Differentially Private Bayesian Optimization

Matt J. Kusner MKUSNER@WUSTL.EDU
Jacob R. Gardner GARDNER.JAKE@WUSTL.EDU
Roman Garnett GARNETT@WUSTL.EDU
Kilian Q. Weinberger KILIAN@WUSTL.EDU
Washington University in St. Louis, 1 Brookings Dr., St. Louis, MO 63130

Abstract

Bayesian optimization is a powerful tool for fine-tuning the hyper-parameters of a wide variety of machine learning models. The success of machine learning has led practitioners in diverse real-world settings to learn classifiers for practical problems. As machine learning becomes commonplace, Bayesian optimization becomes an attractive method for practitioners to automate the process of classifier hyper-parameter tuning. A key observation is that the data used for tuning models in these settings is often sensitive. Certain data such as genetic predisposition, personal email statistics, and car accident history, if not properly private, may be at risk of being inferred from Bayesian optimization outputs. To address this, we introduce methods for releasing the best hyper-parameters and classifier accuracy privately. Leveraging the strong theoretical guarantees of differential privacy and known Bayesian optimization convergence bounds, we prove that under a GP assumption these private quantities are often near-optimal. Finally, even if this assumption is not satisfied, we can use different smoothness guarantees to protect privacy.

1. Introduction

Machine learning is increasingly used in application areas with sensitive data. For example, hospitals use machine learning to predict if a patient is likely to be readmitted soon (Yu et al., 2013), webmail providers classify spam emails from non-spam (Weinberger et al., 2009), and insurance providers forecast the extent of bodily injury in car crashes (Chong et al., 2005). In these scenarios data cannot be shared legally, but companies and hospitals may want to share hyper-parameters and validation accuracies through publications or other means. However, data-holders must be careful, as even a small amount of information can compromise privacy.

Which hyper-parameter setting yields the highest accuracy can reveal sensitive information about individuals in the validation or training data set, reminiscent of reconstruction attacks described by Dwork & Roth (2013) and Dinur & Nissim (2003). For example, imagine updated hyper-parameters are released right after a prominent public figure is admitted to a hospital. If a hyper-parameter is known to correlate strongly with a particular disease the patient is suspected to have, an attacker could make a direct correlation between the hyper-parameter value and the individual.

To prevent these sorts of attacks, we develop a set of algorithms that automatically fine-tune the hyper-parameters of a machine learning algorithm while provably preserving differential privacy (Dwork et al., 2006b). Our approach leverages recent results on Bayesian optimization (Snoek et al., 2012; Hutter et al., 2011; Bergstra & Bengio, 2012; Gardner et al., 2014), training a Gaussian process (GP) (Rasmussen & Williams, 2006) to accurately predict and maximize the validation gain of hyper-parameter settings. We show that the GP model in Bayesian optimization allows us to release noisy final hyper-parameter settings to protect against aforementioned privacy attacks, while only sacrificing a tiny, bounded amount of validation gain.

Our privacy guarantees hold for releasing the best hyper-parameters and best validation gain. Specifically our contributions are as follows: 1. We derive, to the best of our knowledge, the first framework for Bayesian optimization with differential privacy guarantees, with/without oberservation noise, 2. We show that even if our validation gain is not drawn from a Gaussian process, we can guarantee differential privacy under different smoothness assumptions.

We begin with background on Bayesian optimization and differential privacy we will use to prove our guarantees.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

2. Background

In general, our aim will be to protect the privacy of a validation dataset of sensitive records $\mathcal{V} \subseteq \mathcal{X}$ (where $\mathcal{X}$ is the collection of all possible records) when the results of Bayesian optimization depends on $\mathcal{V}$.

Bayesian optimization. Our goal is to maximize an unknown function $f_{\mathcal{V}}\colon \Lambda \rightarrow \mathbb{R}$ that depends on some validation dataset $\mathcal{V} \subseteq \mathcal{X}$:

$$\max_{\lambda \in \Lambda} f_{\mathcal{V}}(\lambda). \tag{1}$$

It is important to point out that all of our results hold for the general setting of eq. (1), but throughout the paper, we use the vocabulary of a common application: that of machine learning hyper-parameter tuning. In this case $f_{\mathcal{V}}(\lambda)$ is the gain of a learning algorithm evaluated on validation dataset $\mathcal{V}$ that was trained with hyper-parameters $\lambda \in \Lambda \subseteq \mathbb{R}^d$.

As evaluating $f_{\mathcal{V}}$ is expensive (e.g., each evaluation requires training a learning algorithm), Bayesian optimization gives a procedure for selecting a small number of locations to sample $f_{\mathcal{V}}$: $[\lambda_1, \ldots, \lambda_T] = \boldsymbol{\lambda}_T \in \mathbb{R}^{d \times T}$. Specifically, given a current sample λ_t, we observe a validation gain v_t such that $v_t = f_{\mathcal{V}}(\lambda_t) + \alpha_t$, where $\alpha_t \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise with possibly non-zero variance σ^2. Then, given v_t and previously observed values $v_1, \ldots, v_{t-1}$, Bayesian optimization updates its belief of $f_{\mathcal{V}}$ and samples a new hyper-parameter λ_{t+1}. Each step of the optimization proceeds in this way.

To decide which hyper-parameter to sample next, Bayesian optimization places a prior distribution over $f_{\mathcal{V}}$ and updates it after every (possibly noisy) function observation. One popular prior distribution over functions is the Gaussian process $\mathcal{GP}\big(\mu(\cdot), k(\cdot,\cdot)\big)$ (Rasmussen & Williams, 2006), parameterized by a mean function $\mu(\cdot)$ (we set $\mu = 0$, w.l.o.g.) and a kernel covariance function $k(\cdot,\cdot)$. Functions drawn from a Gaussian process have the property that any finite set of values of the function are normally distributed. Additionally, given samples $\boldsymbol{\lambda}_T = [\lambda_1, \ldots, \lambda_T]$ and observations $\mathbf{v}_T = [v_1, \ldots, v_T]$, the GP posterior mean and variance has a closed form:

$$\begin{aligned} \mu_T(\lambda) &= k(\lambda, \boldsymbol{\lambda}_T)(\mathbf{K}_T + \sigma^2 \mathbf{I})^{-1} \mathbf{v}_T \\ k_T(\lambda, \lambda') &= k(\lambda, \lambda') - k(\lambda, \boldsymbol{\lambda}_T)(\mathbf{K}_T + \sigma^2 \mathbf{I})^{-1} k(\boldsymbol{\lambda}_T, \lambda') \\ \sigma_T^2(\lambda) &= k_T(\lambda, \lambda), \end{aligned} \tag{2}$$

where $k(\lambda, \boldsymbol{\lambda}_T) \in \mathbb{R}^{1 \times T}$ is evaluated element-wise on each of the T columns of $\boldsymbol{\lambda}_T$, and $\lambda \in \Lambda$ is any hyper-parameter. As well, $\mathbf{K}_T = k(\boldsymbol{\lambda}_T, \boldsymbol{\lambda}_T) \in \mathbb{R}^{T \times T}$ is evaluated on all hyper-parameter pairs from $\boldsymbol{\lambda}_T$. As more samples are observed, the mean function $\mu_T(\lambda)$ approaches $f_{\mathcal{V}}(\lambda)$.

One well-known method to select hyper-parameters λ maximizes the *upper-confidence bound* (UCB) of the posterior GP model of $f_{\mathcal{V}}$ (Auer et al., 2002; Srinivas et al., 2010):

$$\lambda_{t+1} \triangleq \arg\max_{\lambda \in \Lambda} \mu_t(\lambda) + \sqrt{\beta_{t+1}} \sigma_t(\lambda), \tag{3}$$

where β_{T+1} is a parameter that trades off the *exploitation* of maximizing $\mu_t(\lambda)$ and the *exploration* of maximizing $\sigma_t(\lambda)$. Srinivas et al. (2010) proved that given certain assumptions on $f_{\mathcal{V}}$ and fixed, non-zero observation noise ($\sigma^2 > 0$), selecting hyper-parameters λ to maximize eq. (3) is a no-regret Bayesian optimization procedure: $\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} f_{\mathcal{V}}(\lambda^*) - f_{\mathcal{V}}(\lambda_t) = 0$, where $f_{\mathcal{V}}(\lambda^*)$ is the maximizer of eq. (1). For the no-noise setting, de Freitas et al. (2012) give a UCB-based no-regret algorithm.

