential random variable, and in linear and logistic models (Sec. 4.2), showing the consequences of (locally) private estimation in one dimension. Our last two sections develop locally private algorithms for achieving the local minimax risk. The first of these (Sec. 4.3) describes private stochastic gradient algorithms and their (locally uniform) asymptotics, while the last (Sec. 4.4) develops a new locally private algorithm based on Fisher scoring to achieve the L^1 information in one-dimensional exponential families.

4.1 Private analogues of the Fisher Information

Our first set of results builds off of Theorem 1 by performing asymptotic approximations to the variation distance for regular parametric families of distributions. One major consequence of our results is that, under the notions of locally private estimation we consider, the classical Fisher information is *not* the right notion of complexity in estimation, though an analogy is possible. Again we emphasize that we hope to characterize complexity only to numerical constant factors, seeking the problem-dependent terms that analogize the classical information.

We begin by considering parametric families that allow analogues of Le Cam's quadratic mean differentiability (QMD) [52, Ch. 7]. Consider a 1-dimensional parametric collection $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ with dominating measure μ and densities $p_\theta = dP_\theta/d\mu$. Analogizing the QMD definition (8) from the Hellinger to the variation distance, we say $\mathcal{P}$ *is L^1-differentiable at θ_0* with score $\dot{\ell}_{\theta_0} : \mathcal{X} \to \mathbb{R}$ if

$$\int |p_{\theta_0 + h} - p_{\theta_0} - h\dot{\ell}_{\theta_0} p_{\theta_0}| d\mu = o(|h|). \tag{19}$$

For QMD families, L^1-differentiability is automatic (see Appendix C.1 for a proof).

Lemma 2. *Let the family $\mathcal{P} := \{P_\theta\}_{\theta\in\Theta}$ be QMD (8) at the point θ_0. Then $\mathcal{P}$ is L^1-differentiable at θ_0 with identical score $\dot\ell_\theta$ to the QMD case.*

Recalling (as in Sec. 2.2) that for QMD families (8), the Fisher information is $I_\theta = \mathbb{E}_{P_\theta}[(\dot\ell_\theta)^2]$, and $d^2_{\rm hel}(P_{\theta+h}, P_\theta) = \frac{1}{8} I_\theta h^2 + o(h^2)$, by analogy, we define the *L^1-information* as

$$J_{\theta_0} := \mathbb{E}_{P_\theta}[|\dot\ell_\theta|] = \int \left|\dot\ell_{\theta_0}(x)\right| dP_{\theta_0}(x). \tag{20}$$

We can then locally approximate the total variation distance by the L^1-information:

$$\|P_{\theta+h} - P_\theta\|_{\rm TV} = \frac{1}{2} J_\theta |h| + o(|h|).$$

We consider a somewhat general setting in which we wish to estimate the value $\psi(\theta)$ of a functional $\psi : \Theta \to \mathbb{R}$, where ψ is $\mathcal{C}^1$ near θ_0. We measure our error by $L(\psi(\theta) - \psi(\theta_0))$, and give a short proof of the next proposition via Theorem 1 in Appendix B.3.

Proposition 4. *Let $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$ be L^1-differentiable at θ_0 with score $\dot\ell_{\theta_0}$, and $\mathbb{E}_{\theta_0}[|\dot\ell_{\theta_0}|] > 0$. Let $\mathcal{Q}_\varepsilon$ be the family of $(2,\varepsilon^2)$-Rényi locally private sequentially interactive channels. Then for an $N = N(\psi, \theta_0, \mathcal{P}, \varepsilon)$ depending only on ψ, θ_0, the family $\mathcal{P}$, and privacy level ε, for all $n \ge N$*

$$\mathfrak{M}_n^{\rm loc}(P_{\theta_0}, L, \mathcal{P}, \mathcal{Q}_\varepsilon) \ge \frac{1}{8} \cdot L\left(\frac{1}{5\sqrt{2n\varepsilon^2}} \cdot J_{\theta_0}^{-1} \psi'(\theta_0)\right).$$

To obtain a matching upper bound we require the identifiability assumption A1. We make a simplifying assumption that the loss L is reasonably behaved, in that there exists a numerical constant $C < \infty$ and $\beta \in \mathbb{R}_+$ such that $L(at) \le C a^\beta L(t)$ for all $a \ge 1$. Then, even when $\mathcal{Q}_\varepsilon$ is the collection of ε-locally differentially private non-interactive channels (which, by the discussion following Definition 2, is more limiting than channels in Proposition 4), we can upper bound the local minimax risk.

Corollary 6. *Let the family $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$ be L^1 differentiable at θ_0 with score $\dot\ell_{\theta_0}$ and $\mathbb{E}_{\theta_0}[|\dot\ell_{\theta_0}|] > 0$ and additionally let the above assumptions hold. Let $\varepsilon \le 2$. Then there exists a numerical constant $C < \infty$ and a $\delta_0 = \delta_0(\psi, \theta_0, \mathcal{P})$ depending only on ψ, θ_0, and the family $\mathcal{P}$ such that*

$$\mathfrak{M}_n^{\rm loc}(P_{\theta_0}, L, \mathcal{P}, \mathcal{Q}_\varepsilon) \le C \max\left\{(\beta/2e)^{\beta/2} L\left(\frac{\psi'(\theta_0)}{J_{\theta_0}} \frac{1}{\sqrt{n\varepsilon^2}}\right), L(\operatorname{diam}(\psi(\Theta))) e^{-\delta_0^2 n\varepsilon^2}\right\}.$$

The proof (see Appendix B.2) is a straightforward modification of that of Claim 2.1. Proposition 4 and Corollary 6 show that for a (one-dimensional) parametric family $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$, the L^1 information describes the local modulus to within numerical constants for small δ: for the modulus (15), there are numerical constants $0 < c_{\rm low} \le c_{\rm high} < \infty$ such that

$$\omega_{\rm TV}(\delta; P_{\theta_0}, \{P_\theta\}_{\theta\in\Theta}) \in [c_{\rm low}, c_{\rm high}] \cdot \frac{\delta}{J_{\theta_0}} \quad \text{for all small } \delta > 0.$$

(A more general result holds; see Theorem 2 to come.) In analogy with Claim 2.1, where the Fisher information I_{θ_0} characterizes the local minimax squared error in non-private estimation, the L^1 information is an alternative characterization—to within numerical constants—of the local minimax risk in the locally private case.

As an alternative way to understand the proposition and corollary, we can rescale the losses (in analogy with the local asymptotic approach [52, Ch. 7]), and consider the sequence $L_n(t) = L(\sqrt{n} \cdot t)$, where for simplicity we take $L(t) = \min\{t^k, B\}$ for some $k, B < \infty$ (more generally, we could allow L to be bounded and nondecreasing). Then under the conditions of Corollary 6, for $\mathcal{Q}$ a collection of ε-locally private channels,

$$L\left(\frac{\psi'(\theta_0)}{J_{\theta_0}\varepsilon}\right) \lesssim \mathfrak{M}_n^{\rm loc}(P_{\theta_0}, L_n, \mathcal{P}, \mathcal{Q}) \lesssim L\left(\frac{\psi'(\theta_0)}{J_{\theta_0}\varepsilon}\right)$$

for all large n. The analogous bounds in the non-private case are $L(I_{\theta_0}^{-1/2}\psi'(\theta_0))$, the local asymptotic complexity for one-dimensional functionals [52, Ch. 7]. Using Lemma 2, we have

$$J_{\theta_0} = \mathbb{E}_{\theta_0}[|\dot\ell_{\theta_0}|] \le \mathbb{E}_{\theta_0}[\dot\ell_{\theta_0}^2]^{1/2} = I_{\theta_0}^{1/2},$$

so the L^1 information is at most the Fisher information. In some cases, as we shall see in Sec. 4.2, it can be much smaller, while in others the information measures are equal to numerical constants.

4.2 Examples and attainment

We consider the local minimax complexity and L^1-information in four different examples—estimation of Bernoulli and logistic the scale of an exponential random variable, and a 1-dimensional linear regression problem—which are particularly evocative. In each, we derive the L^1-information, applying Proposition 4 and Corollary 6 to characterize the private local minimax complexity. Throughout this section, we let $\mathcal{Q}_\varepsilon$ be the collection of ε-locally differentially private channels, where $\varepsilon = O(1)$ for simplicity. To keep the examples short, we do not always provide algorithms, but we complete the picture via a private stochastic gradient method in Section 4.3.

Example 6 (Example 3 continued)**:** For $P_\theta = \mathsf{Bern}(\theta)$, the score is $\dot\ell_\theta(x) = \frac{x-\theta}{\theta(1-\theta)}$, giving L^1-information $J_\theta = \frac{1}{2}\mathbb{E}_\theta[|\dot\ell_\theta|] = 1$ for all θ and Fisher information $I_\theta = \mathbb{E}_\theta[\dot\ell_\theta^2] = \frac{1}{\theta(1-\theta)}$. Thus

$$\mathfrak{M}_n^{\rm loc}(P_{\theta_0}, L_{\rm sq}, \mathcal{P}, \mathcal{Q}_\varepsilon) \asymp \frac{1}{n\varepsilon^2}$$

for $\varepsilon = O(1)$. The lower bound is Proposition 4, For the upper bound, consider the randomized-response mechanism [55] that releases $Z_i = X_i$ with probability $\frac{e^\varepsilon}{1+e^\varepsilon}$ and $Z_i = 1 - X_i$ otherwise, which is ε-differentially private. The plug-in estimate $\widehat\theta_n = \frac{(1+e^\varepsilon)\overline{Z}_n - 1}{e^\varepsilon - 1}$ is unbiased for θ_0 and has

$$\mathbb{E}_0[(\widehat\theta_n - \theta_0)^2] = \mathrm{Var}_0(\widehat\theta_n) = \left(\frac{1+e^\varepsilon}{e^\varepsilon - 1}\right)^2 \mathrm{Var}(\overline{Z}_n) \le \frac{(1+e^\varepsilon)^2}{4n(e^\varepsilon-1)^2} \lesssim \frac{1}{n\varepsilon^2}.$$

The sample mean achieves risk $\frac{\theta_0(1-\theta_0)}{n}$, so the gap in efficiency between private and non-private estimation grows when $\theta_0(1-\theta_0) \to 0$. Roughly, the noise individual randomization introduces (the statistical cost of privacy) dominates the non-private (classical) statistical cost. $\diamond$

As a brief remark, the paper [41] gives optimal asymptotics for Bernoulli problems with randomized response channels; such precise calculations are challenging when allowing arbitrary channels $\mathcal{Q}_\varepsilon$.

Example 7 (Private one-dimensional logistic regression)**:** A similar result to Bernoulli estimation that may be more striking holds for logistic regression, which is relevant for modern privacy applications, such as learning a classifier from (privately shared) user data [30, 1, 3]. To see this, let P_0 be the distribution on pairs $(x, y) \in \{-1, 1\}^2$ satisfying the logistic regression model

$$P_0(y \mid x) = \frac{1}{1 + e^{-y\theta_0 x}} \quad \text{and} \quad P_0(x = \pm 1) = \frac{1}{2}. \tag{21}$$

Here we wish to construct a classifier that provides good confidence estimates $p(y \mid x)$ of a label y given covariates x. We expect in the logistic regression model (21) that large parameter values θ_0 should make estimating a classifier easier, as is the case without privacy. To make this concrete, we measure the error in estimating the conditional probability $p_\theta(y \mid x)$,

$$L_{\rm pred}(\theta, \theta_0) := \mathbb{E}_{P_0}\left[|p_\theta(Y \mid X) - p_{\theta_0}(Y \mid X)|\right].$$

A calculation gives $L_{\rm pred}(\theta, \theta_0) = |\phi(\theta) - \phi(\theta_0)|$, where $\phi(t) = 1/(1+e^t)$ is the logistic function. The Fisher information for the parameter θ in this model is $I_\theta = \phi(\theta)\phi(-\theta)$, so a change of variables gives $I_{\phi(\theta)} = I_\theta/(\phi'(\theta))^2$, and as $0 \le L_{\rm pred} \le 1$, the non-private local minimax complexity is thus

$$\mathfrak{M}_n^{\rm loc}(P_0, L_{\rm pred}, \mathcal{P}, \{\mathrm{id}\}) \asymp \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{2 + e^{\theta_0} + e^{-\theta_0}}} \asymp \frac{1}{\sqrt{n}} e^{-|\theta_0|/2} \tag{22}$$

by Proposition 1 (and an analogous calculation to Claim 2.1). The delta method shows that the standard maximum likelihood estimator asymptotically achieves this risk.

The private complexity is qualitatively different. Noting that $XY \in \{\pm 1\}$ is a sufficient statistic for the logistic model (21), then applying Example 6 via Proposition 4 and Corollary 6 to the loss $L_{\rm pred}$, we obtain numerical constants $0 < c_0 \le c_1 < \infty$ such that for all large enough n,

$$\frac{c_0}{\sqrt{n\varepsilon^2}} \le \mathfrak{M}_n^{\rm loc}(P_0, L_{\rm pred}, \mathcal{P}, \mathcal{Q}_\varepsilon) \le \frac{c_1}{\sqrt{n\varepsilon^2}}.$$

By comparing this private local minimax complexity with the classical complexity (22), we see there is an exponential gap (in the parameter $|\theta_0|$) between the prediction risk achievable in private and non-private estimators—a non-trivial statistical price to pay for privacy. $\diamond$

While these examples have parameter-independent L^1-information, this is not always the case.

Example 8 (Exponential scale, Example 5 continued)**:** Let $p_\theta(x) = \frac{1}{\theta}\exp(-\frac{x}{\theta})\mathbf{1}\{x \ge 0\}$ be the density of an exponential distribution with scale θ, and $\mathcal{P} = \{P_\theta\}_{a \le \theta \le b}$, where $a < b$ are any finite positive constants. The standard score is $\dot\ell_\theta(x) = \frac{\partial}{\partial\theta}\log p_\theta(x) = -\frac{1}{\theta} + \frac{x}{\theta^2}$, yielding Fisher information $I_\theta = \frac{1}{\theta^2}$, so that the classical local minimax complexity for the squared error (recall Claim 2.1) is $\mathfrak{M}_n^{\rm loc}(P_\theta, L_{\rm sq}, \mathcal{P}, \{\mathrm{id}\}) \asymp \frac{\theta^2}{n}$. In this case, the private local minimax complexity satisfies

$$\mathfrak{M}_n^{\rm loc}(P_\theta, L_{\rm sq}, \mathcal{P}, \mathcal{Q}_\varepsilon) \asymp \frac{\theta^2}{n\varepsilon^2}.$$

To see this, note that the L^1-information (20) is $J_\theta = \mathbb{E}_\theta[|X/\theta^2 - 1/\theta|]$, so $\frac{1}{\theta} \lesssim J_\theta \lesssim \frac{1}{\theta}$. Proposition 4 and Corollary 6 then give the bounds. Thus, the private and non-private local minimax complexities differ (ignoring numerical constants) by the factor $1/\varepsilon^2$. In distinction from Examples 6 and 7, problems that are relatively easy in the classical setting (θ near 0) continue to be easy. $\diamond$

Example 9 (One-dimensional linear regression)**:** Consider a linear regression model where the data come in independent pairs (X_i, Y_i) satisfying

$$Y_i = \theta X_i + W_i \quad \text{where} \quad W_i \sim \mathsf{N}(0, \sigma^2)$$

and the target is to estimate $\theta \in \mathbb{R}$. Fixing the distribution of X and letting $\Theta \subset \mathbb{R}$ be a compact interval, we let $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$. We have negative log-likelihood $\ell_\theta(x, y) = \frac{1}{2\sigma^2}(x\theta - y)^2$ and score $\dot\ell_\theta(x, y) = \frac{1}{\sigma^2}x(x\theta - y)$, so $\dot\ell_{\theta_0}(x, y) = \frac{xw}{\sigma^2}$ for the noise $w = y - x\theta_0$. Calculating the L^1 information $J_\theta = \sqrt{2/\pi}\mathbb{E}[|X|]/\sigma$ and applying Proposition 4 and Corollary 6 yields

$$\mathfrak{M}_n^{\rm loc}(P_{\theta_0}, L_{\rm sq}, \mathcal{P}, \mathcal{Q}_\varepsilon) \asymp \frac{\sigma^2}{n\varepsilon^2\mathbb{E}[|X|]^2}. \tag{23}$$

Comparing the rates (23) with the non-private local minimax rate is instructive. Claim 2.1 shows that $\mathfrak{M}_n^{\mathrm{loc}}(P_{\theta_0}, L_{\mathrm{sq}}, \mathcal{P}, \{\mathrm{id}\}) \asymp \frac{\sigma^2}{n\mathbb{E}[X^2]}$. The local private minimax complexity (23) depends on X through $\mathbb{E}[|X|]^2$, while the non-private complexity above depends inversely on $\mathbb{E}[X^2]$. Thus, as X becomes more dispersed in that the ratio $\mathbb{E}[X^2]/\mathbb{E}[|X|]^2$ grows, the gap between private and non-private rates similarly grows. Intuitively, a dispersed X requires more individual randomization to protect private information in X, increasing the statistical price of privacy. $\diamond$

4.3 Attainment by stochastic gradient methods

The second of our major desiderata is to (locally) uniformly achieve the local minimax risk, and to that end we develop results on a private (noisy) stochastic gradient method, which rely on Polyak and Juditsky [46]; recall also Duchi et al.'s (minimax-optimal) private stochastic gradient method [25]. We prefer (for brevity) to avoid the finest convergence conditions, instead giving references as possible; we show how to attain the rates in Examples 8 and 9.

We wish to minimize a risk $R_P(\theta) := \mathbb{E}_P[\ell(\theta, X)]$, where $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$ is convex in its first argument. A noisy stochastic gradient algorithm iteratively updates a parameter θ for $i = 1, 2, \ldots$,

$$\theta^{i+1} = \theta^i - \eta_i(\nabla \ell(\theta^i, X_i) + \xi^i), \tag{24}$$

where ξ^i is i.i.d. zero-mean noise, we assume $X_i \stackrel{\mathrm{iid}}{\sim} P$, and $\eta_i = \eta_0 i^{-\beta}$ are stepsizes with $\beta \in (\frac{1}{2}, 1)$ and $\eta_0 > 0$. Then under appropriate conditions [46, Thm. 2 and Lem. 2],[3] for $\theta^\star = \operatorname{argmin}_\theta R_P(\theta)$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\theta^i - \theta^\star) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla^2 R_P(\theta^\star)^{-1} (\nabla \ell(\theta^\star, X_i) + \xi^i) + o_P(1). \tag{25}$$

The average $\overline{\theta}^n = \frac{1}{n}\sum_{i=1}^n \theta^i$ is asymptotically linear (13), so it satisfies the regularity properties we outline in Lemma 1. The key is that $\nabla^2 R_P(\theta^\star)^{-1} \nabla \ell(\theta^\star, X_i)$ is typically the influence function for the parameter θ (see [23, Proposition 1 and Lemma 8.1]), and we thus call this case regular:

Definition 3. Let $\mathcal{P}_{\mathrm{sub},0} = \{P_h\}_{h \in \mathbb{R}^d}$ be a sub-model of $\mathcal{P}$ around P_0, quadratic mean differentiable (8) with score $g : \mathcal{X} \to \mathbb{R}^d$ at P_0. Let $R_h(\theta) = \mathbb{E}_{P_h}[\ell(\theta, X)]$ and define $\theta_h = \operatorname{argmin}_{\theta \in \Theta} R_h(\theta)$. The parameter θ_h is *regular* if it has influence function $\dot{\theta}_0(x) = \nabla^2 R_0(\theta_0)^{-1} \nabla \ell(\theta_0, x)$, equivalently,

$$\theta_h = \theta_0 + \nabla^2 R_0(\theta_0)^{-1} \mathrm{Cov}_0(\nabla \ell(\theta_0, X), g(X)) h + o(\|h\|).$$

By combining Lemma 1 with the convergence guarantee (25), we obtain the following result.

Proposition 5. *Let θ^i follow the noisy stochastic gradient iteration (24) and satisfy the convergence (25). Let $\mathcal{P}_{\mathrm{sub},0}$ be a sub-model for which the risk R_P is regular (Def. 3) and let*

$$Z \sim \mathsf{N}\left(0, \nabla^2 R_{P_0}(\theta_0)^{-1} \left(\mathrm{Cov}_0(\nabla \ell(\theta_0, X)) + \mathrm{Cov}(\xi)\right) \nabla^2 R_{P_0}(\theta_0)^{-1}\right).$$

Then for any bounded sequence $h_n \in \mathbb{R}^d$, $\sqrt{n}(\overline{\theta}^n - \theta_{h_n/\sqrt{n}}) \stackrel{d}{\to} Z$ under $X_i \stackrel{\mathrm{iid}}{\sim} P_{h_n/\sqrt{n}}$, and for any bounded continuous L and $c < \infty$,

$$\lim_{n \to \infty} \sup_{\|h\| \le c/\sqrt{n}} \mathbb{E}_{P_h}\left[L(\sqrt{n}(\overline{\theta}^n - \theta_h))\right] = \mathbb{E}[L(Z)].$$

[3] The following suffice: (i) R_P is $\mathcal{C}^2$ near $\theta^\star = \operatorname{argmin}_\theta R_P(\theta)$ with $\nabla^2 R_P(\theta^\star) \succ 0$, (ii) there is some finite C such that $\mathbb{E}_P[\|\nabla \ell(\theta, X)\|^2] \le C(1 + \|\theta - \theta^\star\|^2)$ and $\|\nabla R_P(\theta)\| \le C(1 + \|\theta - \theta^\star\|)$ for all θ, and (iii) $\limsup_{\theta \to \theta^\star} \mathbb{E}[\|\nabla \ell(\theta, X) - \nabla \ell(\theta^\star, X)\|^2] = 0$.

