Learner-Private Convex Optimization

Jiaming Xu, Kuang Xu, and Dana Yang

***Abstract*— Convex optimization with feedback is a framework where a learner relies on iterative queries and feedback to arrive at the minimizer of a convex function. It has gained considerable popularity thanks to its scalability in large-scale optimization and machine learning. The repeated interactions, however, expose the learner to privacy risks from eavesdropping adversaries that observe the submitted queries. In this paper, we study how to optimally obfuscate the learner's queries in convex optimization with first-order feedback, so that their learned optimal value is provably difficult to estimate for an eavesdropping adversary. We consider two formulations of learner privacy: a Bayesian formulation in which the convex function is drawn randomly, and a maximin formulation in which the function is fixed and the adversary's probability of error is measured with respect to a minimax criterion. Suppose that the learner wishes to ensure the adversary cannot estimate accurately with probability greater than $1/L$ for some $L > 0$. Our main results show that the query complexity overhead is additive in L in the maximin formulation, but multiplicative in L in the Bayesian formulation. Compared to existing learner-private sequential learning models with binary feedback, our results apply to the significantly richer family of general convex functions with full-gradient feedback. Our proofs rely on tools from the theory of Dirichlet processes, as well as a novel strategy designed for measuring information leakage under a full-gradient oracle.**

***Index Terms*— First-order convex optimization, privacy, sequential learning, Dirichlet process.**

I. Introduction

CONVEX optimization with feedback is a framework in which a learner repeatedly queries an external data source in order to identify the minimizer of a convex function. This interactive nature of the framework, however, is a double-edged sword. On the one hand, iterative optimization methods offer inherent scalability since the learner is not required to access or store the entire function. As such, it has found applications in large-scale distributed machine learning systems, such as Federated Learning [1], [2], where a learner interacts with millions of individual users (data providers) in order to perform training. On the other hand, the repeated interactions with external entities exposes the learner to potential adversaries who may steal the learned model by eavesdropping on the queries exchanged during the training process, a woe especially poignant when the system involves a large number of data providers, many of which could be eavesdroppers in disguise ([3], [4, Section 4.3]).

To formulate the model stealing attacks and quantify the learner's privacy, we adopt the framework of Private Sequential Learning proposed in a recent line of research, aimed at quantifying the extra query complexities the learner has to suffer in order to ensure the submitted queries provably conceal the learned value [5], [6], [7]. The model is centered around a binary search problem where a learner tries to estimate an unknown value $X^* \in [0, 1]$ by sequentially submitting queries and receiving binary responses, indicating the position of X^* relative to the queries. Meanwhile, an adversary observes all of the learner's queries but not responses, and tries to use this information to estimate X^*. The learner's goal is to design a querying strategy with a minimal number of queries so that she can accurately estimate X^* while ensuring that the eavesdropping adversary cannot reliably estimate X^*. Progress has been made towards understanding the optimal querying strategies in this problem, and upper and lower bounds on the query complexity have been developed that differ by additive constants in the case where the learner's queries are noiseless [5], [7], and are order-wise optimal in the case of noisy queries [7].

While the original binary search formulation provides valuable insights, its assumption that the learner only has access to binary feedback is a severe restriction when it comes to modeling convex optimization. Indeed, most real-world applications provide the learner access to significantly richer feedback such as a full gradient (e.g., model training in machine learning). We elaborate further on the potential applications of our model in Section I-A.

The main purpose of the present paper is to take a step towards closing this gap by studying learner-private optimization with general convex functions and a full-gradient oracle. In a nutshell, our results demonstrate that the most prominent features of the query complexity in the binary search model extend gracefully to the general convex optimization setting. However, to establish that this is the case is far from trivial. A major difficulty stems from the significantly enriched functional class: unlike in a binary search problem where the ground truth is fully described by a scalar (location of X^*), we will see that the private query complexity crucially depends on the shapes of the convex functions in a family, and not just the locations of their minimizers.

Manuscript received 24 October 2021; revised 17 May 2022; accepted 18 August 2022. Date of publication 5 September 2022; date of current version 22 December 2022. This work was supported in part by the NSF under Grant IIS-1838124, Grant CCF-1850743, and Grant CCF-1856424; and in part by the NSF CAREER Award CCF-2144593. An earlier version of this paper was presented in part at the International Conference on Machine Learning, 2021. *(Corresponding author: Dana Yang.)*

Jiaming Xu is with The Fuqua School of Business, Duke University, Durham, NC 27708 USA (e-mail: jx77@duke.edu).

Kuang Xu is with the Stanford Graduate School of Business, Stanford University, Stanford, CA 94305 USA (e-mail: kuangxu@stanford.edu).

Dana Yang is with the Department of Statistics and Data Science, Cornell University, Ithaca, NY 14850 USA (e-mail: dana.yang@cornell.eu).

Communicated by A. Sarwate, Associate Editor for Security and Privacy.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3203989.

Digital Object Identifier 10.1109/TIT.2022.3203989

This added richness necessitates the development of both new problem formulations and analytical techniques. We propose in this paper two new learner-privacy frameworks: a new maximin formulation, as well as a Bayesian formulation that generalizes earlier Bayesian private sequential learning to a full-gradient oracle. A number of new techniques are developed to analyze query complexity under these formulations: we introduce tools from the theory of Dirichlet processes to construct priors that convey the richness of the model. Tools from nonparametric Bayes theory are deployed for the analysis under such prior distributions. In addition to an enriched functional class, another fundamental challenge lies in the richness of the feedback. Unlike the binary search model, the responses align with the location of the query and the shape of the unobserved convex function to a great extent. In the face of a more powerful learner equipped with a full-gradient oracle, we rely on a more sophisticated line of analysis to gauge the amount of information the responses reveal. We will discuss in more detail these ramifications in Section IV.

A. Motivating Examples

A learner naturally suffers from privacy breaches if the learning process involves interactions with third-party users.

1) Federated Learning: Federated Learning (FL) is an emerging model training paradigm, wherein the parameter server (viewed as learner) trains a model by communicating with distributed users while keeping the training data stored locally at the users. Suppose there are M users and each user u has access to a local dataset $\{Z_j, j \in S_u\}$. The learner aims to minimize the global empirical risk function $L(\theta) = \frac{1}{M}\sum_{u=1}^{M} \ell_u(\theta)$ where θ is the model parameter and $\ell_u(\theta) = \sum_{j\in S_u} \ell(Z_j, \theta)$ is the local empirical risk function of user u. The learner trains a model through sequential interactions with the users. For example in the *FederatedSGD* algorithm [1], in the i'th iteration of training, the learner broadcasts the current model parameter θ_i to users. The user then computes the local gradient $\nabla \ell_u(\theta_i)$ at the broadcasted model parameter θ_i and transmits it back to the learner. By taking an average of the received local gradients, the learner generates the next iteration of model parameter θ_{i+1}.

In typical FL systems, when training with thousands of users, the learner lacks enough administrative power over those external users. Thus, an adversary can participate in the training stage by pretending to be an authentic user and eavesdrop the sequence of broadcasted model iterates. By taking the last iteration of model parameters, the adversary can easily steal the trained model. Sophisticated models can be worth millions. The eavesdropper can use the stolen model to make money or even leverage them for illicit purposes [3]. Therefore, it is of paramount importance to protect the learner's privacy from the model stealing attacks [4], [8]. For example, when Google trains the Gboard for next-word prediction in FL systems, it is critical to prevent adversaries from stealing the Gboard model.

There are several potential techniques to conceal the model parameters from the users in Federated Learning, such as restricting each user to run the local computation inside a Trusted Executation Envrionments (TEE) [9] or encrypting the model parameters under a homomorphic encryption scheme before broadcasting it to the users [10]. Unfortunately, as pointed out by the recent survey [4, Section 4.3.3], TEEs and homomorphic encryption are often costly to implement and incur large overhead, especially when these users represent end-devices such as smartphones. This consideration prompts us to investigate whether we can offer provable privacy guarantees on the learner's model in FL.

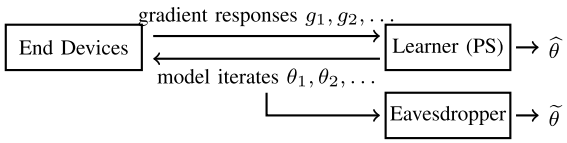

Fig. 1. An abstraction of a Federated Learning system with an eavesdropper. The end devices represent the entire population of the distributed users.

Our model can be viewed as an abstraction of the FL system, as illustrated in Fig. 1. The model iterates $\{\theta_i\}$ (viewed as queries) are observed by all users, including the eavesdropping adversaries. The response g_i corresponds to the average of the local gradients which is equal to the global gradient $\nabla L(\theta_i)$. It is reasonable to assume that the response is only observed by the learner but not the adversary. This is because the size of the adversarial users' local data is often negligible compared to the entire population. Thus in order to observe the response, the adversary would have to access the updates generated by all users in the system, which is not realistic for an adversary that only controls up to a small subset of the users. Since the communication bandwidth is a scarce resource in many FL applications, determining the minimum number of iterations needed is of fundamental importance in both theory and practice, which we will address by investigating the optimal query complexity.

In our current formulation, we assume that the learner is trusted by the users and that the users transmit honest, noiseless responses to the learner. A more realistic representation of FL would allow for noisy or even adversarial responses. We leave that for future investigations. We also note that private FL has been studied in the context of protecting each individual user's local data privacy by adding randomization or noises to local gradient responses and/or model iterates [11], [12]. In sharp contrast, our work focuses on preventing the adversary inferring the learned model by carefully designing the querying strategy with obsfusction.

2) Additional Applications in Medicine and Business: Given the close connection between convex and monotone functions, our work can also be applied to learning monotone functions, for example to clinical dose-response studies [13], [14]. In dose-response analysis, the potency curve $\mu(x)$ is a monotone function that models the treatment effectiveness as a function of the dosage. An incredily important problem is to estimate the minimum effective dose (MED)

$$\mathsf{MED} = \min_x \{x : \mu(x) > \mu(0) + \Delta\}$$

for some threshold Δ. Note that the MED is the minimizer X^* of some unknown convex function f^* (e.g. $f^*(x) = \int_0^t \mu(t)dt - [\mu(0) + \Delta]x$). In new drug development,

a pharmaceutical company (viewed as learner) often estimates MED of a drug via running adpative clinical trials. In particular, the recruiting participants will take the drug according to a given dosage (viewed as query). Based on their responses, the company then measures the treatment effectiveness and determines the next dosage to query. The adversary, who may be a competing company, can pretend to a participant and hearby eavesdrop the queried dosage. Due to the critical importance of MED, it is of great interest for the company to design an adaptive querying scheme to estimate the MED, while preventing the adversary from gaining information on the MED. We also remark that the Dirichlet process is widely used in isotonic regression for modeling monotone functions [14], [15], as we will do when modeling the gradient of the convex function under the Bayesian framework.

Another potential application is pricing optimization, where the goal is to learn the optimal release price of a product by conducting market experiments at test price points (queries). See [5], [7] for more detailed discussions on the pricing optimization example.

B. Related Work

Private information retrieval (PIR) and private function retrieval (PFR) Our model formulation bears some similarities with the PIR [16], [17], [18] and PFR [19] framework. However, there are major distinctions which result in completely different dynamics between the learner and the adversary. In PIR, the database is assumed to contain a vector $(x_i)_{i\leq N}$. The learner's goal is to learn the evaluation x_i at some index i by querying the database, while preventing the database (adversary) from learning the value of i. The PFR problem is formulated similarly, except that the database is indexed by functions. Note that in PIR/PFC, the private index is assumed to be known to the learner a priori. In contrast, in our framework, the private information X^* is something the learner herself is in the process of discovering. As a result, our problem is posed as a sequential learning problem. It has natural applications in model stealing attack prevention, where eavesdropping adversaries attempt to steal the model parameters by participating in the model training process. The fundamental difference between the two settings also leads to completely different techniques for analysis. For us, privacy is ensured by utilizing the adversary's lack of knowledge on the responses, which is not the case in PIR/PFC.

Data-owner privacy models Similar to Private Sequential Learning, the private convex optimization problem we consider diverges significantly from the existing literature on differentially private (DP) iterative learning [20], [21], [22], [23], [24], a key difference being that the latter focuses on protecting data owners' privacy rather than learner's privacy. To protect data owners' privacy, the notion of differential privacy [25] is often adopted and privacy is often achieved by injecting calibrated noise at each iteration of the learning algorithms. In contrast, our work focuses on preventing the adversary inferring the learned model, which is conceptually closer to recent studies of information-theoretically sound obfuscation in sequential decision-making problems [26], [27], [28], [29], [30]. See [7] for a comprehensive discussion on the distinction between data-owner privacy models and this line of work.

More recently, variants of differential privacy, such as local differential privacy (LDP) [11], [12], [31], have been applied to a learning context, in ways that are closer to the spirit of models considered here. However, we should note that in addition to the aforementioned distinction between a learner- vs. data-owner-centric focus, the formulation of private sequential learning also diverges from that of DP and its variants in other important aspects. The approach here tends to have a more limited privacy scope, aiming to protect the learner against a specific inferential goal of the adversary, whereas the more conservative differential privacy aims to protect against a much wider range of possible attacks. As a result, more randomization and obfuscation is generally required under a DP setup than the one considered here, and possibly at the cost of more efficiency loss. For instance, Appendix B of [32] includes an example showing that while a certain Replicated Bisection search strategy is provably private under the Private Sequential Learning framework, it is in fact never differentially private, thus showing that differential privacy is a strictly more restrictive notion. We also refer the reader to a discussion on the comparison of scope with DP in [5]. Another consideration is that, as mentioned in the Introduction, we are generally concerned with settings where the learner's actions can be observed exactly by the adversary, in contrast to typical applications of DP and LDP where the decision maker has the extra degree of freedom of adding additional noise to the data or summary statistic before releasing it for public scrutiny (cf. [11]). While the learner in our model can also obfuscate by injecting randomness into their actions, such obfuscation has more immediate consequences, because it would directly impact the resulting information the learner collects. The same consideration also arises in other models of privacy where the agent's actions are observable and therefore she can only obfuscate by "doing" rather than "hiding" [26], [28], [29]; we refer the reader to [28] for a more elaborate discussion on this distinction and the references therein.

Strategic learning In aiming to prevent modeling stealing, our work aligns with a growing literature on strategic learning and prediction [33], [34], [35], [36], [37]. These papers consider strategic learners who have gained access to their competitor's predicted samples, or even the competitor's entire predictive model. Then, they artificially adjust their own predictive model in order to outperform those of their competitors. In general, in equilibrium such competition could not only harm utilities for the learners involved, but also lead to lower overall social welfare, as defined by the prediction quality experienced by end consumers. Our work thus helps to preempt such pitfalls by providing a theft-proof framework for models training and adaptive data collection.

II. The Model: Learner-Private Convex Optimization

We now introduce our model, dubbed Learner-Private Convex Optimization. The emphasis on the learner's privacy here is to distinguish our model from other forms of private

sequential learning, especially those that focus on protecting the privacy of data owners (See preceding discussion in the Introduction).

A. Learner

Let $\mathcal{F}$ be a family of $\mathbb{R}$-valued convex functions with domain $[0,1]$, such that all elements in $\mathcal{F}$ admit a unique minimizer. Suppose there is an unknown *truth* $f^* \in \mathcal{F}$ with the minimizer $X^* := \arg\min_x f^*(x)$. Fix $n \in \mathbb{N}$. Our decision maker is a *learner* who wants to identify X^* by sequentially submitting a total of n queries in $[0,1]$ to an oracle. For the ith query, q_i, the oracle returns a response r_i that is equal to the gradient of f^* at q_i:

$$r_i = (f^*)'(q_i). \tag{1}$$

If f^* is not differentiable at q_i, then r_i is an arbitrary subgradient of f^* at q_i. We assume that the learner is allowed to introduce outside randomness, in the form of a random seed Y that takes value in a finite discrete alphabet. Formally, we denote by ϕ the *learner's strategy*, which consists of a sequence of mappings $\phi_0, \phi_1, \ldots, \phi_{n-1}$ such that the ith query is generated as a function of all previous responses and the random seed:

$$q_i = \phi_{i-1}(r_1, \ldots, r_{i-1}, Y). \tag{2}$$

Once the querying process is terminated, the learner constructs an estimator of the optimizer X^*, $\widehat{X} = \widehat{X}(q, r, Y)$. We say that the learner strategy ϕ is ϵ-accurate, if

$$\mathbb{P}\left\{\left|\widehat{X} - X^*\right| \le \epsilon/2\right\} = 1, \quad \forall f^* \in \mathcal{F}, \tag{3}$$

where X^* is the minimizer of f^* and the probability averages over the randomness in the random seed, Y.