Contributions. Alongside maximizing $f_{\mathcal{V}}$, we would like to guarantee that if $f_{\mathcal{V}}$ depends on (sensitive) validation data, we can release information about $f_{\mathcal{V}}$ so that the data $\mathcal{V}$ remains private. Specifically, we may wish to release (a) our best guess $\hat{\lambda} \triangleq \arg\max_{t \leq T} f_{\mathcal{V}}(\lambda_t)$ of the true (unknown) maximizer λ^* and (b) our best guess $f_{\mathcal{V}}(\hat{\lambda})$ of the true (also unknown) maximum objective $f_{\mathcal{V}}(\lambda^*)$. The primary question this work aims to answer is: how can we release private versions of $\hat{\lambda}$ and $f_{\mathcal{V}}(\hat{\lambda})$ that are close to their true values, or better, the values λ^* and $f_{\mathcal{V}}(\lambda^*)$? We give two answers to these questions. The first will make a Gaussian process assumption on $f_{\mathcal{V}}$, which we describe immediately below. The second, described in Section 5, will utilize Lipschitz and convexity assumptions to guarantee privacy in the event the GP assumption does not hold.

Setting. For our first answer to this question, let us define a Gaussian process over hyper-parameters $\lambda, \lambda' \in \Lambda$ *and* datasets $\mathcal{V}, \mathcal{V}' \subseteq \mathcal{X}$ as follows: $\mathcal{GP}\big(0, k_1(\mathcal{V}, \mathcal{V}') \otimes k_2(\lambda, \lambda')\big)$. A prior of this form is known as a multi-task Gaussian process (Bonilla et al., 2008). Many choices for k_1 and k_2 are possible. The function $k_1(\mathcal{V}, \mathcal{V}')$ defines a set kernel (e.g., a function of the number of records that differ between $\mathcal{V}$ and $\mathcal{V}'$). For k_2, we focus on either the squared exponential: $k_2(\lambda, \lambda') = \exp\big(-\|\lambda - \lambda'\|_2^2 / (2\ell^2)\big)$ or Matérn kernels: (e.g., for $\nu = 5/2$, $k_2(\lambda, \lambda') = (1 + \sqrt{5}r/\ell + (5r^2)/(3\ell^2)) \exp(-\sqrt{5}r/\ell)$, for $r = \|\lambda - \lambda'\|_2$), for a fixed ℓ, as they have known bounds on the maximum information gain (Srinivas et al., 2010). Note that as defined, the kernel k_2 is normalized (i.e., $k_2(\lambda, \lambda) = 1$).

Assumption 1. *We have a problem of type (1), where all possible dataset functions* $[f_1, \ldots, f_{2^{|\mathcal{X}|}}]$ *are GP distributed* $\mathcal{GP}\big(0, k_1(\mathcal{V}, \mathcal{V}') \otimes k_2(\lambda, \lambda')\big)$ *for known kernels* k_1, k_2*, for all* $\mathcal{V}, \mathcal{V}' \subseteq \mathcal{X}$ *and* $\lambda, \lambda' \in \Lambda$*, where* $|\Lambda| < \infty$.

Similar Gaussian process assumptions have been made in previous work (Srinivas et al., 2010). For a result in the no-noise observation setting, we will make use of the assumptions of de Freitas et al. (2012) for our privacy guarantees, as described in Section 4.

2.1. Differential Privacy

One of the most widely accepted frameworks for private data release is *differential privacy* (Dwork et al., 2006b), which has been shown to be robust to a variety of privacy attacks (Sweeney, 1997; Dinur & Nissim, 2003; Ganta et al., 2008; Narayanan & Shmatikov, 2008). Given an algorithm $\mathcal{A}$ that outputs a value λ when run on dataset $\mathcal{V}$, the goal of differential privacy is to 'hide' the effect of a small change in $\mathcal{V}$ on the output of $\mathcal{A}$. Equivalently, an attacker should not be able to tell if a single private record was changed in $\mathcal{V}$ just by looking at the output of $\mathcal{A}$. If two datasets $\mathcal{V}, \mathcal{V}'$ differ in the value of a single individual, we will refer to them as *neighboring* datasets. Note that any non-trivial algorithm (i.e., an algorithm $\mathcal{A}$ that outputs different values on $\mathcal{V}$ and $\mathcal{V}'$ for some pair $\mathcal{V}, \mathcal{V}' \subseteq \mathcal{X}$) must include some amount of randomness to guarantee such a change in $\mathcal{V}$ is 'unobservable' in the output λ of $\mathcal{A}$ (Dwork & Roth, 2013). The level of privacy we wish to guarantee decides the amount of randomness we need to add to λ (better privacy requires increased randomness). Formally, the definition of differential privacy is stated below.

Definition 1. *A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-**differentially private** for $\epsilon, \delta \geq 0$ if for all $\lambda \in \mathrm{Range}(\mathcal{A})$ and for all neighboring datasets $\mathcal{V}, \mathcal{V}'$ (i.e., such that $\mathcal{V}$ and $\mathcal{V}'$ differ in the value of one record) we have that*

$$\Pr\big[\mathcal{A}(\mathcal{V}) = \lambda\big] \leq e^{\epsilon} \Pr\big[\mathcal{A}(\mathcal{V}') = \lambda\big] + \delta. \tag{4}$$

The parameters ϵ, δ guarantee how private $\mathcal{A}$ is; the smaller, the more private. The maximum privacy is $\epsilon = \delta = 0$ in which case eq. (4) holds with equality. This can be seen by the fact that $\mathcal{V}$ and $\mathcal{V}'$ can be swapped in the definition, and thus the inequality holds in both directions. If $\delta = 0$, we say the algorithm is simply ϵ-differentially private. For a survey on differential privacy we refer the interested reader to Dwork & Roth (2013).

There are two popular methods for making an algorithm ϵ-differentially private: (a) the Laplace mechanism (Dwork et al., 2006b), in which we add random noise to λ and (b) the exponential mechanism (McSherry & Talwar, 2007), which draws a random output $\tilde{\lambda}$ such that $\tilde{\lambda} \approx \lambda$. For each mechanism we must define an intermediate quantity called the *global sensitivity* describing how much $\mathcal{A}$ changes when $\mathcal{V}$ changes.

Definition 2. *(Laplace mechanism) The **global sensitivity** of an algorithm $\mathcal{A}$ over all neighboring datasets $\mathcal{V}, \mathcal{V}'$ (i.e., $\mathcal{V}, \mathcal{V}'$ differ by the value of one record) is*

$$\Delta_{\mathcal{A}} \triangleq \max_{\mathcal{V}, \mathcal{V}' \subseteq \mathcal{X}} \|\mathcal{A}(\mathcal{V}) - \mathcal{A}(\mathcal{V}')\|_1.$$

*(Exponential mechanism) The **global sensitivity** of a function $q : \mathcal{X} \times \Lambda \rightarrow \mathbb{R}$ over all neighboring datasets $\mathcal{V}, \mathcal{V}'$ is*

$$\Delta_q \triangleq \max_{\substack{\mathcal{V}, \mathcal{V}' \subseteq \mathcal{X} \\ \lambda \in \Lambda}} \|q(\mathcal{V}, \lambda) - q(\mathcal{V}', \lambda)\|_1.$$

The Laplace mechanism hides the output of $\mathcal{A}$ by perturbing its output with some amount of random noise.

Definition 3. *Given a dataset $\mathcal{V}$ and an algorithm $\mathcal{A}$, the **Laplace mechanism** returns $\mathcal{A}(\mathcal{V}) + \omega$, where ω is a noise variable drawn from $\mathrm{Lap}(\Delta_{\mathcal{A}}/\epsilon)$, the Laplace distribution with scale parameter $\Delta_{\mathcal{A}}/\epsilon$ (and location parameter 0).*

The exponential mechanism draws a slightly different $\tilde{\lambda}$ that is 'close' to λ, the output of $\mathcal{A}$.