We complete Examples 8 and 9 using Proposition 5:

Example (Example 8 continued)**:** We return to the shape parameter in the exponential family. As the median of X is $\log 2 \cdot \theta_0$, estimating θ_0 is equivalent to estimating $\mathrm{med}(X)$, or solving

$$\mathop{\mathrm{minimize}}_{\theta} R_{P_0}(\theta) := \mathbb{E}_{P_0}[\ell(\theta, X)] \quad \text{for} \quad \ell(\theta, x) = |\theta - x|.$$

The stochastic gradient iteration (24) is $\theta^{i+1} = \theta^i - \eta_i Z_i$ where $Z_i = \mathrm{sign}(X_i - \theta^i) + \frac{1}{2\varepsilon} \cdot \mathsf{Lap}(1)$ is ε-differentially private as and $\mathsf{Lap}(1)$ is the standard Laplacian distribution. For $P = \mathsf{Exp}(\theta)$, $R'(t) = P_\theta(t > X) - P_\theta(t < X) = 1 - 2e^{-t/\theta}$ and $R''(t) = \frac{2}{\theta} e^{-t/\theta}$, at $t = \mathrm{med}(X) = \log 2 \cdot \theta$ we obtain $R''(\mathrm{med}(X)) = \frac{1}{\theta}$. For any symmetric quasiconvex $L : \mathbb{R}_+ \to \mathbb{R}_+$, define $L_n(t) = L(\sqrt{n} \cdot t)$, and let $\widehat{\theta}_n = \overline{\theta}^n / \log 2$. Applying Example 8 and Proposition 5 yields

$$\sup_{c < \infty} \limsup_{n \to \infty} \sup_{|\theta - \theta_0| \le c/\sqrt{n}} \frac{\mathbb{E}_\theta[L(\sqrt{n}(\widehat{\theta}_n - \theta))]}{\mathfrak{M}_n^{\mathrm{loc}}(P_{\theta_0}, L_n, \mathcal{P}, \mathcal{Q}_\varepsilon)} \le \frac{\mathbb{E}[L(C_1 \frac{\theta_0}{\varepsilon} W)]}{L(C_0 \frac{\theta_0}{\varepsilon})}$$

where $W \sim \mathsf{N}(0, 1)$ is standard normal and C_i are numerical constants. Whenever L is such that $\mathbb{E}[L(C_1 \sigma W)] \lesssim L(C_0 \sigma)$—for example, $L(t) = \min\{t^2, B\}$—the private stochastic gradient method is local minimax rate optimal. $\diamond$

Example (Example 9 continued)**:** We have $Y = X\theta_0 + \sigma W$ for $W \sim \mathsf{N}(0, 1)$, where for simplicity we assume σ is known. We transform the problem into a stochastic optimization problem, taking care to choose the correct objective for privacy and efficiency. Let $\varphi : \mathbb{R} \to \mathbb{R}_+$ be any 1-Lipschitz symmetric convex function with Lipschitzian gradient; for example, the Huber loss $\varphi(t) = \frac{1}{2}t^2$ for $|t| \le 1$ and $\varphi(t) = |t| - \frac{1}{2}$ satisfies these conditions. Choosing loss $\ell(\theta, x, y) = \frac{\sigma}{|x|} \varphi(\frac{x\theta - y}{\sigma})$, our problem is to minimize the risk $R_P(\theta) := \mathbb{E}_P[\ell(\theta, X, Y)]$. Evidently ℓ is also 1-Lipschitz with respect to θ, with $\ell'(\theta, x, y) = \mathrm{sign}(x)\varphi'(\sigma^{-1}(x\theta - y))$ and $\ell''(\theta, x, y) = \sigma^{-1}\varphi''(\sigma^{-1}(x\theta - y))$. The private stochastic gradient iteration (24) is then

$$\theta^{i+1} = \theta^i - \eta_i Z_i \quad \text{where} \quad Z_i = \mathrm{sign}(X_i)\varphi'\left(\frac{1}{\sigma}(X_i\theta^i - Y_i)\right) + \frac{1}{2\varepsilon}\mathsf{Lap}(1).$$

Letting $Z_\infty = \varphi'(W_i) + \frac{1}{2\varepsilon}\mathsf{Lap}(1)$, a calculation shows that for P_0 corresponding to θ_0,

$$R''_{P_0}(\theta_0) = \frac{\mathbb{E}[|X|]}{\sigma}\mathbb{E}[\varphi''(W)] \quad \text{and} \quad \mathrm{Var}(Z_\infty) = \mathbb{E}[\varphi'(W)^2] + \frac{1}{2\varepsilon^2}.$$

We apply Proposition 5 to obtain that along any convergent sequence $h_n \to h$,

$$\sqrt{n}\left(\overline{\theta}^n - (\theta_0 + h_n/\sqrt{n})\right) \xrightarrow[P_{\theta_0 + h_n/\sqrt{n}}]{d} \mathsf{N}\left(0, \frac{\sigma^2}{\mathbb{E}[|X|]^2 \mathbb{E}[\varphi''(W)]^2}\left(\mathbb{E}[\varphi'(W)^2] + \frac{1}{2\varepsilon^2}\right)\right).$$

Notably $\mathbb{E}[\varphi'(W)^2] \le 1$ as φ is 1-Lipschitz, and whenever φ is strongly convex near 0, then $\mathbb{E}[\varphi''(W)] > 0$ is a positive constant. In particular, the stochastic gradient estimator is local minimax rate optimal, achieving asymptotic variance $O(1)\frac{\sigma^2}{\varepsilon^2 \mathbb{E}[|X|]^2}$ uniformly near θ_0, which (as in Example 9) is unimprovable except by numerical constants. $\diamond$

4.4 Asymptotic achievability in one-parameter exponential families

Proposition 4 shows an instance-specific lower bound of $(n\varepsilon^2 J_{\theta_0}^2)^{-1}$, where $J_{\theta_0} = \mathbb{E}_{\theta_0}[|\dot{\ell}_{\theta_0}|]$ is the L^1 information, for the estimation of a single parameter. This section develops a novel locally private estimation scheme to achieve the lower bound for general one-parameter exponential family models. Subtleties in the construction make showing that the estimator is regular or uniform challenging, though we conjecture that it is locally uniform. Let $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ be a one parameter exponential family, so that for a base measure μ on $\mathcal{X}$, each distribution P_θ has density

$$p_\theta(x) := \frac{dP_\theta}{d\mu}(x) = \exp\left(\theta T(x) - A(\theta)\right),$$

where $T(x)$ is the sufficient statistic and $A(\theta) = \log \int e^{\theta T(x)} d\mu(x)$ is the log partition function.[4] It is well known (cf. [7, 40, Ch. 2.7]) that A satisfies $A'(\theta) = \mathbb{E}_\theta[T(X)]$ and $A''(\theta) = \mathrm{Var}_\theta(T(X))$. In this case, the L^1-information (20) is the mean absolute deviation

$$J_\theta = \mathbb{E}_\theta[|T(X) - A(\theta)|] = \mathbb{E}_\theta[|T(X) - \mathbb{E}_\theta[T(X)]|].$$

We provide a procedure asymptotically achieving mean square error scaling as $(n\varepsilon^2 J_\theta^2)^{-1}$, which Proposition 4 shows is optimal. Our starting point is the observation that for a one-parameter exponential family, $\theta \mapsto P_\theta(T(X) \geq t)$ is strictly increasing in θ for any fixed $t \in \mathrm{supp}\{T(X)\}$ [40, Lemma 3.4.2]. A natural idea is to first estimate $P_\theta(T(X) \geq t)$ and invert to estimate θ. To that end, we develop a private two-sample procedure, where with the first we estimate $\hat{t} \approx \mathbb{E}[T(X)]$, using the second sample to approximate and invert $P_\theta(T(X) \geq \hat{t})$. Now, define $\Psi : \mathbb{R}^2 \to \mathbb{R}_+$ by

$$\Psi(t, \theta) := P_\theta(T(X) \geq t) = \int \mathbf{1}\{T(x) \geq t\} \exp\left(\theta T(x) - A(\theta)\right) d\mu(x). \tag{26}$$

The private two stage algorithm we develop splits a total sample of size $2n$ in half, using the first half of the sample to construct a consistent estimate $\widehat{T}_n$ of the value $A'(\theta) = \mathbb{E}_\theta[T]$ (Duchi et al.'s ε-differentially private mean estimators provide consistent estimates of $\mathbb{E}[T(X)]$ so long as $\mathbb{E}[|T(X)|^k] < \infty$ for some $k > 1$ [25, Corollary 1].) In the second stage, the algorithm uses $\widehat{T}_n$ and the second half of the sample in a randomized response procedure: construct V_i and private Z_i as

$$V_i = \mathbf{1}\{T(X_i) \geq \widehat{T}_n\}, \quad Z_i = \frac{e^\varepsilon + 1}{e^\varepsilon - 1} \cdot \left[\left\{ \begin{array}{ll} V_i & \text{w.p. } \frac{e^\varepsilon}{e^\varepsilon + 1} \\ 1 - V_i & \text{w.p. } \frac{1}{e^\varepsilon + 1} \end{array} \right\} - \frac{1}{e^\varepsilon + 1} \right].$$

By inspection, this is ε-differentially-private and $\mathbb{E}[Z_i \mid V_i] = V_i$. Now, define the inverse function

$$H(p, t) := \inf\left\{\theta \in \mathbb{R} \mid P_\theta(T(X) \geq t) \geq p\right\} = \inf\left\{\theta \in \mathbb{R} \mid \Psi(t, \theta) \geq p\right\}.$$

Setting $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$, our final ε-differentially private estimator is

$$\widehat{\theta}_n = H(\overline{Z}_n, \widehat{T}_n). \tag{27}$$

We then have a convergence result showing that the estimator (27) has asymptotic variance within a constant factor of the local minimax bounds. We defer the (involved) proof to Appendix C.2.

[4] Writing the family this way is no loss of generality. While typically one writes $p_\theta(x) = h(x)\exp(\theta^T T(x) - A(\theta))$, we can always include h in the base measure μ and push-forward through the statistic T.

Proposition 6. *Assume that* $\mathrm{Var}_\theta(T(X)) > 0$ *and* $\widehat{T}_n \xrightarrow{p} t_0 := \mathbb{E}_{\theta_0}[T(X)]$. *Define* $\delta_\varepsilon^2 = \frac{e^\varepsilon}{(e^\varepsilon-1)^2}$. *Then there exist random variables* $G_n = \Psi(\widehat{T}_n, \theta_0) \in [0,1]$, $\mathcal{E}_{n,1}$, *and* $\mathcal{E}_{n,2}$ *such that under* P_{θ_0},

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = 2J_{\theta_0}^{-1}(\mathcal{E}_{n,1} + \mathcal{E}_{n,2}) + o_P(1)$$

where

$$\left(\mathcal{E}_{n,1}, \frac{1}{G_n(1-G_n)}\mathcal{E}_{n,2}\right) \xrightarrow{d} \mathsf{N}\left(0, \mathrm{diag}(\delta_\varepsilon^{-2}, 1)\right). \tag{28}$$

The complexity of the statement arises because the distribution of $T(X)$ may be discontinuous, including at $\mathbb{E}_{\theta_0}[T(X)]$, necessitating the random variables $\mathcal{E}_{n,1}, \mathcal{E}_{n,2}$, and G_n for the limit.

5 Private local minimax theory for more general functionals

We broaden our investigation to consider the local minimax approach in semi- or nonparametric problems with high or infinite-dimensional parameters, but where the target of interest is one-dimensional. We first present analogues, to within numerical constants, of classical semi-parametric information lower bounds (Section 5.1). We illustrate the bounds for estimating a functional of an exponential family parameter, where subtleties distinguish the problem from the non-private case. Most saliently, as we show, efficiency depends strongly on the model assumed by the statistician: while in the non-private case, parametric and nonparametric models yield the same efficiency bounds (as we will revisit), the private parametric and semi-parametric cases are quite different.

5.1 Private information, influence functions, and tangent spaces

Our goal here is to generalize the results in Section 4 to provide private information lower bounds for semi-parametric estimation problems. Similar to what we did in Section 4, our development builds off Theorem 1 and Proposition 2 by performing a local expansion of the variation distance. We parallel some of the classical development in Section 2.3, presenting one-dimensional submodels, tangent spaces, and an L^1-influence function, after which we derive our information bounds.

We begin as usual with a family $\mathcal{P}$ of distributions, and we consider one-dimensional submodels $\mathcal{P}_{\mathrm{sub},0} \subset \mathcal{P}$ indexed by $h \in \mathbb{R}$. In analogy with quadratic mean differentiability (8) and our treatment in Section 4, we say $h \mapsto P_h$ is L^1-*differentiable at* P_0 with score $g : \mathcal{X} \to \mathbb{R}$ if

$$\int |dP_h - dP_0 - hg dP_0| = o(|h|) \tag{29}$$

as $h \to 0$. As in Section 2.3, we let $h \mapsto P_h$ range over (a collection of) possible submodels to obtain a collection of score functions $\{g\}$, and we define the L^1-*tangent space* $\dot{\mathcal{P}}_{L^1(P_0)}$ to be the closed linear span of these scores. In contrast to the tangent space from quadratic-mean-differentiability (11), which admits Hilbert-space geometry, L^1-differentiability gives a different duality. Moreover, Lemma 2 states that QMD families must be L^1-differentiable, so that the L^1-tangent space $\dot{\mathcal{P}}_{L^1(P_0)}$ always contains the classical tangent space. An example may be clarifying:

Example 10 (Fully nonparametric L^1 tangents)**:** In the fully nonparametric case—where $\mathcal{P}$ consists of all distributions supported on $\mathcal{X}$—we can identify $\dot{\mathcal{P}}_{L^1(P_0)}$ with mean-zero $g \in L^1(P_0)$. Indeed, for any such g, define the models $dP_h = [1+hg]_+ dP_0/C_h$, where $C_h = \int [1+hg]_+ dP_0$. Then $|[1+hg]_+ - 1| \le h|g|$, and so by dominated convergence we have $1 \le C_h = 1 + o(h)$, as

$$0 \le \frac{1}{h}(C_h - 1) = \int \frac{[1+hg]_+ - 1}{h} dP_0 \underset{h\to 0}{\longrightarrow} \int g dP_0 = 0.$$

Similarly $\{dP_h\}_{h\in\mathbb{R}}$ has score g at P_0 because $\lim_{h\to 0}\int \frac{1}{h}|\frac{[1+hg]_+ - 1}{C_h} - hg|dP_0 = 0$ by the dominated convergence theorem as well. Conversely, for any $g \in \dot{\mathcal{P}}_{L^1(P_0)}$, we have $\int |dP_h - dP_0| \le 2$, while $o(h) = \int |dP_h - dP_0 - hgdP_0| \ge |h| \int |g|dP_0 - 2$, so that $g \in L^1(P_0)$. That g is mean zero is immediate, as $\int gdP_0 = \frac{1}{h}\int (dP_0 + hgdP_0 - dP_h) = o(1)$ as $h \to 0$.

In contrast, in the fully nonparametric case for quadratic mean differentiability [52, Example 25.16], the tangent set at P_0 is all mean zero $g \in L^2(P_0)$, a smaller set of potential tangents. $\diamond$

To give a private information for estimating a function $\theta : \mathcal{P} \to \mathbb{R}^d$, we consider submodels $\{P_h\}_{h\in\mathbb{R}}$ where $h \mapsto \theta(P_h)$ is suitably smooth at $h = 0$. We say $\theta(\cdot)$ is *differentiable at P_0 relative to $\dot{\mathcal{P}}_{L^1(P_0)}$* if there exists a continuous linear mapping $\varphi_{P_0} : \dot{\mathcal{P}}_{L^1(P_0)} \to \mathbb{R}^d$ such that for any L^1-differentiable submodel $\mathcal{P}_{\sub,0}$ with score g at P_0,

$$\theta(P_h) - \theta(P_0) = h\varphi(g) + o(h).$$

As $g \mapsto \varphi(g)$ is continuous for $g \in \dot{\mathcal{P}}_{L^1(P_0)}$, it has a continuous extension to all of $L^1(P_0)$ and so by duality there exists $\dot{\theta}_0 : \mathcal{X} \to \mathbb{R}^d$, with coordinate functions in $L^\infty(P_0)$, such that

$$\varphi(g) = \mathbb{E}_{P_0}[\dot{\theta}_0(X)g(X)] = \int \dot{\theta}_0(x)g(x)dP_0(x).$$

We call this $\dot{\theta}_0$ the *private influence function*. Again, contrast with the classical approach is instructive: there (recall Eq. (12)), the Hilbert space structure of the tangent sets allows one to use the Riesz representation theorem to guarantee the existence of an influence function $\dot{\theta}_0 \in L^2(P_0)$.

The main result of this section gives an information-type lower bound for general estimation problems where we wish to estimate a functional $\psi(\theta(P))$, where $\psi : \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{C}^1$. We measure error by a symmetric quasiconvex $L : \mathbb{R} \to \mathbb{R}_+$, suffering loss $L(\widehat{\psi} - \psi(\theta(P)))$ for an estimate $\widehat{\psi}$. We then obtain the following generalization of Proposition 4. (See Appendix B.4 for a proof.)

Theorem 2. *Let $\mathcal{P}_{\sub,0} = \{P_h\}_{h\in\mathbb{R}} \subset \mathcal{P}$ be an L^1-differentiable submodel at P_0 with score g and let $\theta : \mathcal{P} \to \mathbb{R}^d$ have L^1-influence function $\dot{\theta}_0$ at P_0. Let $\mathcal{Q}_\varepsilon$ be the family of $(2, \varepsilon^2)$-locally Rényi private channels (Def. 2). Then for an $N = N(\psi, \varepsilon, \theta, \mathcal{P}_{\sub,0})$ independent of loss L, for all $n \ge N$*

$$\mathfrak{M}_n^{\loc}(P_0, L, \mathcal{P}_{\sub,0}, \mathcal{Q}_\varepsilon) \ge \frac{1}{8} \cdot L\left(\frac{1}{6\sqrt{n\varepsilon^2}} \frac{\nabla\psi(\theta_0)^T \mathbb{E}_0[\dot{\theta}_0(X)g(X)]}{\mathbb{E}_0[|g(X)|]}\right).$$

The quantity $J_{g,0} := \mathbb{E}_0[|g(X)|] = \int |g|dP_0$ is the nonparametric analogue of the private information (20) in Proposition 4, as the score function g is completely parallel to the parametric case. Additional remarks show how this result parallels and complements classical local minimax theory.

Recovering the parametric case Theorem 2 specializes to Proposition 4 for one-dimensional parametric families. Let the family be $\mathcal{P}_{\text{par}} = \{P_\theta\}_{\theta\in\Theta}$ and L^1 differentiable at P_{θ_0}. Then the private tangent space $\dot{\mathcal{P}}_{L^1(P_0)}$ is then the linear space spanned by the score $\dot{\ell}_{\theta_0}$, and the influence function for θ is $\dot{\theta}_0 = I_{\theta_0}^{-1}\dot{\ell}_{\theta_0}$, where the Fisher information is $I_{\theta_0} = \mathbb{E}_0[\dot{\ell}_{\theta_0}(X)^2]$. Specializing Theorem 2 gives $\nabla\psi(\theta_0)^T\mathbb{E}_0[\dot{\theta}_0(X)g(X)]/\mathbb{E}_0[|g(X)|] = \psi'(\theta_0)/\mathbb{E}_0[|\dot{\ell}_{\theta_0}(X)|]$, recovering Proposition 4.

Dualities and classical information bounds As a corollary of the L^1/L^∞ duality that privacy evidently entails in Theorem 2 and Lemma 2, we have the following lower bound.

Corollary 7. *Let the conditions of Theorem 2 hold, and additionally let $\mathcal{P}_{\text{all},0}$ be a collection of QMD sub-models with scores g that are dense in $L^1(P_0)$. Then there exists an N independent of the loss L such that for all $n \ge N$,*

$$\mathfrak{M}^{\text{loc}}(P_0, L, \mathcal{P}_{\text{all},0}, \mathcal{Q}_\varepsilon) \ge \frac{1}{8} \cdot L\left(\frac{1}{8\sqrt{n\varepsilon^2}} \operatorname*{ess\,sup}_x |\nabla\psi(\theta_0)^T \dot\theta_0(x)|\right).$$

The local minimax lower bound *necessarily* depends on the (essential) supremum of the influence function $\nabla\psi(\theta_0)^T\dot\theta_0(x)$ over $x \in \mathcal{X}$; notably, this occurs even when the tangent set $\dot{\mathcal{P}}_0$ is dense in $L^2(P_0)$. We may compare this with classical (nonparametric) information bounds, which rely on the Hilbert-space structure of quadratic-mean-differentiability and are thus smaller and qualitatively different. In the classical setting [52, Ch. 25.3] (and Sec. 2.3), we recall that we may identify $\dot{\mathcal{P}}_0$ with mean-zero functions $g \in L^2(P_0)$, and we obtain the analogous lower bound

$$L\left(\frac{1}{\sqrt{n}} \cdot \sup_{g \in \dot{\mathcal{P}}_0} \frac{\nabla\psi(\theta_0)^T \mathbb{E}_0[\dot\theta_0(X) g(X)]}{\mathbb{E}_0[g(X)^2]^{1/2}}\right) = L\left(\frac{1}{\sqrt{n}} \mathbb{E}_0\left[\left(\nabla\psi(\theta_0)^T\dot\theta_0(X)\right)^2\right]^{1/2}\right),$$

where we have used that $\mathbb{E}_0[\dot\theta_0(X)] = 0$. This information bound is always (to numerical constants) smaller than the private information bound in Corollary 7.