B. Adversary

Meanwhile, an adversary is trying to learn X^* by eavesdropping on the learner's queries: we assume that the adversary observes all n queries submitted by the learner, but not their responses. Denote by $\widetilde{X} = \widetilde{X}(q)$ the adversary's estimator, which is a (possibly random) function of $(q_i)_{i=1,\ldots,n}$. Wary of such an adversary, the high-level objective of the learner are to (1) generate a query sequence that is largely "uninformative" towards X^*, and (2) at the same time minimizing the number of queries needed, n.

We next formalize in what sense a learner's strategy can be private. Generally speaking, a learner strategy is private if we can ensure that the adversary's estimator $\widetilde{X}$ is *not accurate*. Importantly, different definitions of the adversary's accuracy will lead to drastically different definitions of privacy, and consequently, distinct algorithms, guarantees and domains of applications. In this paper, we will analyze two privacy metrics, Bayesian and maximin, that parallel the two paradigms in the statistics literature. The Bayesian formulation extends the Bayesian private learning model in [5], while the maximin formulation is new.

C. Maximin

The truth f^* is a deterministic but unknown function in $\mathcal{F}$. We say that a learner strategy ϕ is (δ, L)-private if

$$\sup_{\widetilde{X}} \inf_{f^* \in \mathcal{F}} \mathbb{P}\left\{\left|\widetilde{X} - X^*\right| \le \delta/2\right\} \le 1/L, \tag{4}$$

where the probability is measured with respect to the internal randomness employed by the learner's querying strategy and that used in the adversary's estimator. In other words, the learner strategy is considered private if the adversary's minimax accuracy is low.

D. Bayesian

The truth f^* is drawn from a prior distribution π, a probability distribution over $\mathcal{F}$. We say that a learner strategy ϕ is (δ, L)-private if

$$\sup_{\widetilde{X}} \mathbb{P}\left\{\left|\widetilde{X} - X^*\right| \le \delta/2\right\} \le 1/L, \tag{5}$$

where the probability is measured with respect to all randomness in the system, including the prior π and any internal randomness employed by the learner's querying strategy and the adversary's estimator. We note that our Bayesian notion of privacy bears some similarity with the reconstruction privacy proposed in [11, Definition 2.3]. One notable difference is that the probability in (5) is conditional on the query sequence q and thus the reconstruction privacy is more stringent, requiring the adeversary to fail for every possible realization of query sequence q. Moreover, despite the similarity in definitions, the reconstruction privacy is adopted to protect the data owner's privacy as opposed to the learner's privacy.

E. Private Query Complexity

Finally, we have come to the main metric of interest. In both the maximin and the Bayesian formulations, we define the optimal query complexity, $N(\epsilon, \delta, L)$, as the least number of queries necessary for there to exist an ϵ-accurate learner strategy that is also (δ, L)-private:

$$\begin{aligned} N(\epsilon, \delta, L) = \min\{n : \exists \phi \text{ with at most } n \text{ queries,} \\ \text{that is } \epsilon\text{-accurate and } (\delta, L)\text{-private}\}. \end{aligned}$$

The Bayesian and maximin formulations translate to lower bounds on the adversary's Bayes risk and minimax risk respectively. To break the maximin privacy, the adversary needs to perform well in the worst case. Thus the maximin privacy criteria is a lot harder on the adversary and easier on the learner. Hence, the optimal query complexity is significantly lower under the minimax formulation than the Bayesian one, as our results will show next.

The Bayesian formulation is more intuitive and we include the maximin formulation primarily for completeness to provide results for both estimation frameworks. The maximin formulation is much more stringent on the adversary. One possible interpretation is that the adversary is very risk-averse. A potential application is autonomous driving [38], where the goal is to protect the privacy of a flagship manufacturer (learner) from model stealing attacks of competing

companies (eavesdropping adversary). The risk-averse nature of autonomous driving algorithms forces the adversary to ensure that the stolen model performs reliably under all circumstances, or, minimax-accurate in the traditional statistical sense. Without a worst-case guarantee, the adversary cannot act upon the stolen model. Thus a maximin-private learner's strategy renders the adversary powerless.

III. Main Results

A. Maximin Formulation

We will assume that the function class $\mathcal{F}$ satisfies the following assumption:

Assumption 1 (Complexity of $\mathcal{F}$): Fix $f \in \mathcal{F}$ and interval $I \subset [0,1]$ that contains the minimizer of f. Then, for every $x \in I$, there exists $g \in \mathcal{F}$ such that g is minimized at x, and the gradient of f and g coincide outside of I.

Assumption 1 is needed to rule out trivial cases where a learner may exactly pinpoint the location of the minimizer solely by looking at far-away gradients. We show in Section V that this richness assumption on $\mathcal{F}$ is in some sense necessary. Examples of function classes that satisfy Assumption 1 include the set of all convex functions on $[0,1]$, the set of all convex functions in $C^1([0,1])$, and the set of all piecewise-linear convex functions on $[0,1]$. The next theorem is our main result for the maximin formulation:

Theorem 1 (Maximin Query Complexity): Assume that $\mathcal{F}$ satisfies Assumption 1. If $2\epsilon \le \delta \le 1/L$, then[1]

$$N(\epsilon,\delta,L) \le \begin{cases} 2L + \log\frac{\delta}{\epsilon} & \text{if } L \ge \log\frac{1}{\delta} \\ L + \log\frac{1}{\epsilon} & \text{o.w.} \end{cases}.$$

Furthermore, if all functions in $\mathcal{F}$ are differentiable on $[0,1]$, then

$$N(\epsilon,\delta,L) \ge 2L + \log\frac{\delta}{\epsilon} - 2.$$

Note that if there were no privacy consideration, the maximin optimal query complexity would be $\log(1/\epsilon)$. Thus under the maximin formulation, a higher level of privacy L leads to an *additive* overhead in the optimal query complexity, that is at most about $2L$.

Remark 1: In the proof of Theorem 1, we only use convexity of the functions to ensure that the sign of the gradient is consistent with the direction of X^*. Therefore assuming $\mathcal{F}$ satisfies Assumption 1, the results of Theorem 1 hold true for more general classes of unimodal (but not necessarily convex) functions, i.e. for all $f^* \in \mathcal{F}$, $(f^*)'(x) < 0$ for $x < X^*$ and $(f^*)'(x) > 0$ for $x > X^*$.

Remark 2 (Multidimensional Extensions): By considering a separable class of functions, and using the ℓ_∞ norm to measure the error of the learner and the adversary's estimators, Theorem 1 can be extended to d dimensions. The upper and lower bounds of the query complexity take the same form, with L replaced with $L^{1/d}$. See the supplementary material for the precise statement and proof.

[1]Here and subsequently log refers to logarithm with base 2.

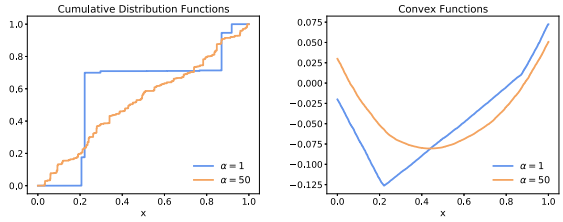

Fig. 2. The left figure exemplifies realizations of F following the Dirichlet Process with base function $\lambda_{[0,1]}$ and different concentration parameters α. The right figure shows the corresponding convex functions f^*, with $\gamma_+ = 0.5$ and $\gamma_- = -0.5$.

B. Bayesian Formulation

In the Bayesian formulation, we seek a function class and prior distribution that are sufficiently rich to capture real-world data, while at the same time amenable to analysis. A good candidate in this respect is the so-called Dirichlet process, a family of measures over non-decreasing functions, which we will use to model the gradient function of f^*. Dirichlet processes are fundamental objects in nonparametric Bayes theory and widely used in Bayesian isotonic regression for modeling monotone functions [14], [15], [39]. We begin by defining a Dirichlet process:

Definition 1 (Dirichlet Process): Given a base probability measure μ_0 on $\mathcal{X}$ and a concentration parameter $\alpha > 0$, a random probability measure μ over $\mathcal{X}$ is said to follow the Dirichlet process $\mathrm{DP}(\mu_0,\alpha)$, if for any finite partition of $\mathcal{X} = \cup_{i\le n}\mathcal{X}_i$,

$$(\mu(\mathcal{X}_1),\ldots,\mu(\mathcal{X}_n)) \sim \mathrm{Dir}((\alpha\mu_0(\mathcal{X}_1),\ldots,\alpha\mu_0(\mathcal{X}_n))),$$

where $\mathrm{Dir}(c)$ denotes the Dirichlet distribution over the n-dimensional simplex Δ^{n-1} with density

$$g_{\mathrm{Dir}(c)}(x_1,\ldots,x_n) \propto \prod_{i=1}^{n} x_i^{c_i-1}, \quad x \in \Delta^{n-1}. \tag{6}$$

We now construct the prior distribution of f^* using a Dirichlet process. The prior is parameterized by two quantities:

1) a concentration parameter $\alpha > 0$, which controls the dispersion of the distribution of the minimizer;
2) a probability distribution η on $\mathbb{R}_+$, which captures the range of gradients of f^*. We assume that $\eta(\{0\}) = 0$ to ensure uniqueness of the minimizer X^*.

Definition 2 (Bayesian Prior using Dirichlet Process): Fix α and η. Denote by $\lambda_{[0,1]}$ the Lebesgue measure restricted to $[0,1]$. Then, the prior π corresponds to the following procedure for generating f^*:[2]

1) Sample γ_+ from η. Set $\gamma_- = -\gamma_+$.
2) Sample μ from the Dirichlet process with concentration parameter α and base distribution $\lambda_{[0,1]}$. Let F be the cumulative distribution function of μ.
3) Set $f^*(x) = \gamma_- x + \int_0^x (\gamma_+ - \gamma_-)F(t)dt$, for $x \in [0,1]$.

[2]Note that in this definition we have restricted the function f^* to have zero intercept. This restriction is without loss of generality, since any constant offset will not change the location of a minimizer and similarly our results will carry through if one wishes to incorporate a different intercept.

Note that $(f^*(x))' = \gamma_+ (2F(x) - 1)$ and thus the minimizer X^* of f^* corresponds to the median of F, or more precisely the smallest x for which $F(x) \geq 1/2$. By construction, F is a monotone simple function that consists of countably many points of discontinuity that are dense on $[0, 1]$. Its level of discreteness is modeled through the concentration parameter α. For a small α, the increase of F from 0 to 1 is mostly from a few abrupt jumps, and the convex function f^* resembles a piece-wise linear function with finitely many pieces; as α grows, the increase of F becomes more gradual, and f^* starts to concentrate around a smooth quadratic function. See Figure 2 for some realizations of the distribution function F and the corresponding convex function f^* for different value of α.[3]

The following theorem is our main result for the Bayesian formulation.

Theorem 2 (Bayesian Query Complexity): Fix $\alpha > 0$. Suppose that $2\epsilon \leq \delta < \frac{1}{2LH_\alpha}$, with $H_\alpha = (3 + 2e^{-1})\alpha + 14$. Then

$$c_1 L \log \frac{\delta}{\epsilon} \leq N(\epsilon, \delta, L) \leq L \log \frac{\delta}{\epsilon} + c_2 L + \log \frac{1}{\delta L},$$

where c_1,c_2 are positive constants that only depend on α such that $c_1 \to 1$ as $\alpha \to 0$.

The above theorem shows that, in the Bayesian formulation, the query complexity overhead due to privacy constraints scales *multiplicatively* with respect to the privacy level L. Note that this is substantially higher than the maximin setting where such overhead is only additive in L. When $\alpha \to 0$, F converges to a step function and our query complexity bounds recover the existing ones in the binary search problem [6], showing that $N(\epsilon, \delta, L) \sim L \log \frac{1}{\epsilon}$ as $\epsilon \to 0$ for fixed δ, L.

IV. Discussion

In this section, we discuss some of the most salient features of our main results and modeling assumptions.

A. Applying the Bayesian and Maximin Privacy Criteria

Our results show that the two privacy criteria lead to distinct query complexity scalings, so it would be instructive to understand in what application domain each metric is most applicable. The Bayesian formulation is more straightforward: both the adversary and the learner are assumed to have access to the historical data that forms the prior distribution, and all probabilities in various guarantees are measured with respect to such shared common knowledge. The emphasis of this paper is on the Bayesian formulation: it requires more novel analysis techniques, and we expect the Bayesian formulation to be most relevant in data-driven machine learning and online optimization such as in Federated Learning and pricing optimization; the aforementioned dose-response analysis is also a natural application of the Bayesian formulation due to the close connection between potency curves and convex functions.

[3]To plot the convex functions together, we shift them by some constants on the y-axis. This shift is irrelevant to the optimization task since the response only contains gradient information.

The maximin formulation is a new metric proposed in this paper, and we discuss here some nuances with this definition. Note that minimax guarantees in traditional statistical learning are typically the strongest, since they hold over any adversary choice of problem parameter. However, in our setting, the maximin formulation provides arguably the weakest privacy guarantee due to the negation inherent in its definition: a learner strategy is maximin-private as long as there does *not* exist a minimax-accurate adversary estimator. For instance, even if an adversary is able to accurately predict X^* under the majority of functions in $\mathcal{F}$, failing only over a small subset, the learner can still proclaim its strategy to be private under the maximin formulation.

We tend to believe that the maximin formulation is more applicable when the adversary is very risk-averse. One interesting example is in law and criminal justice. Here, a prosecutor should have to prove that the accuracy of any conclusion drawn from evidence holds up *regardless* of the value of a certain hidden parameter. Indeed, many legal systems currently require that criminal convictions be reached only if the evidence can prove guilt "beyond reasonable doubt" (cf. [40], [41], [42]). Any supposed prior on crucial, unobserved parameters can be ill-defined and potentially discriminatory. Other potential applications include the aforementioned autonomous driving [38] application, where the performance guarantee of an estimator needs to be valid in the worst case, for the sake of public safety. In these examples, a maximin-private learner strategy will effectively prevent the adversary from coming up with *any* viable estimator, thus render the adversary powerless.

B. Comparisons With Private Sequential Learning

As mentioned in the Introduction, our convex optimization framework generalizes the Private Sequential Learning (PSL) model. As such, the two settings share similarities (as one would expect), as well as some marked differences. Recall that in the PSL framework, the responses are binary and only indicate whether the minimizer is to the left or right of a given query; this is equivalent, in our setting, to returning only the sign of the gradient. The maximin and Bayesian formulations proposed in this paper parallel the deterministic and Bayesian formulations in PSL, respectively. Our maximin formulation is new, but it does have a fundamental connection to the deterministic formulation of PSL, where a learner strategy is considered private if its queries are guaranteed to generate a large set of "plausible" targets (information set), with a large covering number[4]; we explore this formally in Section V-B. Our Bayesian formulation is a natural generalization of the Bayesian PSL model: we now assign a prior over the entire function, as opposed to only the location of the minimizer. Notably, our Bayesian formulation recovers the original Bayesian PSL problem in the limit where the concentration parameter α in the Dirichlet prior approaches 0. As such, our Bayesian formulation includes the original Bayesian PSL model as a special case.

[4]The δ-covering number of a set $A \subset \mathbb{R}$ is the size of the smallest set $\mathcal{N}$, such that $\cup_{r \in \mathcal{N}} [r - \delta/2, r + \delta/2] \supseteq A$.

Our main results recover similar dependencies on the level of privacy, with overheads that are additive and multiplicative in L in the maximin and Bayesian formulations, respectively. The upshot in our setting is that the results are established in a substantially more general setting of convex optimization.

There are several major differences that distinguish our private convex optimization framework from the PSL model. First and foremost, the learner now has access to the entire gradient instead of only its sign. A most direct implication of this enriched information structure is that, when analyzing the amount of information leakage of a learner strategy, we will have to keep track of the distributions over target functions, as opposed to only the minimizers, as was the case in PSL. Moreover, when the learner has access to full gradients, it is in principle possible for the learner to gather information about the minimizer's precise location even from queries that are submitted far away from the minimizer, which was not possible within bisection search. For instance, if the underlying target function is known to be quadratic, then two queries placed anywhere are sufficient to uncover the minimizer. To address these complexities, our goal is to precisely measure the amount of information about the minimizer that the learner and adversary may obtain from a given sequence of queries. We will do so both by developing more sophisticated information theoretic arguments, and by exploiting structural properties of the Dirichlet process.