Definition 4. *Given a dataset $\mathcal{V}$ and an algorithm $\mathcal{A}(\mathcal{V}) = \arg\max_{\lambda \in \Lambda} q(\mathcal{V}, \lambda)$, the **exponential mechanism** returns $\tilde{\lambda}$, where $\tilde{\lambda}$ is drawn from the distribution $\frac{1}{Z} \exp\big(\epsilon q(\mathcal{V}, \lambda)/(2\Delta_q)\big)$, and Z is a normalizing constant.*

Given Λ, a possible set of hyper-parameters, we derive methods for privately releasing the best hyper-parameters and the best function values $f_{\mathcal{V}}$, approximately solving eq. (1). We first address the setting with observation noise ($\sigma^2 > 0$) in eq. (2) and then describe small modifications for the no-noise setting. For each setting we use the UCB sampling technique in eq. (3) to derive our private results.

3. With observation noise

In general cases of Bayesian optimization, observation noise occurs in a variety of real-world modeling settings such as sensor measurements (Krause et al., 2008). In hyper-parameter tuning, noise in the validation gain may be as a result of noisy validation or training features.

Algorithm 1 Private Bayesian Opt. (noisy observations)

```
Input: $\mathcal{V}$; $\Lambda \subseteq \mathbb{R}^d$; $T$; $(\epsilon, \delta)$; $\sigma^2_{\mathcal{V},0}$; $\gamma_T$
$\mu_{\mathcal{V},0} = 0$
for $t = 1 \ldots T$ do
    $\beta_t = 2\log(|\Lambda| t^2 \pi^2 / (3\delta))$
    $\lambda_t \triangleq \arg\max_{\lambda \in \Lambda} \mu_{\mathcal{V},t-1}(\lambda) + \sqrt{\beta_t}\sigma_{\mathcal{V},t-1}(\lambda)$
    Observe validation gain $v_{\mathcal{V},t}$, given $\lambda_t$
    Update $\mu_{\mathcal{V},t}$ and $\sigma^2_{\mathcal{V},t}$ according to (2)
end for
$c = 2\sqrt{\big(1 - k(\mathcal{V}, \mathcal{V}')\big) \log\big(3|\Lambda|/\delta\big)}$
$q = \sigma\sqrt{8\log(3/\delta)}$
$C_1 = 8/\log(1 + \sigma^{-2})$
Draw $\tilde{\lambda} \in \Lambda$ w.p. $\Pr[\lambda] \propto \exp\Big(\frac{\epsilon \mu_{\mathcal{V},T}(\lambda)}{2(2\sqrt{\beta_{T+1}} + c)}\Big)$
$v^* = \max_{t \leq T} v_{\mathcal{V},t}$
Draw $\theta \sim \mathrm{Lap}\Big[\frac{\sqrt{C_1 \beta_T \gamma_T}}{\epsilon\sqrt{T}} + \frac{c}{\epsilon} + \frac{q}{\epsilon}\Big]$
$\tilde{v} = v^* + \theta$
Return: $\tilde{\lambda}, \tilde{v}$
```

In the sections that follow, although the quantities f, μ, σ, v all depend on the validation dataset $\mathcal{V}$, for notational simplicity we will occasionally omit the subscript $\mathcal{V}$. Similarly, for $\mathcal{V}'$ we will often write: f', μ', σ'^2, v'.

3.1. Private near-maximum hyper-parameters

In this section we guarantee that releasing $\tilde{\lambda}$ in Algorithm 1 is private (Theorem 1) and has high utility (Theorem 2). Our proof strategy is as follows: we will first demonstrate the global sensitivity of $\mu_T(\lambda)$ with probability at least $1-\delta$. Then we will show that releasing $\tilde{\lambda}$ via the exponential mechanism is (ϵ, δ)-differentially private. Finally, we show that $\mu_T(\tilde{\lambda})$ is close to $\max_{\lambda \in \Lambda} \mu_T(\lambda)$.

Global sensitivity. As a first step we bound the global sensitivity of $\mu_T(\lambda)$ as follows:

Theorem 1. *Given Assumption 1, for any two neighboring datasets $\mathcal{V}, \mathcal{V}'$ and for all $\lambda \in \Lambda$ with probability at least $1-\delta$ there is an upper bound on the global sensitivity (in the exponential mechanism sense) of μ_T:*

$$|\mu'_T(\lambda) - \mu_T(\lambda)| \leq 2\sqrt{\beta_{T+1}} + \sigma_1\sqrt{2\log(3|\Lambda|/\delta)},$$

for $\sigma_1 = \sqrt{2\big(1-k_1(\mathcal{V},\mathcal{V}')\big)}$, $\beta_t = 2\log\Big(|\Lambda|t^2\pi^2/(3\delta)\Big)$.

Proof. Note that, by applying the triangle inequality twice, for all $\lambda \in \Lambda$,

$$\begin{aligned}|\mu'_T(\lambda) - \mu_T(\lambda)| \leq\ & |\mu'_T(\lambda) - f'(\lambda)| + |f'(\lambda) - \mu_T(\lambda)| \\ \leq\ & |\mu'_T(\lambda) - f'(\lambda)| + |f'(\lambda) - f(\lambda)| + |f(\lambda) - \mu_T(\lambda)|.\end{aligned}$$

We can now bound each one of the terms in the summation on the right hand side (RHS) with probability at least $\frac{\delta}{3}$. According to Srinivas et al. (2010), Lemma 5.1, we obtain $|\mu'_T(\lambda) - f'(\lambda)| \leq \sqrt{\beta_{T+1}}\sigma'_T(\lambda)$. The same can be applied to $|f(\lambda) - \mu_T(\lambda)|$. As $\sigma'_T(\lambda) \leq 1$, because $k(\lambda,\lambda) = 1$, we can upper bound both terms by $2\sqrt{\beta_{T+1}}$. In order to bound the remaining (middle) term on the RHS recall that for a random variable $Z \sim \mathcal{N}(0,1)$ we have: $\Pr\big[|Z| > \gamma\big] \leq e^{-\gamma^2/2}$. For variables $Z_1, \ldots Z_n \sim \mathcal{N}(0,1)$, we have, by the union bound, that $\Pr\big[\forall i,\ |Z_i| \leq \gamma\big] \geq 1 - ne^{-\gamma^2/2} \triangleq 1 - \frac{\delta}{3}$. If we set $Z = \frac{|f(\lambda) - f'(\lambda)|}{\sigma_1}$ and $n = |\Lambda|$, we obtain $\gamma = \sqrt{2\log\big(3|\Lambda|/\delta\big)}$, which completes the proof. ■

We remark that all of the quantities in Theorem 1 are either given or selected by the modeler (e.g, δ, T). Given this upper bound we can apply the exponential mechanism to release $\tilde{\lambda}$ privately, as per Definition 1:

Corollary 1. *Let $\mathcal{A}(\mathcal{V})$ denote Algorithm 1 applied on dataset $\mathcal{V}$. Given Assumption 1, $\tilde{\lambda}$ is (ϵ, δ)-differentially private, i.e.,* $\Pr\big[\mathcal{A}(\mathcal{V}) = \tilde{\lambda}\big] \leq e^{\epsilon}\Pr\big[\mathcal{A}(\mathcal{V}') = \tilde{\lambda}\big] + \delta$, *for any pair of neighboring datasets $\mathcal{V}, \mathcal{V}'$.*

We leave the proof of Corollary 1 to the supplementary material. As described in (Srinivas et al., 2010), lines 3-8 (UCB) of Algorithm 1 is a no-regret Bayesian optimization procedure: $\lim_{T\to\infty}\sum_{t=1}^{T}(f(\lambda^*) - f(\lambda_t))/T = 0$. Further, with high probability (as given by the Gaussian tail probability bounds) for selected hyperparameters $\alpha_i \in \{\alpha_1, \ldots, \alpha_T\}$ we have that $\mu_T(\alpha_i) \approx f(\lambda_i)$. Therefore, it is reasonable to select $\lambda = \arg\max_{\lambda\in\Lambda}\mu_T(\lambda)$, to optimize (1). McSherry & Talwar (2007) show that exponential mechanism (line 12) selects a noisy maximum as follows.