5.2 Nonparametric modeling with exponential families

While Section 4 characterizes local minimax complexities for several one-dimensional problems, treating one parameter exponential families in Section 4.4, it relies on the model's correct specification. Here, we consider estimating functionals of a potentially mis-specified exponential family model. To formally describe the setting, we start with a d-parameter exponential family $\{P_\theta\}_{\theta\in\Theta}$ with densities $p_\theta(x) = \exp(\theta^T x - A(\theta))$ with respect to some base measure μ, where for simplicity we assume that the exponential family is regular and minimal, meaning that $\nabla^2 A(\theta) = \text{Cov}_\theta(X) \succ 0$ for all $\theta \in \operatorname{dom} A$, and the log partition function $A(\theta)$ is analytic on the interior of its domain [40, Thm. 2.7.1]. We record a few standard facts on the associated convex analysis (for more, see the books [7, 54, 36]). Recall the conjugate $A^*(x) := \sup_\theta\{\theta^T x - A(\theta)\}$. Then [cf. 36, Ch. X]

$$\nabla A^*(x) = \theta_x \quad \text{for the unique } \theta_x \text{ such that } \quad \mathbb{E}_{\theta_x}[X] = x. \tag{30}$$

In addition, ∇A^* is continuously differentiable, one-to-one, and

$$\operatorname{dom} A^* \supset \text{Range}(\nabla A(\cdot)) = \{\mathbb{E}_\theta[X] \mid \theta \in \operatorname{dom} A\}.$$

Moreover, by the inverse function theorem, we also have that on the interior of $\operatorname{dom} A^*$,

$$\nabla^2 A^*(x) = (\nabla^2 A(\theta_x))^{-1} = \text{Cov}_{\theta_x}(X)^{-1} \quad \text{for the unique } \theta_x \text{ s.t. } \mathbb{E}_{\theta_x}[X] = x. \tag{31}$$

The uniqueness follows because ∇A^* is one-to-one, as the exponential family is minimal and $\nabla^2 A(\theta) \succ 0$. For a distribution P with mean $\mathbb{E}_P[X]$, so long as the mean belongs to the range of $\nabla A(\theta) = \mathbb{E}_\theta[X]$ as θ varies, the minimizer of the log loss $\ell_\theta(x) = -\log p_\theta(x)$ is

$$\theta(P) := \operatorname*{argmin}_\theta \mathbb{E}_P[\ell_\theta(X)] = \nabla A^*(\mathbb{E}_P[X]).$$

We consider estimation of smooth functionals $\psi : \mathbb{R}^d \to \mathbb{R}$ of the parameters θ, measuring the loss of an estimated value $\widehat\psi$ by

$$L(\widehat\psi - \psi(\theta(P))),$$

where $L : \mathbb{R} \to \mathbb{R}_+$ as usual is quasi-convex and symmetric. In the sequel, we show local lower bounds on estimation, develop a (near) optimal regular estimator, and contrast our results and the possibilities of adaptation in private and non-private cases with somewhat striking differences.

5.2.1 Private estimation rates

We begin with a local minimax lower bound that almost immediately follows Theorem 2.

Corollary 8. *Let P_0 be such that $\mathbb{E}_{P_0}[X] \in \operatorname{int}(\operatorname{Range}(\nabla A))$, and let $\mathcal{P}_{\text{all},0}$ be a collection of sub-models with scores g dense in $L^1(P_0)$ at P_0. Let $\mathcal{Q}_\varepsilon$ denote the collection of all $(2, \varepsilon^2)$-locally Rényi private sequentially interactive channels. Then there exists $N = N(\mathcal{P}_{\text{all},0}, \psi)$ independent of the loss L such that $n \ge N$ implies*

$$\mathfrak{M}_n^{\text{loc}}(P_0, L, \mathcal{P}_{\text{all},0}, \mathcal{Q}_\varepsilon) \ge \frac{1}{8} \cdot L\left(\frac{1}{5\sqrt{2n\varepsilon^2}} \cdot \operatorname*{ess\,sup}_x \left|\nabla\psi(\theta_0)^T \nabla^2 A(\theta_0)^{-1}(\mathbb{E}_{P_0}[X] - x)\right|\right).$$

Proof. The exponential family influence function is $\dot\theta_0(x) = \nabla^2 A(\theta_0)^{-1}(x - \mathbb{E}_0[X])$ [52, Ch. 25.3]. Take g with $\frac{\nabla\psi(\theta_0)^T \mathbb{E}_0[\dot\theta_0(X) g(X)]}{\mathbb{E}_0[|g(X)|]} \ge \frac{3}{4} \operatorname{ess\,sup}_x |\nabla\psi(\theta_0)^T \nabla^2 A(\theta_0)^{-1}(x - \mathbb{E}_0[X])|$ in Theorem 2. □

Before we turn to private estimation, we compare Corollary 8 to the non-private case. The maximum likelihood estimator takes the sample mean $\widehat\mu_n = \frac{1}{n}\sum_{i=1}^n X_i$ and sets $\widehat\theta_n = \nabla A^*(\widehat\mu_n)$. Letting $\theta_0 = \nabla A^*(\mathbb{E}_{P_0}[X])$, Taylor expansion arguments and the delta-method [52, Chs. 3–5] yield

$$\sqrt{n}(\widehat\theta_n - \theta_0) \xrightarrow{d} \mathsf{N}\left(0, \nabla^2 A(\theta_0)^{-1} \operatorname{Cov}_0(X) \nabla^2 A(\theta_0)^{-1}\right)$$

and

$$\sqrt{n}(\psi(\widehat\theta_n) - \psi(\theta_0)) \xrightarrow{d} \mathsf{N}\left(0, \nabla\psi(\theta_0)^T \nabla^2 A(\theta_0)^{-1} \operatorname{Cov}_0(X) \nabla^2 A(\theta_0)^{-1} \nabla\psi(\theta_0)\right),$$

and these estimators are regular (and hence locally uniform). The lower bound in Corollary 8 is always larger than this classical limit. In this sense, the private lower bounds exhibit both the importance of local geometry—via $\nabla^2 A(\theta_0)^{-1}\nabla\psi(\theta_0)$—and the challenge of privacy in addressing "extraneous" noise that must be privatized. We will discuss this more in Section 5.3.

5.2.2 An optimal one-step procedure

An optimal procedure for functionals of (possibly) mis-specified exponential family models is similar to classical one-step estimation procedures [e.g. 52, Ch. 5.7]. To motivate the approach, let us assume we have a "good enough" estimate $\widetilde\mu_n$ of $\mu_0 := \mathbb{E}_P[X]$. Then if $\widetilde\theta_n = \nabla A^*(\widetilde\mu_n)$, we have

$$\begin{aligned}
\psi(\theta_0) &= \psi(\widetilde\theta_n) + \nabla\psi(\widetilde\theta_n)^T(\theta_0 - \widetilde\theta_n) + O(\|\theta_0 - \widetilde\theta_n\|^2) \\
&= \psi(\widetilde\theta_n) + \nabla\psi(\widetilde\theta_n)^T(\nabla A^*(\mu_0) - \nabla A^*(\widetilde\mu_n)) + O(\|\mu_0 - \widetilde\mu_n\|^2) \\
&= \psi(\widetilde\theta_n) + \nabla\psi(\widetilde\theta_n)^T \nabla^2 A(\widetilde\theta_n)^{-1}(\mu_0 - \widetilde\mu_n) + O(\|\mu_0 - \widetilde\mu_n\|^2),
\end{aligned}$$

where each equality freely uses the duality relationships (30) and (31). In this case, if $\widetilde\mu_n - \mu_0 = o_P(n^{-1/4})$ and we have an estimator T_n satisfying

$$\sqrt{n}\left(T_n - \nabla\psi(\widetilde\theta_n)^T \nabla^2 A(\widetilde\theta_n)^{-1}\mu_0\right) \xrightarrow{d} \mathsf{N}(0, \sigma^2),$$

then the estimator

$$\widehat\psi_n := \psi(\widetilde\theta_n) + T_n - \nabla\psi(\widetilde\theta_n)^T \nabla^2 A(\widetilde\theta_n)^{-1}\widetilde\mu_n \tag{32}$$

satisfies $\sqrt{n}(\widehat\psi_n - \psi(\theta_0)) \xrightarrow{d} \mathsf{N}\left(0, \sigma^2\right)$ by Slutsky's theorems.

We now exhibit such an estimator. To avoid some of the difficulties associated with estimation from unbounded data [25], we assume the domain $\mathcal{X} \subset \mathbb{R}^d$ is the norm ball $\{x \in \mathbb{R}^d \mid \|x\| \le 1\}$. For dual norm $\|z\|_* = \sup_{\|x\| \le 1} x^T z$, the essential supremum in Corollary 8 thus has bounds

$$\operatorname*{ess\,sup}_x \left|\nabla\psi(\theta_0)^T \nabla^2 A(\theta_0)^{-1} \left(\mathbb{E}_{P_0}[X] - x\right)\right| \in \left[\frac{1}{2}, 2\right] \cdot \left\|\nabla^2 A(\theta_0)^{-1} \nabla\psi(\theta_0)\right\|_* . \tag{33}$$

Let us split the sample of size n into two sets of size $n_1 = \lceil n^{2/3} \rceil$ and $n_2 = n - n_1$. For the first set, let Z_i be any ε-locally differentially private estimate of X_i satisfying $\mathbb{E}[Z_i \mid X_i] = X_i$ and $\mathbb{E}[\|Z_i\|^2] < \infty$, so that the Z_i are i.i.d.; for example, $X_i + W_i$ for a random vector of appropriately large Laplace noise suffices [29, 25]. Define $\widetilde{\mu}_n = \frac{1}{n_1} \sum_{i=1}^{n_1} Z_i$, in which case $\widetilde{\mu}_n - \mu_0 = O_P(n^{-1/3})$, and let $\widetilde{\theta}_n = \nabla A^*(\widetilde{\mu}_n)$. Now, for $i = n_1 + 1, \ldots, n$, define the ε-differentially private quantity

$$Z_i := \nabla\psi(\widetilde{\theta}_n)^T \nabla^2 A(\widetilde{\theta}_n)^{-1} X_i + \frac{\|\nabla^2 A(\widetilde{\theta}_n)^{-1} \nabla\psi(\widetilde{\theta}_n)\|_*}{\varepsilon} W_i \quad \text{where} \quad W_i \stackrel{\text{iid}}{\sim} \mathsf{Lap}(1).$$

Letting $\overline{X}_{n_2} = \frac{1}{n_2} \sum_{i=n_1+1}^{n} X_i$ and similarly for $\overline{W}_{n_2}$ and $\overline{Z}_{n_2}$, we find that

$$\begin{aligned}
&\sqrt{n}\left(\overline{Z}_{n_2} - \nabla\psi(\widetilde{\theta}_n)^T \nabla^2 A(\widetilde{\theta}_n)^{-1} \mu_0\right) \\
&= \sqrt{n}\left[\nabla\psi(\widetilde{\theta}_n)^T \nabla^2 A(\widetilde{\theta}_n)^{-1} \left(\overline{X}_{n_2} - \mu_0\right) + \frac{\|\nabla^2 A(\widetilde{\theta}_n)^{-1} \nabla\psi(\widetilde{\theta}_n)\|_*}{\varepsilon} \overline{W}_{n_2}\right] \stackrel{d}{\to} \mathsf{N}(0, \sigma^2(P, \psi, \varepsilon))
\end{aligned}$$

by Slutsky's theorem, where for $\theta_0 = \nabla A^*(\mathbb{E}_P[X])$ we define

$$\sigma^2(P, \psi, \varepsilon) := \nabla\psi(\theta_0)^T \nabla^2 A(\theta_0)^{-1} \mathrm{Cov}_P(X) \nabla^2 A(\theta_0)^{-1} \nabla\psi(\theta_0) + \frac{2}{\varepsilon^2} \left\|\nabla^2 A(\theta_0)^{-1} \psi(\theta_0)\right\|_*^2 . \tag{34}$$

Moreover, the difference above is asymptotically linear (13), so by continuity we have

$$\sqrt{n}(\widehat{\psi}_n - \psi(\theta_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla\psi(\theta_0)^T \nabla^2 A(\theta_0)^{-1} (X_i - \mu_0) + \frac{\left\|\nabla^2 A(\theta_0)^{-1} \nabla\psi(\theta_0)\right\|_*}{\varepsilon} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i + o_{P_0}(1).$$

Summarizing, we can apply Lemma 1, because the smoothness of $A(\cdot)$ means that the parameter θ_0 is regular in that it has influence function $\dot{\theta}_0(x) = \nabla^2 A(\theta_0)^{-1}(x - \mu_0)$ (recall also Definition 3). Recalling the equivalence (33) between the dual norm measures and essential supremum, we have thus shown that the two-step estimator (32) is locally minimax rate optimal.

Proposition 7. *Let $\widehat{\psi}_n$ be the estimator (32), $\{P_h\}$ be quadratic mean differentiable at P_0, $\theta_h = \operatorname{argmax}_\theta \mathbb{E}_{P_h}[\log p_\theta(X)]$, and $\sigma^2(P_0, \psi, \varepsilon)$ be as in (34). Let $Z \sim \mathsf{N}(0, \sigma^2(P_0, \psi, \varepsilon))$ and h_n be a bounded sequence. Then $\sqrt{n}(\widehat{\psi}_n - \psi(\theta_{h_n/\sqrt{n}})) \stackrel{d}{\to} Z$ under $X_i \stackrel{\text{iid}}{\sim} P_{h_n/\sqrt{n}}$, and for any bounded continuous L and $c < \infty$,*

$$\lim_{n\to\infty} \sup_{\|h\| \le c/\sqrt{n}} \mathbb{E}_{P_h}\left[L(\sqrt{n}(\widehat{\psi}_n - \psi(\theta_h)))\right] = \mathbb{E}[L(Z)].$$

5.2.3 An extension to functionals of GLM parameters

In our experiments, we will investigate the behavior of locally private estimators for generalized linear models on a variable Y conditioned on X, where the model has the form

$$p_\theta(y \mid x) = \exp\left(T(x, y)^T \theta - A(\theta \mid x)\right), \tag{35}$$

where $A(\theta \mid x) = \int e^{T(x,y)^T\theta} d\mu(y)$ for some base measure μ and $T : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ is the sufficient statistic. We assume the distribution P_{x} on X is known. This assumption is strong, but may (approximately) hold in practice; in biological applications, for example, we may have covariate data and wish to estimate the conditional distribution of $Y \mid X$ for a new outcome Y [e.g. 11]. For a distribution P on the pair (X, Y), let P_{x} denote the marginal over X, which we assume is fixed and known, $P_{\mathsf{y}|\mathsf{x}}$ be the conditional distribution over Y given X, and $P = P_{\mathsf{y}|\mathsf{x}} P_{\mathsf{x}}$ for shorthand. Define the population risk using the log loss $\ell_\theta(y \mid x) = -\log p_\theta(y \mid x)$, by

$$R_P(\theta) = \mathbb{E}_P[\ell_\theta(Y \mid X)] = \mathbb{E}_P[-T(X,Y)]^T\theta + \mathbb{E}_{P_{\mathsf{x}}}[A(\theta \mid X)] = -\mathbb{E}_P[T(X,Y)]^T\theta + A_{P_{\mathsf{x}}}(\theta),$$

where we use the shorthand $A_{P_{\mathsf{x}}}(\theta) := \mathbb{E}_{P_{\mathsf{x}}}[A(\theta \mid X)]$. Let $\mathcal{P}_{\mathsf{y}}$ be a collection of conditional distributions of $Y \mid X$, and for $P_{\mathsf{y}|\mathsf{x}} \in \mathcal{P}_{\mathsf{y}}$, we analogize the general exponential family case to define

$$\theta(P_{\mathsf{y}|\mathsf{x}}) := \operatorname*{argmin}_\theta R_{P_{\mathsf{y}|\mathsf{x}} P_{\mathsf{x}}}(\theta) = \nabla A^*_{P_{\mathsf{x}}}(\mathbb{E}_{P_{\mathsf{y}|\mathsf{x}} P_{\mathsf{x}}}[T(X,Y)]).$$

Considering again the loss $L(\widehat{\psi} - \psi(\theta(P_{\mathsf{y}|\mathsf{x}}))$ for a smooth functional ψ, Corollary 8 implies

Corollary 9. *Let $\mathcal{P}_{\mathsf{y}}$ be a collection of conditional distributions on $Y \mid X$, $P_0 \in \mathcal{P}_{\mathsf{y}}$, and $\mathcal{Q}_\varepsilon$ be the collection of $(2, \varepsilon^2)$-Rényi-private channels (Def. 2). Then for numerical constants $c_0, c_1 > 0$ there exists $N = N(\mathcal{P}_{\mathsf{y}}, \psi)$ independent of the loss L such that $n \ge N$ implies*

$$\mathfrak{M}_n^{\mathrm{loc}}(P_0, L, \mathcal{P}_{\mathsf{y}}, \mathcal{Q}_\varepsilon) \ge c_0 \sup_{P_{\mathsf{y}|\mathsf{x}} \in \mathcal{P}_{\mathsf{y}}} L\left(c_1 \frac{\nabla\psi(\theta_0)^T \nabla^2 A_{P_{\mathsf{x}}}(\theta_0)^{-1} (\mathbb{E}_{P_0 P_{\mathsf{x}}}[T(X,Y)] - \mathbb{E}_{P_{\mathsf{y}|\mathsf{x}} P_{\mathsf{x}}}[T(X,Y)])}{\sqrt{n\varepsilon^2}}\right).$$

If the set $\mathcal{P}_{\mathsf{y}}$ and distribution P_{x} are such that $\{\mathbb{E}_{P_{\mathsf{y}|\mathsf{x}} P_{\mathsf{x}}}[T] \mid P_{\mathsf{y}|\mathsf{x}} \in \mathcal{P}_{\mathsf{y}}\} \supset \{t \in \mathbb{R}^d \mid \|t\| \le r\}$, then we have the simplified lower bound

$$\mathfrak{M}_n^{\mathrm{loc}}(P_0, L, \mathcal{P}_{\mathsf{y}}, \mathcal{Q}_\varepsilon) \ge c_0 L\Big(c_1 \frac{r \left\|\nabla^2 A_{P_{\mathsf{x}}}(\theta_0)^{-1} \nabla\psi(\theta_0)\right\|_*}{\sqrt{n\varepsilon^2}}\Big).$$

An optimal estimator parallels Section 5.2.2. Split a non-private sample $\{(X_i, Y_i)\}_{i=1}^n$ into samples of size $n_1 = \lceil n^{2/3} \rceil$ and $n_2 = n - n_1$. For $i = 1, \ldots, n_1$, let Z_i be any ε-locally differentially private estimate of $T(X_i, Y_i)$ with $\mathbb{E}[Z_i \mid X_i, Y_i] = T(X_i, Y_i)$ and $\mathbb{E}[\|Z_i\|^2] < \infty$, and define $\widetilde{\mu}_n = \overline{Z}_{n_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} Z_i$ and $\widetilde{\theta}_n = \nabla A^*_{P_{\mathsf{x}}}(\widetilde{\mu}_n) = \operatorname{argmin}_\theta\{-\widetilde{\mu}_n^T\theta + A_{P_{\mathsf{x}}}(\theta)\}$. Then, for $i = n_1 + 1, \ldots, n$, let

$$Z_i = \nabla\psi(\widetilde{\theta}_n)^T \nabla^2 A_{P_{\mathsf{x}}}(\widetilde{\theta}_n)^{-1} T(X_i, Y_i) + \frac{r\|\nabla^2 A_{P_{\mathsf{x}}}(\widetilde{\theta}_n)^{-1} \nabla\psi(\widetilde{\theta}_n)\|_*}{\varepsilon} W_i \quad \text{where} \quad W_i \overset{\mathrm{iid}}{\sim} \mathsf{Lap}(1),$$

The Z_i are evidently ε-differentially private, and we then define the private estimator

$$\widehat{\psi}_n := \overline{Z}_{n_2} + \nabla\psi(\widetilde{\theta}_n)^T \left(\widetilde{\theta}_n - \nabla^2 A_{P_{\mathsf{x}}}(\widetilde{\theta}_n)^{-1} \widetilde{\mu}_n\right). \tag{36}$$

An identical analysis to that we use to prove Proposition 7 then gives the following corollary, which shows a locally uniform optimal rate of convergence. (We use the shorthand $\|x\|_C^2 = x^T C x$.)

Corollary 10. *Let $\widehat{\psi}_n$ be the estimator (36) and $\theta_0 = \nabla A^*_{P_{\mathsf{x}}}(\mathbb{E}_P[T(X,Y)]) = \operatorname{argmin}_\theta R_P(\theta)$. Then*

$$\sqrt{n}(\widehat{\psi}_n - \psi(\theta_0)) \xrightarrow{d} \mathsf{N}\left(0, \left\|\nabla^2 A_{P_{\mathsf{x}}}(\theta_0)^{-1} v\right\|^2_{\mathrm{Cov}(T(X,Y))} + \frac{2}{\varepsilon^2} \left\|\nabla^2 A_{P_{\mathsf{x}}}(\theta_0)^{-1} v\right\|_*^2\right),$$

and convergence is locally uniform over QMD submodels.