C. Open Questions

Our results leave open a number of questions. For the Bayesian query complexity in one dimension, there remains a gap between the leading constants in the upper and lower bounds, in the regime where α is bounded away from zero. Generalizing the main theorems to a multi-dimensional setting, where $x \in \mathbb{R}^d, d \geq 2$, is also interesting and practically relevant. We take a first step in this direction by extending our results to multi-dimensional separable functions (see supplementary material), while the general case with non-separable objective functions remains open and appears to be challenging. Our problem formulation only considers first-order feedback. An interesting direction is to consider convex optimization with more general types of feedback, e.g., bandit feedback [43].

D. A Different Notion of Privacy in [44]

A recent work [44] also aims to extend the private sequential learning model of [5] to convex optimization. They use a different notion of privacy criteria that bear some superficial similarities to ours. However, the definition of privacy in [44] contains crucial errors that render it vacuous, in the sense that there cannot exist any private learner strategy satisfying that definition. To be precise, here is Definition 2 of [44]: fix $\epsilon, \delta \in (0,1)$. A learner strategy is said to be (ϵ, δ)-private if for any adversary estimator $\widetilde{X}$ *and* any truth $f \in \mathcal{F}$,

$$\mathbb{P}(\text{err}(\widetilde{X}, f) \leq \epsilon) \leq \delta, \tag{7}$$

where $\text{err}(\cdot,\cdot)$ is a certain error function which measures the discrepancy between the adversary estimator and the true minimizer. For instance, in our example $\text{err}(\widetilde{X}, f) = |\widetilde{X} - \arg\min f(x)|$.

The problem with this privacy definition is that it can never be satisfied by any learner strategy. Indeed, consider an adversary that simply sets $\widetilde{X} = x^*$ for some fixed x^* without even taking into account the queries. Under this trivial estimator, we automatically have $\mathbb{P}(\text{err}(\widetilde{X}, f) = 0) = 1$ for any $f \in \mathcal{F}$ that is minimized at x^*. So (7) cannot possibly hold uniformly across all adversary estimators and all f.

This faulty formulation should lead to an infinite optimal query complexity, suggesting that the upper bound proof contains error. Upon a closer look, we find that in the upper bound analysis, the authors make the following assumption: *"Without loss of generality, assume the adversary is endowed with an uniform prior knowledge on where X^* is and assume the maximum uncertainty for X^*"* [44, p19]. This assumption specifies the behavior of the adversary, and is therefore inconsistent with the minimax formulation which requires the adversary to fail for *any* adversary's estimator.

V. Proof of Main Results

We present in this section the proofs of our main results. We begin by giving an overview of the key steps.

A. Overview of Main Ideas

1) Maximin Setting: Since the response contains the full gradient information, the key challenge in the analysis is to track the amount of information available to the learner. Note that aside from the directional information $\mathbb{1}\{X^* \geq q_i\}$, the response for a query q_i contains additional information on $(f^*)'(q_i)$. The key message in the proof under the maximin setting, is that under the Assumption 1 on the richness of the family of functions, only the directional information is relevant to the learning task. Therefore, it suffices to only track the learner's knowledge with the directional information from the responses.

Starting with the upper bound, we design a querying strategy that is ϵ-accurate, (δ, L)-private, and submits at most $\max\{2L + \log(\delta/\epsilon), L + \log(1/\epsilon)\}$ queries. In particular, our querying strategy only utilizes the directional information of the gradient responses. Firstly, note that since the gradient responses contain the binary directional information, the learner can always check whether an interval contains X^* by querying the two endpoints: $X^* \in [a,b]$ if and only if $(f^*)'(a) \leq 0$ and $(f^*)'(b) \geq 0$. All intervals refer to closed intervals unless otherwise specified. We refer to a pair of queries at q and $q + \epsilon$ as a *guess*. The key privacy-ensuring mechanism is to check L guesses that are δ apart from each other. By doing so, the learner manually plants L possible locations for X^* that an adversary cannot rule out without observing the responses, thus achieving (δ, L)-privacy.

To prove the lower bound, we need to show that a querying strategy that only utilizes the directional information can be optimal. Firstly, let us give a heuristic argument of why only the gradient information is relevant to learning X^* under

Assumption 1. Given $(f^*)'(a) < 0$ and $(f^*)'(b) > 0$, under Assumption 1, X^* can be anywhere between a and b regardless of the value of the gradients $(f^*)'(a)$, $(f^*)'(b)$. We should point out that the richness assumption is necessary. For example suppose $\mathcal{F}$ is the family of convex polynomial functions with fixed degree d. Then the learner can solve for the X^* by submitting d distinct queries at arbitrary locations, making both learning and obfuscation trivial.

The lower bound proof contains two main ingredients.

(a) Step 1: Rigorously justify the claim that under Assumption 1, the learner does not benefit from the additional gradient information aside from the one-bit directional response. In particular, we show that the learner cannot search faster than the bisection method on any interval $I \subset [0,1]$. Therefore, for each interval of length δ, it takes at least $\log(\delta/\epsilon)$ queries in I to achieve ϵ-accuracy, in the worst case.

(b) Step 2: Relate the adversary's statistical performance to the size of the information set [5] of a query sequence q, defined as

$$\mathcal{I}(q) = \{x \in [0,1] : \exists f \in \mathcal{F} \text{ and } y, \\ s.t.\ \ x = \arg\min f, \text{ and } \mathsf{q}(f,y) = q\},$$

where $\mathsf{q}(f,y)$ refers to the realization of the query sequence when $f^* = f, Y = y$. The information set contains all possible values of X^* that could lead to the query sequence q. We show that to ensure the adversary achieves δ-accuracy with probability at most $1/L$, there must be some q for which the δ-covering number of $\mathcal{I}(q)$ is at least L. Note that from the ϵ-accuracy requirement, each member of $\mathcal{I}(q)$ is sandwiched between a pair of queries in q that are at most ϵ-apart. Therefore, q contains at least L such pairs of queries, contributing a total of $2L$ queries.

After performing these two steps, some challenges remain. The functions associated with q (in step 2) may not coincide with the worst-case instances that arise from step 1. Therefore, the remaining task is to combine the two lower bounds $\log(\delta/\epsilon)$ and $2L$. For this step, we show the existence of some interval I, such that for some f minimized in I, the learner must pay not only the $\log(\delta/\epsilon)$ queries for accuracy, but also the $2L$ queries for privacy. The high-level idea behind the proof is to divide q into two sub-sequences q_{before}, q_{after}, before and after the $2L$ queries (in step 2) are submitted. The key observation is that q_{before} is shared by a large class of functions whose minimizers lie in some δ-length interval I. For all these functions, the cost of $2L$ queries would have been committed in q_{before}. For at least one of them, an extra cost of $\log(\delta/\epsilon)$ queries must be paid in q_{after}.

2) Bayesian Setting: Similar to the maximin setting, the upper bound here is also established by analyzing a constructive algorithm. The key challenge in designing a private learning algorithm in the Bayesian setting arises from the fact that the prior distribution on X^* is always non-uniform under the Dirichlet process model. In particular, we can no longer simply apply the replicated search strategy from [7], since the non-uniform distribution of X^* provides the adversary with additional prior information.

To address this difficulty, our key algorithmic idea is to find L intervals that occupy the same prior mass, while at the same time are at least δ-separated from each other. One of these intervals contains the true value X^*. On each of the other $L-1$ intervals, we sample a proxy for X^* according to the conditional distribution of X^* restricted to the interval. Via a genie-aided reduction argument, we show that the adversary cannot perform better than a random guess among the L candidates: the truth and the $L-1$ proxies. By construction of the intervals, the L candidates are at least δ apart. Therefore the adversary cannot achieve an additive error of $\delta/2$ with probability higher than $1/L$.

For the lower bound, the challenge again lies in tracking and quantifying the amount of information the learner gains from the responses. Compared to the binary search model, the full gradient responses can potentially reveal too much information to the learner. To tackle this challenge, our key proof strategy is to find a event on which the learner cannot gather information on X^* too rapidly. The proof follows the following main steps.

(a) Step 1: quantify the learner's information. We adopt the notion of "learner's intervals", $I_0, I_1, \ldots$. Here, $I_0 = [0,1]$ and I_i is the smallest interval that the learner knows to contain X^* after the first i queries.

(b) Step 2: analyze the conditional distribution of X^* over the learner's interval. This is the key step of the proof. We want to find a "good" event on which the learner does not possess too much information on the location of X^*. In this step, we construct an event $\mathcal{B}$, such that

$$X^* \mid \mathcal{B} \sim \mathrm{Unif}[I_i \cap J], \tag{8}$$

where J is an arbitrary subinterval of $[0,1]$. Here, $\mathcal{B}$ is an event that encodes all the information available to the learner up to time i, the assumption that $X^* \in J$, and some other desirable feature of the unknown convex function f^*. The construction of $\mathcal{B}$ crucially utilizes the stick-breaking characterization of the Dirichlet Process, and the proof of (8) heavily relies on the self-similarity property of the Dirichlet Process. We discuss the self-similarity property in detail in Section A. It helps control the amount of information about the shape of f^* inside the learner's interval, given all the queries and responses outside.

(c) Step 3: control the speed at which the learner's interval shrinks. Divide $[0,1]$ into $2/\delta$ subintervals $J_1, \ldots, J_{2/\delta}$ of length $\delta/2$, and let J^* denote the subinterval that contains X^*. In this step, from (8), by integrating over instances of $\mathcal{B}$, and letting J range over the $2/\delta$ subintervals, we show that for some suitably-defined event $\mathcal{A}$,

$$\begin{aligned} &\mathbb{E}\left(\log \frac{|I_{i+1} \cap J^*|}{|I_i \cap J^*|} \,\middle|\, \mathcal{A}\right) \\ &\geq -\mathbb{P}\left\{q_{i+1} \in I_i \cap J^* \mid \mathcal{A}\right\} \\ &\geq -\mathbb{P}\left\{q_{i+1} \in J^* \mid \mathcal{A}\right\}. \end{aligned} \tag{9}$$

(d) Step 4: from (9), via a simple telescoping sum and an application of Jensen's inequality, we can deduce that

$$\begin{aligned}&\mathbb{E}\left(\text{number of queries in } J^* \mid \mathcal{A}\right)\\ &\geq \log\frac{\delta}{2} - \log\mathbb{P}\left(|I_n \cap J^*| \mid \mathcal{A}\right) \geq \log\frac{\delta}{\epsilon},\end{aligned}$$

where the second inequality follows from the ϵ-accuracy requirement. By considering an adversary who adopts the *proportional-sampling* strategy [6], we have for any querying strategy that is (δ, L)-private,

$$\begin{aligned}n =& \text{total number of queries}\\ \geq& L\mathbb{E}\left[\text{total number of queries in } J^*\right]\\ \geq& \mathbb{P}\left(\mathcal{A}\right)\cdot L\log\frac{\delta}{\epsilon}.\end{aligned}$$

B. Proof Under the Maximin Setting

Proof of the Upper Bound in Theorem 1: Define a guess at q as a pair of queries placed at q and $q+\epsilon$. The guess allows the learner to test whether X^* is contained in the ϵ-length interval $[q, q+\epsilon]$. To ensure privacy, we create L potential locations for X^* that are at least δ-separated but induce the same querying sequence. That is achieved by submitting L *guesses* that are δ-separated. Once guessed correctly, the learner's accuracy requirement is automatically fulfilled and the remaining queries can be used to conceal X^* from the adversary. We consider the cases $\delta \leq 2^{-L}$ and $\delta > 2^{-L}$ separately. The querying strategy is contained in Algorithm 1.

We first prove the upper bound in the case $\delta \leq 2^{-L}$. In the first stage, the learner submits L guesses, each consisting of two queries. In the second stage, the learner runs a bisection on an interval of length 2^{-L} up to accuracy ϵ. In total, $2L + \log(2^{-L}/\epsilon) = L + \log(1/\epsilon)$ queries are submitted. The strategy is clearly ϵ-accurate. To see that it is also (δ, L)-private, note that all f^* whose minimizer lies in one of the L intervals $[1/2, 1/2+\epsilon]$, $[3/4, 3/4+\epsilon]$, …, $[1-2^{-L}, 1-2^{-L}+\epsilon]$ share exactly the same query sequence. Under Assumption 1, for each i there exists at least one function f_i minimized at some $x_i \in [1-2^{-i}, 1-2^{-i}+\epsilon]$. When $\delta \leq 2^{-L}$, the x_i's are at least δ apart from each other. Therefore no adversary can achieve $\inf_{f\in\{f_1,\ldots,f_L\}}\mathbb{P}\{|\widetilde{X} - X^*| \leq \delta/2\} > 1/L$.

When $\delta > 2^{-L}$, the learner again submits L guesses in the first stage, totaling up to $2L$ queries. In the second stage, the length of I is at most 2δ. Thus bisection on I takes at most $\log(\delta/\epsilon) + 1$ queries. Note that the first guess at 0 always contains a trivial query at 0. Removing the trivial query yields a total query complexity of $\log(\delta/\epsilon) + 2L$. To prove (δ, L)-privacy, note that for if f^* is minimized in one of the L intervals $[0, \epsilon]$, $[1-2^{-i}, 1-2^{-i}+\epsilon]$ for $i \leq K$, or $[1-2^{-K}+i\ell_K, 1-2^{-K}+i\ell_K+\epsilon]$ for $i \leq L-K-1$, then they induce the same query sequence. This completes the proof of the upper bound. □

We now turn to the lower bound. As a first step, we prove that if $\mathcal{F}$ satisfies Assumption 1, then the learner cannot search faster than the bisection method on any interval $I \subset [0,1]$. The lemma below contains a formal statement of this claim. Note that by taking $I = [0,1]$, Lemma 1 immediately implies a lower bound of $\log(1/\epsilon)$ on the optimal query complexity.

Algorithm 1 Querying Strategy Under the Maximin Setting

1: Let $I = [0,1]$.
2: **if** $\delta \leq 2^{-L}$ **then**
3: Submit the first guess at $1/2$.
4: Recursively submit the remaining $L-1$ guesses via bisection: if none of the submitted guesses is correct, update $I = [a,b]$ according the gradient $(f^*)'(q)$ at the previous guess q. If $(f^*)'(q) \leq 0$, then $X^* \geq q$, so we let the updated I be $[q,b]$; otherwise update I to be $[a,q]$. Submit the next guess at the midpoint of the updated I.
5: Once a guess is found to be correct, always (do this also for all the remaining guesses) update I to be the right half of I, and submit the next guess at the midpoint of the updated I.
6: **else**
7: Submit the first guess at 0.
8: Let K be an integer solution in $\{0, 1, \ldots, L-1\}$ such that $\ell_K := 2^{-K}/(L-K) \in [\delta, 2\delta]$. When $\delta > 2^{-L}$, a solution always exists.
9: Submit the next K guesses via bisection. Update I accordingly. As in the $\delta \leq 2^{-L}$ case, once any guess is found to be correct, always update I to its right half.
10: Divide I into $L-K$ equal length subintervals. Submit the next $L-K-1$ guesses at the endpoints of the subintervals (excluding the 2 endpoints of I).
11: **if** none of the guesses is correct **then**
12: Run bisection search on I until reaching ϵ-accuracy.
13: **else**
14: If the total number of queries is below $n = \max\{2L + \log(\delta/\epsilon), L+\log(1/\epsilon)\}$, fill the remaining query sequence with trivial queries at 1.

Lemma 1: Suppose $\mathcal{F}$ satisfies Assumption 1. Let ϕ be an ϵ-accurate querying strategy. Then for each $f \in \mathcal{F}$, each interval $I \subset [0,1]$ that contains the minimizer of f, and each realization of the random seed y, there exists $\widetilde{f} \in \mathcal{F}$, such that

(1) under ϕ, the query sequence $\mathsf{q}(\widetilde{f}, y)$ contains at least $\log(|I|/\epsilon)$ queries in I;
(2) the gradient of $\widetilde{f}$ and f coincide outside of I.

Next, we prove the lower bound in Theorem 1 assuming correctness of Lemma 1. The proof of Lemma 1 is deferred to the end of this subsection.

Proof of the Lower Bound in Theorem 1: A key step in this proof is to connect definition of (δ, L)-privacy with the covering numbers of the information sets. We claim that for a strategy to be (δ, L)-private in the maximin sense, there must be one information set with a large covering number.