Theorem 2. *(McSherry & Talwar, 2007) The exponential mechanism selects $\tilde{\lambda}$ that has value $\mu_T(\tilde{\lambda})$ that is close to the maximum $\max_{\lambda\in\Lambda}\mu_T(\lambda)$,*

$$\max\nolimits_{\lambda\in\Lambda}\mu_T(\lambda) - \mu_T(\tilde{\lambda}) \leq \tfrac{2\Delta}{\epsilon}(\log|\Lambda| + a), \qquad (5)$$

w.p. $\geq 1 - (\delta + e^{-a})$, where $\Delta = 2\sqrt{\beta_{T+1}} + c$ (for β_{T+1} and c defined as in Algorithm 1).

Notice that as T increases, so does Δ in the above inequality, as the distribution in line 12 becomes more uniform. However, because Δ increases much more slowly than T (i.e. Δ is asymptotically $O(\sqrt{\log(T+1)^2})$), T should be selected as large as possible to achieve small average regret.

3.2. Private near-maximum validation gain

In this section we demonstrate releasing the validation gain $\tilde{v}$ in Algorithm 1 is private (Theorem 3) and that the noise we add to ensure privacy is bounded with high probability (Theorem 4). As in the previous section our approach will be to first derive the global sensitivity of the maximum v found by Algorithm 1. Then we show releasing $\tilde{v}$ is (ϵ, δ)-differentially private via the Laplace mechanism. Surprisingly, we can also show that $\tilde{v}$ is close to $f(\lambda^*)$.

Global sensitivity. We bound the global sensitivity of the maximum v found with Bayesian optimization and UCB:

Theorem 3. *Given Assumption 1, and neighboring $\mathcal{V}, \mathcal{V}'$, we have the following global sensitivity bound (in the Laplace mechanism sense) for the maximum v, w.p. $\geq 1-\delta$*

$$|\max_{t\leq T} v'_t - \max_{t\leq T} v_t| \leq \frac{\sqrt{C_1\beta_T\gamma_T}}{\sqrt{T}} + c + q.$$

(for c, q, β_T in Algorithm 1) where the maximum GP information gain γ_T is bounded above for the squared exponential and Matérn kernels (Srinivas et al., 2010).

Proof. For notational simplicity let us denote the regret term as $\Omega \triangleq \sqrt{C_1 T\beta_T\gamma_T}$. Then from Theorem 1 in Srinivas et al. (2010) we have that

$$\frac{\Omega}{T} \geq f(\lambda^*) - \frac{1}{T}\sum_{t=1}^{T} f(\lambda_t) \geq f(\lambda^*) - \max_{t\leq T} f(\lambda_t). \qquad (6)$$

This implies $f(\lambda^*) \leq \max_{t\leq T} f(\lambda_t) + \frac{\Omega}{T}$ with probability at least $1 - \frac{\delta}{3}$ (with appropriate choice of β_T).

Recall that in the proof of Theorem 1 we showed that $|f(\lambda) - f'(\lambda)| \leq c$ with probability at least $1 - \frac{\delta}{3}$ (for

c given in Algorithm 1). This along with the above expression imply the following two sets of inequalities with probability greater than $1-\frac{2\delta}{3}$:

$$f'(\lambda^*)-c \le f(\lambda^*) < \max_{t\le T} f(\lambda_t)+\tfrac{\Omega}{T};$$
$$f(\lambda^*)-c \le f'(\lambda^*) < \max_{t\le T} f'(\lambda_t)+\tfrac{\Omega}{T}.$$

These, in turn, imply the two sets of inequalities:

$$\max_{t\le T} f'(\lambda_t) \le f'(\lambda^*) < \max_{t\le T} f(\lambda_t)+\tfrac{\Omega}{T}+c;$$
$$\max_{t\le T} f(\lambda_t) \le f(\lambda^*) < \max_{t\le T} f'(\lambda_t)+\tfrac{\Omega}{T}+c.$$

This implies $|\max_{t\le T} f'(\lambda_t)-\max_{t\le T} f(\lambda_t)| \le \frac{\Omega}{T}+c$. That is, the global sensitivity of $\max_{t\le T} f(\lambda_t)$ is bounded. Given the sensitivity of the maximum f, we can readily derive the sensitivity of maximum v. First note that we can use the triangle inequality to derive

$$\begin{aligned}|\max_{t\le T} v'_t - \max_{t\le T} v_t| \le{}& |\max_{t\le T} v_t - \max_{t\le T} f(\lambda_t)| \\ &+ |\max_{t\le T} v'_t - \max_{t\le T} f'(\lambda_t)| \\ &+ |\max_{t\le T} f'(\lambda_t) - \max_{t\le T} f(\lambda_t)|.\end{aligned}$$

We can immediately bound the final term on the right hand side. For the first term, let $t_{v,\max}=\arg\max_{t\le T} v_t$ be the index of the best observed validation gain. Let $t_{f,\max}=\arg\max_{t\le T} f(\lambda_t)$ be the index of the best true validation gain. The first term is upper bounded by $|\alpha| \triangleq \max\{|\alpha_{t_{v,\max}}|,|\alpha_{t_{f,\max}}|\}$ (this can be easily seen by enumerating the cases: (a) $t_{v,\max}=t_{f,\max}$ and (b) $t_{v,\max}\neq t_{f,\max}$). Similarly, $|\alpha'|$ upper bounds the second term (defined in the same way). Let $\hat{\alpha}=\max\{\alpha,\alpha'\}$ so we have,

$$|\max_{t\le T} v'_t - \max_{t\le T} v_t| \le \tfrac{\Omega}{T}+c+|2\hat{\alpha}|.$$

Although $|\hat{\alpha}|$ can be arbitrarily large, recall that for $Z\sim\mathcal{N}(0,1)$ we have: $\Pr[|Z|\le\gamma]\ge 1-e^{-\gamma^2/2}\triangleq 1-\frac{\delta}{3}$. Therefore if we set $Z=\frac{\hat{\alpha}}{\sigma}$ we have $\gamma=\sqrt{2\log(3/\delta)}$. This implies that $|2\hat{\alpha}|\le\sigma\sqrt{8\log(3/\delta)}=q$ with probability at least $1-\frac{\delta}{3}$. Therefore, if Theorem 1 from Srinivas et al. (2010) and the bound on $|f(\lambda)-f'(\lambda)|$ hold together with probability at least $1-\frac{2\delta}{3}$ as described above, the theorem follows directly. ■

As in Theorem 1 each quantity in the above bound is given in Algorithm 1 (β, c, q), given in previous results (Srinivas et al., 2010) (γ_T, C_1) or specified by the modeler (T, δ). Now that we have a bound on the sensitivity of the maximum v we will use the Laplace mechanism to prove our privacy guarantee (proof in supplementary material):

Corollary 2. *Let $\mathcal{A}(\mathcal{V})$ denote Algorithm 1 run on dataset $\mathcal{V}$. Given Assumption 1, releasing $\tilde{v}$ is (ϵ,δ)-differentially private, i.e.,* $\Pr[\mathcal{A}(\mathcal{V})=\tilde{v}]\le e^{\epsilon}\Pr[\mathcal{A}(\mathcal{V}')=\tilde{v}]+\delta$.

Further, as the Laplace distribution has exponential tails, the noise we add to obtain $\tilde{v}$ is small and shrinks as T increases. In fact, we can show that this noisy validation error is close to the optimal noise-free validation error.