5.3 Model adaptation in locally private exponential family estimation

We conclude this section by highlighting a phenomenon that distinguishes locally private estimation from non-private estimation, focusing especially on exponential families as in Section 5.2. We recall Stein [48], who roughly asks the following: given a parameter θ of interest and a (potentially) infinite dimensional nuisance G, can we estimate θ asymptotically as well regardless of whether we know G? Here, we consider this in the context of G being the full distribution P_0, and we delineate cases—which depend on the channel set $\mathcal{Q}$ being either the identity (non-private) or a private collection—when for a sub-family $\mathcal{P}_{\sub,0} \subset \mathcal{P}$ containing P_0, we have

$$\mathfrak{M}_n^{\loc}(P_0, L, \mathcal{P}, \mathcal{Q}) \asymp \mathfrak{M}_n^{\loc}(P_0, L, \mathcal{P}_{\sub,0}, \mathcal{Q}). \tag{37}$$

For exponential families, the non-private local minimax risk is (up to constants) independent of whether the containing family $\mathcal{P}$ of distributions is parametric or non-parametric, while the private local minimax risk is larger in the non-parametric than parametric settings, necessitating the construction of distinct private estimators with different optimality properties that depend on the overall model the statistician is willing to assume.

For simplicity we study one-dimensional potentially misspecified models with densities $p_\theta(x) = \exp(\theta x - A(\theta))$ and base measure μ. We consider nonparametric and parametric families, making an assumption (for convenience) that the first has uniformly bounded (arbitrary) fourth moment:

$$\mathcal{P}_{\text{non-par}} := \left\{P : \mathbb{E}_P[X] \in \Range(\nabla A), \mathbb{E}_P[|X|^4] \le M < \infty\right\} \quad \text{and} \quad \mathcal{P}_{\text{par}} := \{P_\theta\}_{\theta \in \Theta}.$$

To avoid issues of infinite loss, we use the truncated squared error $L_{\wedge B}(\theta - \theta(P)) = (\theta - \theta(P))^2 \wedge B$, where $0 < B < \infty$ is otherwise arbitrary.

To compare the private and non-private cases, we evaluate their local minimax risks. In the non-private case, the model class is immaterial, as the efficient influence and score functions for exponential families are identical in both parametric and nonparametric cases [52, Ch. 25.3], so we have the equivalence (37) when $\mathcal{Q} = \{\id\}$. We prove the following characterization in Section A.4.1.

Claim 5.1. *Let $P_0 = P_{\theta_0}$ belong to the exponential family above. Then for large enough n,*

$$\mathfrak{M}_n^{\loc}(P_0, L_{\wedge B}, \mathcal{P}_{\text{non-par}}, \{\id\}) \asymp \mathfrak{M}_n^{\loc}(P_0, L_{\wedge B}, \mathcal{P}_{\text{par}}, \{\id\}) \asymp \frac{1}{n \Var_0(X)}.$$

The risks, by comparison, have different behavior, as the discussion below shows.

Claim 5.2. *Let $P_0 = P_{\theta_0}$ belong to the exponential family above. Let $\mathcal{Q}_\varepsilon$ be the collection of $(2, \varepsilon^2)$-Rényi differentially private channels and $\mathcal{P}_{\text{all},0} \subset \mathcal{P}_{\text{non-par}}$ be a collection of sub-models with scores g dense in $L^1(P_0)$ at P_0. Then there exist numerical constants $0 < c_0 \le c_1 < \infty$ such that for large enough n,*

$$\mathfrak{M}_n^{\loc}(P_0, L_{\wedge B}, \mathcal{P}_{\text{all},0}, \mathcal{Q}_\varepsilon) \ge c_0 \cdot \frac{1}{n\varepsilon^2} \operatorname*{ess\,sup}_x \frac{(\mathbb{E}_{P_0}[X] - x)^2}{\Var_0(X)^2} \tag{38a}$$

and

$$\mathfrak{M}_n^{\loc}(P_0, L_{\wedge B}, \mathcal{P}_{\text{par}}, \mathcal{Q}_\varepsilon) \in [c_0, c_1] \cdot \frac{1}{n\varepsilon^2} \cdot \frac{1}{\mathbb{E}_0[|X - \mathbb{E}_0[X]|]^2}. \tag{38b}$$

Additionally, $1/\mathbb{E}_0[|X - \mathbb{E}_0[X]|]^2 \le \operatorname{ess\,sup}_x (\mathbb{E}_0[X] - x)^2 / \Var_0(X)^2$.

An alternative way to view (and prove) the right-hand (variance) quantities in the claims is via influence and score functions. The efficient influence function in exponential families [52, Ch. 25.3]

is $\dot{\theta}_0(x) = (x - \mathbb{E}_0[X])/\mathrm{Var}_0(X) = (x - A'(\theta_0))/A''(\theta_0)$, with the second equality following because $P_0 = P_{\theta_0}$ by assumption. The asymptotic variance [52, Ex. 25.16] in the non-private case becomes

$$\sup_{g \in L^2(P_0)} \frac{\mathbb{E}_0[\dot{\theta}_0(X) g(X)]^2}{\mathbb{E}_0[g(X)^2]} = \left\|\dot{\theta}_0(X)\right\|_{L^2(P_0)}^2 = \frac{1}{\mathrm{Var}_0(X)},$$

attaining the supremum at the parametric score $\dot{\ell}_{\theta_0}(x) = x - \mathbb{E}_0[X]$. In the private case, we have

$$\sup_g \frac{\mathbb{E}_0[\dot{\theta}_0(X) g(X)]^2}{\mathbb{E}_0[|g(X)|]^2} = \operatorname*{ess\,sup}_x \frac{(\mathbb{E}_0[X] - x)^2}{\mathrm{Var}_0(X)^2} \quad \text{while} \quad \frac{\mathbb{E}_0[\dot{\theta}_0(X) \dot{\ell}_{\theta_0}(X)]^2}{\mathbb{E}_0[|\dot{\ell}_\theta(X)|]^2} = \frac{1}{\mathbb{E}_0[|X - \mathbb{E}_0[X]|]^2}$$

as in inequalities (38a) and (38b), respectively. (Applying Corollary 8 thus demonstrates the lower bound (38a), and the preceding display also gives the final result in Claim 5.2. For the bounds (38b), use Proposition 4 and Corollary 6 with score $\dot{\ell}_\theta(x) = x - A'(\theta)$.) This contrast shows how the worst score g in the nonparametric case depends strongly on whether we have privacy or not; in the latter, it is simply $g = \dot{\ell}_{\theta_0}$, while in the former, the structure is quite different.

6 Experiments on a flow cytometry dataset

We perform experiments investigating the behavior of our proposed locally optimal estimators, comparing their performance both to non-private estimators and to minimax optimal estimators developed by Duchi et al. [25] for locally private estimation. We consider the generalized linear model (35) and estimating the linear functional $\psi(\theta) = v^T\theta$. As motivation, consider the problem of testing whether a covariate X_j is relevant to a binary outcome $Y \in \{-1, 1\}$. In this case, the logistic GLM model (35) is $p_\theta(y \mid x) = \exp(yx^T\theta)/(1 + \exp(yx^T\theta))$, and using the standard basis vectors $v = e_j$, estimating $v^T\theta$ corresponds to testing $\theta_j \lesseqgtr 0$ while controlling for the other covariates.

We investigate the performance of the locally private one-step estimator (36) on a flow-cytometry dataset for predicting protein expression [35, Ch. 17], comparing against (global) minimax optimal stochastic gradient estimators [25]. The flow-cytometry dataset contains expression level measurements of $d = 11$ proteins on $n = 7466$ cells, and the goal is to understand the network structure linking the proteins: how does protein j's expression level depend on the remaining proteins. As the raw data is heavy-tailed and skewed, we perform an inverse tangent transformation $x_{ij} \mapsto \tan^{-1}(x_{ij})$. Letting $X \in \mathbb{R}^{n \times d}$ be the data matrix, to compare the methods and to guarantee a ground truth in our experiments, we treat X as the *full population*, so each experiment consists of sampling rows of X with replacement.

Let $x \in \mathbb{R}^d$ denote a row of X. For $i \in [d]$, we wish to predict $y = \mathrm{sign}(x_i)$ based on $x_{-i} \in \mathbb{R}^{d-1}$, the remaining covariates, and we use the logistic regression model

$$\log \frac{P_\theta(\mathrm{sign}(x_i) = 1 \mid x_{-i})}{P_\theta(\mathrm{sign}(x_i) = -1 \mid x_{-i})} = \theta^T x_{-i} + \theta_{\mathrm{bias}},$$

so that $T(x_{-i}, y) = y[x_{-i}^T \; 1]^T$ and $A(\theta \mid x_{-i}) = \log(e^{\theta^T x_{-i} + \theta_{\mathrm{bias}}} + e^{-\theta^T x_{-i} - \theta_{\mathrm{bias}}})$, where $y = \mathrm{sign}(x_i)$ is the sign of the expression level of protein i. We let $\theta_{\mathrm{ml}}^{(i)} \in \mathbb{R}^d$ be the parameter (including the bias) maximizing the likelihood for this logistic model of predicting x_i using the full data X.

We perform multiple experiments, where each is as follows. We sample N rows of X uniformly (with replacement) and vary the privacy parameter in $\varepsilon \in \{1, 4\}$. We perform perform two private procedures (and one non-private procedure) on the resampled data $X_{\mathrm{new}} \in \mathbb{R}^{N \times d}$:

(i) The non-private maximum likelihood estimator (MLE) on the resampled data of size N.

Sample size	$N = 2n$		$N = 8n$		$N = 40n$	
Privacy ε	$\varepsilon = 1$	$\varepsilon = 4$	$\varepsilon = 1$	$\varepsilon = 4$	$\varepsilon = 1$	$\varepsilon = 4$
vs. initializer	0.501	0.82	0.791	0.848	0.825	0.852
vs. minimax (stochastic gradient)	0.321	0.677	0.659	0.79	0.777	0.817

Table 1. Frequency with which the one-step estimator outperforms initialization and minimax (stochastic-gradient-based) estimator over $T = 100$ tests, all coordinates j of the parameter and proteins $i = 1, \ldots, d$ for the flow-cytometry data.

(ii) The minimax optimal stochastic gradient procedure of Duchi et al. [25, Secs. 4.2.3 & 5.2]. In brief, this procedure begins from $\theta^0 = 0$, and at iteration k draws a pair (x, y) uniformly at random, then uses a carefully designed ε-locally private version Z^k of $T = T(x, y)$ with the property that $\mathbb{E}[Z \mid x, y] = T(x, y)$ and $\sup_k \mathbb{E}[\|Z^k\|^2] < \infty$, updating

$$\theta^{k+1} = \theta^k - \eta_k \left(\nabla A_{P_{\mathsf{x}}}(\theta^k) - Z^k \right),$$

where $\eta_k > 0$ is a stepsize sequence. (We use optimal the ℓ_∞ sampling mechanism [25, Sec. 4.2.3] to construct Z_i.) We use stepsizes $\eta_k = 1/(20\sqrt{k})$, which gave optimal performance over many choices of stepsize and power $k^{-\beta}$. We perform N steps of this stochastic gradient method, yielding estimator $\widehat{\theta}^{(i)}_{\mathrm{sg}}$ for prediction of protein i from the others.

(iii) The one-step corrected estimator (36). To construct the initial $\widetilde{\theta}_n$, we use Duchi et al.'s ℓ_∞ sampling mechanism to construct the approximation $\widetilde{\mu}_n = \frac{1}{n_1} \sum_{i=1}^n Z_i$ and let $\widehat{\theta}^{(i)}_{\mathrm{init}} = \widetilde{\theta}_n = \nabla A^*(P_{\mathsf{x}})(\widetilde{\mu}_n)$. For coordinates $i = 1, \ldots, d$, we set $\psi(\theta) = v^T \theta$ for $v = e_1, \ldots, e_d$ as in (36).

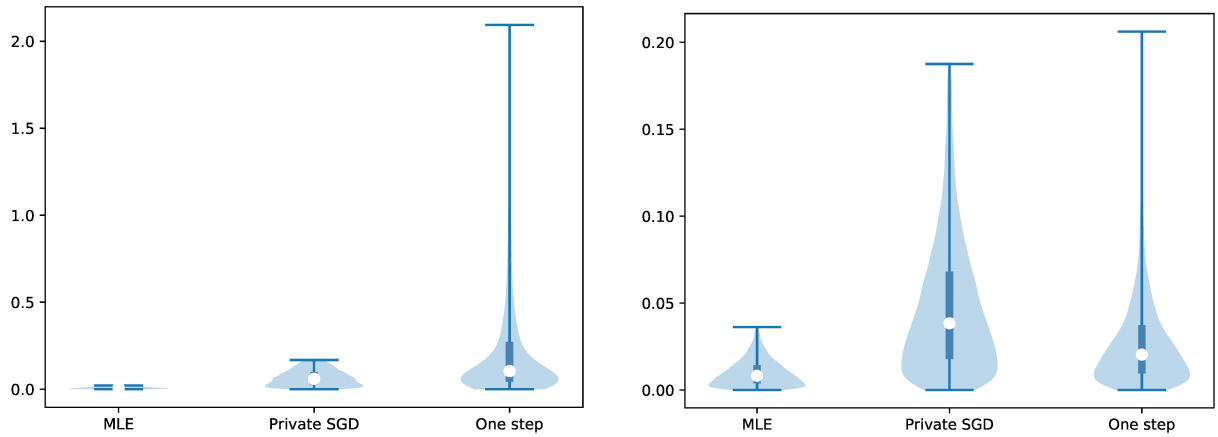
We perform each of these three-part tests $T = 100$ times, where within each test, each method uses an identical sample (the samples are of course independent across tests).

We summarize our results in Figure 1 and Table 1. Figure 1 plots the errors across all coordinates of $\theta^{(i)}_{\mathrm{ml}}$, $i = 1, \ldots, d$, and all $T = 100$ tests of the three procedures, with top whisker at the 99th percentile error for each. We vary sample sizes $N \in \{2n, 8n, 40n\}$ and privacy level $\varepsilon \in \{1, 4\}$; results remain consistent for other sample sizes. As the sample size (or ε) grows, the one-step estimator converges more quickly than the minimax stochastic gradient procedure, though for the smaller sample size the private SGD method exhibits better performance.

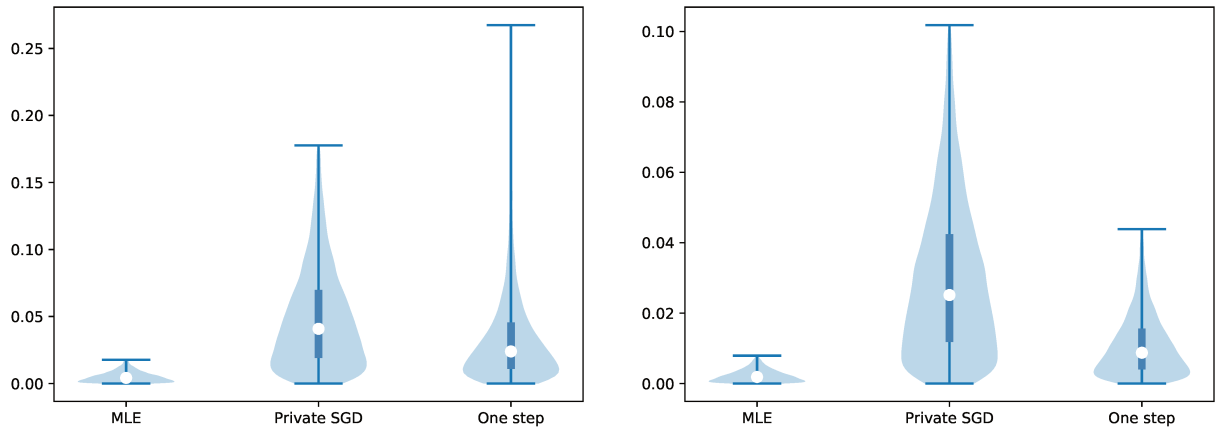
In Table 1, we compare the estimators $\widehat{\theta}^{(i)}_{\mathrm{init}}$, $\widehat{\theta}^{(i)}_{\mathrm{sg}}$, and $\widehat{\theta}^{(i)}_{\mathrm{os}}$ of the true parameter $\theta^{(i)}_{\mathrm{ml}}$ more directly. For each, we count the number of experiments (of T) and indices $j = 1, \ldots, d$ for which

$$\left| [\widehat{\theta}^{(i)}_{\mathrm{os}}]_j - [\theta^{(i)}_{\mathrm{ml}}]_j \right| < \left| [\widehat{\theta}^{(i)}_{\mathrm{init}}]_j - [\theta^{(i)}_{\mathrm{ml}}]_j \right| \quad \text{and} \quad \left| [\widehat{\theta}^{(i)}_{\mathrm{os}}]_j - [\theta^{(i)}_{\mathrm{ml}}]_j] \right| < \left| [\widehat{\theta}^{(i)}_{\mathrm{sg}}]_j - [\theta^{(i)}_{\mathrm{ml}}]_j \right|,$$

that is, the number of experiments in which the one-step estimator provides a better estimate than its initializer or the minimax stochastic gradient-based procedure. Table 1 shows these results, displaying the proportion of experiments in which the one-step method has higher accuracy than the other procedures. For large sample sizes, the asymptotic optimality of the one-step appears to be salient, as its performance relative to the other methods improves. Based on additional simulations, it appears that the initializer $\widehat{\theta}^{(i)}_{\mathrm{init}}$ is inaccurate for small sample sizes, so the one-step correction has poor Hessian estimate and performs poorly. The full minimax procedure [25] adds more noise than is necessary, as it privatizes the entire statistic xy in each iteration—a necessity because it iteratively builds the estimates $\widehat{\theta}^{(\cdot)}_{\mathrm{sg}}$—causing an increase in sample complexity.



(a) $N = 2n$, $\varepsilon = 1$ (b) $N = 2n$, $\varepsilon = 4$



(c) $N = 8n$, $\varepsilon = 1$ (d) $N = 40n$, $\varepsilon = 1$

Figure 1. Errors $|\widehat{\psi}_N - v^T\theta^{(i)}_{\rm ml}|$ across all experiments, for $v = e_1, \ldots, e_d$ and $i = 1, \ldots, d$, in the logistic regression model, with medians and interquartile ranges marked.

The one-step correction typically outperforms alternative approaches in large-sample regimes, and such large samples may be more effectively achievable than is *prima facie* obvious, as locally private procedures can guarantee strong central differential privacy. Erlingsson et al. [31] consider privacy amplification in the *shuffle model*, where the data $\{X_i\}$ are permuted before the sampling $Z_i \sim Q(\cdot \mid X_i, Z_{1:i-1})$; other variants [4] randomize and then permute the Z_i into $Z_{\pi(1:n)} \in \mathcal{Z}^n$. The permuted vector $Z_{\pi(1:n)}$ then achieves $(\varepsilon_{\rm cen}, \delta)$-differential privacy [4, Corollary 5.3.1] for

$$\varepsilon_{\rm cen} = O(1)e^{\varepsilon}\sqrt{\frac{\min\{1, \varepsilon^2\}\log\frac{1}{\delta}}{n}}.$$

Applying this randomize-then-shuffle approach $K \ll n$ distinct times, whenever $\varepsilon = O(1)$, composition bounds for differential privacy [26, Ch. 3.5.2] guarantee $(\varepsilon_{\rm cen}, \delta)$-central differential privacy for $\varepsilon_{\rm cen} = O(1)\sqrt{\frac{K\varepsilon^2 \log \delta^{-1}}{n}}$. The one-step estimator (36) falls in this framework, and (via a calculation) achieves $\varepsilon_{\rm cen} \le 1$ for $N = 40n$, $\varepsilon = 1$. Consequently, this type of behavior may be acceptable in natural local privacy applications: situations (such as web-scale data) with large sample sizes or where resampling is possible, as we may achieve both strong privacy and reasonable performance.

7 Proofs of main results

We collect the proofs of our main results in this section, as they are reasonably brief and (we hope) elucidating. The main technical tool underpinning our lower bounds is that our definitions of privacy imply strong contractions on the space of probability measures. Such contractive properties have been important in the study of information channels and *strong data processing* [14, 16] and in the mixing properites of Markov chains under so-called *strong mixing conditions*, such as the Dobrushin condition [17]. Consequently, before turning to the main proofs, we first present a few results on contractions of probability measures, as they underly our subsequent development.

7.1 Contractions of probability measures

We provide our contractions using f-divergences. For a convex function $f : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$ with $f(1) = 0$, the f-divergence between distributions P and Q is

$$D_f\left(P \| Q\right) := \int f\left(\frac{dP}{dQ}\right) dQ,$$

which is non-negative and strictly positive when $P \neq Q$ and f is strictly convex at the point 1. We typically consider f-divergences parameterized by $k \in [1, \infty)$ of the form

$$f_k(t) := |t - 1|^k.$$

Given a channel Q, for $a \in \{0, 1\}$, define the marginal distributions

$$M_a(S) := \int Q(S \mid x) dP_a(x).$$

The goal is then to provide upper bounds on the f-divergence $D_f\left(M_0 \| M_1\right)$ in terms of the channel Q; the standard data-processing inequality [15, 42] guarantees $D_f\left(M_0 \| M_1\right) \leq D_f\left(P_0 \| P_1\right)$. Dobrushin's celebrated ergodic coefficient $\alpha(Q) := 1 - \sup_{x,x'} \|Q(\cdot \mid x) - Q(\cdot \mid x')\|_{\mathrm{TV}}$ guarantees that for any f-divergence (see [14, 16]),

$$D_f\left(M_0 \| M_1\right) \leq \sup_{x,x'} \left\|Q(\cdot \mid x) - Q(\cdot \mid x')\right\|_{\mathrm{TV}} D_f\left(P_0 \| P_1\right). \tag{39}$$

Thus, as long as the Dobrushin coefficient is strictly positive, one obtains a strong data processing inequality. In our case, our privacy guarantees provide a stronger condition than the positivity of the Dobrushin coefficient. Consequently, we are able to provide substantially stronger data processing inequalities: we can even show that it is possible to modify the underlying f-divergence.