Let ϕ be a querying strategy that is both ϵ-accurate and (δ, L)-private. Define the information set of a query sequence q as

$$\begin{aligned}\mathcal{I}(q) = \{x \in [0,1] : \exists f \in \mathcal{F} \text{ and } y,\\ s.t.\ \ x = \arg\min f,\ \text{and } \mathsf{q}(f,y) = q\}.\end{aligned}$$

Denote the $\delta/2$-covering number of $\mathcal{I}(q)$ as $N_c(\mathcal{I}(q), \delta/2)$. Fix the adversary's strategy to be one that samples uniformly from a δ-covering set of $\mathcal{I}(q)$. Since ϕ is (δ, L)-private, there must exist some f minimized at x, for which

$$\begin{aligned} 1/L >& \mathbb{P}\left\{\left|\widetilde{X} - x\right| \le \delta/2\right\} \\ =& \mathbb{E}\left[\mathbb{P}\left\{\left|\widetilde{X} - x\right| \le \delta/2 \;\middle|\; q\right\}\right], \end{aligned}$$

where the first integration is over q and the second is over the randomness from the adversary's estimation scheme conditional on q. Since x is in $\mathcal{I}(q)$, it must be $\delta/2$-close to at least one of the points in the covering set. Therefore for all q,

$$\mathbb{P}\left\{\left|\widetilde{X} - x\right| \le \delta/2 \;\middle|\; q\right\} \ge \frac{1}{N_c(\mathcal{I}(q), \delta/2)}.$$

Taking expected value over q on both sides, we have $\mathbb{E}(1/N_c(\mathcal{I}(q), \delta/2)) < 1/L$. Hence there must exist some query sequence $\bar{q}$ for which $N_c(\mathcal{I}(\bar{q}), \delta/2) > L$. As a result, $\mathcal{I}(\bar{q})$ contains L points $x_1, \ldots, x_L$ that are at least $\delta/2$-apart.

By definition of the information set, there exist $f_1, \ldots, f_L \in \mathcal{F}$ and $y_1, \ldots, y_L \in [0, 1]$, such that f_i is minimized at x_i, and $\mathsf{q}(f_i, y_i) = \bar{q}$ for all i. Notice that for each i, $\bar{q}$ must contain a pair of queries at most ϵ-apart that sandwiches x_i. Otherwise suppose the closest pair of queries in $\bar{q}$ that contains x_i forms an interval I of size larger than ϵ. Under Assumption 1, for each $x \in I$, there exists $f \in \mathcal{F}$ for which f is minimized at x and $\mathsf{q}(f, y_i)$ is also $\bar{q}$. By taking x to be arbitrarily close to the endpoints of I, the ϵ-accuracy requirement is violated since no estimator $\widehat{X}$ can ensure $|\widehat{X} - x| \le \epsilon/2$ for all $x \in I$. Therefore, the length of I is at most ϵ. Combined with the fact that $x_1, \ldots, x_L$ are δ-separated, and the assumption $\delta \ge 2\epsilon$, we have shown that $\bar{q}$ contains L pairs of distinct queries. Thus the optimal query complexity is lower bounded by $2L$.

To improve the lower bound to the desired $2L + \log(\delta/\epsilon)$, we would like to argue that aside from the L pairs queries in $\bar{q}$, the learner must submit enough queries elsewhere to search for X^* in order to fulfill the accuracy requirement. Indeed, the worst-case query complexity is lower bounded by $\log(1/\epsilon)$ for any strategy that is ϵ-accurate. However, the worst-case instance may not be one of $f_1, \ldots, f_L$. To combine the $2L$ queries used to ensure privacy with the queries used to ensure accuracy therefore becomes the main challenge of the lower bound proof. To address this difficulty, we will again utilize Assumption 1 on the richness of $\mathcal{F}$. On a high level, Assumption 1 allows us to find a large class of functions in $\mathcal{F}$ which can also lead to the query sequence $\bar{q}$. Out of these functions, we show that for at least one of them it takes $\log(\delta/\epsilon)$ extra queries to search for its minimizer. Next we give the rigorous proof of the existence of such a function.

Firstly, note that $\bar{q}$ contains L pairs of ϵ-close queries that sandwich $x_1, \ldots, x_L$. Since $\delta \ge \epsilon$, we have that for all i, $\bar{q}$ contains at least one query in $[x_i - \delta/2]$, and one query in $[x_i + \delta/2]$. Once at least one query has appeared in each of $[x_i - \delta/2, x_i]$ and $[x_i, x_i + \delta/2]$, we say x_i is "$\delta/2$-localized". Let x_j be the last one to be $\delta/2$-localized out of $x_1, \ldots, x_L$, and suppose it is $\delta/2$-localized at time T. Without loss of generality, assume a query in $[x_j - \delta/2, x_j]$ appears first, so that $\bar{q}_T \in [x_j, x_j + \delta/2]$. Let $I = [a, b]$ with a defined as the query in $\bar{q}_1, \ldots, \bar{q}_T$ to the left of x_j that is the closest to x_j, and $b = x_j + \delta/2$. See Figure 3 for an illustration.

Fig. 3. An illustration of the lower bound argument with $L = 3$. The ticks represent all queries in $\bar{q}$. The L pairs of ϵ-close queries that sandwich $x_1, \ldots, x_L$ are colored red. Suppose x_2 is the last one out of $x_1, \ldots, x_L$ to be $\delta/2$-localized, and the query in $[x_2 - \delta/2, x_2]$ appears before the one in $[x_2, x_2 + \delta/2]$, then I is defined as the shaded interval. Note that until all of $x_1, \ldots, x_L$ are $\delta/2$-localized, no query is submitted in I.

Apply Lemma 1 with $I = [a, b]$, $f = f_j$ and $y = y_j$. We can find some $\widetilde{f} \in \mathcal{F}$ that satisfies the two criteria in the statement of Lemma 1. Criterion (2) ensures that the gradient of $\widetilde{f}$ and f_j coincide outside of I. Since all functions in $\mathcal{F}$ are assumed to be differentiable, any query outside of I leads to identical responses for the two functions $\widetilde{f}$ and f_j. Since x_j is $\delta/2$-localized at time T, $\bar{q}_1, \ldots, \bar{q}_{T-1}$ do not contain any queries between a and b. Thus $\mathsf{q}(\widetilde{f}, y_j)$ and $\mathsf{q}(f_j, y_j) = \bar{q}$ agree completely up to time $T - 1$, and contain at least the $2L - 1$ queries outside of I used to sandwich $x_1, \ldots, x_L$. The reason we need to subtract 1 is because the T'th queries in $\bar{q}$ is in I.

By criterion (1) in the statement of Lemma 1, $\mathsf{q}(\widetilde{f}, y_j)$ contains at least $\log(|I|/\epsilon) \ge \log(\delta/(2\epsilon))$ queries in I. Combined with the $2L - 1$ queries outside of I, we arrive at the desired lower bound $2L + \log(\delta/\epsilon) - 2$. □

Proof of Lemma 1: The lemma is proved by constructing an $\widetilde{f}$ that satisfies both criteria. Our construction scheme is inspired by that of Nemirovski's (See Lemma 1.1.1 in lecture notes by Nemirovski [45]). With the querying strategy ϕ fixed, we will construct a sequence of functions $(g_1, \ldots, g_K)$ in $\mathcal{F}$ adapted to the queries and the responses and set $\widetilde{f} = g_K$. The value of K will be specified later. The construction ensures that for each $i \ge 0$, there is an interval $\Delta_i \subset I$ with $|\Delta_i| \ge |I|/2^i$, such that

1) g_i is minimized at the midpoint of Δ_i;
2) in the query sequence $\mathsf{q}(g_i, y)$, the first i queries in I are outside of Δ_i.

By Assumption 1, there exists a function in $\mathcal{F}$ whose gradient of f agrees with that of f outside of I, and is minimized at the midpoint of I. Let this function be g_0 and let $\Delta_0 = I$.

Inductively construct the rest of $\{g_i\}$. Given $g_0, \ldots, g_i$, by the induction hypothesis in $\mathsf{q}(g_i, y)$, the first i queries in I are all outside of $\Delta_i = [a_i, b_i]$. Let q be the $(i+1)$'th query of $\mathsf{q}(g_i, y)$ in I. If q is not in Δ_i, then we can simply let $g_{i+1} = g_i$ and $\Delta_{i+1} = \Delta_i$ to complete the $(i+1)$'th step of the induction. If $q \in \Delta_i$, depending on whether q lands to the left or right of the midpoint of Δ_i, let Δ_{i+1} be either $[q, b_i]$ or $[a_i, q]$, so that $|\Delta_{i+1}| \ge |\Delta_i|/2$. Let $g_{i+1} \in \mathcal{F}$ be a function whose gradient agrees with g_i outside of Δ_i, and is minimized at the midpoint of Δ_{i+1}. By Assumption 1 such a g_{i+1} always exists.

The construction can be carried out until for some integer K, we cannot find the $(K+1)$'th query of $\mathsf{q}(g_K, y)$ in I.

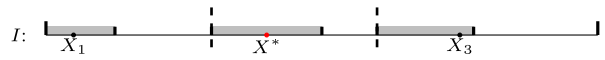

Fig. 4. Example of phases 2 to 4 of the querying strategy under the Bayesian setting with $L = 3$. In phase 2, the learner queries the $1/3$ and $2/3$ quantile of ν_I (represented by the dashed lines), and learns that $X^* \in I_2$. In phase 3, she queries the medians $m_1, \dots m_L$, and learns that X^* is to the left of m_2. Therefore $J_1, \dots, J_L$ are defined to be the shaded intervals. In phase 4, X_1 and X_3 are sampled from ν_{J_1} and ν_{J_3} respectively and X_2 is defined to be X^*. Note that the separation of $X_1, \dots, X_L$ are guaranteed by the separation of $J_1, \dots, J_L$.

That is, $\mathsf{q}(g_K, y)$ contains only K queries in I. By construction, $\mathsf{q}(g_K, y)$ does not contain any queries in Δ_K. Therefore under Assumption 1, the learner cannot rule out any member of Δ_K being X^*. For the strategy to be ϵ-accurate, we must have $|\Delta_K| < \epsilon$; hence $K > \log(|I|/\epsilon)$. Taking $\widetilde{f} = g_K$ finishes the proof of the lemma. □

C. Proof Under the Bayesian Setting

Proof of the Upper Bound in Theorem 2: Let ν denote the distribution of X^*. For an interval $I \subset [0,1]$, write ν_I for the probability distribution of ν conditioned on I, *i.e.*, $\frac{d\nu_I}{d\nu}(x) = \mathbb{1}\{x \in I\}/\nu(I)$. We design the following multi-phase querying strategy to attain the desired upper bound.

Algorithm 2 Querying Strategy Under the Bayesian Setting

1: Recursively query the median of the posterior distribution of X^*, until it is supported on an interval I with $\nu(I) \in [2\delta L H_\alpha, 4\delta L H_\alpha]$.
2: Let κ_j be the j/L quantile of ν_I for $j = 0, 1, \dots, L$ and let $I_j = [\kappa_{j-1}, \kappa_j]$ for $j \in [L]$. Query $\kappa_1, \dots, \kappa_{L-1}$ and identify j^* for $f'(\kappa_{j^*-1}) \le 0$ and $f'(\kappa_{j^*}) > 0$ so that I_{j^*} contains X^*.
3: Query the median m_j of ν_{I_j} for $j \in [L]$. If $f'(m_{j^*}) > 0$, let $J_j = [\kappa_{j-1}, m_j]$ for all j; otherwise let $J_j = [m_j, \kappa_j]$.
4: For all $j \neq j^*$, sample $X_j \sim \nu_{J_j}$ independently. Denote $X_{j^*} = X^*$. For $j = 1, \dots, L$, run the regular bisection search on J_j to locate X_j up to ϵ-accuracy.

Phase 1 runs the median-based bisection search, which is equivalent to the regular bisection search on $U = F_\nu(X^*) \sim \text{Unif}[0,1]$, where F_ν is the CDF of ν. Note that this step is always possible under the assumption $2\delta L H_\alpha \le 1$. Phase 2 divides I into L subintervals $I_1, \dots, I_L$ with equal ν-probability and determines I_{j^*} containing X^*. Phase 3 is the key to ensure adequate separation between the subintervals $\{J_j\}_{j \in [L]}$. Phase 4 serves to achieve the ϵ-accuracy while obfuscating the adversary. See Figure 4 for an illustration of phases 2 to 4.

The querying strategy outlined in Algorithm 2 is clearly ϵ-accurate by design. We now show that it is also (δ, L)-private. The high-level proof idea is to consider an adversary who has access to $X_1, \dots, X_L$. Using a genie-aided argument, we argue that this adversary is stronger than the one who only has access to the query sequence. We then establish that the conditional distribution of X^* given $X_1, \dots, X_L$ is uniform on the X_j's. Moreover, phase 3 of the querying strategy ensures that the X_j's are all δ-separated. Therefore even with the additional knowledge of $X_1, \dots, X_L$, the adversary cannot estimate X^* accurately with probability higher than $1/L$.

Proof of Privacy: Since the adversary only has access to the query sequence q, any adversary's estimator $\widetilde{X}$ must be a (random) function of q, that is $\widetilde{X} \equiv \widetilde{X}(q)$. Meanwhile by the design of our querying strategy, q can be completely reconstructed from $X_1, \dots, X_L$. To see that, note that $I, \{I_j\}, \{J_j\}$ and all the queries in phase 4 are deterministic functions of $X_1, \dots, X_L$. Therefore there is a mapping $\widetilde{\psi}$ such that $\widetilde{X}(q) = \widetilde{\psi}(X_1, \dots, X_L)$. Thus,

$$\begin{aligned}
&\mathbb{P}\left\{\left|\widetilde{X} - X^*\right| \le \frac{\delta}{2}\right\} \qquad (10)\\
=&\mathbb{E}\left[\mathbb{P}\left\{\left|\widetilde{X}(q) - X^*\right| \le \frac{\delta}{2} \;\middle|\; q\right\}\right]\\
\le&\mathbb{E}\left[\sup_{\widetilde{\psi}} \mathbb{P}\left\{\left|\widetilde{\psi}(X_1, \dots, X_L) - X^*\right| \le \frac{\delta}{2} \;\middle|\; X_1, \dots, X_L\right\}\right]\\
\le&\mathbb{E}\left[\sup_{\tilde{x} \in [0,1]} \mathbb{P}\left\{|\tilde{x} - X^*| \le \frac{\delta}{2} \;\middle|\; X_1, \dots, X_L\right\}\right]. \qquad (11)
\end{aligned}$$

We claim that

(i) $X^* \mid X_1, \dots, X_L \sim \text{Unif}\{X_1, \dots, X_L\}$.
(ii) With probability 1, $|X_i - X_j| > \delta$ for all $i \neq j$.

Assuming the two claims hold,

$$\begin{aligned}
&\sup_{\tilde{x} \in [0,1]} \mathbb{P}\left\{|\tilde{x} - X^*| \le \frac{\delta}{2} \;\middle|\; X_1, \dots, X_L\right\}\\
&= \sup_{\tilde{x} \in [0,1]} \frac{1}{L} \sum_{j \le L} \mathbb{1}\left\{|\tilde{x} - X_j| \le \frac{\delta}{2}\right\} \le \frac{1}{L},
\end{aligned}$$

where the equality is from (i) and the inequality is from (ii). Continuing (11), we have $\mathbb{P}\{|\widetilde{X} - X^*| \le \delta/2\} \le 1/L$. Thus our strategy is (δ, L)-private. It remains to prove claims (i), (ii).

Proof of (i): Recall that the index of the subinterval containing X^* is j^*. Since $\nu(I_j)$ are equal for all j, j^* is distributed uniformly in $\{1, \dots, L\}$. Therefore the desired claim $X^* \mid X_1, \dots, X_L \sim \text{Unif}\{X_1, \dots, X_L\}$ is equivalent to j^* and $(X_1, \dots, X_L)$ being independent.

To show $j^* \perp\!\!\!\perp (X_1, \dots, X_L)$, first note that $j^* \perp\!\!\!\perp (J_1, \dots, J_L)$, because conditional on j^*, either $J_j = [\kappa_{j-1}, m_j]$ for all j or $J_j = [m_j, \kappa_j]$ for all j, with equal probability. Second, for any fixed j', conditional on $j^* = j'$ and $(J_1, \dots, J_L)$, the X_j's are independent and X_j is distributed ν_{J_j}. Thus $j^* \perp\!\!\!\perp (X_1, \dots, X_L)$ given $J_1, \dots, J_L$. Combined with $j^* \perp\!\!\!\perp (J_1, \dots, J_L)$, we arrive at the conclusion $j^* \perp\!\!\!\perp (X_1, \dots, X_L)$.