Theorem 4. *Given Assumption 1 we have,*

$$|\tilde{v}-f(\lambda^*)| \le \sigma\sqrt{8\log(1/\delta)}+\tfrac{\Omega}{T}+a\Big(\tfrac{\Omega}{\epsilon T}+\tfrac{c}{\epsilon}+\tfrac{q}{\epsilon}\Big),$$

with probability at least $1-(\delta+e^{-a})$ *for* $\Omega=\sqrt{C_1 T\beta_T\gamma_T}$.

Proof. Let Z be a Laplace random variable with scale parameter b and location parameter 0; $Z\sim Lap(b)$. Then $\Pr\big[|Z|\le ab\big]=1-e^{-a}$. Thus, in Algorithm 1, $|\tilde{v}-\max_{t\le T} v_t|\le ab$ for $b=\frac{\Omega}{\epsilon T}+\frac{c}{\epsilon}+\frac{q}{\epsilon}$ with probability at least $1-e^{-a}$. Let $t_{f,\max}=\arg\max_{t\le T} f(\lambda_t)$. Further observe,

$$\begin{aligned} ab &\ge \max_{t\le T} v_t-\tilde{v} \ge (\max_{t\le T} f(\lambda_t)-\alpha_{t_{f,\max}})-\tilde{v}\\ &\ge f(\lambda^*)-\tfrac{\Omega}{T}-\alpha_{t_{f,\max}}-\tilde{v} \end{aligned} \tag{7}$$

where the second and third inequality follow from the proof of Theorem 3 (using the regret bound of Srinivas et al. (2010): Theorem 1). Note that the third inequality holds with probability greater than $1-\frac{\delta}{2}$ (given β_t in Algorithm 1). The final inequality implies $f(\lambda^*)-\tilde{v}\le\alpha_{t_{f,\max}}+\frac{\Omega}{T}+ab$. Let $t_{v,\max}=\arg\max_{t\le T} v_t$. Also note,

$$\begin{aligned} ab &\ge \tilde{v}-\max_{t\le T} v_t \ge \tilde{v}-(\max_{t\le T} f(\lambda_t)+\alpha_{t_{v,\max}})\\ &\ge \tilde{v}-f(\lambda^*)-\tfrac{\Omega}{T}-\alpha_{t_{v,\max}} \end{aligned} \tag{8}$$

This implies that $f(\lambda^*)-\tilde{v}\ge-\alpha_{t_{v,\max}}-\frac{\Omega}{T}-ab$. Thus we have that $|\tilde{v}-f(\lambda^*)|\le\alpha_{t_{f,\max}}+\alpha_{t_{v,\max}}+\frac{\Omega}{T}+ab$. Finally, because $\alpha_{t_{f,\max}}$ and $\alpha_{t_{v,\max}}$ could be arbitrarily large we give a high probability upper bound on these quantities. Let $\hat{\alpha}\triangleq\max\{\alpha_{t_{f,\max}},\alpha_{t_{v,\max}}\}$. Again recall that for $Z\sim\mathcal{N}(0,1)$, we have the tail probability bound $\Pr\big[Z\le\gamma\big]\ge 1-(1/2)e^{-\gamma^2/2}\triangleq 1-\frac{\delta}{2}$. Therefore if we set $Z=\frac{\hat{\alpha}}{\sigma}$ we arrive at $\gamma=\sigma\sqrt{2\log(1/\delta)}$. This entails that $2\hat{\alpha}\le\sigma\sqrt{8\log(1/\delta)}$. Therefore, our proposed bound in Theorem 4 holds. ■

The right side of Theorem 4 becomes smaller as T increases, so T should be set as large as possible. We note that, because releasing either $\tilde{\lambda}$ or $\tilde{v}$ is (ϵ,δ)-differentially private, by Corollaries 1 and 2, releasing both private quantities in Algorithm 1 guarantees $(2\epsilon,2\delta)$-differential privacy for validation dataset $\mathcal{V}$. This is due to the composition properties of (ϵ,δ)-differential privacy (Dwork et al., 2006a) (in fact stronger composition results can be demonstrated, (Dwork & Roth, 2013)).

4. Without observation noise

In hyper-parameter tuning it may be reasonable to assume that we can observe function evaluations exactly: $v_{\mathcal{V},t} = f_{\mathcal{V}}(\lambda_t)$. First note that we can use the same algorithm to report the maximum λ in the no-noise setting. Indeed, Theorems 1 and 2 still hold. However, we cannot readily report a private maximum f as the information gain γ_T in Theorems 3 and 4 approaches infinity as $\sigma^2 \to 0$. Therefore, we extend results from the previous section to the exact observation case via the regret bounds of de Freitas et al. (2012). In this setting, the authors show that they can achieve exponentially-decreasing regret *for each sample*: $f(\lambda^*) - f(\lambda_t) < O(e^{-t/\log(t)})$. Therefore our strategy will be to select the last sample and add enough noise to cover a change in a validation record. Algorithm 2 demonstrates this strategy for privatizing the maximum f.

Algorithm 2 Private Bayesian Opt. (noise free obs.)

```
Input: $\mathcal{V}$; $\Lambda \subseteq \mathbb{R}^d$; $T$; $(\epsilon, \delta)$; $A, \tau$; assumptions on $f_{\mathcal{V}}$ in de Freitas et al. (2012)
Observe noise-free samples: $f_{\mathcal{V}}(\lambda_1), \ldots, f_{\mathcal{V}}(\lambda_T)$ using method of de Freitas et al. (2012)
$c = 2\sqrt{\big(1 - k(\mathcal{V}, \mathcal{V}')\big) \log(2|\Lambda|/\delta)}$
Draw $\theta \sim \mathrm{Lap}\Big[\frac{A}{\epsilon} e^{-\frac{T\tau}{(\log T)^{d/4}}} + \frac{c}{\epsilon}\Big]$
Return: $\tilde{f} = f_{\mathcal{V}}(\lambda_T) + \theta$
```

4.1. Private near-maximum validation gain

We demonstrate that releasing $\tilde{f}$ in Algorithm 2 is private (Corollary 3) and that a small amount of noise is added to make $\tilde{f}$ private (Theorem 6). To do so, we derive the global sensitivity of $f_{\mathcal{V}}(\lambda_T)$ in Algorithm 2 independent of the maximum information gain γ_T via de Freitas et al. (2012). Then we prove releasing $\tilde{f}$ is (ϵ, δ)-differentially private and that $\tilde{f}$ is close to the optimal $f(\lambda^*)$.

Global sensitivity. The following Theorem gives a bound on the global sensitivity of the maximum f.

Theorem 5. *Given Assumption 1 and the assumptions in Theorem 2 of de Freitas et al. (2012), for neighboring datasets $\mathcal{V}, \mathcal{V}'$ we have the following global sensitivity bound (in the Laplace mechanism sense),*

$$|f'(\lambda_T) - f(\lambda_T)| \le Ae^{-\frac{T\tau}{(\log T)^{d/4}}} + c$$

w.p. at least $1 - \delta$ for $c = 2\sqrt{\big(1 - k(\mathcal{V}, \mathcal{V}')\big) \log(2|\Lambda|/\delta)}$, given constants A and τ in de Freitas et al. (2012).

We leave the proof to the supplementary material.

Given this sensitivity, we may apply the Laplace mechanism to release $\tilde{f}$.

Corollary 3. *Let $\mathcal{A}(\mathcal{V})$ denote Algorithm 2 run on dataset $\mathcal{V}$. Given Assumption 1 and that f satisfies the assumptions of de Freitas et al. (2012), $\tilde{f}$ is (ϵ, δ)-differentially private, with respect to any neighboring dataset $\mathcal{V}'$,* i.e.,

$$\Pr\big[\mathcal{A}(\mathcal{V}) = \tilde{f}\big] \le e^{\epsilon} \Pr\big[\mathcal{A}(\mathcal{V}') = \tilde{f}\big] + \delta.$$

Even though we must add noise to the maximum f we show that $\tilde{f}$ is still close to the optimal $f(\lambda^*)$.

Theorem 6. *Given the assumptions of Corollary 3, we have the utility guarantee for Algorithm 2:*

$$|\tilde{f} - f(\lambda^*)| \le \Omega + a\Big(\frac{\Omega}{\epsilon} + \frac{c}{\epsilon}\Big)$$

w.p. at least $1 - (\delta + e^{-a})$ for $\Omega = Ae^{-\frac{T\tau}{(\log T)^{d/4}}}$.