We have the following proposition, which provides a strong data processing inequality for all channels that are uniformly close under the polynomial f-divergences with f_k.

Proposition 8. *Let $f_k(t) = |t - 1|^k$ for some $k > 1$, and let P_0 and P_1 be arbitrary distributions on a common space $\mathcal{X}$. Let Q be a Markov kernel from $\mathcal{X}$ to $\mathcal{Z}$ satisfying*

$$D_{f_k}\left(Q(\cdot \mid x) \| Q(\cdot \mid x')\right) \leq \varepsilon^k$$

for all $x, x' \in \mathcal{X}$ and $M_a(\cdot) = \int Q(\cdot \mid x) dP_a(x)$. Then

$$D_{f_k}\left(M_0 \| M_1\right) \leq (2\varepsilon)^k \left\|P_0 - P_1\right\|_{\mathrm{TV}}^k.$$

See Section 7.1.1 for a proof.

Jensen's inequality implies that $2^k \|P_0 - P_1\|_{\mathrm{TV}}^k \le D_{f_k}(P_0 \| P_1)$, so Proposition 8 provides a stronger guarantee than the classical bound (39) for the specific divergence associated with $f_k(t) = |t-1|^k$. Because $\|P_0 - P_1\|_{\mathrm{TV}} \le 1$ for all P_0, P_1, it is possible that the f_k-divergence is infinite, while the marginals are much closer together. It is this transfer from power divergence to variation distance, that is, f_k to $f_1(t) = |t-1|$, that allows us to prove the strong localized lower bounds depending on variation distance such as Theorem 1.

We may parallel the proof of [25, Theorem 1] to obtain a tensorization result. In this context, the most important divergence for us is the Rényi 2-divergence (Def. 2), which corresponds to the case $k = 2$ (i.e. the χ^2-divergence) in Proposition 8, $f(t) = (t-1)^2$, and $D_{\chi^2}(P\|Q) = \exp(D_2(P\|Q)) - 1$. Recall the sequentially interactive formulation (1) and let

$$Q^n(S \mid x_{1:n}) := \int_{z_{1:n} \in S} \prod_{i=1}^n dQ(z_i \mid x_i, z_{1:i-1}).$$

Now, let $P_a, a = 0, 1$ be product distributions on $\mathcal{X}$, where we say that the distribution of X_i either follows $P_{0,i}$ or $P_{1,i}$, and define $M_a^n(\cdot) = \int Q^n(\cdot \mid x_{1:n}) dP_a(x_{1:n})$, noting that $dP_a(x_{1:n}) = \prod_{i=1}^n dP_{a,i}(x_i)$ as P_a is a product distribution. We have the following corollary.

Corollary 11. *Let Q be sequentially interactive and satisfy $(2, \varepsilon^2)$-Rényi privacy (Def. 2). Then*

$$D_{\chi^2}(M_0^n \| M_1^n) \le \prod_{i=1}^n \left(1 + 4\varepsilon^2 \|P_{0,i} - P_{1,i}\|_{\mathrm{TV}}^2\right) - 1.$$

See Section 7.1.2 for a proof. An immediate consequence of Corollary 11 and the fact [50, Lemma 2.7] that $D_{\mathrm{kl}}(P_0 \| P_1) \le \log(1 + D_{\chi^2}(P_0 \| P_1))$ yields

$$D_{\mathrm{kl}}(M_0^n \| M_1^n) \le \sum_{i=1}^n \log\left(1 + 4\varepsilon^2 \|P_{0,i} - P_{1,i}\|_{\mathrm{TV}}^2\right) \le 4\varepsilon^2 \sum_{i=1}^n \|P_{0,i} - P_{1,i}\|_{\mathrm{TV}}^2. \tag{40}$$

The tensorization (40) is the key to our results, as we see in the later sections.

7.1.1 Proof of Proposition 8

Let p_0 and p_1 be the densities of P_0, P_1 with respect to some base measure μ dominating P_0, P_1. Without loss of generality, we may assume that $\mathcal{Z}$ is finite, as all f-divergences are approximable by finite partitions [51]; we let m_a denote the associated p.m.f. For $k > 1$, the function $t \mapsto t^{1-k}$ is convex on $\mathbb{R}_+$. Thus, applying Jensen's inequality, we may bound $D_{f_k}(M_0 \| M_1)$ by

$$\begin{aligned} D_{f_k}(M_0 \| M_1) = \sum_z \frac{|m_0(z) - m_1(z)|^k}{m_1(z)^{k-1}} &\le \sum_z \int \frac{|m_0(z) - m_1(z)|^k}{q(z \mid x_0)^{k-1}} p_1(x_0) d\mu(x_0) \\ &= \int \underbrace{\left(\sum_z \frac{|m_0(z) - m_1(z)|^k}{q(z \mid x_0)^{k-1}}\right)}_{=: W(x_0)} p_1(x_0) d\mu(x_0). \end{aligned} \tag{41}$$

It thus suffices to upper bound $W(x_0)$. To do so, we rewrite $m_0(z) - m_1(z)$ as

$$m_0(z) - m_1(z) = \int q(z \mid x)(dP_0(x) - dP_1(x)) = \int (q(z \mid x) - q(z \mid x_0))(dP_0(x) - dP_1(x)),$$

where we have used that $\int (dP_0 - dP_1) = 0$. Now define the function

$$\Delta(z \mid x, x_0) := \frac{q(z \mid x) - q(z \mid x_0)}{q(z \mid x_0)^{1-1/k}}.$$

By Minkowski's integral inequality, we have the upper bound

$$W(x_0)^{1/k} = \left(\sum_z \left| \int \Delta(z \mid x, x_0)(p_0(x) - p_1(x)) d\mu(x) \right|^k \right)^{1/k} \tag{42}$$

$$\leq \int \left(\sum_z \left| \Delta(z \mid x, x_0)(p_0(x) - p_1(x)) \right|^k \right)^{1/k} d\mu(x) = \int \left(\sum_z |\Delta(z \mid x, x_0)|^k \right)^{\frac{1}{k}} |dP_0(x) - dP_1(x)|.$$

Now we compute the inner summation: we have that

$$\sum_z |\Delta(z \mid x, x_0)|^k = \sum_z \left| \frac{q(z \mid x)}{q(z \mid x_0)} - 1 \right|^k q(z \mid x_0) = D_{f_k} \left(Q(\cdot \mid x) \| Q(\cdot \mid x_0) \right).$$

Substituting this into our upper bound (42) on $W(x_0)$, we obtain that

$$W(x_0) \leq \sup_{x \in \mathcal{X}} D_{f_k} \left(Q(\cdot \mid x) \| Q(\cdot \mid x_0) \right) 2^k \left\| P_0 - P_1 \right\|_{\mathrm{TV}}^k,$$

as $\int |dP_0 - dP_1| = 2 \left\| P_0 - P_1 \right\|_{\mathrm{TV}}$. Substitute this upper bound into inequality (41).

7.1.2 Proof of Corollary 11

We use an inductive argument. The base case in which $n = 1$ follows immediately by Proposition 8. Now, suppose that Corollary 11 holds at $n - 1$; we will show that the claim holds for $n \in \mathbb{N}$. We use the shorthand $m_a(z_{1:k})$ for the density of the measure M_a^k, $a \in \{0, 1\}$ and $k \in \mathbb{N}$, which we may assume exists w.l.o.g. Then, by definition of the χ^2-divergence, we have

$$D_{\chi^2} \left(M_0^n \| M_1^n \right) + 1 = \mathbb{E}_{M_1} \left[\frac{m_0^2(Z_{1:n})}{m_1^2(Z_{1:n})} \right] = \mathbb{E}_{M_1} \left[\frac{m_0^2(Z_{1:n-1})}{m_1^2(Z_{1:n-1})} \mathbb{E}_{M_1} \left[\frac{m_0^2(Z_n \mid Z_{1:n-1})}{m_1^2(Z_n \mid Z_{1:n-1})} \mid Z_{1:n-1} \right] \right].$$

Noting that the kth marginal distributions $M_{a,k}(\cdot \mid z_{1:k-1}) = \int Q(\cdot \mid x, z_{1:k-1}) dP_{a,i}(x)$ for $a \in \{0, 1\}$, we see that for any $z_{1:n-1} \in \mathcal{Z}^{n-1}$,

$$\begin{aligned}
\mathbb{E}_{M_1} \left[\frac{m_0^2(Z_n \mid z_{1:n-1})}{m_1^2(Z_n \mid z_{1:n-1})} \mid z_{1:n-1} \right] &= 1 + D_{\chi^2} \left(M_{0,n}(\cdot \mid z_{1:n-1}) \| M_{1,n}(\cdot \mid z_{1:n-1}) \right) \\
&\leq 1 + 4\varepsilon^2 \left\| P_{0,n}(\cdot \mid z_{1:n-1}) - P_{1,n}(\cdot \mid z_{1:n-1}) \right\|_{\mathrm{TV}}^2 \\
&= 1 + 4\varepsilon^2 \left\| P_{0,n} - P_{1,n} \right\|_{\mathrm{TV}}^2,
\end{aligned}$$

where the inequality is Proposition 8 and the final equality follows because X_n is independent of $Z_{1:n-1}$. This yields the inductive step and completes the proof once we recall the inductive hypothesis and that $\mathbb{E}_{M_1}[\frac{m_0^2(Z_{1:n-1})}{m_1^2(Z_{1:n-1})}] = D_{\chi^2}(M_0^{n-1} \| M_1^{n-1}) + 1$.

7.2 Proof of Theorem 1

We follow the typical reduction of estimation to testing, common in the literature on lower bounds [2, 25, 50, 57]. For shorthand, let $\theta_v = \theta(P_v)$ for $v = 0, 1$ throughout the proof. Define the "distance"

$$d_L(P_0, P_1) := \inf_\theta \left\{L(\theta - \theta(P_0)) + L(\theta - \theta(P_1))\right\},$$

which satisfies $d_L(P_0, P_1) = 2L(\frac{\theta_0 - \theta_1}{2})$ when L is convex and (by quasi-convexity and symmetry) satisfies $d_L(P_0, P_1) \ge L(\frac{\theta_0 - \theta_1}{2})$. By definition of d_L, we have the mutual exclusion that for any θ,

$$L(\theta - \theta_0) < \frac{1}{2} d_L(P_0, P_1) \;\; \text{implies} \;\; L(\theta - \theta_1) \ge \frac{1}{2} d_L(P_0, P_1). \tag{43}$$

Let M_0^n and M_1^n be the marginal probabilities over observations $Z_{1:n}$ under P_0 and P_1 for a channel $Q \in \mathcal{Q}$. Using Markov's inequality, we have for any estimator $\widehat{\theta}$ based on $Z_{1:n}$ and any $\delta \ge 0$ that

$$\begin{aligned}
\mathbb{E}_{M_0^n}\left[L(\widehat{\theta} - \theta_0)\right] + \mathbb{E}_{M_1^n}\left[L(\widehat{\theta} - \theta_1)\right] &\ge \delta\left[M_0^n(L(\widehat{\theta} - \theta_0) \ge \delta) + M_1^n(L(\widehat{\theta} - \theta_1) \ge \delta)\right] \\
&= \delta\left[1 - M_0^n(L(\widehat{\theta} - \theta_0) < \delta) + M_1^n(L(\widehat{\theta} - \theta_1) \ge \delta)\right].
\end{aligned}$$

Setting $\delta = \delta_{01} := \frac{1}{2} d_L(P_0, P_1)$ and using the implication (43), we obtain

$$\begin{aligned}
\mathbb{E}_{M_0^n}\left[L(\widehat{\theta} - \theta_0)\right] + \mathbb{E}_{M_1^n}\left[L(\widehat{\theta} - \theta_1)\right] &\ge \delta_{01}\left[1 - M_0^n(L(\widehat{\theta} - \theta_0) < \delta) + M_1^n(L(\widehat{\theta} - \theta_1) \ge \delta)\right] \\
&\ge \delta_{01}\left[1 - M_0^n(L(\widehat{\theta} - \theta_1) \ge \delta) + M_1^n(L(\widehat{\theta} - \theta_1) \ge \delta)\right] \\
&\ge \delta_{01}\left[1 - \|M_0^n - M_1^n\|_{\mathrm{TV}}\right],
\end{aligned} \tag{44}$$

where in the last step we used the definition of the variation distance.

Now we make use of the contraction inequality of Corollary 11 and its consequence (40) for KL-divergences. By Pinsker's inequality and the corollary, we have

$$2\left\|M_0^n - M_1^n\right\|_{\mathrm{TV}}^2 \le D_{\mathrm{kl}}\left(M_0^n \| M_1^n\right) \le \log(1 + D_{\chi^2}\left(M_0^n \| M_1^n\right)) \le n \log\left(1 + 4\varepsilon^2 \left\|P_0 - P_1\right\|_{\mathrm{TV}}^2\right).$$

Substituting this into our preceding lower bound (44) and using that $\widehat{\theta}$ is arbitrary and $\delta_{01} = \frac{1}{2} d_L(P_0, P_1)$, we have that for any distributions P_0 and P_1,

$$\inf_{\widehat{\theta}} \inf_{Q \in \mathcal{Q}} \max_{P \in \{P_0, P_1\}} \mathbb{E}_P\left[L(\widehat{\theta} - \theta(P))\right] \ge \frac{1}{4} d_L(P_0, P_1)\left[1 - \sqrt{\frac{n}{2} \log\left(1 + 4\varepsilon^2 \left\|P_0 - P_1\right\|_{\mathrm{TV}}^2\right)}\right].$$

Now, for any $\delta \ge 0$, if $\frac{n}{2}\log(1 + 4\varepsilon^2\delta^2) \le \frac{1}{4}$, or equivalently, $\delta^2 \le \frac{1}{4\varepsilon^2}(\exp(\frac{1}{2n}) - 1)$, then $1 - \sqrt{\frac{n}{2}\log(1 + 4\varepsilon^2\delta^2)} \ge \frac{1}{2}$. Applying this to the bracketed term in the preceding display, we obtain

$$\begin{aligned}
\mathfrak{M}_n^{\mathrm{loc}}(P_0, L, \mathcal{P}, \mathcal{Q}) &\ge \frac{1}{8} \sup_{P_1 \in \mathcal{P}} \left\{d_L(P_0, P_1) \mid \left\|P_0 - P_1\right\|_{\mathrm{TV}}^2 \le \frac{1}{4\varepsilon^2}\left[e^{\frac{1}{2n}} - 1\right]\right\} \\
&\ge \frac{1}{8} \sup_{P_1 \in \mathcal{P}} \left\{d_L(P_0, P_1) \mid \left\|P_0 - P_1\right\|_{\mathrm{TV}}^2 \le \frac{1}{8n\varepsilon^2}\right\}
\end{aligned}$$

because $e^x - 1 \ge x$ for all x. When L is convex, this is precisely $\frac{1}{4}\omega_{L,\mathrm{TV}}(\frac{1}{\sqrt{8n\varepsilon^2}}; P_0, \mathcal{P})$, while in the quasi-convex case, it is at least $\frac{1}{8}\omega_{L,\mathrm{TV}}(\frac{1}{\sqrt{8n\varepsilon^2}}; P_0, \mathcal{P})$.

7.3 Proof of Proposition 3

Our starting point is a lemma extending [8, Thm. 1]. In the lemma and the remainder of this section, for measures P_0 and P_1 we define the 2-affinity

$$\rho\left(P_0 \| P_1\right) := D_{\chi^2}\left(P_0 \| P_1\right) + 1 = \mathbb{E}_{P_1}\left[\frac{dP_0^2}{dP_1^2}\right] = \mathbb{E}_{P_0}\left[\frac{dP_0}{dP_1}\right],$$

which measures the similarity between distributions P_0 and P_1. With these definitions, we have the following constrained risk inequality.

Lemma 3 ([22], Theorem 1)**.** *Let $\theta_0 = \theta(P_0)$, $\theta_1 = \theta(P_1)$, and define $\Delta = \Phi(\frac{1}{2}\left\|\theta_0 - \theta_1\right\|_2)$. If the estimator $\widehat{\theta}$ satisfies $R(\widehat{\theta}, \theta_0, P_0) \le \delta$ for some $\delta \ge 0$, then*

$$R(\widehat{\theta}, \theta_1, P_1) \ge \left[\Delta^{1/2} - \left(\rho(P_1 \| P_0) \cdot \delta\right)^{1/2}\right]_+^2.$$

The lemma shows that if an estimator has small risk under distribution P_0, then its risk for a nearby distribution P_1 must be nearly the distance between the associated parameters θ_0 and θ_1.

With Lemma 3 in hand, we can prove Proposition 3. For shorthand let $R_a(\widehat{\theta}) = R(\widehat{\theta}, \theta_a, M_a^n)$ denote the risk under the marginal M_a^n. By Lemma 3, for any distributions P_0 and P_1, we have

$$R_1(\widehat{\theta}) \ge \left[\Phi\left(\frac{1}{2}\left\|\theta_0 - \theta_1\right\|_2\right) - \left(\rho\left(M_1^n \| M_0^n\right) R(\widehat{\theta}, M_0^n)\right)^{1/2}\right]_+^2,$$

and by Corollary 11 we have

$$\rho\left(M_1^n \| M_0^n\right) \le \left(1 + 4\varepsilon^2 \left\|P_0 - P_1\right\|_{\mathrm{TV}}^2\right)^n \le \exp\left(4n\varepsilon^2 \left\|P_0 - P_1\right\|_{\mathrm{TV}}^2\right).$$

Let $\omega_L(\delta; P_0) = \omega_{L,\mathrm{TV}}(\delta; P_0, \mathcal{P})$ for shorthand. For $t \in [0,1]$, let $\mathcal{P}_t$ be the collection of distributions

$$\mathcal{P}_t := \left\{P \in \mathcal{P} \mid \left\|P_0 - P_1\right\|_{\mathrm{TV}}^2 \le t\frac{\log\frac{1}{\eta}}{4n\varepsilon^2}\right\},$$

so that under the conditions of the proposition, any distribution $P_1 \in \mathcal{P}_t$ satisfies

$$R_1(\widehat{\theta}) \ge \left[\Phi\left(\frac{1}{2}\left\|\theta_0 - \theta_1\right\|_2\right) - \eta^{\frac{(1-t)}{2}}\omega_L\left((4n\varepsilon^2)^{-1/2}; P_0\right)^{1/2}\right]_+^2. \tag{45}$$

As $L(\frac{1}{2}(\theta_0 - \theta(P_1))) = \Phi(\frac{1}{2}\left\|\theta_0 - \theta(P_1)\right\|_2)$, inequality (45) implies that for all $t \in [0,1]$, there exists $P_1 \in \mathcal{P}_t$ such that

$$R(\widehat{\theta}, M_1^n) \ge \left[\omega_L\left(\frac{\sqrt{t\log\frac{1}{\eta}}}{\sqrt{4n\varepsilon^2}}; P_0\right)^{1/2} - \eta^{\frac{(1-t)}{2}}\omega_L\left(\frac{1}{\sqrt{4n\varepsilon^2}}; P_0\right)^{1/2}\right]_+^2.$$

Because $\delta \mapsto \omega_L(\delta)$ is non-decreasing, if $t \in [0,1]$ we may choose $P_1 \in \mathcal{P}_t$ such that

$$R(\widehat{\theta}, M_1^n) \ge \left[1 - \eta^{(1-t)/2}\right]_+^2 \omega_L\left(\frac{\sqrt{t\log\frac{1}{\eta}}}{\sqrt{4n\varepsilon^2}}; P_0\right). \tag{46}$$

Lastly, we lower bound the modulus of continuity at P_0 by a modulus at P_1. We claim that under Condition C.1, for all $\delta > 0$, if $\|P_0 - P_1\|_{\rm TV} \le \delta$ then

$$\omega_L(2\delta; P_0) \ge \frac{1}{2\gamma}\omega_L(\delta; P_1). \tag{47}$$

Deferring the proof of this claim, note that by taking $\delta^2 = t \log \frac{1}{\eta}/(16n\varepsilon^2)$ in inequality (47), Eq. (46) implies that there exists $P_1 \in \mathcal{P}_t$ such that

$$R(\widehat{\theta}, M_1^n) \ge \left[1 - \eta^{(1-t)/2}\right]_+^2 \omega_L\left(2\delta; P_0\right) \ge \frac{1}{2\gamma}\left[1 - \eta^{(1-t)/2}\right]_+^2 \omega_L\left(\frac{1}{4}\frac{\sqrt{t\log\frac{1}{\eta}}}{\sqrt{n\varepsilon^2}}; P_1\right).$$