Proof of (ii): It suffices to show that the intervals $J_1, \dots, J_L$ are δ-separated, or equivalently, $|I_j \backslash J_j| \ge \delta$ for all $j \le L$. Since phase 2 of the querying strategies queries all the medians of $I_1, \dots, I_L$, we have $\nu(I_j \backslash J_j) = \nu(I_j)/2 = \nu(I)/(2L) \ge \delta H_\alpha$. Let $\mathbf{m} = d\nu/d\lambda$ be the density of ν. Then

$$|I_j \backslash J_j| \ge \frac{\nu(I_j \backslash J_j)}{\sup_t \mathbf{m}(t)} \ge \frac{\delta H_\alpha}{\sup_t \mathbf{m}(t)}. \qquad (12)$$

To finish the proof of this claim, we only need to bound the density of ν from above. Recall that ν is the distribution of

X^*, which is the median of F. Thus the distribution function of ν has the form

$$\nu([0,t]) = \mathbb{P}\{X^* \le t\} = \mathbb{P}\{F(t) \ge 1/2\}.$$

Since $F \sim \mathrm{DP}(\alpha, \lambda_{[0,1]})$, we have $(F(t), 1-F(t)) \sim \mathrm{Dir}(\alpha t, \alpha(1-t))$. Therefore $F(t) \sim \mathrm{Beta}(\alpha t, \alpha(1-t))$. We will use the following Lemma 2 to bound the density of ν. The proof of Lemma 2 is deferred to the end of this subsection.

Lemma 2: Suppose $X \sim \mathrm{Beta}(\alpha t, \alpha(1-t))$ for some $\alpha > 0$, then for all $t \in (0,1)$,

$$h_\alpha \le \frac{d}{dt}\mathbb{P}\{X \ge 1/2\} \le H_\alpha,$$

where $h_\alpha = \frac{1}{3}2^{-\alpha-2}$ and $H_\alpha = (3+2e^{-1})\alpha + 14$.

By Lemma 2,

$$\mathbf{m}(t) = \frac{d}{dt}\mathbb{P}\{F(t) \ge 1/2\} \le H_\alpha, \tag{13}$$

for all $t \in [0,1]$. Combining (12) and (13) yields that

$$|I_j \backslash J_j| \ge \frac{\delta H_\alpha}{H_\alpha} \ge \delta.$$

We have shown that $\nu_{j_1}, \ldots \nu_{j_L}$ are continuous distributions supported on L intervals that are δ-separated from each other. Therefore $|X_i - X_j| > \delta$ for all $i \ne j$ with probability 1.

Query Complexity: The number of queries submitted in phase 1 is at most $\log(1/(2\delta L H_\alpha))$. Phase 2 and phase 3 involve $L-1$ and L queries respectively. The number of queries submitted in phase 4 equals

$$\begin{aligned}
&\sum_{j\le L}\left\lceil \log\frac{|J_j|}{\epsilon}\right\rceil \\
\le& L + \sum_{j\le L}\log\frac{|J_j|}{\epsilon} \\
=& L + \log\left(\prod_{j\le L}|J_j|\right) + L\log\frac{1}{\epsilon},
\end{aligned}$$

To bound the above, note that from Lemma 2 we have

$$\sum_{j\le L}|J_j| \le \frac{\nu(\cup_{j\le L}J_j)}{h_\alpha} \le \frac{2\delta L H_\alpha}{h_\alpha}.$$

Therefore $\prod_{j\le L}|J_j| \le (2\delta H_\alpha/h_\alpha)^L$. Thus the total number of queries submitted by the learner is at most

$$\begin{aligned}
&\log\frac{1}{2\delta L H_\alpha} + (L-1) + L + L\left(\log\frac{\delta}{\epsilon} + \log\frac{4H_\alpha}{h_\alpha}\right) \\
=& L\left(\log\frac{\delta}{\epsilon} + \log\frac{16H_\alpha}{h_\alpha}\right) + \log\frac{1}{\delta L} + \log\frac{1}{4H_\alpha} \\
\le& L\left(\log\frac{\delta}{\epsilon} + c_2\right) + \log\frac{1}{\delta L}
\end{aligned}$$

for $c_2 = \log(16H_\alpha/h_\alpha)$. The inequality is from $H_\alpha > 14$ for all $\alpha > 0$. □

Proof of the Lower Bound in Theorem 2: Let ϕ be a querying strategy that is both ϵ-accurate and (δ, L)-private. By definition of (δ, L)-privacy, we must have for any adversary's estimator $\widetilde{X}$,

$$\frac{1}{L} \ge \mathbb{P}\left\{\widetilde{X} \in [X^* - \delta/2, X^* + \delta/2]\right\}.$$

For the purpose of the lower bound, we can assume without loss of generality that the learner always submits a fixed n number of queries under strategy ϕ. If the lengths of the query sequences $\mathsf{q}(f^*, Y)$ depend on f^* and Y, the learner can always fill the short sequences with $n - |\mathsf{q}(f^*, Y)|$ trivial queries at 0 without hurting the accuracy or the privacy of learning.

Next we complete the lower bound proof following the outline given in Section V-A.

Step 1: Quantify the learner's information using learner's intervals. Recall that the i'th learner's interval I_i denotes the smallest interval that the learner knows to contain X^*.

Step 2: Analyze the conditional distribution of X^* over the learner's interval. To find a "good" event $\mathcal{B}$ on which the conditional distribution is uniform, we heavily rely on the stick-breaking characterization of the Dirichlet Process. Namely, the event $\mathcal{B}$ is associated with the length of the longest stick in the stick-breaking process. For completeness, we shall include a brief description of the stick-breaking process here.

Given base distribution μ_0 and scaling parameter $\alpha > 0$, draw $\{X_k\}_{k=1}^\infty$ i.i.d. from μ_0, and independently draw $\{V_k\}_{k=1}^\infty$ i.i.d. from $\mathrm{Beta}(1,\alpha)$. From a stick of unit length, break off the first stick of length V_1; break off V_2 fraction of the remaining stick and repeat. In other words, denote by β_k the length of the k'th stick. We have

$$\beta_k = V_k \cdot \prod_{j\le k-1}(1-V_k)$$

and $\sum_{k=1}^\infty \beta_k = 1$. Let $\mu = \sum_{k\ge1}\beta_k\delta_{X_k}$ be the discrete distribution supported on $\{X_k\}_{k=1}^\infty$, where δ_{X_k} denotes the point mass distribution at X_k. Then the distribution μ which corresponds to a cumulative distribution function of F follows the Dirichlet process $\mathrm{DP}(\mu_0, \alpha)$.

Here is a heuristic argument on how the stick-breaking process helps us prove the uniformity of the conditional distribution of X^*. Under our prior construction, X^* is at the median of $F \sim \mathrm{DP}(\lambda_{[0,1]}, \alpha)$, where we recall that $\lambda_{[0,1]}$ is the Lebesgue measure on $[0,1]$. Therefore, X^* occurs at one of the stick-breaking locations X_k. Even though the X_k's are distributed *i.i.d.* uniformly in $[0,1]$, X^* itself does not follow the uniform distribution since the index i that corresponds to X^* is random. The key observation is that the conditional distribution of X^* is uniform conditional on the event $\mathcal{A}$ where the length of the longest stick is at least $1/2$. To prove uniformity, we first show that on the event $\mathcal{A}$, the median X^* must occur at the X_k that corresponds to the longest stick. Moreover, by independence of the stick lengths $\{\beta_k\}_{k\ge1}$ and the locations $\{X_k\}_{k\ge1}$, the distribution of the location corresponding to the longest stick is uniform in $[0,1]$. Furthermore, the posterior distribution of X^* remains uniform as queries are sequentially submitted. The following Lemma 3 contains the precise statement on uniformity.

Some notation is necessary before stating Lemma 3. Firstly, denote by $\beta_{(1)}, \beta_{(2)}, \ldots$ the order statistics of the lengths of

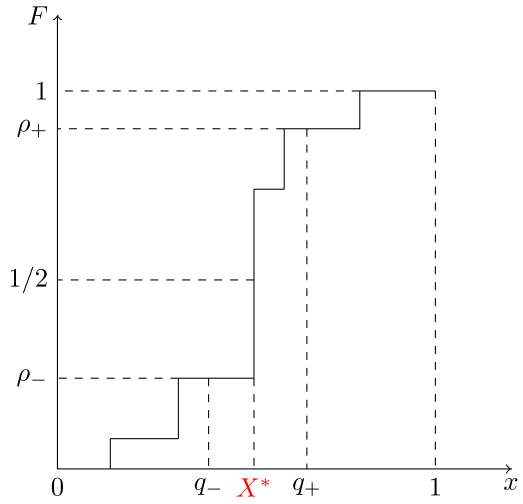

Fig. 5. An illustration of the quantities in (14). Conditional on $X^* \in J$ and the responses to the first i queries, the range of X^* is narrowed down to $I_i \cap J = [q_-, q_+]$. Further conditioning on $F(q_-) = \rho_-$ and $F(q_+) = \rho_+$, we show that F restricted to $[q_-, q_+]$ also follows a Dirichlet process after appropriate scaling.

the sticks in the stick-breaking process corresponding to F. Let

$$\mathcal{A} = \left\{\beta_{(1)} \geq 1/2\right\} = \cup_{z \geq 1/2} \mathcal{A}_z,$$

where $\mathcal{A}_z = \{\beta_{(1)} = z\}$. Let $J \subset [0,1]$ be an arbitrary fixed interval. Write $[q_-, q_+] = I_i \cap J$. Let the event $\mathcal{B} = \mathcal{B}(z, J, y, i, \rho^{(i)}, \rho_-, \rho_+)$ encode the random instances of F, Y and the first i responses, defined as

$$\begin{aligned}\mathcal{B} = \{&\mathcal{A}_z,\ X^* \in J,\ Y = y,\ r^{(i)} = \rho^{(i)},\\ &F(q_-) = \rho_-,\ F(q_+) = \rho_+\}.\end{aligned} \tag{14}$$

See Figure 5 for an example of F and the quantities in (14).

Lemma 3: For all $z \geq 1/2$, J, y, i, $\rho^{(i)}$, $\rho_- < 1/2$, $\rho_+ > 1/2$, we have for $\mathcal{B}$ defined in (14),

$$\mathcal{L}\left(X^* \mid \mathcal{B}\right) = \mathrm{Unif}[q_-, q_+],$$

where $\mathcal{L}(\cdot)$ denotes the (conditional) distribution.

The proof of Lemma 3 is deferred to the end of this subsection. It utilizes the self-similarity property of the Dirichlet process. See Section A in the appendix for a description and proof of the self-similarity property. In short, it ensures that the values of F inside of $[q_-, q_+]$ conditional on information outside of $[q_-, q_+]$ also follows a scaled Dirichlet process. Thus the learner cannot gain too much information about the location of X^* in $[q_-, q_+]$.

Step 3: Control the speed at which the learner's interval shrinks. Heuristically, since the conditional distribution of X^* stays uniform over the learner's interval in view of Lemma 3, the learner cannot search faster than the bisection method, and the learner's interval cannot shrink faster than $1/2$ each time a query is submitted.

Recall that $[0,1]$ is divided into $2/\delta$ subintervals $J_1, \ldots, J_{2/\delta}$ of length $\delta/2$, and J^* denotes the subinterval that contains X^*. In this step, by integrating over instances of $\mathcal{B}$, and letting J range over the $2/\delta$ subintervals, we prove the following lemma.

Lemma 4: For all i, we have that

$$\mathbb{E}\left(\log \frac{|I_{i+1} \cap J^*|}{|I_i \cap J^*|}\Big|\mathcal{A}\right) \geq -\mathbb{P}\left\{q_{i+1} \in J^* \mid \mathcal{A}\right\}. \tag{15}$$

The proof of Lemma 4 is deferred to the end of this subsection.

Step 4: In this step, we apply Lemma 4 to obtain the desired lower bound on the optimal query complexity. By writing $\log|I_n \cap J^*|$ as a telescoping sum, we have that

$$\begin{aligned}&\mathbb{E}\left(\log|I_n \cap J^*| \mid \mathcal{A}\right)\\ =&\log|I_0 \cap J^*| + \sum_{i=0}^{n-1}\mathbb{E}\left(\log\frac{|I_{i+1}\cap J^*|}{|I_i \cap J^*|}\Big|\mathcal{A}\right)\\ =&\log\frac{\delta}{2} + \sum_{i=0}^{n-1}\mathbb{E}\left(\log\frac{|I_{i+1}\cap J^*|}{|I_i \cap J^*|}\Big|\mathcal{A}\right)\\ \geq&\log\frac{\delta}{2} - \mathbb{E}\left(\text{number of queries in } J^* \mid \mathcal{A}\right).\end{aligned}$$

Therefore, on the one hand, by Jensen's inequality,

$$\begin{aligned}&\log\mathbb{E}(|I_n \cap J^*| \mid \mathcal{A})\\ \geq&\mathbb{E}\left(\log|I_n \cap J^*| \mid \mathcal{A}\right)\\ \geq&\log\frac{\delta}{2} - \mathbb{E}\left(\text{number of queries in } J^* \mid \mathcal{A}\right).\end{aligned} \tag{16}$$

On the other hand, from the accuracy requirement, we must have $|I_n| \leq \epsilon$ with probability 1. Therefore

$$\mathbb{E}\left(|I_n \cap J^*| \;\Big|\; \mathcal{A}\right) \leq \mathbb{E}\left(|I_n| \;\Big|\; \mathcal{A}\right) \leq \epsilon/2. \tag{17}$$

Combining (16), (17) yields

$$\mathbb{E}\left(\text{number of queries in } J^* \mid \mathcal{A}\right) \geq \log\frac{\delta}{\epsilon}. \tag{18}$$

Consider an adversary who adopts the *proportional-sampling* strategy [6]. That is, suppose the adversary's estimator $\widetilde{X}$ is sampled from the empirical distribution of the queries. For this particular $\widetilde{X}$,

$$\begin{aligned}&\mathbb{P}\left\{\widetilde{X} \in [X^* - \delta/2, X^* + \delta/2]\right\}\\ =&\frac{\mathbb{E}(\text{number of queries in } [X^* - \delta/2, X^* + \delta/2])}{n},\end{aligned}$$

which gives a lower bound on the total number of queries:

$$n \geq L\mathbb{E}(\text{number of queries in } [X^* - \delta/2, X^* + \delta/2]). \tag{19}$$

Since $J^* \subset [X^* - \delta/2, X^* + \delta/2]$, it follows from (19) and (18) that

$$\begin{aligned}N(\epsilon, \delta, L) \geq&L\mathbb{E}(\text{number of queries in } J^*)\\ \geq&\mathbb{P}\left(\mathcal{A}\right) L\log\frac{\delta}{\epsilon}.\end{aligned} \tag{20}$$

We have thus arrived at the desired query complexity lower bound with

$$c_1 = \mathbb{P}(\mathcal{A}) = \mathbb{P}\left\{\beta_{(1)} > 1/2\right\} \geq \mathbb{P}\left\{\beta_1 > 1/2\right\},$$

where $\beta_1 \sim \mathrm{Beta}(1, \alpha)$ is the length of the first stick fom the stick-breaking characterization of the Dirichlet process. That completes the proof of the Bayesian lower bound. □

Proof of Lemma 3: Since the gradient of the convex function f^* is defined with $(f^*)' = \gamma_- + (\gamma - \gamma_-)F$, the minimizer of f^* is at the median of F, i.e.,

$$X^* = \inf\left\{x : F(x) \geq \frac{-\gamma_-}{\gamma_+ - \gamma_-} = \frac{1}{2}\right\}.$$

Under our prior construction, the distribution of F follows a Dirichlet process with the uniform base distribution on $[0,1]$ and scale parameter α. Therefore with probability 1, F is a distribution function with countably many points of discontinuity, which we will refer to as *jumps*. If we characterize F with the stick breaking process, then the locations of the jumps are at $X_1, X_2, \ldots$ where the X_k's are independently and uniformly distributed on $[0,1]$. The sizes of the jumps $\beta_1, \beta_2, \ldots$ correspond to the lengths of the sticks from the stick-breaking process. We have $\sum \beta_k = 1$, and the two sequences $\{X_k\}_{k\geq 1}$ and $\{\beta_k\}_{k\geq 1}$ are independent.

To proceed, we first show that if the size of the largest jumps is larger than $1/2$, then X^* must occur at the largest jump. That is,

$$\mathcal{A} \subset \cup_{k\geq 1}\left\{X^* = X_k, \beta_{(1)} = \beta_k\right\}. \tag{21}$$

To see why, recall that X^* is the median of F. Thus $F(X^*) \geq 1/2$ and $\sup_{x<X^*} F(x) \leq 1/2$. Suppose $\beta_{(1)} = \beta_k$. We consider two cases:

1) if $X^* < X_k$, then $F(X_k) \geq F(X^*) + \beta_{(1)} > 1$;
2) if, on the other hand, $X^* > X_k$, then $F(X_k) \leq \sup_{x<X^*} F(x) - \beta_{(1)} \leq 1/2 - \beta_{(1)} < 0$.