We prove Corollary 3 and Theorem 6 in the supplementary material. As in the noisy setting, selecting T as large as possible is the best strategy, as this reduces the right side of Theorem 6 (less Laplacian noise is required). We have thus shown that in noisy and noise-free settings it is possible to release private, high-quality hyper-parameter settings $\tilde{\lambda}$ as well as private, near-optimal function evaluations $\tilde{v}, \tilde{f}$.

5. Without the GP assumption

Even if our our true validation score f is not drawn from a Gaussian process (Assumption 1), we can still guarantee differential privacy for releasing its value after Bayesian optimization $f^{\text{BO}} = \max_{t \le T} f(\lambda_t)$. In this section we describe a different functional assumption on f that also yields differentially private Bayesian optimization for the case of machine learning hyper-parameter tuning.

Assume we have a (nonsensitive) training set $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, which, given a hyperparameter λ produces a model $\mathbf{w}(\lambda)$ from the following optimization,

$$\mathbf{w}_\lambda = \arg\min_{\mathbf{w}} \overbrace{\frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}, \mathbf{x}_i, y_i)}^{O_\lambda(\mathbf{w})}, \tag{9}$$

The function ℓ is a training loss function (e.g., logistic loss, hinge loss). Given a (sensitive) validation set $\mathcal{V} = \{(\overline{\mathbf{x}}_i, \overline{y}_i)\}_{i=1}^m \subseteq \mathcal{X}$ we would like to use Bayesian optimization to maximize a validation score $f_{\mathcal{V}}$.

Assumption 2. *Our true validation score $f_{\mathcal{V}}$ is*

$$f_{\mathcal{V}}(\mathbf{w}_\lambda) = -\frac{1}{m}\sum_{i=1}^{m} g(\mathbf{w}_\lambda, \overline{\mathbf{x}}_i, \overline{y}_i),$$

where $g(\cdot)$ is a validation loss function that is L-Lipschitz in $\mathbf{w}_\lambda$ (e.g., ramp loss, normalized sigmoid (Huang et al., 2014)). Additionally, the training model $\mathbf{w}_\lambda$ is the minimizer of eq. (9) for a training loss $\ell(\cdot)$ that is 1-Lipschitz in $\mathbf{w}_\lambda$ and convex (e.g., logistic loss, hinge loss).

Algorithm 3 describes a procedure for privately releasing the best validation accuracy f^{BO} given Assumption 2. Different from previous algorithms, we may run Bayesian optimization in Algorithm 3 with any acquisition function (e.g., expected improvement (Mockus et al., 1978), UCB) and privacy is still guaranteed.

Algorithm 3 Private Bayesian Opt. (Lipschitz and convex)

```
**Input:** $\mathcal{T}$ size $n$; $\mathcal{V}$ size $m$; $\Lambda$; $\lambda_{\min}$; $\lambda_{\max}$; $\epsilon$; $T$; $L$; $d$
Run Bayesian optimization for $T$ timesteps, observing:
    $f_{\mathcal{V}}(\mathbf{w}_{\lambda_1}), \ldots, f_{\mathcal{V}}(\mathbf{w}_{\lambda_T})$ for $\{\lambda_1, \ldots, \lambda_T\} = \Lambda_{\mathcal{V},T} \subseteq \Lambda$
$f^{\mathrm{BO}} = \max_{t \le T} f_{\mathcal{V}}(\mathbf{w}_{\lambda_t})$
$g^* = \max_{(\mathbf{x},y)\in\mathcal{X},\mathbf{w}\in\mathcal{R}^p} g(\mathbf{w}, \mathbf{x}, y)$
Draw $\theta \sim \mathrm{Lap}\left[\frac{1}{\epsilon}\min\{\frac{g^*}{m}, \frac{L}{m\lambda_{\min}}\} + \frac{(\lambda_{\max}-\lambda_{\min})L}{\epsilon\lambda_{\max}\lambda_{\min}}\right]$
**Return:** $\tilde{f}_L = f^{\mathrm{BO}} + \theta$
```

Similar to Algorithms 1 and 2 we use the Laplace mechanism to mask the possible change in validation accuracy when a single record is changed in $\mathcal{V}$. Different from the related work of Chaudhuri & Vinterbo (2013) changing $\mathcal{V}$ to neighboring dataset $\mathcal{V}'$ may also lead to Bayesian optimization searching different hyper-parameters, $\Lambda_{\mathcal{V},T}$ vs. $\Lambda_{\mathcal{V}',T}$. Therefore, we must bound the *total global sensitivity* of f with respect to $\mathcal{V}$ *and* λ,

Definition 5. *The* ***total global sensitivity*** *of f over all neighboring datasets $\mathcal{V}, \mathcal{V}'$ is*

$$\Delta_f \triangleq \max_{\substack{\mathcal{V},\mathcal{V}'\subseteq\mathcal{X}\\ \lambda,\lambda'\in\Lambda}} |f_{\mathcal{V}}(\mathbf{w}_\lambda) - f_{\mathcal{V}'}(\mathbf{w}_{\lambda'})|.$$

In the following theorem we demonstrate that we can bound the change in f for arbitrary $\lambda < \lambda'$ (w.l.o.g.).

Theorem 7. *Given Assumption 2, for neighboring $\mathcal{V}, \mathcal{V}'$ and arbitrary $\lambda < \lambda'$ we have that,*

$$|f_{\mathcal{V}}(\mathbf{w}_\lambda) - f_{\mathcal{V}'}(\mathbf{w}_{\lambda'})| \le \tfrac{(\lambda'-\lambda)L}{\lambda'\lambda} + \min\{\tfrac{g^*}{m}, \tfrac{L}{m\lambda_{\min}}\}$$

where L is the Lipschitz constant of f, m is the size of $\mathcal{V}$, and $g^* = \max_{(\mathbf{x},y)\in\mathcal{X},\mathbf{w}\in\mathcal{R}^p} g(\mathbf{w}, \mathbf{x}, y)$.

Proof. Applying the triangle inequality yields

$$\begin{aligned}|f_{\mathcal{V}}(\mathbf{w}_\lambda) - f_{\mathcal{V}'}(\mathbf{w}_{\lambda'})| \le\ & |f_{\mathcal{V}}(\mathbf{w}_\lambda) - f_{\mathcal{V}}(\mathbf{w}_{\lambda'})| \\ & + |f_{\mathcal{V}}(\mathbf{w}_{\lambda'}) - f_{\mathcal{V}'}(\mathbf{w}_{\lambda'})|.\end{aligned}$$

This second term is bounded by Chaudhuri & Vinterbo (2013) in the proof of Theorem 4. The only difference is, as we are not adding random noise to $\mathbf{w}_{\lambda'}$ we have that $|f_{\mathcal{V}}(\mathbf{w}_{\lambda'}) - f_{\mathcal{V}'}(\mathbf{w}_{\lambda'})| \le \min\{g^*/m, L/(m\lambda_{\min})\}$.