Let us return to the claim (47). For distributions P_0, P_1, P_2 with parameters $\theta_a = \theta(P_a)$,

$$L\left(\frac{\theta_1 - \theta_2}{2}\right) \le L(\theta_0 - \theta_1) + L(\theta_0 - \theta_2) \le \gamma L\left(\frac{\theta_0 - \theta_1}{2}\right) + \gamma L\left(\frac{\theta_0 - \theta_2}{2}\right)$$

by Condition C.1. Then for any $\delta \ge 0$ and P_1 with $\|P_1 - P_0\|_{\rm TV} \le \delta$, we have

$$\begin{aligned}
\omega_L(2\delta; P_0) &= \sup_{\|P_0 - P\|_{\rm TV} \le 2\delta} L\left(\frac{\theta_0 - \theta(P)}{2}\right) \ge \sup_{\|P_1 - P\|_{\rm TV} \le \delta} L\left(\frac{\theta_0 - \theta(P)}{2}\right) \\
&\ge \sup_{\|P - P_1\|_{\rm TV} \le \delta} \left\{\gamma^{-1} L\left(\frac{\theta_1 - \theta(P)}{2}\right) - L\left(\frac{\theta_0 - \theta_1}{2}\right)\right\} \ge \gamma^{-1}\omega_L(\delta; P_1) - \omega_L(\delta; P_0).
\end{aligned}$$

Rearranging, we have inequality (47), as for any distribution P_1 such that $\|P_0 - P_1\|_{\rm TV} \le \delta$,

$$2\omega_L(2\delta; P_0) \ge \omega_L(\delta; P_0) + \omega_L(2\delta; P_0) \ge \gamma^{-1}\omega_L(\delta; P_1).$$

8 Discussion

By the careful construction of locally optimal and adaptive estimators, as well as our local minimax lower bounds, we believe results in this paper indicate more precisely the challenges associated with locally private estimation. To illustrate this, let us reconsider the estimation of a linear functional $v^T\theta$ in a classical statistical problem. Let $\{P_\theta\}$ be a family with Fisher information matrices $\{I_\theta\}$ and score $\dot\ell_\theta : \mathcal{X} \to \mathbb{R}^d$. Then a classical estimators $\widehat{\theta}_n$ of the parameter θ_0 is efficient [52, Sec. 8.9] among regular estimators if and only if

$$\widehat{\theta}_n - \theta_0 = \frac{1}{n}\sum_{i=1}^n -I_{\theta_0}^{-1}\dot\ell_{\theta_0}(X_i) + o_P(1/\sqrt{n}),$$

and an efficient estimator $\widehat{\psi}_n$ of $v^T\theta$ satisfies $\widehat{\psi}_n = v^T\theta_0 - n^{-1}\sum_{i=1}^n v^T I_{\theta_0}^{-1}\dot\ell_{\theta_0}(X_i) + o_P(n^{-1/2})$. In constrast, in the private case, our rate-optimal estimators (recall Section 5.2) in the nonparametric case have the asymptotic form

$$\widehat{\psi}_{{\rm priv},n} = v^T\theta_0 - v^T\left(\frac{1}{n}\sum_{i=1}^n I_{\theta_0}^{-1}\dot\ell_{\theta_0}(X_i)\right) + \frac{1}{n}\sum_{i=1}^n W_i + o_P(1/\sqrt{n}),$$

where the random variables W_i must add noise of a magnitude scaling as $\frac{1}{\varepsilon}\sup_x |v^T I_{\theta_0}^{-1}\dot\ell_{\theta_0}(x)|$, because otherwise it is possible to distinguish examples for which $v^T I_{\theta_0}^{-1}\dot\ell_{\theta_0}(X_i)$ is large from those

for which it has small magnitude. This enforced lack of distinguishability of "easy" problems (those for which the scaled score $I_{\theta_0}^{-1}\dot{\ell}_{\theta_0}(X_i)$ is typically small) from "hard" problems (for which it is large) is a feature of local privacy schemes, and it helps to explain the difficulty of estimation, as well as to illustrate the more nuanced scaling of the best possible estimators with problem parameter θ_0, when $\sup_x |v^T I_{\theta_0}^{-1}\dot{\ell}_{\theta_0}(x)|$ may be similar to $\mathbb{E}_0[(v^T I_{\theta_0}^{-1}\dot{\ell}_{\theta_0}(X))^2]^{1/2}$, the optimal non-private asymptotic variance.

We thus believe it prudent to more carefully explore feasible definitions of privacy, especially in local senses. Regulatory decisions and protection against malfeasance may require less stringent notions of privacy than pure differential privacy, but local notions of privacy—where no sensitive non-privatized data leaves the hands of a sample participant—are desirable. The asymptotic expansions above suggest a notion of privacy that allows some type of *relative* noise addition, to preserve the easiness of "easy" problems, will help. Perhaps large values of ε, at least for high-dimensional problems, may still provide acceptable privacy protection, at least in concert with centralized privacy guarantees. We look forward to continuing study of these fundamental limitations and acceptable tradeoffs between data utility and protection of study participants.

A Proofs of non-private minimax results

In this appendix, we collect the (more or less standard) proofs of the results in Section 2.2.

A.1 Proof of Proposition 1

The lower bound follows the typical reduction of estimation to testing commin in the literature on lower bounds [2, 50, 57]. Fix any distribution $P_1 \in \mathcal{P}$, let $\theta_v = \theta(P_v)$ for shorthand, and define $\delta = |\theta_0 - \theta_1|/2$. Then for any $\theta \in \mathbb{R}$, that $|\theta - \theta_0| < \delta$ implies $|\theta - \theta_1| \ge \delta$. Thus we have

$$\begin{aligned}
\mathbb{E}_{P_0^n}\left[(\widehat{\theta} - \theta_0)^2\right] + \mathbb{E}_{P_1^n}\left[(\widehat{\theta} - \theta_1)^2\right] &\stackrel{(i)}{\ge} \delta^2\left[P_0^n\left(|\widehat{\theta} - \theta_0| \ge \delta\right) + P_1^n\left(|\widehat{\theta} - \theta_1| \ge \delta\right)\right] \\
&= \delta^2\left[1 - P_0^n\left(|\widehat{\theta} - \theta_0| < \delta\right) + P_1^n\left(|\widehat{\theta} - \theta_1| \ge \delta\right)\right] \\
&\stackrel{(ii)}{\ge} \delta^2\left[1 - P_0^n\left(|\widehat{\theta} - \theta_1| \ge \delta\right) + P_1^n\left(|\widehat{\theta} - \theta_1| \ge \delta\right)\right],
\end{aligned}$$

where inequality (i) is Markov's inequality, and the second is the implication preceding the display. By the definition of variation distance and that $\|P - Q\|_{\rm TV} \le \sqrt{2} d_{\rm hel}(P, Q)$ for any P, Q, we obtain

$$\mathbb{E}_{P_0^n}\left[(\widehat{\theta} - \theta_0)^2\right] + \mathbb{E}_{P_1^n}\left[(\widehat{\theta} - \theta_1)^2\right] \ge \delta\left[1 - \|P_0^n - P_1^n\|_{\rm TV}\right]. \ge \delta\left[1 - \sqrt{2} d_{\rm hel}(P, Q)\right]. \tag{48}$$

The tensorization properties of the Hellinger distance imply that

$$d_{\rm hel}^2(P_0^n, P_1^n) = \left[1 - \left(1 - d_{\rm hel}^2(P_0, P_1)\right)^n\right] \le n d_{\rm hel}^2(P_0, P_1),$$

and substituting this into the bound (48) gives that for any $P_1 \in \mathcal{P}$,

$$\mathfrak{M}_n^{\rm loc}(P_0, \mathcal{P}) \ge \frac{1}{2}\mathbb{E}_{P_0^n}\left[(\widehat{\theta} - \theta_0)^2\right] + \frac{1}{2}\mathbb{E}_{P_1^n}\left[(\widehat{\theta} - \theta_1)^2\right] \ge \frac{1}{8}\sup_{P_1 \in \mathcal{P}} (\theta(P_0) - \theta(P_1))^2\left[1 - \sqrt{2n d_{\rm hel}^2(P_0, P_1)}\right]_+.$$

Taking a supremum over all $P_1 \in \mathcal{P}$ satisfying $d_{\rm hel}^2(P_0, P_1) \le \frac{1}{4n}$ then implies

$$\mathfrak{M}_n^{\rm loc}(P_0, \mathcal{P}) \ge \frac{\sqrt{2} - 1}{8\sqrt{2}} \omega_{\rm hel}^2(n^{-1/2}/2; P_0, \mathcal{P}).$$

To prove the upper bound, we exhibit an estimator. Let $\theta_v = \theta(P_v)$ as above, and assume w.l.o.g. that P_a have densities p_a (take base measure $\mu = P_0 + P_1$). Define the acceptance set $A := \{x \in \mathcal{X}^n \mid \prod_{i=1}^n \frac{p_0(x_i)}{p_1(x_i)} \ge 1\}$ and estimator $\widehat{\theta}_n = \theta_0 1_A + \theta_1 1_{A^c}$. It is then immediate that

$$\max_{P \in \{P_0, P_1\}} \mathbb{E}_{P^n}\left[(\widehat{\theta}_n - \theta(P))^2\right] = (\theta_0 - \theta_1)^2 \max\left\{P_0^n(A^c), P_1^n(A)\right\} \le (\theta_0 - \theta_1)^2 \left[1 - \|P_0^n - P_1^n\|_{\mathrm{TV}}\right].$$

Using the tensorization properties of Hellinger distance and that $\|P - Q\|_{\mathrm{TV}} \ge d_{\mathrm{hel}}^2(P, Q)$ for any distributions P and Q, we obtain

$$\|P_0^n - P_1^n\|_{\mathrm{TV}} \ge d_{\mathrm{hel}}^2(P_0^n, P_1^n) = \left[1 - \left(1 - d_{\mathrm{hel}}^2(P_0, P_1)\right)^n\right] \ge 1 - \exp\left(-n d_{\mathrm{hel}}^2(P_0, P_1)\right),$$

so that

$$\max_{P \in \{P_0, P_1\}} \mathbb{E}_{P^n}\left[(\widehat{\theta}_n - \theta(P))^2\right] \le (\theta_0 - \theta_1)^2 \exp\left(-n d_{\mathrm{hel}}^2(P_0, P_1)\right).$$

Taking a supremum over P_1 gives the claimed upper bound.

Finally, we turn to the bound (7). We have by assumption that for any $\delta \ge n^{-1/2}/2$,

$$\begin{aligned}\left\{\omega_{\mathrm{hel}}(\delta; P_0, \mathcal{P}) \exp(-n\delta^2)\right\} &\le \omega_{\mathrm{hel}}(n^{-1/2}/2; P_0, \mathcal{P}) \cdot B(4n\delta^2)^{\beta/2} \exp(-n\delta^2) \\ &\le \omega_{\mathrm{hel}}(n^{-1/2}/2; P_0, \mathcal{P}) B \beta^{\beta/2} e^{-\beta/2},\end{aligned}$$

where the supremum is attained at $\delta^2 = \frac{\beta}{2n}$.

A.2 Proof of Claim 2.1

We use the shorthand $\omega_{\mathrm{hel}}(\delta; \theta_0, \Theta) = \sup_{\theta \in \Theta}\{|v^T(\theta_0 - \theta)| \mid d_{\mathrm{hel}}(P_\theta, P_{\theta_0}) \le \delta\}$. The lower bound is nearly immediate via Proposition 1: by the QMD assumption there exists $\delta > 0$ such that $\|h\| \le \delta$ implies $\frac{1}{9} h^T I_{\theta_0} h \le d_{\mathrm{hel}}^2(P_{\theta_0 + h}, P_\theta) \le \frac{1}{7} h^T I_{\theta_0} h$. Thus we obtain for all $n \gtrsim \frac{1}{\lambda_{\min}(I_{\theta_0})\delta^2}$ that

$$\begin{aligned}\omega_{\mathrm{hel}}(n^{-1/2}/2; \theta_0, \Theta) &= \sup_h \left\{|h^T v| \mid d_{\mathrm{hel}}^2(P_{\theta_0 + h}, P_{\theta_0}) \le \frac{1}{4n}\right\} \\ &\ge \sup_{\|h\| \le \delta} \left\{h^T v \mid h^T I_{\theta_0} h \le \frac{7}{4n}\right\} = \sup\left\{h^T v \mid h^T I_{\theta_0} h \le \frac{7}{4n}\right\} = \frac{\sqrt{7}}{2\sqrt{n}} \|I_{\theta_0}^{-1/2} v\|_2.\end{aligned}$$

For the upper bound, choose $\delta > 0$ such that $\|h\| \le \delta$ implies that $d_{\mathrm{hel}}^2(P_{\theta_0 + h}, P_{\theta_0}) \ge \frac{1}{9} h^T I_{\theta_0} h$, while $\|h\| > \delta$ implies that $d_{\mathrm{hel}}^2(P_{\theta_0 + h}, P_{\theta_0}) > \gamma > 0$; such a pair of δ and γ exist by Assumption A1 and quadratic mean differentiability. There thus exists $r_0 = r_0(\delta, \theta_0)$ such that $d_{\mathrm{hel}}(P_{\theta_0 + h}, P_{\theta_0}) \le r_0$ implies $\|h\| \le \delta$, and so for any $r \le r_0$, we have that $d_{\mathrm{hel}}^2(P_{\theta_0 + h}, P_{\theta_0}) \le r^2$ implies

$$\frac{1}{9} h^T I_{\theta_0} h \le d_{\mathrm{hel}}^2(P_{\theta_0 + h}, P_{\theta_0}) \le r^2.$$

Using this in the definition of the modulus of continuity yields

$$\omega_{\mathrm{hel}}(r; \theta_0, \Theta) = \sup_h \left\{|v^T h| \mid d_{\mathrm{hel}}^2(P_{\theta_0 + h}, P_{\theta_0}) \le r^2\right\} \le \sup_h \left\{v^T h \mid h^T I_{\theta_0} h \le 9r^2\right\} = 3r \|I_{\theta_0}^{-1/2} v\|_2$$

for all $r \le r_0$. Noting that $\omega_{\mathrm{hel}} \le \mathrm{diam}(\Theta)$ regardless, we apply Proposition 1 and observe

$$\begin{aligned}\sup_{r \ge 0}\left\{\omega_{\mathrm{hel}}^2(r; \theta_0, \Theta) \exp(-nr^2)\right\} &\le \max\left\{\sup_{0 \le r \le r_0} \omega_{\mathrm{hel}}^2(r; \theta_0, \Theta) \exp(-nr^2), \sup_{r > r_0} \omega_{\mathrm{hel}}^2(r; \theta_0, \Theta) \exp(-nr^2)\right\} \\ &\le \max\left\{\sup_{r \ge 0} 9r^2 v^T I_{\theta_0}^{-1} v \exp\left(-nr^2\right), \mathrm{diam}^2(\Theta) \exp(-nr_0^2)\right\} \\ &= \max\left\{\frac{9}{en} v^T I_{\theta_0}^{-1} v, \mathrm{diam}^2(\Theta) \exp(-nr_0^2)\right\}.\end{aligned}$$

A.3 Proof of Lemma 1

The proof is essentially [52, Lemma 8.14]. Letting $I_0 = \mathbb{E}_{P_0}[gg^T]$, we have under $P_0^n \times P_{\text{aux}}^n$ that

$$\log \frac{dP_{h_n/\sqrt{n}}^n \times dP_{\text{aux}}^n}{dP_0^n \times dP_{\text{aux}}^n}(X_{1:n}, \xi_{1:n}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h^T g(X_i) - h^T I_0 h + o_{P_0}(1)$$

(recall [52, Theorem 7.2]). Thus we have

$$\begin{bmatrix} \sqrt{n}(\widehat{\theta}_n - \theta(P_0)) \\ \log \frac{dP_{h_n/\sqrt{n}}^n \times dP_{\text{aux}}^n}{dP_0^n \times dP_{\text{aux}}^n} \end{bmatrix} \xrightarrow[P_0 \times P_{\text{aux}}]{d} \mathsf{N}\left(\begin{bmatrix} 0 \\ -\frac{1}{2} h^T I_0 h \end{bmatrix}, \begin{bmatrix} \Sigma_0 + \Sigma_{\text{aux}} & \mathbb{E}[\dot{\theta}_0(X) g(X)^T] h \\ h^T \mathbb{E}[g(X) \dot{\theta}_0(X)^T] & h^T I_0 h \end{bmatrix} \right).$$

Applying the delta method and Le Cam's third lemma [52, Example 6.7] gives that

$$\sqrt{n}(\widehat{\theta}_n - \theta(P_0)) \xrightarrow[P_{h_n/\sqrt{n}} \times P_{\text{aux}}]{d} \mathsf{N}\left(\mathbb{E}[\dot{\theta}_0(X) g(X)^T] h, \Sigma_0 + \Sigma_{\text{aux}} \right).$$

The differentiability of $h \mapsto \theta(P_h)$ at $h = 0$ then gives the first result.

The second limiting result follows by a standard compactness argument.

A.4 Inequality (17): bounds on the Hellinger modulus

For the lower bound on $\omega_{\text{hel}}(\delta)$, we use techniques from semiparametric inference [e.g. 52, Ch. 25]. Let $\theta_0 = \mathbb{E}_{P_0}[X]$ and define the function $g(x) = (x - \theta_0)$. Define the distribution $dP_t = [1 + tg]_+ \, dP_0 / C_t$, where $C_t = \int [1 + tg]_+ \, dP_0$. Then we have

$$\begin{aligned} 1 \le C_t &\le \int (1 + tg) dP_0 + t^2 \int \frac{[1 + tg]_+ - (1 + tg)}{t^2} dP_0 \\ &= 1 - t^2 \int \frac{1}{t^2} (1 + tg) \mathbf{1}\{g \le -1/t\} \, dP_0 \\ &\le 1 + t^2 \left[\frac{P_0(g \le -1/t)}{t^2} + \frac{\sqrt{\text{Var}_0(g) P_0(g \le -1/t)}}{t} \right] \le 1 + 2t^2 \text{Var}_0(g) \end{aligned}$$

by Chebyshev's inequality. A standard calculation [52, Ch. 25.3] via the dominated convergence theorem—with the observation that the influence function of the mean is $\dot{\theta}_0(x) = (x - \theta_0)$—yields

$$\mathbb{E}_{P_t}[X] = \theta_0 + t \text{Var}_{P_0}(X)(1 + o(1)) \quad \text{and} \quad d_{\text{hel}}^2(P_t, P_0) = \frac{1}{8} t^2 \text{Var}_{P_0}(X)(1 + o(1))$$

as $t \to 0$. Let t_0 be small enough that $|o(1)| \le \frac{1}{7}$ for $t \le t_0$. Then for $\delta^2 \le \text{Var}_0(X) t_0 / 7$,

$$\omega_{\text{hel}}(\delta) \ge \sup_{t \le t_0} \left\{ \frac{7}{8} t \text{Var}_{P_0}(X) \mid \frac{1}{7} t^2 \text{Var}_0(X) \le \delta^2 \right\} = \frac{7\sqrt{7}}{8} \sqrt{\text{Var}_{P_0}(X)},$$

while (as $o(1) \to 0$)

$$\liminf_{\delta \downarrow 0} \frac{\omega_{\text{hel}}(\delta)}{\sqrt{8 \text{Var}_{P_0}(X) \delta}} \ge 1.$$

For the upper bound on $\omega_{\rm hel}$, we require a few more steps. Let $P_1 \in \mathcal{P}$ be an arbitrary distribution, where we assume that $d_{\rm hel}^2(P_0, P_1) \le \frac{1}{4}$, and use the shorthand and $\theta_1 = \theta(P_1)$. Then

$$\begin{aligned}
\theta_1 - \theta_0 &= \int (x - \theta_0)(dP_1 - dP_0) = \int (x - \theta_0)(\sqrt{dP_1} + \sqrt{dP_0})(\sqrt{dP_1} - \sqrt{dP_0}) \\
&\le \left(\int (x - \theta_0)^2 (\sqrt{dP_1} + \sqrt{dP_0})^2 \right)^{1/2} \sqrt{2} d_{\rm hel}(P_0, P_1) \\
&\le \left(2\mathbb{E}_0[(X - \theta_0)^2] + 2\mathbb{E}_1[(X - \theta_0)^2] \right)^{1/2} \sqrt{2} d_{\rm hel}(P_0, P_1) \\
&= 2\sqrt{\mathrm{Var}_0(X) + \mathrm{Var}_1(X) + (\theta_0 - \theta_1)^2} \cdot d_{\rm hel}(P_0, P_1) \qquad (49)
\end{aligned}$$

by the Cauchy-Schwarz inequality and definition of Hellinger distance. Noting that $\mathbb{E}[(X - \theta)^4] \le 2^3(\mathbb{E}[X^4] + \theta^4)$, we may assume there exists some $M_4 < \infty$ such that $M_4^4 \ge \mathbb{E}_P[(X - \theta_0)^4]$ for all $P \in \mathcal{P}$. We now bound the variance $\mathrm{Var}_1(X)$ in terms of $\mathrm{Var}_0(X)$ and $d_{\rm hel}(P_0, P_1)$. Using that

$$(x - \theta_0)^2 - (x - \theta_1)^2 = -2x(\theta_0 - \theta_1) + \theta_0^2 - \theta_1^2,$$

we obtain

$$\begin{aligned}
\mathrm{Var}_0(X) - \mathrm{Var}_1(X) &= \int (x - \theta_0)^2 (dP_0 - dP_1) - 2\theta_1(\theta_0 - \theta_1) + \theta_0^2 - \theta_1^2 \\
&= \int (x - \theta_0)^2 (\sqrt{dP_0} + \sqrt{dP_1})(\sqrt{dP_0} - \sqrt{dP_1}) + (\theta_0 - \theta_1)^2.
\end{aligned}$$