In neither case can F be a distribution function. Therefore we must have $X^* = X_k$ is the location of the largest jump.

For $z \geq 1/2$, conditional on $\mathcal{A}_z$ and $X^* \in [q_-, q_+]$, we know that X^* is at the largest jump in $[q_-, q_+]$. Moreover, since the learner would not have submitted any queries between q_- and q_+ at time i, the events conditioned on do not contain any information on the location of the largest jump. Therefore the conditional distribution of X^* is uniform. To prove the claim rigorously, we need to invoke the self-similarity property of the Dirichlet process.

Recall that F follows a Dirichlet Process is supported on $[0,1]$ with base distribution $\lambda_{[0,1]}$. The self-similarity property asserts that for any finite partition $0 = x_0 \leq x_1 \leq \ldots \leq x_{n-1} \leq x_n = 1$ of $[0,1]$, conditional on the realization of F on $x_1, \ldots, x_n$, the restriction of F onto each subinterval is also a Dirichlet process scaled. In particular, for each $j \leq n$, we have

$$\begin{aligned}&\mathcal{L}\left(\frac{[F]_{[x_j,x_{j+1}]} - t_j}{t_{j+1} - t_j} \;\middle|\; F(x_1) = t_1, \ldots, F(x_{n-1}) = t_{n-1}\right)\\ &= \mathrm{DP}\left(\lambda_{[x_j,x_{j+1}]}, \alpha\lambda\left([x_j, x_{j+1}]\right)\right),\end{aligned}$$

where $[F]_I$ denotes the function F restricted to interval I, λ_I denotes the uniform probability measure on I, and $\lambda(I)$ denotes the Lebesgue measure of I. This property follows from the definition of the Dirichlet process. See Section A in the appendix for a proof.

Importantly, the following is a direct consequence of the self-similarity property. For each interval $[a,b] \subset [0,1]$, conditional on the value of $F(a)$ and $F(b)$, the distribution of F restricted to $[a,b]$ is independent of the realization of F outside of $[a,b]$. As a result, for each interval $I \subset [0,1]$, given $X^* \in I$, the learner cannot gain any additional information on X^* without querying in I. This property ensures that the posterior distribution of X^* conditional on $\mathcal{A}$ and the responses is uniform between the two closest queries that sandwich X^*. Therefore, the learner cannot beat the bisection search on the event $\mathcal{A}$.

By definition of the learner's interval I_i, none of the first i queries $q_1, \ldots, q_i$ can be in $I_i \cap J = [q_-, q_+]$. Since X^* is determined by the values of F inside $[q_-, q_+]$, by the self-similarity property of the Dirichlet process, X^* is independent of the responses to the first i queries conditioning on the values of $F(q_-)$ and $F(q_+)$. Therefore the event $\{r^{(i)} = \rho^{(i)}\}$ can be dropped from $\mathcal{B}$ without changing the conditional distribution of X^*. The indicator $\mathbb{1}\{X^* \in J\}$ is completely determined by whether ρ_- and ρ_+ are above or below $1/2$; and the outside randomness Y is independent of F. Therefore we can drop both events $\{X^* \in J\}$ and $\{Y = y\}$, and obtain

$$\mathcal{L}\left(X^* \mid \mathcal{B}\right) = \mathcal{L}\left(X^* \mid \mathcal{A}_z, F(q_-) = \rho_-, F(q_+) = \rho_+\right).$$

By the self-similarity property of the Dirichlet process, given $F(q_-) = \rho_-$ and $F(q_+) = \rho_+$,the conditional distribution of $(F - \rho_-)/(\rho_+ - \rho_-)$ restricted to $[q_-, q_+]$ is also a Dirichlet process with the uniform base distribution on $[q_-, q_+]$ and scaling parameter $\alpha' = \alpha(q_+ - q_-)$. In other words, there exist ancillary random vectors $\{X'_k\}_{k\geq 1}$, $\{\beta'_k\}_{k\geq 1}$ generated from a stick-breaking process that characterize the distribution function

$$\widetilde{F} = (F - \rho_-)/(\rho_+ - \rho_-)$$

on $[q_-, q_+]$. In addition, $X'_k \overset{i.i.d.}{\sim} \mathrm{Unif}[q_-, q_+]$, and $(\{X'_k\}_{k\geq 1}, \{\beta'_k\}_{k\geq 1})$ is independent of $(F(q_-), F(q_+))$.

We claim that for all $z \geq 1/2$, the event $\mathcal{A}_z = \{\beta_{(1)} = z\}$ is equivalent to $\{\beta'_{(1)} = z/(\rho_+ - \rho_-)\}$. Suppose $\mathcal{A}_z$ holds, and say $\beta_{(1)} = \beta_j$. Then by (21), $X^* = X_j$. Thus $[q_-, q_+]$ contains the largest jump in F. Since $\widetilde{F}$ is a scaled version of F restricted to $[q_-, q_+]$, the largest jump of $\widetilde{F}$ must be of size $z/(\rho_+ - \rho_-)$. Conversely, if $\beta'_{(1)} = z/(\rho_+ - \rho_-)$, then F contains a jump of size z. When $z \geq 1/2$, this must be the largest jump in F, i.e. $\beta_{(1)} = z$.

Note that conditional on $\mathcal{A}_z$ for $z \geq 1/2$, X^* can be written as the location of the largest jump in $\widetilde{F}$. We have shown that X^* and $\mathcal{A}_z$ can both be expressed as functions that only depend on $\{X'_k, \beta'_k\}$. As a result,

$$\begin{aligned}&\mathcal{L}(X^* \mid \mathcal{A}_z, F(q_-) = \rho_-, F(q_+) = \rho_+)\\ =&\mathcal{L}\Big(\text{location of the largest jump in } \widetilde{F} \;\Big|\\ &\qquad \beta'_{(1)} = \frac{z}{\rho_+ - \rho_-}, F(q_-) = \rho_-, F(q_+) = \rho_+\Big)\\ \overset{(a)}{=}&\mathcal{L}\Big(\text{location of the largest jump in } \widetilde{F} \;\Big|\\ &\qquad \beta'_{(1)} = \frac{z}{\rho_+ - \rho_-}\Big)\\ \overset{(b)}{=}&\mathcal{L}\left(\text{location of the largest jump in } \widetilde{F}\right)\\ \overset{(c)}{=}&\mathcal{L}(X'_1) = \mathrm{Unif}[q_-, q_+],\end{aligned}$$

where (a) is from the independence between $(\{X'_k\}_{k\geq 1}, \{\beta'_k\}_{k\geq 1})$ and $(F(q_-), F(q_+))$; (b) holds because by the stick-breaking characterization of the Dirichlet process, the locations of the jumps $\{\beta_k\}_{k\geq 1}$ and the sizes of the jumps $\{X_k\}_{k\geq 1}$ are independent. More specifically, let j be the index of the largest jump, i.e., $\beta'_{(1)} = \beta'_j$. Then j is only a function of $\{\beta_k\}_{k\geq 1}$ and is therefore independent of $\{X'_k\}_{k\geq 1}$. We have X'_j is independent of $\{\beta'_k\}_{k\geq 1}$, thus we can drop the conditional event which only depends on $\{\beta'_k\}_{k\geq 1}$; (c) is again from the independence of j and $\{\beta'_k\}_{k\leq 1}$. Since $\{X'_k\}_{k\geq 1}$ are distributed *i.i.d.* Unif q_-, q_+, we have $\mathcal{L}(X'_j) = \mathcal{L}(X'_1) = \text{Unif}[q_-, q_+]$. □

Proof of Lemma 4: From Lemma 3, we have $\mathcal{L}(X^* \mid \mathcal{B}) = \text{Unif}[I_i \cap J]$. We first claim that as a consequence,

$$\begin{aligned}&\mathbb{E}\left(\log \frac{|I_{i+1}\cap J|}{|I_i \cap J|}\Big|\mathcal{B}\right)\\ \geq &-\mathbb{1}\left\{q_{i+1} = \phi_i(\rho^{(i)}, y) \in I_i \cap J\right\}. \end{aligned} \tag{22}$$

The inequality (22) can be interpreted as follows. Firstly, the interval $I_i \cap J^*$ is only shortened when querying within $I_i \cap J^*$. Secondly, conditional on all instances of the behavior of F outside of $I_i \cap J^*$, on average, no query can reduce the length of $I_i \cap J^*$ by more than a half.

By taking the union of the events $\mathcal{B}$ over all the variables $z > 1/2$, $y \in [0,1]$, $\rho_- < 1/2$, $\rho_+ > 1/2$, $\rho^{(i)}$, and J ranging over $J_1, \ldots, J_{2/\delta}$, we arrive at the event $\mathcal{A}$. Therefore, integrating (22) over these variables yields that

$$\begin{aligned}&\mathbb{E}\left(\log \frac{|I_{i+1}\cap J^*|}{|I_i \cap J^*|}\,\Big|\,\mathcal{A}\right)\\ \geq &-\mathbb{P}\left\{q_{i+1} \in I_i \cap J^* \mid \mathcal{A}\right\}\\ \geq &-\mathbb{P}\left\{q_{i+1} \in J^* \mid \mathcal{A}\right\}.\end{aligned}$$

It remains to verify (22). If $q_{i+1} \notin I_i \cap J$, then $I_{i+1} \cap J = I_i \cap J$ and the claim (22) trivially holds. If $q_{i+1} \in I_i \cap J$, we have

$$\begin{aligned}&\log \frac{|I_{i+1}\cap J|}{|I_i \cap J|}\\ =&\mathbb{1}\{X^* \leq q_{i+1}\}\log\frac{q_{i+1}-q_-}{q_+ - q_-} +\\ &\mathbb{1}\{X^* > q_{i+1}\}\log\frac{q_+ - q_{i+1}}{q_+ - q_-}.\end{aligned}$$

Since the conditional distribution of X^* is uniform, we have

$$\begin{aligned}&\mathbb{E}\left(\log \frac{|I_{i+1}\cap J|}{|I_i \cap J|}\Big|\mathcal{B}\right)\\ \geq &\inf_{t\in[0,1]}[t\log t + (1-t)\log(1-t)] = -1.\end{aligned}$$

We have finished the proof of (22) and, by consequence, Lemma 4. □

VI. Extension to Multidimensions

In this section we extend our results under the maximin setting to optimization of convex separable functions in $\mathbb{R}^d$. Separable convex optimization arises in a variety applications such as inventory control in operation research, resource allocation in networking, and distributed optimization in multi-agent networks [46], [47], [48], when the global objection function is a sum of the local objective functions and each local objective function depends only on one component of the decision variable. Here, separability ensures that there is no cross-coordinate information leakage. Further generalizing our result to allow for general (non-separable) functions in $\mathbb{R}^d$ is left as future work.

Suppose the true function $f^* : [0,1]^d \to \mathbb{R}$ belongs to a family of convex separable functions

$$\mathcal{F} = \left\{f : f(x) = \sum_{i=1}^d f_i(x_i),\ f_i \in \mathcal{F}_i\right\},$$

where each $\mathcal{F}_i$ is a family of one-dimensional convex functions. For each query $q_j \in [0,1]^d$ submitted, the learner receives the gradient vector $\nabla f(q_j) = (f'_1(q_{j,1}), \ldots, f'_d(q_{j,d}))$ as the response. We say a querying strategy is ϵ-accurate if

$$\inf_{f^*\in\mathcal{F}} \mathbb{P}\left\{\left\|\widehat{X} - X^*\right\|_\infty \leq \epsilon/2\right\} = 1,$$

We say ϕ is (δ, L)-private if

$$\sup_{\widetilde{X}} \inf_{f^*\in\mathcal{F}} \mathbb{P}\left\{\left\|\widetilde{X} - X^*\right\|_\infty \leq \delta/2\right\} \leq 1/L.$$

In other words, we declare privacy breach if the adversary's estimator is within a $\delta/2$-neighborhood around the true minimizer with probability higher than $1/L$. As in the one-dimensional case, we need to impose some assumption on the complexity of the function class $\mathcal{F}$. Since $\mathcal{F}$ contains only separable functions, we can simply impose the one-dimensional assumption onto each of the d one-dimensional function classes $\mathcal{F}_1, \ldots, \mathcal{F}_d$. Below is the extension of our one-dimensional result to d dimensions.

Theorem 3: Let $N_d(\epsilon, \delta, L)$ denote the optimal query complexity in dimension d under the maximin setting. Suppose $\mathcal{F}_i$ all satisfy Assumption 1 for all $i = 1, \ldots, d$. If $2\epsilon \leq \delta \leq L^{-1/d}$, then

$$\begin{aligned}&2L^{1/d} + \log\frac{\delta}{\epsilon} - 2\\ \leq &N_d(\epsilon, \delta, L)\\ \leq &\begin{cases}2L^{1/d} + \log\frac{\delta}{\epsilon} & \text{if } L^{1/d} \geq \log\frac{1}{\delta}\\ L^{1/d} + \log\frac{1}{\epsilon} & \text{o.w.}\end{cases}.\end{aligned}$$

Remark 3: We choose to quantify the error of the learner and the adversary with respect to the $\|\cdot\|_\infty$ norm because $\|x - y\|_\infty \leq \epsilon/2$ is equivalent to $|x_i - y_i| \leq \epsilon/2$ for all $i \leq d$, so the analysis can be elegantly reduced to the one-dimensional case. However our result does not crucially depend on the choice of the norm. From the basic inequality $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{d}\|x\|_\infty$, we have that the optimal query complexity can differ by at most a d-dependent additive constant if the Euclidian norm were used instead.

Proof of the Upper Bound: Under the maximin privacy framework, to make a strategy private, we only need to find L functions $f^{(1)}, .., f^{(L)} \in \mathcal{F}$ whose minimizers are δ-apart, such that the query sequence for $f^{(1)}, \ldots, f^{(L)}$ are identical. That would ensure that the adversary who only observes

the query sequence cannot succeed with probability higher than $1/L$.

To construct such L functions, we design a querying strategy that submits $L^{1/d}$ *guesses* δ-apart along each dimension. To recap, in Section V-A we defined a guess at x to be a pair of ϵ-apart queries $(x, x+\epsilon)$. The guesses across the d dimensions intersect with each other in $[0,1]^d$ to create $(L^{1/d})^d = L$ cubes of diameter ϵ that potentially contain the minimizer of the true function f^*. The guesses are submitted following the same algorithm as in the one-dimensional case (see the upper bound proof of Theorem 1), except with L replaced by $L^{1/d}$.

Note that since each query is a d-dimensional vector and the function f^* is separable, we can run the search algorithms along the d directions in parallel. More concretely, write $f^*(x) = \sum_{i\le d} f_i^*(x_i)$, and let $q = (q_1, q_2, \ldots, q_n)$ be the query sequence where $q_j = (q_{j,1}, \ldots, q_{j,d}) \in [0,1]^d$. Each time the learner submits a query q_j, she receives the gradient vector

$$\nabla f^*(q_j) = ((f_1^*)'(q_{j,1}), \ldots, (f_d^*)'(q_{j,d}))\,.$$

For each dimension i, the learner leverages the gradient information $(f_i^*)'(q_{j,i})$ and constructs the next query $q_{j+1,i}$ in dimension i, as if she were learning the minimizer of f_i^* in one-dimension.

In particular, fix any dimension $1 \le i \le d$. The first $2L^{1/d}$ queries $q_{1,i}, \ldots, q_{2L^{1/d},i}$ consist of $L^{1/d}$ pairs of queries (guesses) that are δ-apart. When $\delta \le 2^{-L^{1/d}}$, these guesses are submitted along the bisection search path:

1) The first guess is at $1/2$, i.e., $q_{1,i} = 1/2$ and $q_{2,i} = 1/2+\epsilon$. The learner's interval I is initialized to be $[0,1]$.
2) For each $1 \le j \le L^{1/d}-1$, submit the $(j+1)$'th guess at follows: if none of the previous guesses is correct, then inspect the gradient $(f_i^*)'(q_{2j-1,i})$ from the j'th guess to deduce which half of I contains the minimizer X_i^* of f_i^*. Update the learner's interval I accordingly so that it contains X_i^*. Submit the $(j+1)$'th guess at the midpoint of the updated I. If one of the first j guesses is correct, then update I to its right half, and submit the $(j+1)$'th guess at its midpoint.

When $\delta > 2^{-L^{1/d}}$, only the first K guesses are submitted along the bisection path, and the remaining $L^{1/d}-K$ guesses are submitted via a grid search on the interval I generated from the first K guesses. Here K is the largest integer for which all the guesses are δ-apart. Under the assumption $\delta \le L^{-1/d}$ such a K always exists.