To bound the first term, let $O_\lambda(\mathbf{w})$ be the value of the objective in eq. (9) for a particular λ. Note that $O_\lambda(\mathbf{w})$ and $O_{\lambda'}(\mathbf{w})$ are λ and λ'-strongly convex. Define

$$h(\mathbf{w}) = O_{\lambda'}(\mathbf{w}) - O_\lambda(\mathbf{w}) = \tfrac{\lambda'-\lambda}{2}\|\mathbf{w}\|_2^2. \tag{10}$$

Further, define the minimizers $\mathbf{w}_\lambda = \arg\min_{\mathbf{w}} O_\lambda(\mathbf{w})$ and $\mathbf{w}_{\lambda'} = \arg\min_{\mathbf{w}}[O_\lambda(\mathbf{w}) + h(\mathbf{w})]$. This implies that

$$\nabla O_\lambda(\mathbf{w}_\lambda) = \nabla O_\lambda(\mathbf{w}_{\lambda'}) + \nabla h(\mathbf{w}_{\lambda'}) = 0. \tag{11}$$

Given that O_λ is λ-strongly convex (Shalev-Shwartz, 2007), and by the Cauchy-Schwartz inequality,

$$\begin{aligned}&\lambda\,\|\mathbf{w}_\lambda - \mathbf{w}_{\lambda'}\|_2^2 \le \Big[\nabla O_\lambda(\mathbf{w}_\lambda) - \nabla O_\lambda(\mathbf{w}_{\lambda'})\Big]^\top \Big[\mathbf{w}_\lambda - \mathbf{w}_{\lambda'}\Big] \\ &\le \nabla h(\mathbf{w}_{\lambda'})^\top \Big[\mathbf{w}_\lambda - \mathbf{w}_{\lambda'}\Big] \le \|\nabla h(\mathbf{w}_{\lambda'})\|_2 \|\mathbf{w}_\lambda - \mathbf{w}_{\lambda'}\|_2.\end{aligned}$$

Rearranging,

$$\frac{1}{\lambda}\|\nabla h(\mathbf{w}_{\lambda'})\|_2 = \left\|\frac{\lambda'-\lambda}{2}\nabla\|\mathbf{w}_{\lambda'}\|_2^2\right\|_2 \ge \|\mathbf{w}_\lambda - \mathbf{w}_{\lambda'}\|_2 \tag{12}$$

Now as $\mathbf{w}_{\lambda'}$ is the minimizer of $O_{\lambda'}$ we have,

$$\nabla\|\mathbf{w}_{\lambda'}\|_2^2 = \tfrac{2}{\lambda'}\big[-\tfrac{1}{n}\textstyle\sum_{i=1}^n \nabla\ell(\mathbf{w}_{\lambda'}, \mathbf{x}_i, y_i)\big].$$

Substituting this value of $\mathbf{w}_{\lambda'}$ into eq. (12) and noting that we can pull the positive constant term $(\lambda'-\lambda)/2$ out of the norm and drop the negative sign in the norm gives us

$$\tfrac{1}{\lambda}\|\nabla h(\mathbf{w}_{\lambda'})\|_2 = \tfrac{\lambda'-\lambda}{\lambda\lambda'}\left\|\tfrac{1}{n}\textstyle\sum_{i=1}^n \nabla\ell(\mathbf{w}_{\lambda'}, \mathbf{x}_i, y_i)\right\|_2 = \tfrac{\lambda'-\lambda}{\lambda\lambda'}.$$

The last equality follows from the fact that the loss ℓ is 1-Lipschitz by Assumption 2 and the triangle inequality. Thus, along with eq. (12), we have

$$\|\mathbf{w}_\lambda - \mathbf{w}_{\lambda'}\|_2 \le \tfrac{1}{\lambda}\|\nabla h(\mathbf{w}_{\lambda'})\|_2 \le \tfrac{\lambda'-\lambda}{\lambda\lambda'}.$$

Finally, as f is L-Lipschitz in $\mathbf{w}$,

$$|f_{\mathcal{V}}(\mathbf{w}_\lambda) - f_{\mathcal{V}}(\mathbf{w}_{\lambda'})| \le L\|\mathbf{w}_\lambda - \mathbf{w}_{\lambda'}\|_2 \le L\tfrac{\lambda'-\lambda}{\lambda\lambda'}$$

Combining the result of Chaudhuri & Vinterbo (2013) with the above expression completes the proof. ■

Given a finite set of hyperparameters Λ, in order to bound the total global sensitivity of f note that, by Theorem 7,

$$\Delta_f \le \tfrac{(\lambda_{\max}-\lambda_{\min})L}{\lambda_{\max}\lambda_{\min}} + \min\Big\{\tfrac{g^*}{m}, \tfrac{L}{m\lambda_{\min}}\Big\},$$

as $(\lambda'-\lambda)/(\lambda'\lambda)$ is strictly increasing in λ' and is strictly decreasing in λ. Given the total global sensitivity of f we can use the Laplace mechanism to 'hide' the validation set.

Corollary 4. *Let $\mathcal{A}(\mathcal{V})$ denote Algorithm 3 applied on dataset $\mathcal{V}$. Given Assumption 2, $\tilde{f}_L$ is ϵ-differentially private, i.e.,* $\Pr\big[\mathcal{A}(\mathcal{V}) = \tilde{f}_L\big] \le e^\epsilon \Pr\big[\mathcal{A}(\mathcal{V}') = \tilde{f}_L\big]$

We leave the proof to the supplementary material. Further, by the exponential tails of the Laplace mechanism we have the following utility guarantee,

Theorem 8. *Given the assumptions of Theorem 7, we have the following utility guarantee for $\tilde{f}_L$ w.r.t. f^{BO},*

$$|\tilde{f}_L - f^{BO}| \le a\left[\frac{1}{\epsilon m}\min\{g^*, \frac{L}{\lambda_{\min}}\} + \frac{(\lambda_{\max}-\lambda_{\min})L}{\epsilon\lambda_{\max}\lambda_{\min}}\right]$$

with probability at least $1 - e^{-a}$.

Proof. This follows exactly from the tail bound on Laplace random variables, given in the proof of Theorem 6. ■

One should set T as large as possible as Bayesian optimization may only possibly find better parameter settings when T is increased, and the noise added is independent of T.

6. Related work

There has been much work towards differentially private convex optimization (Chaudhuri et al., 2011; Kifer et al., 2012; Duchi et al., 2013; Song et al., 2013; Jain & Thakurta, 2014; Bassily et al., 2014). The work of Bassily et al. (2014) established upper and lower bounds for the excess empirical risk of ϵ and (ϵ, δ)-differentially private algorithms for many settings including convex and strongly convex risk functions that may or may not be smooth. There is also related work towards private high-dimensional regression, where the dimensions outnumber the number of instances (Kifer et al., 2012; Smith & Thakurta, 2013a). In such cases the Hessian becomes singular and so the loss is nonconvex. However, it is possible to use the restricted strong convexity of the loss in the regression case to guarantee privacy.

Differential privacy has been shown to be achievable in online and interactive kernel learning settings (Jain et al., 2012; Smith & Thakurta, 2013b; Jain & Thakurta, 2013; Mishra & Thakurta, 2014). In general, non-private online algorithms are closest in spirit to the methods of Bayesian optimization. However, all of the previous work in differentially private online learning represents a dataset as a sequence of bandit arm pulls (the equivalent notion in Bayesian optimization is function evaluations $f(\lambda_t)$). Instead, we consider functions in which changing a single dataset entry possibly affects *all future function evaluations*. Closest to our work is that of Chaudhuri & Vinterbo (2013), who show that given a fixed set of hyper-parameters which are always evaluated for any validation set, they can return a private version of the index of the best hyper-parameter, as well as a private model trained with that hyper-parameter. In one sense, their setting is more difficult as they guarantee privacy for the training and validation sets (we assume the training set is fixed). On the other hand, our selection strategy is more general in that, if the validation set changes, Bayesian optimization could search completely different hyper-parameters.

Bayesian optimization, largely due to its principled handling of the exploration/exploitation trade-off of global, black-box function optimization, is quickly becoming the global optimization paradigm of choice. Alongside promising empirical results there is a wealth of recent work on convergence guarantees for Bayesian optimization, similar to those used in this work (Srinivas et al., 2010; de Freitas et al., 2012). Vazquez & Bect (2010) and Bull (2011) give regret bounds for optimization with the expected improvement acquisition function. BayesGap (Hoffman et al., 2014) gives a convergence guarantee for Bayesian optimization with budget constraints. Bayesian optimization has also been extended to multi-task optimization (Bardenet et al., 2013; Swersky et al., 2013), the parallel experiment setting (Azimi et al., 2012; Snoek et al., 2012), and to constrained optimization (Gardner et al., 2014).

7. Conclusion

We have introduced methods for privately releasing the best hyper-parameters and validation accuracies in the case of exact and noisy observations. Our work makes use of the differential privacy framework, which has become commonplace in private machine learning (Dwork & Roth, 2013). We believe we are the first to demonstrate differentially private quantities in the setting of global optimization of expensive (possibly nonconvex) functions, through the lens of Bayesian optimization.