Again applying Cauchy-Schwarz, we observe that

$$|\mathrm{Var}_0(X) - \mathrm{Var}_1(X)| \le 2M_4^2 d_{\rm hel}(P_0, P_1) + (\theta_0 - \theta_1)^2.$$

Substituting this bound into inequality (49) and squaring yields

$$(\theta_1 - \theta_0)^2 \le 4 \left(2\mathrm{Var}_0(X) + 2M_4^2 d_{\rm hel}(P_0, P_1) + 2(\theta_0 - \theta_1)^2 \right) d_{\rm hel}^2(P_0, P_1),$$

or

$$(\theta_1 - \theta_0)^2 \le \frac{8\mathrm{Var}_0(X)}{1 - 8d_{\rm hel}^2(P_0, P_1)} d_{\rm hel}^2(P_0, P_1) + \frac{8M_4^2}{1 - 8d_{\rm hel}^2(P_0, P_1)} d_{\rm hel}^3(P_0, P_1).$$

In particular, as soon as $d_{\rm hel}^2(P_0, P_1) \le \frac{1}{16}$ and $d_{\rm hel}(P_0, P_1) \le (\mathrm{Var}_0(X)/M_4^2)^{1/3}$,

$$(\theta_1 - \theta_0)^2 \le 32\mathrm{Var}_0(X) d_{\rm hel}^2(P_0, P_1).$$

Solving for the modulus (5) gives the result, and eliminating higher order terms yields

$$\limsup_{\delta \downarrow 0} \sup \left\{ \frac{|\theta(P_0) - \theta(P_1)|}{d_{\rm hel}(P_0, P_1)} \mid d_{\rm hel}(P_0, P_1) \le \delta \right\} \le \sqrt{8\mathrm{Var}_{P_0}(X)}.$$

A.4.1 Proof of Claim 5.1

A minor extension of Proposition 1 shows there exist numerical constants $0 < c_0, c_1 < \infty$ such that

$$c_0(\omega_{\rm hel}^2(c_0 n^{-1/2}; P_0, \mathcal{P}) \wedge B) \le \mathfrak{M}_n^{\rm loc}(P_0, L_{\wedge B}, \mathcal{P}, \{\mathrm{id}\}) \le c_1 \sup_{r \ge 0} \left\{ \left(\omega_{\rm hel}^2(r; P_0, \mathcal{P}) \wedge B \right) e^{-nr^2} \right\} \quad (50)$$

for all $n \in \mathbb{N}$ and any family $\mathcal{P}$. For the influence function $\dot{\theta}_0(x) = (x - \mathbb{E}_0[X])/\mathrm{Var}_0(X)$ following the claim, by appropriate renormalization, we may apply the limiting equality (17) to obtain

$$\omega_{\mathrm{hel}}(\delta, P_0, \mathcal{P}_{\text{non-par}}) = \frac{\sqrt{8}}{\sqrt{\mathrm{Var}_0(X)}}\delta(1 + o(1)) = \sqrt{8\delta^2\mathbb{E}_0[\dot{\theta}_0(X)^2]}(1 + o(1)).$$

The lower bound $\mathfrak{M}_n^{\mathrm{loc}}(P_0, L_{\wedge B}, \mathcal{P}_{\text{non-par}}, \{\mathrm{id}\}) \gtrsim \frac{1}{n\mathrm{Var}_0(X)}$ for large n then follows by inequality (50). The matching upper bound similarly follows, as for all large enough n, if $r \le 1/n^{1/4}$ we have

$$\omega^2_{\mathrm{hel}}(r; P_0, \mathcal{P}_{\text{non-par}}) \le 16r^2\mathbb{E}_0[\dot{\theta}_0(X)^2],$$

and so

$$\begin{aligned}\sup_{r\ge 0}\left\{\left(\omega^2_{\mathrm{hel}}(r; P_0, \mathcal{P}_{\text{non-par}}) \wedge B\right)e^{-nr^2}\right\} &\le \max\left\{\sup_{r\ge 0} 16r^2\mathbb{E}_0[\dot{\theta}_0(X)^2]e^{-nr^2}, \sup_{r\ge n^{-1/4}} Be^{-nr^2}\right\} \\ &= O(1)\max\left\{\frac{1}{n}\mathbb{E}_0[\dot{\theta}_0(X)^2], Be^{-\sqrt{n}}\right\}.\end{aligned}$$

A derivation *mutatis mutandis* identical to that for Claim 2.1 gives the parametric result, as the exponential family has score $\dot{\ell}_{\theta_0}(x) = x - \mathbb{E}_{\theta_0}[X]$ and Fisher information $I_{\theta_0} = \mathrm{Var}_0(X)$.

B Deferred main proofs

B.1 Proof of Proposition 2

Let P_0 and P_1 be distributions on $\mathcal{X}$, each with densities p_0, p_1 according to some base measure μ. Let $\theta_a = \theta(P_a)$, and consider the problem of privately collecting observations and deciding whether $\theta = \theta_0$ or $\theta = \theta_1$. We define a randomized-response estimator for this problem using a simple hypothesis test. Define the acceptance set $A := \{x \in \mathcal{X} \mid p_0(x) > p_1(x)\}$, so $P_0(A) - P_1(A) = \|P_0 - P_1\|_{\mathrm{TV}}$. Now, consider the following estimator: for each X_i, define

$$T_i = \mathbf{1}\{X_i \in A\} \quad \text{and} \quad Z_i \mid \{T_i = t\} = \begin{cases} 1 & \text{with probability } (e^\varepsilon + 1)^{-1}(e^\varepsilon t + 1 - t) \\ 0 & \text{with probability } (e^{-\varepsilon} + 1)^{-1}(e^{-\varepsilon} t + 1 - t)\end{cases}$$

Then the channel $Q(\cdot \mid X_i)$ for $Z_i \mid X_i$ is ε-differentially-private, and setting $\delta_\varepsilon = \frac{e^\varepsilon}{1+e^\varepsilon} - \frac{1}{2}$, we have

$$\mathbb{E}_0[Z_i] = \frac{1+\delta_\varepsilon}{2}P_0(A) + \frac{1-\delta_\varepsilon}{2}P_0(A^c) = \frac{1-\delta_\varepsilon}{2} + \delta_\varepsilon P_0(A) \quad \text{and} \quad \mathbb{E}_1[Z_i] = \frac{1-\delta_\varepsilon}{2} + \delta_\varepsilon P_1(A)$$

while $Z_i \in \{0, 1\}$. Define the statistic

$$K_n := \frac{1}{\delta_\varepsilon}\left(\frac{1}{n}\sum_{i=1}^n Z_i - \frac{1-\delta_\varepsilon}{2}\right),$$

so that $\mathbb{E}_0[K_n] = P_0(A)$ and $\mathbb{E}_1[K_n] = P_1(A)$. We define the estimator

$$\widehat{\theta} := \theta_0\mathbf{1}\left\{K_n \ge \frac{P_0(A) + P_1(A)}{2}\right\} + \theta_1\mathbf{1}\left\{K_n < \frac{P_0(A) + P_1(A)}{2}\right\}.$$

We now analyze the performance of $\widehat{\theta}$. By construction of the acceptance set A,

$$\frac{P_0(A)+P_1(A)}{2} = P_0(A) + \frac{P_1(A)-P_0(A)}{2} = P_0(A) - \frac{1}{2}\|P_1-P_0\|_{\mathrm{TV}} = P_1(A) + \frac{1}{2}\|P_1-P_0\|_{\mathrm{TV}},$$

so by Hoeffding's inequality, we have

$$\max\left\{P_0\left(K_n \le \frac{P_0(A)+P_1(A)}{2}\right), P_1\left(K_n \ge \frac{P_0(A)+P_1(A)}{2}\right)\right\} \le \exp\left(-\frac{n\delta_\varepsilon^2\|P_0-P_1\|_{\mathrm{TV}}^2}{2}\right).$$

In particular, we have

$$\mathbb{E}_0[L(\widehat{\theta}-\theta_0)] + \mathbb{E}_1[L(\widehat{\theta}-\theta_1)] \le [L(\theta_1-\theta_0)+L(\theta_0-\theta_1)]\exp\left(-\frac{n\delta_\varepsilon^2\|P_0-P_1\|_{\mathrm{TV}}^2}{2}\right).$$

Using the growth condition C.1, we obtain

$$\begin{aligned}
\mathbb{E}_0[L(\widehat{\theta}-\theta_0)] + \mathbb{E}_1[L(\widehat{\theta}-\theta_1)] &\le 2\gamma L\left(\frac{\theta_0-\theta_1}{2}\right)\exp\left(-\frac{n\delta_\varepsilon^2\|P_0-P_1\|_{\mathrm{TV}}^2}{2}\right) \\
&\le 2\gamma \sup_{P\in\mathcal{P}} L\left(\frac{\theta_0-\theta(P)}{2}\right)\exp\left(-\frac{n\delta_\varepsilon^2\|P_0-P\|_{\mathrm{TV}}^2}{2}\right) \\
&= 2\gamma \sup_{r\ge 0}\left\{\omega_{L,\mathrm{TV}}(r;P_0)\exp\left(-\frac{n\delta_\varepsilon^2 r^2}{2}\right)\right\}.
\end{aligned}$$

B.2 Proof of Corollary 6

Define the shorthand

$$\omega_{\mathrm{TV}}(\delta) := \sup_h\left\{|\psi(\theta_0+h)-\psi(\theta_0)| \text{ s.t. } \|P_{\theta_0+h}-P_{\theta_0}\|_{\mathrm{TV}} \le \delta\right\}.$$

We first apply Proposition 2. As in this setting the constant $\gamma = O(1)$ from Condition C.1 is automatically a universal constant, we obtain for numerical constants $C_0, C_1 < \infty$ that

$$\mathfrak{M}^{\mathrm{loc}}(P_{\theta_0}, L, \mathcal{P}, \mathcal{Q}) \le C_0 \sup_{\tau\ge 0} L\left(\omega_{\mathrm{TV}}(C_1\tau)\right) e^{-n\tau^2\varepsilon^2}. \tag{51}$$

We bound $\omega_{\mathrm{TV}}(\delta)$ for small δ. Let r_ψ and r_0 be remainders as in the proof of Proposition 4, so that $\|P_{\theta_0+h}-P_\theta\|_{\mathrm{TV}} = \frac{1}{2}J_{\theta_0}|h| + r_0(h)$ and $\psi(\theta_0+h) = \psi(\theta_0)+\psi'(\theta_0)h + r_\psi(h)$, where both are $o(h)$ as $|h|\to 0$. Let $h_0 > 0$ be such that $|r_0(h)| \le |J_{\theta_0}h|/4$ for $|h| \le h_0$. Choose $\delta_0 > 0$ so that $\|P_{\theta_0+h}-P_{\theta_0}\|_{\mathrm{TV}} \le \delta_0$ implies $|h| \le h_0$, which is possible by Assumption A1, and $|h| \le \frac{4\delta}{J_{\theta_0}}$ implies $|r_\psi(h)| \le |\psi'(\theta_0)h|$. Then for all $\delta \le \delta_0$, if $\|P_{\theta_0+h}-P_{\theta_0}\|_{\mathrm{TV}} \le \delta$, we have $|h| \le h_0$ and consequently

$$\delta \ge \|P_{\theta_0+h}-P_{\theta_0}\|_{\mathrm{TV}} = \frac{1}{2}J_{\theta_0}|h| + r_0(h) \ge \frac{1}{4}J_{\theta_0}|h|,$$

or $|h| \le 4\delta/J_{\theta_0}$. Thus we obtain

$$\begin{aligned}
\omega_{\mathrm{TV}}(\delta) &= \sup_h\left\{|\psi(\theta_0+h)-\psi(\theta_0)| \text{ s.t. } \|P_{\theta_0+h}-P_{\theta_0}\|_{\mathrm{TV}} \le \delta\right\} \\
&\le \sup_h\left\{|\psi'(\theta_0)h + r_\psi(h)| \text{ s.t. } |h| \le \frac{4\delta}{J_{\theta_0}}\right\} \le \frac{8|\psi'(\theta_0)|}{J_{\theta_0}}\delta.
\end{aligned}$$

We now return to inequality (51). Substituting the preceding bound, for numerical constants $C_0, C_1 < \infty$ whose values may change from line to line,

$$\mathfrak{M}_n^{\rm loc}(P_{\theta_0}, L, \mathcal{P}, \mathcal{Q}) \le C_0 \max\left\{ \sup_{0 \le \tau \le \delta_0/C_1} L\left(\omega_{\rm TV}\left(C_1\tau\right)\right) e^{-\tau^2 n\varepsilon^2}, L\left(\operatorname{diam}(\psi(\Theta))\right) e^{-\delta_0^2 n\varepsilon^2/C_1^2} \right\}$$
$$\le C_0 \max\left\{ \sup_{0 \le \tau \le \delta_0/C_1} L\left(\frac{C_1|\psi'(\theta_0)|}{J_{\theta_0}}\tau\right) e^{-\tau^2 n\varepsilon^2}, L\left(\operatorname{diam}(\psi(\Theta))\right) e^{-\delta_0^2 n\varepsilon^2/C_1^2} \right\}.$$

Finally, we use the assumption that $L(at) \le Ca^\beta L(t)$ for all $a \ge 1$. We have

$$L\left(\frac{C_1|\psi'(\theta_0)|}{J_{\theta_0}}\tau\right) \le C'(n\varepsilon^2\tau^2)^{\beta/2} L\left(\frac{|\psi'(\theta_0)|}{J_{\theta_0}}\frac{1}{\sqrt{n\varepsilon^2}}\right)$$

and using that $\sup_t t^{\beta/2}e^{-t} = (\beta/2)^{\beta/2}e^{-\beta/2}$ gives the result.

B.3 Proof of Proposition 4

We assume that $\psi'(\theta_0) \neq 0$; the result is otherwise trivial. Applying Theorem 1, we have

$$\mathfrak{M}_n^{\rm loc}(P_{\theta_0}, L, \mathcal{P}, \mathcal{Q}_\varepsilon) \ge \frac{1}{8}\omega_{L,\rm TV}\left(\frac{1}{\sqrt{8n\varepsilon^2}}; P_{\theta_0}, \mathcal{P}\right). \tag{52}$$

Now, we evaluate $\omega_L(\delta) := \omega_{L,\rm TV}(\delta; P_{\theta_0}, \mathcal{P})$ for small $\delta > 0$. By assumption, there exist remainders r_0 and r_ψ, both satisfying $|r(h)/h| \to 0$ as $h \to 0$, such that $\|P_{\theta_0+h} - P_{\theta_0}\|_{\rm TV} = \frac{1}{2}J_{\theta_0}|h| + r_0(h)$ and $\psi(\theta_0 + h) - \psi(\theta_0) = \psi'(\theta_0)h + r_\psi(h)$. Then

$$\omega_L(\delta) \ge \sup_h \left\{ L\left(\frac{1}{2}(\psi(\theta_0+h) - \psi(\theta_0))\right) \mid \|P_{\theta_0+h} - P_{\theta_0}\|_{\rm TV} \le \delta \right\}$$
$$= \sup_h \left\{ L\left(\frac{1}{2}(\psi'(\theta_0)h + r_\psi(h))\right) \mid J_{\theta_0}|h| + 2r_0(h) \le 2\delta \right\}.$$

Choose $h_0 = h(\theta_0, \psi, \mathcal{P}) > 0$ such that $|h| \le h_0$ implies that $|r_0(h)| \le J_{\theta_0}|h|/2$ and $|r_\psi(h)| \le |\psi'(\theta_0)h|/5$. Then evidently

$$\omega_L(\delta) \ge \sup_{|h| \le h_0} \left\{ L\left(\frac{2}{5}|\psi'(\theta_0)h|\right) \mid J_{\theta_0}|h| \le \delta \right\} \stackrel{(\star)}{=} L\left(\frac{2}{5}J_{\theta_0}|\psi'(\theta_0)|\delta\right),$$

where equality $(\star)$ occurs whenever $\delta \le h_0/J_{\theta_0}$. Setting $\delta = \frac{1}{\sqrt{8n\varepsilon^2}}$, letting n grow, and substituting into inequality inequality (52) gives the proposition.

B.4 Proof of Theorem 2

The proof mirrors that of Proposition 4. Let $\theta_0 = \theta(P_0)$ be the desired parameter. Again, we assume that $\nabla\psi(\theta_0)^T\mathbb{E}_0[\dot\theta_0(X)g(X)] \neq 0$, as otherwise the result is trivial. For the L^1-information $J_{g,0} := \int |g| dP_0$, there exist remainders r_0, r_ψ both satisfying $r(h) = o(h)$ and

$$\|P_0 - P_h\|_{\rm TV} = \frac{1}{2}|h|J_{g,0} + r_0(h) \quad \text{and} \quad \psi(\theta(P_h)) = \psi(\theta_0) + \nabla\psi(\theta_0)^T\mathbb{E}_0[\dot\theta_0(X)g(X)]h + r_\psi(h)$$

by the differentiability assumptions. Choose $h_0 > 0$ small enough that $|h| \le h_0$ implies that $|r_0(h)| \le \frac{1}{2}|h|J_{g,0}$ and $|r_\psi(h)| \le |\nabla\psi(\theta_0)^T\mathbb{E}_0[\dot\theta_0(X)g(X)]h|/20$, which depends only on ψ and $\mathcal{P}_{\text{sub},0}$. Then defining $\omega_L(\delta) := \omega_{L,\text{TV}}(\delta; P_0, \mathcal{P})$ for shorthand, we have

$$\begin{aligned}\omega_L(\delta) &\ge \sup_{|h|\le h_0}\left\{L\left(\frac{1}{2}(\psi(\theta_0)-\psi(P_h))\right) \mid \|P_0 - P_h\|_{\text{TV}} \le \delta\right\} \\ &\ge \sup_{|h|\le h_0}\left\{L\left(\frac{9}{20}\nabla\psi(\theta_0)^T\mathbb{E}_0[\dot\theta_0(X)g(X)]h\right) \mid J_{g,0}|h| \le \delta\right\}.\end{aligned}$$

For all $\delta \le h_0/J_{g,0}$, then, we obtain

$$\omega_L(\delta) \ge L\left(\frac{19\delta}{40}\frac{\nabla\psi(\theta_0)^T\mathbb{E}_0[\dot\theta_0(X)g(X)]}{J_{g,0}}\right).$$

Applying Theorem 1 and setting $\delta = \frac{1}{\sqrt{8n\varepsilon^2}}$ gives the result.

C Technical Appendices

C.1 Proof of Lemma 2

By the triangle inequality, we have

$$\begin{aligned}&\int |p_{\theta_0+h} - p_{\theta_0} - h^T\dot\ell_{\theta_0}p_{\theta_0}|d\mu \\ &\le \underbrace{\int\left|p_{\theta_0+h} - p_{\theta_0} - \frac{1}{2}h^T\dot\ell_{\theta_0}\sqrt{p_{\theta_0}}(\sqrt{p_{\theta_0+h}}+\sqrt{p_{\theta_0}})\right|d\mu}_{:=I_1(h;\theta_0)} + \underbrace{\int\left|\frac{1}{2}h^T\dot\ell_{\theta_0}\sqrt{p_{\theta_0}}(\sqrt{p_{\theta_0+h}}-\sqrt{p_{\theta_0}})\right|d\mu}_{:=I_2(h;\theta_0)}.\end{aligned}$$

We show that each of the integral terms I_1 and I_2 are both $o(\|h\|)$ as $h \to 0$. By algebraic manipulation and the Cauchy–Schwarz inequality,

$$\begin{aligned}I_1(h;\theta_0) &= \int |\sqrt{p_{\theta_0+h}}+\sqrt{p_{\theta_0}}|\cdot\left|\sqrt{p_{\theta_0+h}}-\sqrt{p_{\theta_0}} - \frac{1}{2}h^T\dot\ell_{\theta_0}\sqrt{p_{\theta_0}}\right|d\mu \\ &\le \left(\int|\sqrt{p_{\theta_0+h}}+\sqrt{p_{\theta_0}}|^2d\mu\right)^{\frac{1}{2}}\cdot\left(\int\left|\sqrt{p_{\theta_0+h}}-\sqrt{p_{\theta_0}} - \frac{1}{2}h^T\dot\ell_{\theta_0}\sqrt{p_{\theta_0}}\right|^2d\mu\right)^{\frac{1}{2}}\end{aligned}$$

Jensen's inequality gives $\int|\sqrt{p_{\theta_0+h}}+\sqrt{p_{\theta_0}}|^2d\mu \le 2\int(p_{\theta_0+h}+p_{\theta_0})d\mu = 2$. The assumption that $\mathcal{P}$ is QMD at θ_0 immediately yields $I_1(h;\theta_0) = o(\|h\|)$. To bound I_2, we again apply the Cauchy–Schwarz inequality, obtaining

$$2I_2(h;\theta_0) \le \left(\int|h^T\dot\ell_{\theta_0}\sqrt{p_{\theta_0}}|^2d\mu\right)^{\frac{1}{2}}\cdot\left(\int|\sqrt{p_{\theta_0+h}}-\sqrt{p_{\theta_0}}|^2d\mu\right)^{\frac{1}{2}}$$

Since $\mathcal{P}$ is QMD at θ_0, we have $\int|\sqrt{p_{\theta_0+h}}-\sqrt{p_{\theta_0}}|^2d\mu = \int|\frac{1}{2}h^T\dot\ell_{\theta_0}\sqrt{p_{\theta_0}}|^2d\mu + o(\|h\|^2) = O_{\theta_0}(\|h\|^2)$ (see [52, Ch. 7.2]). Thus $I_2(h;\theta_0) = O_{\theta_0}(\|h\|^2)$, giving the lemma.