After all the guesses are submitted, if none of the guesses is correct, the learner runs a simple bisection search on a $\max\{2^{-L^{1/d}}, \delta\}$-length interval until reaching ϵ-accuracy; otherwise the learner simply fills the remaining queries along this dimension with trivial queries $q_{i,j} = 1$ for all $j \ge 2L^{1/d}$. The total number of queries is exactly the desired upper bound $2L^{1/d} + \log(\max\{2^{-L^{1/d}}, \delta\}/\epsilon)$.

Next we show this querying strategy is (δ, L)-private. Here we give the proof in the $\delta \le 2^{-L^{1/d}}$ case. The proof for the $\delta > 2^{-L^{1/d}}$ case follows analogously. For each i, it is easy to see that if

$$X_i^* \in \cup_{j\le L^{1/d}}[1-2^{-j}, 1-2^{-j}+\epsilon]$$

then the queries along the i'th dimension would always be L guesses at $1/2, 3/4, \ldots, 1-2^{L^{1/d}}$, followed by trivial queries at 1. As a result, for all $f^* \in \mathcal{F}$ such that

$$X^* \in \prod_{i\le d}\left(\cup_{j\le L^{1/d}}[1-2^{-j}, 1-2^{-j}+\epsilon]\right) \triangleq J,$$

share the same query sequence. Clearly J contains $(L^{1/d})^d$ members that are separated by at least δ in $\|\cdot\|_\infty$ distance. Hence the strategy is (δ, L)-private. □

Proof of the Lower Bound: Let ϕ be a querying strategy that is ϵ-accurate and (δ, L)-private. Via the same argument in one-dimension, we can show that there is at least one query sequence q whose information set $\mathcal{I}(q)$ has a $\delta/2$-covering number at least L. For each $i = 1, \ldots, d$, let

$$\mathcal{I}_i(q) = \{x_i : x = (x_1, \ldots, x_i, \ldots, x_d) \in \mathcal{I}(q) \text{ for some } x \in [0,1]^d\}$$

be the projection of $\mathcal{I}(q)$ to dimension i. Then we have $\mathcal{I}(q) \subset \prod_{i\le d}\mathcal{I}_i(q)$, thus

$$\begin{aligned} L \le & N_c\left(\mathcal{I}(q), \delta/2, \|\cdot\|_\infty\right) \\ \le & N_c\left(\prod_{i\le d}\mathcal{I}_i(q), \delta/2, \|\cdot\|_\infty\right) \\ = & \prod_{i\le d} N_c\left(\mathcal{I}_i(q), \delta/2, |\cdot|\right). \end{aligned}$$

Therefore for at least one $i \le d$, we must have that the $\delta/2$-covering number of the projection $\mathcal{I}_i(q)$ is no less than $L^{1/d}$. It follows that $\mathcal{I}_i(q)$ contains $x_i^{(1)}, \ldots, x_i^{(L^{1/d})}$ that are at least $\delta/2$-apart. For the strategy to be ϵ-accurate, the queries in q along this dimension i must contain at least $L^{1/d}$ pairs of ϵ-apart queries sandwiching $x_i^{(1)}, \ldots, x_i^{(L^{1/d})}$. The rest of the proof exactly follows the one-dimensional case. □

APPENDIX A
SELF-SIMILARITY PROPERTY OF THE DIRICHLET PROCESS

Proposition 1: Let μ be a random probability measure on $\mathcal{X}$ that follows a Dirichlet Process with base distribution function μ_0 and concentration parameter α. Let $\mathcal{X} = \cup_{i\le n} B_i$ be an arbitrary finite partition of $\mathcal{X}$. Then for all $i \le n$, we have

$$\mu_{B_i} \mid \mu(B_1), \ldots, \mu(B_n) \sim \mathrm{DP}\left(\mu_{0,B_i}, \alpha\mu_0(B_i)\right),$$

where μ_{B_i} and μ_{0,B_i} denote the conditional probability measures of μ and μ_0 respectively, conditioned on B_i.

Proof: For simplicity we present the proof only for $i = 1$. The proof for general i is identical. Let $B_1 = \cup_{j\le m} A_j$ be an arbitrary finite partition of B_1. Then $(A_1, \ldots, A_m, B_2, \ldots, B_n)$ is a partition of $\mathcal{X}$. Therefore from the definition of the Dirichlet Process, we have

$$\begin{aligned} &(\mu(A_1), \ldots, \mu(A_m), \mu(B_2), \ldots, \mu(B_n)) \\ \sim &\mathrm{Dir}\left(\alpha\mu_0(A_1), \ldots, \alpha\mu_0(A_m), \alpha\mu_0(B_2), \ldots, \alpha\mu_0(B_n)\right). \end{aligned}$$

From the density function of the Dirichlet distribution, we can derive that

$$\begin{aligned}&\frac{(\mu(A_1),\ldots,\mu(A_m))}{1-\sum_{i\geq 2}\mu(B_i)}\;\Big|\;\mu(B_2),\ldots,\mu(B_n)\\ &\sim\text{Dir}(\alpha\mu_0(A_1),\ldots,\alpha\mu_0(A_m)).\end{aligned}$$

Again by definition of the Dirichlet Process, we have

$$\begin{aligned}&\mu_{B_1}\mid\mu(B_2),\ldots,\mu(B_n)\\ &\sim\text{DP}\left([\mu_0]_{B_1},\alpha\right)=\text{DP}\left(\mu_{0,B_1},\alpha\mu_0(B_1)\right),\end{aligned}$$

where $[\mu_0]_{B_1}$ denotes the measure μ_0 restricted to B_1, which is not necessarily a probability measure. □

Consider the special case where $\mathcal{X}=[0,1]$. As a corollary of Proposition 1, we have for any finite partition $0=x_0\leq x_1\leq\ldots\leq x_{n-1}\leq x_n=1$ of $[0,1]$,

$$\begin{aligned}&\mathcal{L}\left(\frac{[F]_{[x_i,x_{i+1}]}-t_i}{t_{i+1}-t_i}\;\Big|\;F(x_1)=t_1,\ldots,F(x_{n-1})=t_{n-1}\right)\\ &=\text{DP}\left(\mu_{0,[x_i,x_{i+1}]},\alpha\mu_0[x_i,x_{i+1}]\right).\end{aligned}$$

Appendix B
Proof of Lemma 2

In this section we prove the technical result Lemma 2 on the Beta distribution. The statement of Lemma 2 is repeated below.

Lemma 2: Suppose $X\sim\text{Beta}(\alpha t,\alpha(1-t))$ for some $\alpha>0$, then for all $t\in(0,1)$,

$$h_\alpha\leq\frac{d}{dt}\mathbb{P}\{X\geq 1/2\}\leq H_\alpha,$$

where $h_\alpha=\frac{1}{3}2^{-\alpha-2}$ and $H_\alpha=(3+2e^{-1})\alpha+14$.

Proof: We can assume WOLG that $t\in(0,1/2]$. That is because for $t>1/2$, $1-X\sim\text{Beta}(\alpha(1-t),\alpha t)$ and

$$\frac{d}{dt}\mathbb{P}\{X\geq 1/2\}=\frac{d}{d(1-t)}\mathbb{P}\{1-X\geq 1/2\}.$$

Let $\phi_t(x)=x^{\alpha t-1}(1-x)^{\alpha(1-t)-1}$ be the unnormalized density of the $\text{Beta}(\alpha t,\alpha(1-t))$ distribution. Since $\frac{d}{dt}\phi_t(x)=\alpha\ln\frac{x}{1-x}\phi_t(x)$, we have

$$\begin{aligned}&\frac{d}{dt}\mathbb{P}\{X\geq 1/2\}\\ =&\frac{d}{dt}\frac{\int_{1/2}^1\phi_t(x)dx}{\int_0^1\phi_t(x)dx}\\ =&\alpha\Bigg[\int_{1/2}^1\ln\frac{x}{1-x}\phi_t(x)dx\int_0^1\phi_t(x)dx\\ &-\int_{1/2}^1\phi_t(x)dx\int_0^1\ln\frac{x}{1-x}\phi_t(x)dx\Bigg]\Bigg/\\ &\left(\int_0^1\phi_t(x)dx\right)^2\\ =&\alpha\Bigg[\mathbb{E}\left(\mathbb{1}\{X\geq 1/2\}\ln\frac{X}{1-X}\right)\\ &-\mathbb{P}\{X\geq 1/2\}\mathbb{E}\left(\ln\frac{X}{1-X}\right)\Bigg].\end{aligned}$$

To prove the lemma, we claim that for $t\leq 1/2$,

$$\begin{aligned}2^{-\alpha-2}t\leq&\alpha\mathbb{E}\left(\mathbb{1}\{X\geq 1/2\}\ln\frac{X}{1-X}\right)\\ \leq&\max\{3\alpha,12\};\\ &\left[2^{-\alpha-2}\left(\tfrac{1}{2}-\tfrac{t}{1-t}\right)\right]_+\\ \leq&-\alpha\mathbb{P}\{X\geq 1/2\}\mathbb{E}\left(\ln\frac{X}{1-X}\right)\\ \leq&2e^{-1}\alpha+2,\end{aligned}\tag{23}$$

$$(24)$$

where $[\cdot]_+=\max\{\cdot,0\}$ stands for the positive part.

The upper bound $\frac{d}{dt}\mathbb{P}\{X\geq 1/2\}\leq H_\alpha$ follows easily from adding up the two upper bounds. For the lower bound on the derivative, the two lower bounds in (23) and (24) yield

$$\begin{aligned}\frac{d}{dt}\mathbb{P}\{X\geq 1/2\}\geq&2^{-\alpha-2}\left(t+\left(\frac{1}{2}-\frac{t}{1-t}\right)_+\right)\\ \geq&\frac{1}{3}2^{-\alpha-2}=h_\alpha,\end{aligned}$$

where the last equality is achieved at $t=1/3$.

It remains to prove (23) and (24). Let us start from the cross-product term (23). Since $\mathbb{1}\{X\geq 1/2\}\ln\frac{X}{1-X}\geq 0$, by Tonelli's theorem,

$$\begin{aligned}&\mathbb{E}\left(\mathbb{1}\{X\geq 1/2\}\ln\frac{X}{1-X}\right)\\ =&\int_0^\infty\mathbb{P}\left\{\mathbb{1}\{X\geq 1/2\}\ln\frac{X}{1-X}>s\right\}ds\\ =&\int_0^\infty\mathbb{P}\left\{X\geq\frac{e^s}{1+e^s}\right\}ds.\end{aligned}$$

The density function of X allows us to write

$$\begin{aligned}&\mathbb{E}\left(\mathbb{1}\{X\geq 1/2\}\ln\frac{X}{1-X}\right)\\ =&\frac{\int_0^\infty\int_{\frac{e^s}{1+e^s}}^1 x^{\alpha t-1}(1-x)^{\alpha(1-t)-1}dxds}{B(\alpha t,\alpha(1-t))},\end{aligned}\tag{25}$$

where $B(\alpha,\beta)=\int_0^1 s^{\alpha-1}(1-s)^{\beta-1}ds$ is the Beta function. First we prove the upper bound in (23). For the numerator, since $\alpha t-1>-1$ and $x\geq\frac{e^s}{1+e^s}\geq 1/2$, we have $x^{\alpha t-1}\leq 2$, and

$$\begin{aligned}&\int_{\frac{e^s}{1+e^s}}^1 x^{\alpha t-1}(1-x)^{\alpha(1-t)-1}dx\\ \leq&2\int_{\frac{e^s}{1+e^s}}^1(1-x)^{\alpha(1-t)-1}dx\\ =&\frac{2(1+e^s)^{-\alpha(1-t)}}{\alpha(1-t)}.\end{aligned}$$

Therefore the numerator of (25) is upper bounded by

$$2\int_0^\infty\frac{e^{-\alpha(1-t)s}}{\alpha(1-t)}ds=\frac{2}{\alpha^2(1-t)^2}\leq\frac{8}{\alpha^2}$$

for all $t\leq 1/2$. Moreover,

$$B(\alpha t,\alpha(1-t))=\frac{\Gamma(\alpha t)\Gamma(\alpha(1-t))}{\Gamma(\alpha)}$$

is minimized at $t = 1/2$ by the log-convexity of the Gamma function $\Gamma(z)$ [49], where $\Gamma(z) = \int_0^\infty s^{z-1}e^{-s}ds$ satisifying $\Gamma(z+1) = z\Gamma(z)$ for $z > 0$. Hence it follows from (25) that for all $t \le 1/2$,

$$\alpha\mathbb{E}\left(\mathbb{1}\{X \ge 1/2\}\ln\frac{X}{1-X}\right) \le \frac{8\Gamma(\alpha)}{\alpha\Gamma(\alpha/2)^2}. \quad (26)$$

We claim that the right-hand side of (26) is a non-decreasing function in α on $(0,\infty)$. To see that, let $g(\alpha) = 8\Gamma(\alpha)/(\alpha\Gamma(\alpha/2)^2)$. We have

$$\begin{aligned}\frac{d}{d\alpha}(\ln g(\alpha)) =& \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \frac{1}{\alpha} - \frac{\Gamma'(\alpha/2)}{\Gamma(\alpha/2)}\\ =& \psi(\alpha) - \psi(\alpha/2) - \frac{1}{\alpha}. \end{aligned} \quad (27)$$

Here $\psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ is the digamma function with expansion [50, 6.3.16]

$$\psi(1+z) = -\gamma + \sum_{n=1}^{\infty}\frac{z}{n+z},$$

where γ is the Euler-Mascheroni constant. Applying the expansion on (27) yields

$$\begin{aligned}\frac{d}{d\alpha}(\ln g(\alpha)) =& \sum_{n=1}^{\infty}\left(\frac{\alpha-1}{n+\alpha-1} - \frac{\alpha/2-1}{n+\alpha/2-1}\right) - \frac{1}{\alpha}\\ \ge& \frac{\alpha-1}{1+\alpha-1} - \frac{\alpha/2-1}{1+\alpha/2-1} - \frac{1}{\alpha} = 0.\end{aligned}$$

We have shown that g is a non-decreasing function on $\mathbb{R}^+$. It follows from (26) that for all $\alpha \le 4$, $\alpha\mathbb{E}(\mathbb{1}\{X \ge 1/2\}\ln\frac{X}{1-X}) \le g(4) = 12$.