One key future direction is to design techniques to release each sampled hyper-parameter and validation accuracy privately (during the run of Bayesian optimization). This requires analyzing how the maximum upper-confidence bound changes as the validation dataset changes. Another interesting direction is extending our guarantees in Sections 3 and 4 to other acquisition functions.

For the case of machine learning hyper-parameter tuning our results are designed to guarantee privacy of the validation set only (it is equivalent to guarantee that the training set is never allowed to change). To simultaneously protect the privacy of the training set it may be possible to use techniques similar to the training stability results of Chaudhuri & Vinterbo (2013). Training stability could be guaranteed, for example, by assuming an additional training set kernel that bounds the effect of altering the training set on f. We leave developing these guarantees for future work.

Acknowledgments

KQW, MJK, and JRG are supported by NSF grants IIA-1355406, IIS-1149882, EFRI-1137211. This work was also supported by Yahoo! Labs, where MJK was a research intern working on this project. The authors thank Abhradeep Thakurta and Avishek Saha for helpful guidance, discussions, and suggestions.

References

Auer, Peter, Cesa-Bianchi, Nicolo, and Fischer, Paul. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Azimi, Javad, Jalali, Ali, and Fern, Xiaoli Z. Hybrid batch bayesian optimization. In *ICML 2012*, pp. 1215–1222. ACM, 2012.

Bardenet, Rémi, Brendel, Mátyás, Kégl, Balázs, and Sebag, Michèle. Collaborative hyperparameter tuning. In *ICML*, 2013.

Bassily, Raef, Smith, Adam, and Thakurta, Abhradeep. Private empirical risk minimization, revisited. *arXiv preprint arXiv:1405.7085*, 2014.

Bergstra, James and Bengio, Yoshua. Random search for hyper-parameter optimization. *JMLR*, 13:281–305, 2012.

Bonilla, Edwin, Chai, Kian Ming, and Williams, Christopher. Multi-task gaussian process prediction. In *NIPS*, 2008.

Bull, Adam D. Convergence rates of efficient global optimization algorithms. *JMLR*, 12:2879–2904, 2011.

Chaudhuri, Kamalika and Vinterbo, Staal A. A stability-based validation procedure for differentially private machine learning. In *Advances in Neural Information Processing Systems*, pp. 2652–2660, 2013.

Chaudhuri, Kamalika, Monteleoni, Claire, and Sarwate, Anand D. Differentially private empirical risk minimization. *JMLR*, 12:1069–1109, 2011.

Chong, Miao M, Abraham, Ajith, and Paprzycki, Marcin. Traffic accident analysis using machine learning paradigms. *Informatica (Slovenia)*, 29(1):89–98, 2005.

Cortes, Corinna and Vapnik, Vladimir. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

de Freitas, Nando, Smola, Alex, and Zoghi, Masrour. Exponential regret bounds for gaussian process bandits with deterministic observations. In *ICML*, 2012.

Dinur, Irit and Nissim, Kobbi. Revealing information while preserving privacy. In *Proceedings of the SIGMOD-SIGACT-SIGART symposium on principles of database systems*, pp. 202–210. ACM, 2003.

Duchi, John C, Jordan, Michael I, and Wainwright, Martin J. Local privacy and statistical minimax rates. In *FOCS*, pp. 429–438. IEEE, 2013.

Dwork, Cynthia and Roth, Aaron. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2013.

Dwork, Cynthia, Kenthapadi, Krishnaram, McSherry, Frank, Mironov, Ilya, and Naor, Moni. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006*, pp. 486–503. Springer, 2006a.

Dwork, Cynthia, McSherry, Frank, Nissim, Kobbi, and Smith, Adam. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pp. 265–284. Springer, 2006b.

Ganta, Srivatsava Ranjit, Kasiviswanathan, Shiva Prasad, and Smith, Adam. Composition attacks and auxiliary information in data privacy. In *KDD*, pp. 265–273. ACM, 2008.

Gardner, Jacob, Kusner, Matt, Xu, Zhixiang, Weinberger, Kilian, and Cunningham, John. Bayesian optimization with inequality constraints. In *ICML*, pp. 937–945, 2014.

Hoffman, Matthew, Shahriari, Bobak, and de Freitas, Nando. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *AISTATS*, pp. 365–374, 2014.

Huang, Xiaolin, Shi, Lei, and Suykens, Johan AK. Ramp loss linear programming support vector machine. *The Journal of Machine Learning Research*, 15(1):2185–2211, 2014.

Hutter, Frank, Hoos, H. Holger, and Leyton-Brown, Kevin. Sequential model-based optimization for general algorithm configuration. In *Learning and Intelligent Optimization*, pp. 507–523. Springer, 2011.

Jain, Prateek and Thakurta, Abhradeep. Differentially private learning with kernels. In *ICML*, pp. 118–126, 2013.

Jain, Prateek and Thakurta, Abhradeep Guha. (near) dimension independent risk bounds for differentially private learning. In *ICML*, pp. 476–484, 2014.

Jain, Prateek, Kothari, Pravesh, and Thakurta, Abhradeep. Differentially private online learning. *COLT*, 2012.

Kifer, Daniel, Smith, Adam, and Thakurta, Abhradeep. Private convex empirical risk minimization and high-dimensional regression. *JMLR*, 1:41, 2012.

Krause, Andreas, Singh, Ajit, and Guestrin, Carlos. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *JMLR*, 9:235–284, 2008.

McSherry, Frank and Talwar, Kunal. Mechanism design via differential privacy. In *FOCS*, pp. 94–103. IEEE, 2007.

Mishra, Nikita and Thakurta, Abhradeep. Private stochastic multi-arm bandits: From theory to practice. In *ICML Workshop on Learning, Security, and Privacy*, 2014.

Mockus, Jonas, Tiesis, Vytautas, and Zilinskas, Antanas. The application of bayesian methods for seeking the extremum. *Towards Global Optimization*, 2(117-129):2, 1978.

Narayanan, Arvind and Shmatikov, Vitaly. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pp. 111–125. IEEE, 2008.

Rasmussen, Carl Edward and Williams, Christopher K. I. Gaussian processes for machine learning. 2006.

Schölkopf, Bernhard and Smola, Alexander J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

Shalev-Shwartz, Shai. Online learning: Theory, algorithms, and applications. 2007.

Smith, Adam and Thakurta, Abhradeep Guha. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *COLT*, pp. 819–850, 2013a.

Smith, Adam and Thakurta, Abhradeep Guha. (nearly) optimal algorithms for private online learning in full-information and bandit settings. In *NIPS*, pp. 2733–2741, 2013b.

Snoek, Jasper, Larochelle, Hugo, and Adams, Ryan P. Practical bayesian optimization of machine learning algorithms. In *NIPS*, pp. 2951–2959, 2012.

Song, Shuang, Chaudhuri, Kamalika, and Sarwate, Anand D. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing*, 2013.

Srinivas, Niranjan, Krause, Andreas, Kakade, Sham M, and Seeger, Matthias. Gaussian process optimization in the bandit setting: No regret and experimental design. In *ICML*, 2010.

Sweeney, Latanya. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3):98–110, 1997.

Swersky, Kevin, Snoek, Jasper, and Adams, Ryan P. Multi-task bayesian optimization. In *NIPS*, pp. 2004–2012, 2013.

Vazquez, Emmanuel and Bect, Julien. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and Inference*, 140(11):3088–3095, 2010.

Weinberger, Kilian, Dasgupta, Anirban, Langford, John, Smola, Alex, and Attenberg, Josh. Feature hashing for large scale multitask learning. In *ICML*, pp. 1113–1120. ACM, 2009.

Yu, Shipeng, Esbroeck, Alexander van, Farooq, Faisal, Fung, Glenn, Anand, Vikram, and Krishnapuram, Balaji. Predicting readmission risk with institution specific prediction models. In *IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 415–420. IEEE, 2013.