C.2 Proof of Proposition 6

We require one additional piece of notation before we begin the proof. Let $W_i = Z_i - V_i$ be the error in the private version of the quantity V_i, so that $\mathbb{E}[W_i \mid V_i] = 0$, and

$$W_i = \begin{cases} \frac{2}{e^\varepsilon - 1} V_i - \frac{1}{e^\varepsilon - 1} & \text{w.p. } \frac{e^\varepsilon}{e^\varepsilon + 1} \\ \frac{-2e^\varepsilon}{e^\varepsilon - 1} V_i + \frac{e^\varepsilon}{e^\varepsilon - 1} & \text{w.p. } \frac{1}{e^\varepsilon + 1}. \end{cases}$$

Recall our definitions of $V_i = \mathbf{1}\{T(X_i) \ge \widehat{T}_n\}$ and Z_i as the privatized version of V_i. Letting $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$, and similarly for $\overline{V}_n$ and $\overline{W}_n$, recall also the definition of the random variable $G_n := \Psi(\widehat{T}_n, \theta_0) = P_{\theta_0}(T(X) \ge \widehat{T}_n)$. By mimicking the delta method, we will show that

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = 2J_{\theta_0}^{-1} \cdot \sqrt{n}\left(\overline{V}_n - G_n + \overline{W}_n\right) + o_P(1). \tag{53}$$

Deferring the proof of the expansion (53), let us show how it implies the proposition.

First, with our definition of the W_i, we have

$$\mathrm{Var}(W_i \mid V_i) = \mathbb{E}[W_i^2 \mid V_i] = \frac{e^\varepsilon}{(e^\varepsilon - 1)^2} = \delta_\varepsilon^{-2},$$

so that $\overline{W}_n = \frac{1}{n}\sum_{i=1}^n W_i$ satisfies $\sqrt{n}\overline{W}_n \xrightarrow{d} \mathsf{N}(0, \delta_\varepsilon^{-2})$ by the Lindeberg CLT. Thus, assuming the expansion (53), it remains to show the weak convergence result

$$\frac{\sqrt{n}\left(\overline{V}_n - G_n\right)}{G_n(1 - G_n)} \xrightarrow{d} \mathsf{N}(0, 1). \tag{54}$$

where $G_n = \Psi(\widehat{T}_n, \theta_0)$. By definition, the $\{X_i\}_{i=1}^n$ are independent of $\widehat{T}_n$, and hence

$$\mathbb{E}[V_i \mid \widehat{T}_n] = \Psi(\widehat{T}_n, \theta_0) = G_n \quad \text{and} \quad \mathrm{Var}(V_i \mid \widehat{T}_n) = \Psi(\widehat{T}_n, \theta_0)(1 - \Psi(\widehat{T}_n, \theta_0)) = G_n(1 - G_n).$$

The third central moments of the V_i conditional on $\widehat{T}_n$ have the bound

$$\mathbb{E}\left[\left|V_i - \mathbb{E}[V_i \mid \widehat{T}_n]\right|^3 \mid \widehat{T}_n\right] \le \Psi(\widehat{T}_n, \theta_0)(1 - \Psi(\widehat{T}_n, \theta_0)) = G_n(1 - G_n).$$

Thus, we may apply the Berry-Esseen Theorem [40, Thm 11.2.7] to obtain

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n}\left(\overline{V}_n - G_n\right)}{G_n(1 - G_n)} \le t \mid \widehat{T}_n \right) - \Phi(t) \right| \le U_n := \frac{1}{\sqrt{n G_n(1 - G_n)}} \wedge 2.$$

Jensens's inequality then implies

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n}\left(\overline{V}_n - G_n\right)}{G_n(1 - G_n)} \le t \right) - \Phi(t) \right| \le \mathbb{E}\left[\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n}\left(\overline{V}_n - G_n\right)}{G_n(1 - G_n)} \le t \mid \widehat{T}_n \right) - \Phi(t) \right| \right] \le \mathbb{E}[U_n]$$

To show the convergence (54), it is thus sufficient to show that $\mathbb{E}[U_n] \to 0$ as $n \uparrow \infty$. To that end, the following lemma on the behavior of $\Psi(t, \theta) = P_\theta(T(X) \ge t)$ is useful.

Lemma 4. *Let $t_0 = \mathbb{E}_{\theta_0}[T(X)]$ and assume that $\mathrm{Var}_{\theta_0}(T(X)) > 0$. Then there exist $\epsilon > 0$ and $c \in (0, \frac{1}{2})$ such that if $t \in [t_0 \pm \epsilon]$ and $\theta \in [\theta_0 \pm \epsilon]$, then $\Psi(t, \theta) \in [c, 1 - c]$.*

Proof. By the dominated convergence theorem and our assumption that $\mathrm{Var}_{\theta_0}(T(X)) > 0$, where $t_0 = \mathbb{E}_{\theta_0}[T(X)]$, we have

$$\liminf_{t \uparrow t_0} \Psi(t, \theta_0) = P_{\theta_0}(T(X) \ge t_0) \in (0,1) \quad \text{and} \quad \limsup_{t \downarrow t_0} \Psi(t, \theta_0) = P_{\theta_0}(T(X) > t_0) \in (0,1).$$

The fact that $t \mapsto \Psi(t, \theta_0)$ is non-increasing implies that for some $\epsilon_1 > 0, c \in (0, \frac{1}{4})$, we have $\Psi(t, \theta_0) \in [2c, 1-2c]$ for $t \in [t_0 - \epsilon_1, t_0 + \epsilon_1]$. Fix this ϵ_1 and c. By [40, Thm 2.7.1], we know that any $t \in \mathbb{R}$, the function $\theta \mapsto \Psi(t, \theta)$ is continuous and non-decreasing. Thus for any $\epsilon_2 > 0$, we have

$$\Psi(t_0 + \epsilon_1, \theta_0 - \epsilon_2) \le \Psi(t, \theta) \le \Psi(t_0 - \epsilon_1, \theta_0 + \epsilon_2) \quad \text{for} \quad (t, \theta) \in [t_0 \pm \epsilon_1] \times [\theta_0 \pm \epsilon_2].$$

Using the continuity of $\theta \mapsto \Psi(t, \theta)$, we may choose $\epsilon_2 > 0$ small enough that

$$\Psi(t, \theta) \in [c, 1-c] \quad \text{for} \quad (t, \theta) \in \{t_0 - \epsilon_1, t_0 + \epsilon_1\} \times \{\theta_0 - \epsilon_2, \theta_0 + \epsilon_2\}.$$

The lemma follows by taking $\epsilon = \epsilon_1 \wedge \epsilon_2$. □

As $\mathrm{Var}_{\theta_0}(T(X)) > 0$ by assumption, Lemma 4 and the fact that $\widehat{T}_n \xrightarrow{p} t_0$ imply

$$G_n := \Psi(\widehat{T}_n, \theta_0) = P_{\theta_0}(T(X) \ge \widehat{T}_n) \in [c + o_P(1), 1 - c + o_P(1)]. \tag{55}$$

The bounds (55) imply that $G_n(1 - G_n) \ge c(1-c) + o_P(1)$, so $U_n \xrightarrow{p} 0$. By construction $|U_n| \le 2$ for all n, so the bounded convergence theorem implies $\mathbb{E}[U_n] \to 0$, which was what we required to show the weak convergence result (54). The joint convergence in the proposition follows because $\overline{W}_n$ and $\overline{V}_n - G_n$ are conditionally uncorrelated.

The delta method expansion We now return to demonstrate the claim (53). For $p \in [0,1]$, recall the definition (27) of the function H, and define

$$H_n(p) := H(p, \widehat{T}_n) = \inf\left\{\theta \in \mathbb{R} \mid P_\theta(T(X) \ge \widehat{T}_n) \ge p\right\}, \tag{56}$$

where the value is $-\infty$ or $+\infty$ for p below or above the range of $\theta \mapsto P_\theta(T(X) \ge \widehat{T}_n)$, respectively. Then $\widehat{\theta}_n = H_n(\overline{Z}_n)$ by construction (27). We would like to apply Taylor's theorem and the inverse function theorem to $\widehat{\theta}_n - \theta_0 = H_n(\overline{Z}_n) - \theta_0$, but this requires a few additional steps.

By the inverse function theorem, $p \mapsto H_n(p)$ is $\mathcal{C}^\infty$ on $(\inf_\theta \Psi(\widehat{T}_n, \theta), \sup_\theta \Psi(\widehat{T}_n, \theta))$, and letting

$$\dot{\Psi}_\theta(t, \theta) = \frac{\partial}{\partial \theta} \Psi(t, \theta) = \mathbb{E}_\theta[\mathbf{1}\{T(X) \ge t\}\,(T(X) - A'(\theta))] = \frac{\partial}{\partial \theta} P_\theta(T(X) \ge t),$$

we have $H_n'(p) = \dot{\Psi}_\theta(\widehat{T}_n, H_n(p))^{-1}$ whenever p is interior to the range of $\theta \mapsto P_\theta(T(X) \ge \widehat{T}_n)$. To show that $\overline{Z}_n$ is (typically) in this range, we require a bit of analysis on $\dot{\Psi}_\theta$.

Lemma 5. *The function* $(t, \theta) \mapsto \dot{\Psi}_\theta(t, \theta) = \mathbb{E}_\theta[\mathbf{1}\{T(X) \ge t\}\,(T(X) - A'(\theta))]$ *is continuous at* (t_0, θ_0)*, where* $t_0 = \mathbb{E}_{\theta_0}[T(X)] = A'(\theta_0)$*.*

To avoid disrupting the flow, we defer the proof to Section C.2.1. Now, we have that $\dot{\Psi}_\theta(t_0, \theta_0) = \frac{1}{2}\mathbb{E}_{\theta_0}[|T(X) - t_0|] > 0$, so Lemma 5 implies there exists $\epsilon > 0$ such that

$$\inf_{|t - t_0| \le \epsilon, |\theta - \theta_0| \le \epsilon} \dot{\Psi}_\theta(t, \theta) \ge c > 0 \tag{57}$$

for some constant c. Thus, we obtain that

$$\mathbb{P}\left(\overline{Z}_n \notin \text{Range}(\Psi(\widehat{T}_n, \cdot))\right) \le \mathbb{P}\left(\overline{Z}_n \notin \text{Range}(\Psi(\widehat{T}_n, \cdot)), \widehat{T}_n \in [t_0 \pm \epsilon]\right) + \mathbb{P}\left(\widehat{T}_n \notin [t_0 \pm \epsilon]\right)$$
$$\stackrel{(i)}{\le} \mathbb{P}\left(\overline{Z}_n \notin [\Psi(\widehat{T}_n, \theta_0) \pm c\epsilon]\right) + o(1) \to 0, \tag{58}$$

where inequality (i) follows because $\text{Range}(\Psi(t, \cdot)) \supset [\Psi(t, \theta_0) \pm c\epsilon]$ for all t such that $|t - t_0| \le \epsilon$ by condition (57), and the final convergence because $\overline{Z}_n - \Psi(\widehat{T}_n, \theta_0) \stackrel{p}{\to} 0$ and $\widehat{T}_n \stackrel{p}{\to} t_0$.

We recall that for fixed t, $\theta \mapsto \Psi(t, \theta)$ is analytic on the interior of the natural parameter space and strictly increasing at all θ for which $\Psi(t, \theta) \in (0, 1)$ (cf. [40, Thm. 2.7.1, Thm. 3.4.1]). Thus,

$$H_n(\Psi(\widehat{T}_n, \theta)) = \theta \quad \text{whenever} \quad \Psi(\widehat{T}_n, \theta) \in (0, 1).$$

As $G_n = \Psi(\widehat{T}_n, \theta_0) \in [c + o_P(1), 1 - c + o_P(1)]$ by definition (55) of G_n, we obtain

$$\mathbb{P}\left(H_n(\Psi(\widehat{T}_n, \theta_0)) \ne \theta_0\right) \to 0.$$

By the differentiability of H_n on the interior of its domain (i.e. the range of $\Psi(\widehat{T}_n, \cdot)$), we use the convergence (58) and Taylor's intermediate value theorem to obtain that for some p_n between $\overline{Z}_n$ and $\Psi(\widehat{T}_n, \theta_0)$, we have

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \sqrt{n}(\widehat{\theta}_n - H_n(\Psi(\widehat{T}_n, \theta_0))) + o_P(1) \tag{59}$$
$$= H_n'(p_n)\sqrt{n}\left(\overline{Z}_n - \Psi(\widehat{T}_n, \theta_0)\right) + o_P(1) = \dot{\Psi}_\theta(\widehat{T}_n, H_n(p_n))^{-1}\sqrt{n}\left(\overline{Z}_n - \Psi(\widehat{T}_n, \theta_0)\right) + o_P(1)$$

as $p_n \in \text{int dom}\, H_n$ with high probability by (58).

It remains to show that $H_n(p_n) \stackrel{p}{\to} \theta_0$. When $\widehat{T}_n \in [t_0 \pm \epsilon]$, the growth condition (57) implies

$$\Psi(\widehat{T}_n, \theta_0 + \epsilon) = P_{\theta_0+\epsilon}(T(X) \ge \widehat{T}_n) \ge P_{\theta_0}(T(X) \ge \widehat{T}_n) + c\epsilon = \Psi(\widehat{T}_n, \theta_0) + c\epsilon$$
$$\Psi(\widehat{T}_n, \theta_0 - \epsilon) = P_{\theta_0-\epsilon}(T(X) \ge \widehat{T}_n) \le P_{\theta_0}(T(X) \ge \widehat{T}_n) - c\epsilon = \Psi(\widehat{T}_n, \theta_0) - c\epsilon,$$

and thus

$$\mathbb{P}(|H_n(p_n) - \theta_0| \ge \epsilon) \le \mathbb{P}(|\overline{Z}_n - \Psi(\widehat{T}_n, \theta_0)| \ge c\epsilon) + \mathbb{P}(|\widehat{T}_n - t_0| \ge \epsilon) \to 0.$$

We have the convergence $\dot{\Psi}_\theta(\widehat{T}_n, H_n(p_n)) \stackrel{p}{\to} \frac{1}{2}\mathbb{E}_{\theta_0}[|T(X) - A'(\theta_0)|] = \frac{1}{2}J_{\theta_0}$ by the continuous mapping theorem, and Slutsky's theorem applied to Eq. (59) gives the delta-method expansion (53).

C.2.1 Proof of Lemma 5

We have

$$\dot{\Psi}_\theta(t_0, \theta_0) - \dot{\Psi}_\theta(t, \theta) = \mathbb{E}_{\theta_0}[\mathbf{1}\{T(X) \ge t_0\}(T(X) - A'(\theta_0))] - \mathbb{E}_\theta[\mathbf{1}\{T(X) \ge t\}(T(X) - A'(\theta))]$$
$$\stackrel{(i)}{=} \mathbb{E}_{\theta_0}\left[[T(X) - t_0]_+\right] - \mathbb{E}_\theta[\mathbf{1}\{T(X) \ge t\}(T(X) - t + t - A'(\theta))]$$
$$= \mathbb{E}_{\theta_0}\left[[T(X) - t_0]_+\right] - \mathbb{E}_\theta\left[[T(X) - t]_+\right] + P_\theta(T(X) \ge t)(t - A'(\theta))$$
$$\stackrel{(ii)}{\in} \mathbb{E}_{\theta_0}\left[[T(X) - t_0]_+\right] - \mathbb{E}_\theta\left[[T(X) - t_0]_+\right] \pm |t - t_0| \pm |t - A'(\theta)|,$$

where step (i) follows because $t_0 = A'(\theta_0) = \mathbb{E}_{\theta_0}[T(X)]$, while the inclusion (ii) is a consequence of the 1-Lipschitz continuity of $t \mapsto [t]_+$. Now we use the standard facts that $A(\theta)$ is analytic in θ and that $\theta \mapsto \mathbb{E}_\theta[f(X)]$ is continuous for any f (cf. [40, Thm. 2.7.1]) to see that for any $\epsilon > 0$, we can choose $\delta > 0$ such that $|t - t_0| \le \delta$ and $|\theta - \theta_0| \le \delta$ imply

$$|t - t_0| \le \epsilon, \quad |t - A'(\theta)| \le \epsilon, \quad \text{and} \quad \left|\mathbb{E}_{\theta_0}\left[[T(X) - t_0]_+\right] - \mathbb{E}_\theta\left[[T(X) - t_0]_+\right]\right| \le \epsilon.$$

References

[1] M. Abadi, A. Chu, I. Goodfellow, B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *23rd ACM Conference on Computer and Communications Security (ACM CCS)*, pages 308–318, 2016.

[2] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

[3] Apple Differential Privacy Team. Learning with privacy at scale, 2017. Available at https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html.

[4] B. Balle, J. Bell, A. Gascon, and K. Nissim. The privacy blanket of the shuffle model. In *39th Annual International Cryptology Conference (CRYPTO)*, 2019.

[5] P. Bickel, C. A. J. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer Verlag, 1998.

[6] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwebte Gebiet*, 65:181–238, 1983.

[7] L. D. Brown. *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, Hayward, California, 1986.

[8] L. D. Brown and M. G. Low. A constrained risk inequality with applications to nonparametric functional estimation. *Annals of Statistics*, 24(6):2524–2535, 1996.

[9] M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference (TCC)*, pages 635–658, 2016.

[10] T. Cai and M. Low. A framework for estimating convex functions. *Statistica Sinica*, 25: 423–456, 2015.

[11] E. J. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: Model-X knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society, Series B*, 80 (3):551–577, 2018.

[12] G. Casella and W. Strawderman. Estimating a bounded normal mean. *Annals of Statistics*, 9 (4):870–878, 1981.

[13] S. Chatterjee, J. Duchi, J. Lafferty, and Y. Zhu. Local minimax complexity of stochastic convex optimization. In *Advances in Neural Information Processing Systems 29*, 2016.

[14] J. E. Cohen, J. Kemperman, and G. Zbaganu. *Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics, and Population Sciences*. Birkhäuser, Boston, 1998.

[15] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition*. Wiley, 2006.

[16] P. Del Moral, M. Ledoux, and L. Miclo. On contraction properties of Markov kernels. *Probability Theory and Related Fields*, 126:395–420, 2003.

[17] R. L. Dobrushin. Central limit theorem for nonstationary markov chains. i. *Theory of Probability and Its Applications*, 1(1):65–80, 1956.

[18] D. L. Donoho and R. C. Liu. Geometrizing rates of convergence I. Technical Report 137, University of California, Berkeley, Department of Statistics, 1987.

[19] D. L. Donoho and R. C. Liu. Geometrizing rates of convergence II. *Annals of Statistics*, 19

(2):633–667, 1991.

[20] D. L. Donoho and R. C. Liu. Geometrizing rates of convergence III. *Annals of Statistics*, 19 (2):688–701, 1991.

[21] J. C. Duchi and R. Rogers. Lower bounds for locally private estimation via communication complexity. In *Proceedings of the Thirty Second Annual Conference on Computational Learning Theory*, 2019.

[22] J. C. Duchi and F. Ruan. A constrained risk inequality for general losses. *arXiv:1804.08116 [stat.TH]*, 2018.

[23] J. C. Duchi and F. Ruan. Asymptotic optimality in stochastic optimization. *Annals of Statistics*, To Appear, 2020.

[24] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *54th Annual Symposium on Foundations of Computer Science*, pages 429–438, 2013.

[25] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Minimax optimal procedures for locally private estimation (with discussion). *Journal of the American Statistical Association*, 113 (521):182–215, 2018.

[26] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3 & 4):211–407, 2014.

[27] C. Dwork and G. Rothblum. Concentrated differential privacy. *arXiv:1603.01887 [cs.DS]*, 2016.

[28] C. Dwork and A. Smith. Differential privacy for statistics: what we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2):135–154, 2009.

[29] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference*, pages 265–284, 2006.

[30] U. Erlingsson, V. Pihur, and A. Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 21st ACM Conference on Computer and Communications Security (CCS)*, 2014.

[31] U. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2019.

[32] European Union. 2018 reform of EU data protection rules, 2018. URL https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en. Accessed May 2018.

[33] A. V. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second Symposium on Principles of Database Systems*, pages 211–222, 2003.

[34] M. Hardt and K. Talwar. On the geometry of differential privacy. In *Proceedings of the Forty-Second Annual ACM Symposium on the Theory of Computing*, pages 705–714, 2010. URL http://arxiv.org/abs/0907.3754.

[35] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009.

[36] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II*. Springer, New York, 1993.

[37] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8):e1000167, 2008.

[38] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, 1986.

[39] L. Le Cam and G. L. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2000.

[40] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses, Third Edition*. Springer, 2005.

[41] F. W. Leysieffer and S. L. Warner. Respondent jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association*, 71(355):649–656, 1976.

[42] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

[43] I. Mironov. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017.

[44] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–125. IEEE, 2008.

[45] D. Pollard. Another look at differentiability in quadratic mean. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, chapter 19. Springer, 1997.

[46] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[47] A. Rohde and L. Steinberger. Geometrizing rates of convergence under differential privacy constraints. *arXiv:1805.01422 [stat.ML]*, 2018.

[48] C. Stein. Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 187–195, 1956.

[49] A. Tsybakov. Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Annals of Statistics*, 26(6):2420–2469, 1998.

[50] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[51] I. Vajda. On the f-divergence and singularity of probability measures. *Periodica Mathematica Hungarica*, 2(1–4):223–234, 1972.

[52] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[53] T. van Erven and P. Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[54] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

[55] S. Warner. Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[56] L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[57] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, 1997.