Next we show that for all $\alpha > 4$, the cross-product term in (23) is upper bounded by 3α. By Markov's inequality,

$$\begin{aligned}\mathbb{P}\left\{X \ge \frac{e^s}{1+e^s}\right\} =& \mathbb{P}\left\{1 - X \le \frac{1}{1+e^s}\right\}\\ =& \mathbb{P}\left\{\frac{1}{1-X} \ge 1 + e^s\right\}\\ \le& \frac{1}{1+e^s}\mathbb{E}\left[\frac{1}{1-X}\right].\end{aligned}$$

Since $1 - X \sim \mathrm{Beta}(\alpha(1-t), \alpha t)$, we have

$$\mathbb{E}\left[\frac{1}{1-X}\right] = \frac{\int_0^1 x^{\alpha(1-t)-2}(1-x)^{\alpha t-1}dx}{\int_0^1 x^{\alpha(1-t)-1}(1-x)^{\alpha t-1}dx}.$$

For all $\alpha \ge 4$ and $t \le 1/2$, $\alpha(1-t) - 1 \ge 0$, hence both integrals converge, and

$$\begin{aligned}\mathbb{E}\left[\frac{1}{1-X}\right] =& \frac{B(\alpha(1-t)-1, \alpha t)}{B(\alpha(1-t), \alpha t)}\\ =& \frac{\Gamma(\alpha(1-t)-1)\Gamma(\alpha t)/\Gamma(\alpha-1)}{\Gamma(\alpha(1-t))\Gamma(\alpha t)/\Gamma(\alpha)}\\ =& \frac{\alpha-1}{\alpha(1-t)-1} \le 3\end{aligned}$$

when $\alpha \ge 4$. Therefore

$$\begin{aligned}&\alpha\mathbb{E}\left(\mathbb{1}\{X \ge 1/2\}\ln\frac{X}{1-X}\right)\\ \le& 3\alpha\int_0^\infty \frac{1}{1+e^s}ds \le 3\alpha.\end{aligned}$$

That finishes the proof of the upper bound in (23). Next we prove the lower bound in (23). Since $x^{\alpha t-1} \ge \min\{(1/2)^{\alpha t-1}, 1\}$ for all $x \ge e^s/(1+e^s) \ge 1/2$, we have that the numerator in (25) is lower bounded by

$$\begin{aligned}&\min\left\{\left(\tfrac{1}{2}\right)^{\alpha t-1}, 1\right\}\int_0^\infty\int_{\frac{e^s}{1+e^s}}^1 (1-x)^{\alpha(1-t)-1}dxds\\ =& \frac{\min\left\{\left(\tfrac{1}{2}\right)^{\alpha t-1}, 1\right\}}{\alpha(1-t)}\int_0^\infty\left(\frac{1}{1+e^s}\right)^{\alpha(1-t)}ds\\ \ge& \frac{\min\left\{\left(\tfrac{1}{2}\right)^{\alpha t-1}, 1\right\}\left(\tfrac{1}{2}\right)^{\alpha(1-t)}}{\alpha(1-t)}\int_0^\infty e^{-s\alpha(1-t)}ds\\ =& \frac{\left(\tfrac{1}{2}\right)^{\max\{\alpha-1, \alpha(1-t)\}}}{\alpha^2(1-t)^2} \ge \frac{2^{-\alpha}}{\alpha^2}.\end{aligned} \quad (28)$$

To handle the denominator in (23), note that $(1-x)^{\alpha(1-t)-1} \le 2$ for all $x \le 1/2$ and $x^{\alpha t-1} \le 2$ for all $x \ge 1/2$. Therefore the denominator in (23)

$$\begin{aligned}&B(\alpha t, \alpha(1-t))\\ \le& 2\int_0^{1/2} x^{\alpha t-1}dx + 2\int_{1/2}^1 (1-x)^{\alpha(1-t)-1}dx\\ =& 2\left[\frac{2^{-\alpha t}}{\alpha t} + \frac{2^{-\alpha(1-t)}}{\alpha(1-t)}\right]\\ \le& \frac{2}{\alpha t(1-t)}.\end{aligned} \quad (29)$$

Combining (25), (28) and (29) yields

$$\begin{aligned}&\alpha\mathbb{E}\left(\mathbb{1}\{X \ge 1/2\}\ln\frac{X}{1-X}\right)\\ \ge& \alpha\frac{2^{-\alpha}\alpha t(1-t)}{2\alpha^2} \ge 2^{-\alpha-2}t.\end{aligned}$$

Next let us prove (24). Firstly, write

$$\begin{aligned}\mathbb{E}\left(\ln\frac{X}{1-X}\right) =& \psi(\alpha t) - \psi(\alpha) - (\psi(\alpha(1-t)) - \psi(\alpha))\\ =& \psi(\alpha t) - \psi(\alpha(1-t))\end{aligned}$$

where we recall that $\psi(z) = \frac{d}{dz}\ln\Gamma(z)$ is the digamma function. Since Γ is log-convex on $\mathbb{R}^+$, ψ is non-decreasing. Therefore for all $t \le 1/2$, we have

$$-\alpha\mathbb{P}\{X \ge 1/2\}\mathbb{E}\left(\ln\frac{X}{1-X}\right) \ge 0.$$

Furthermore, it has been shown in [51, Eq (2.2)] that for all $z > 0$, the digamma function satisfies

$$\frac{1}{2z} < \ln z - \psi(z) < \frac{1}{z}. \quad (30)$$

Therefore

$$\begin{aligned}&-\mathbb{E}\left(\ln\frac{X}{1-X}\right)\\ =& \psi(\alpha(1-t)) - \psi(\alpha t)\\ \ge& \ln(\alpha(1-t)) - \frac{1}{\alpha(1-t)} - \ln(\alpha t) + \frac{1}{2\alpha t}\end{aligned}$$

$$\geq \frac{1}{\alpha}\left(\frac{1}{2t}-\frac{1}{1-t}\right) \tag{31}$$

when $t \leq 1/2$.

We still need to bound $\mathbb{P}\{X \geq 1/2\}$ from below. As in the proof of (23), we can write

$$\mathbb{P}\left\{X \geq \tfrac{1}{2}\right\} = \frac{\int_{1/2}^{1} x^{\alpha t-1}(1-x)^{\alpha(1-t)-1}dx}{B(\alpha t, \alpha(1-t))}. \tag{32}$$

Again from $x^{\alpha t-1} \geq \min\{(1/2)^{\alpha t-1}, 1\}$ for all $x \geq 1/2$, we have that the numerator of (32) is bounded from below by

$$\max\left\{\left(\tfrac{1}{2}\right)^{\alpha t-1}, 1\right\}\int_{1/2}^{1}(1-x)^{\alpha(1-t)-1}dx \geq \frac{2^{-\alpha}}{\alpha}.$$

Combining the last displayed equation with (29) and (32) yields that

$$\mathbb{P}\left\{X \geq \tfrac{1}{2}\right\} \geq \frac{2^{-\alpha}}{\alpha} \times \frac{\alpha t(1-t)}{2} \geq 2^{-\alpha-2}t$$

for all $t \leq 1/2$. In view of (31), it follows that

$$-\alpha\mathbb{P}\left\{X \geq \tfrac{1}{2}\right\}\mathbb{E}\left(\ln\frac{X}{1-X}\right) \geq 2^{-\alpha-2}\left(\frac{1}{2}-\frac{t}{1-t}\right).$$

That concludes the proof of the lower bound in (24). Next we move to the upper bound in (24). By Markov's inequality,

$$\mathbb{P}\{X \geq 1/2\} \leq 2\mathbb{E}X = 2t. \tag{33}$$

Again from (30) we have that for all $t \leq 1/2$,

$$\begin{aligned}
&-\mathbb{E}\left(\ln\frac{X}{1-X}\right)\\
=&\psi(\alpha(1-t)) - \psi(\alpha t)\\
\leq&\ln(\alpha(1-t)) - \frac{1}{2\alpha(1-t)} - \left(\ln(\alpha t) - \frac{1}{\alpha t}\right)\\
=&\ln\frac{1-t}{t} + \frac{2-3t}{2\alpha t(1-t)}.
\end{aligned}$$

Combining the last displayed equation with (33) yields that

$$\begin{aligned}
&-\alpha\mathbb{P}\{X \geq 1/2\}\mathbb{E}\left(\ln\frac{X}{1-X}\right)\\
\leq&2\alpha t\left(\ln\frac{1-t}{t} + \frac{2-3t}{2\alpha t(1-t)}\right)\\
\leq&(2t\ln(1/t))\alpha + \frac{2-3t}{1-t} \leq 2e^{-1}\alpha + 2.
\end{aligned}$$

We have thus established the inequalities (23) and (24). □

ACKNOWLEDGMENT

The authors would like to thank Niva Ran and Benjamin Ran for inspiring the algorithm used in the upper bound of the Bayesian formulation of the problem and also would like to thank the Associate Editor and the anonymous reviewers for the comments and suggestions.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.
[2] B. McMahan and D. Ramage. (2017). *Federated Learning: Collaborative Machine Learning Without Centralized Training Data*. Accessed: Jul. 31, 2020. [Online]. Available: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html
[3] M. Juuti, S. Szyller, S. Marchal, and N. Asokan, "PRADA: Protecting against DNN model stealing attacks," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroSP)*, Jun. 2019, pp. 512–527.
[4] P. Kairouz *et al.*, "Advances and open problems in federated learning," 2019, *arXiv:1912.04977*.
[5] J. N. Tsitsiklis, K. Xu, and Z. Xu, "Private sequential learning," 2018, *arXiv:1805.02136*.
[6] K. Xu, "Query complexity of Bayesian private learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 2431–2440.
[7] J. Xu, K. Xu, and D. Yang, "Optimal query complexity for private sequential learning against eavesdropping," 2019, *arXiv:1909.09836*.
[8] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," 2020, *arXiv:2003.02133*.
[9] P. Subramanyan, R. Sinha, I. Lebedev, S. Devadas, and S. A. Seshia, "A formal foundation for secure remote execution of enclaves," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 2435–2450.
[10] P. Mohassel and Y. Zhang, "SecureML: A system for scalable privacy-preserving machine learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 19–38.
[11] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning," 2018, *arXiv:1812.00984*.
[12] S. Truex, L. Liu, K.-H. Chow, M. E. Gursoy, and W. Wei, "LDP-fed: Federated learning with local differential privacy," in *Proc. 3rd ACM Int. Workshop Edge Syst., Anal. Netw.*, Apr. 2020, pp. 61–66.
[13] P. Ramgopal, P. W. Laud, and A. F. M. Smith, "Nonparametric Bayesian bioassay with prior constraints on the shape of the potency curve," *Biometrika*, vol. 80, no. 3, pp. 489–498, 1993.
[14] B. Bornkamp and K. Ickstadt, "Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis," *Biometrics*, vol. 65, no. 1, pp. 198–205, Mar. 2009.
[15] M. Lavine and A. Mockus, "A nonparametric Bayes method for isotonic regression," *J. Stat. Planning Inference*, vol. 46, no. 2, pp. 235–248, Aug. 1995.
[16] M. Abadi, J. Feigenbaum, and J. Kilian, "On hiding information from an Oracle," *J. Comput. Syst. Sci.*, vol. 39, no. 1, pp. 21–50, Aug. 1989.
[17] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *Proc. IEEE 36th Annu. Found. Comput. Sci.*, Oct. 1995, pp. 41–50.
[18] W. Gasarch, "A survey on private information retrieval," *Bull. EATCS*, vol. 82, nos. 72–107, p. 113, 2004.
[19] M. Mirmohseni and M. A. Maddah-Ali, "Private function retrieval," in *Proc. Iran Workshop Commun. Inf. Theory (IWCIT)*, Apr. 2018, pp. 1–6.
[20] S. Song, K. Chaudhuri, and A. D. Sarwate, "Stochastic gradient descent with differentially private updates," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Dec. 2013, pp. 245–248.
[21] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.
[22] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. McMahan, "cpSGD: Communication-efficient and differentially-private distributed SGD," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7564–7575.
[23] P. Jain, P. Kothari, and A. Thakurta, "Differentially private online learning," in *Proc. Conf. Learn. Theory*, 2012, pp. 1–24.
[24] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 691–706.
[25] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*, M. Agrawal, D. Du, Z. Duan, and A. Li, Eds. Berlin, Germany: Springer, 2008, pp. 1–19.
[26] G. Fanti, P. Kairouz, S. Oh, and P. Viswanath, "Spy vs. Spy: Rumor source obfuscation," *SIGMETRICS Perform. Eval. Rev.*, vol. 43, no. 1, pp. 271–284, Jun. 2015.
[27] W. Luo, W. P. Tay, and M. Leng, "Infection spreading and source identification: A hide and seek game," *IEEE Trans. Signal Process.*, vol. 64, no. 16, pp. 4228–4243, Aug. 2016.
[28] J. N. Tsitsiklis and K. Xu, "Delay-predictability trade-offs in reaching a secret goal," *Oper. Res.*, vol. 66, no. 2, pp. 587–596, Apr. 2018.

[29] M. S. Erturk and K. Xu, "Anonymous stochastic routing," *Stochastic Syst.*, vol. 12, no. 1, pp. 92–126, Mar. 2022.
[30] W. Tang, W. Wang, G. Fanti, and S. Oh, "Privacy-utility tradeoffs in routing cryptocurrency over payment channel networks," in *Proc. Abstr. SIGMETRICS/Perform. Joint Int. Conf. Meas. Modeling Comput. Syst.*, Jun. 2020, pp. 81–82.
[31] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy, data processing inequalities, and statistical minimax rates," 2013, *arXiv:1302.3203*.
[32] K. Xu, "Query complexity of Bayesian private learning," 2019, *arXiv:1911.06903*.
[33] G. Aridor, Y. Mansour, A. Slivkins, and Z. S. Wu, "Competing bandits: The perils of exploration under competition," 2020, *arXiv:2007.10144*.
[34] O. Ben-Porat and M. Tennenholtz, "Best response regression," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1498–1507.
[35] O. Ben-Porat and M. Tennenholtz, "Regression equilibrium," in *Proc. ACM Conf. Econ. Comput.*, Jun. 2019, pp. 173–191.
[36] N. Immorlica, A. T. Kalai, B. Lucier, A. Moitra, A. Postlewaite, and M. Tennenholtz, "Dueling algorithms," in *Proc. 43rd Annu. ACM Symp. Theory Comput.*, 2011, pp. 215–224.
[37] Y. Mansour, A. Slivkins, and Z. S. Wu, "Competing bandits: Learning under competition," 2017, *arXiv:1702.08533*.
[38] Y. Ren, J. Duan, S. Eben Li, Y. Guan, and Q. Sun, "Improving generalization of reinforcement learning with minimax distributional soft actor-critic," 2020, *arXiv:2002.05502*.
[39] B. Neelon and D. B. Dunson, "Bayesian isotonic regression and trend analysis," *Biometrics*, vol. 60, no. 2, pp. 398–406, Jun. 2004.
[40] *Miles v. United States, 103 U.S. 304, 26 L. Ed. 481*, Supreme Court United States, Washington, DC, USA, 1881.
[41] *Coffin v. United States, 156 U.S. 432, No. 741*, Supreme Court United States, Washington, DC, USA, 1895.
[42] W. Young, N. Cameron, and Y. Tinsley, "Juries in criminal trials part 2," Section 56, New Zealand Law Commission, Wellington, New Zealand, Tech. Rep., 2001.
[43] A. Agarwal, D. P. Foster, D. Hsu, S. M. Kakade, and A. Rakhlin, "Stochastic convex optimization with bandit feedback," *SIAM J. Optim.*, vol. 23, no. 1, pp. 213–240, 2013.
[44] W. Tang, C.-J. Ho, and Y. Liu, "Optimal query complexity of secure stochastic convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9771–9781.
[45] A. Nemirovski, "Efficient methods in convex programming," Lect. Notes, 1994. [Online]. Available: https://www2.isye.gatech.edu/~nemirovs/Lect_EMCO.pdf
[46] A. Nedic *et al.*, "Cooperative distributed multi-agent optimization," in *Convex Optimization in Signal Processing and Communications*, Y. Eldar and D. Palomar, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[47] A. Padakandla and R. Sundaresan, "Separable convex optimization problems with linear ascending constraints," *SIAM J. Optim.*, vol. 20, no. 3, pp. 1185–1204, Jan. 2010.
[48] S. Boyd, N. Parikh, and E. Chu, *Distributed Optimization and Statistical Learning Via the Alternating Direction Method of Multipliers*. Norwell, MA, USA: Now Publishers, 2011.
[49] E. Artin, *The Gamma Function*. New York, NY, USA: Dover, 2015.
[50] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, vol. 55. Washington, DC, USA: U.S. Government Printing Office, 1948.
[51] H. Alzer, "On some inequalities for the gamma and PSI functions," *Math. Comput. Amer. Math. Soc.*, vol. 66, no. 217, pp. 373–389, 1997.

Jiaming Xu received the B.E. degree in electrical and computer engineering from Tsinghua University in 2009, the M.S. degree in electrical and computer engineering from UT-Austin in 2011, and the Ph.D. degree in electrical and computer engineering from UIUC in 2014. He is currently an Associate Professor with The Fuqua School of Business, Duke University. Before that, he was an Assistant Professor with the Krannert School of Management, Purdue University, from August 2016 to June 2018, a Research Fellow with the Simons Institute for the Theory of Computing, UC Berkeley, from January 2016 to June 2016, and a Post-Doctoral Fellow with the Statistics Department, The Wharton School, University of Pennsylvania, from January 2015 to December 2015. His research interests include data science, high-dimensional statistical inference, information theory, convex and non-convex optimization, queueing theory, and game theory. He was a recipient of Simons-Berkeley Fellowship in 2016 and NSF CAREER award in 2022.

Kuang Xu is currently an Associate Professor with the Stanford Graduate School of Business, Stanford University. His research focuses on fundamental principles for decision-making in a stochastic systems, with applications to operations, experimentation, machine learning, logistics, and privacy. He has received a First Place in the INFORMS George E. Nicholson Student Paper Competition, a Best Paper Award and Outstanding Student Paper Award at ACM SIGMETRICS, and an ACM SIGMETRICS Rising Star Research Award. He currently serves as an Associate Editor for *Operations Research and Management Science*.

Dana Yang received the B.S. degree in mathematics from Tsinghua University in 2013 and the M.A. degree in statistics and the Ph.D. degree in statistics and data science from Yale University in 2014 and 2019, respectively. She is currently an Assistant Professor at the Department of Statistics and Data Science, Cornell University. Before that, she was a Research Fellow with the Simons Institute for the Theory of Computing, UC Berkeley, from August 2021 to December 2021, and a Post-Doctoral Associate with The Fuqua School of Business, Duke University, from August 2019 to July 2021. Her research interests are in high-dimensional statistics and machine learning.