Differentially Private Bayesian Optimization

Matt J. Kusner, Jacob R. Gardner, Roman Garnett, Kilian Q. Weinberger
Computer Science & Engineering
Washington University in St. Louis
{mkusner,gardner.jake,kilian}@wustl.edu, rgarnett@uni-bonn.de

Abstract

As machine learning outside of academic settings becomes commonplace, Bayesian optimization is rapidly becoming an attractive method for practitioners to automate the process of classifier hyper-parameter tuning. Much practical 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 these private quantities are also near-optimal.

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 pool resources and share model specifications, hyper-parameters and validation accuracies through publications or other means. However, data-holders must be careful as even little amounts of information can compromise privacy. Which hyper-parameter setting yields highest accuracy can reveal sensitive information about individuals in the validation or training data set, reminiscent of reconstruction attacks in Dinur & Nissim (2003).

In this paper we develop an algorithm that automatically tunes the hyper-parameters of a machine learning algorithm using Bayesian optimization (Hutter et al., 2011; Bergstra & Bengio, 2012; Snoek et al., 2012; Hoffman et al., 2014; Shah et al., 2014) while provably preserving differential privacy (Dwork et al., 2006b). A practitioner can use our approach to efficiently find near-optimal hyper-parameters and safely share them without compromising sensitive information. Our privacy guarantees hold for releasing the best hyper-parameters and best validation accuracy. In fact, all of our results hold for the general setting of optimizing an expensive (possibly nonconvex) objective function using Bayesian optimization. Specifically our contributions are as follows: (a) we derive, to the best of our knowledge, the first framework for Bayesian optimization with provable differential privacy guarantees; and (b) we develop variations with and without measurement noise.

2 Background

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

Bayesian optimization. Consider the task of maximizing an unknown function $f_{\mathcal{D}} \colon \mathcal{X} \to \mathbb{R}$:

$$\max_{\mathbf{x} \in \mathcal{X}} f_{\mathcal{D}}(\mathbf{x}). \tag{1}$$

that depends on some dataset $\mathcal{D} \subseteq \mathcal{S}$. Throughout, we use the vocabulary of a common application: that of machine learning hyper-parameter tuning. In this case $f_{\mathcal{D}}(\mathbf{x})$ is the accuracy of a learning algorithm evaluated on validation dataset $\mathcal{D}$ that was trained with hyper-parameters $\mathbf{x} \in \mathcal{X} \subseteq \mathcal{R}^d$.

Bayesian optimization selects a small number of locations to sample $f_{\mathcal{D}}$: $[\mathbf{x}_1, \ldots, \mathbf{x}_T] = \mathbf{X}_T$ to optimize (1). Specifically, given a current sample $\mathbf{x}_t$, we observe a function evaluation (accuracy) y_t such that $y_t = f_{\mathcal{D}}(\mathbf{x}_t) + \alpha_t$, where $\alpha_t \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise with possibly non-zero variance σ^2. One well-known procedure to select hyper-parameters $\mathbf{x}$ maximizes the *upper-confidence bound* of a posterior Gaussian process (GP) surrogate of $f_{\mathcal{D}}$ (Auer et al., 2002; Srinivas et al., 2010):

$$\mathbf{x}_{t+1} \triangleq \arg\max_{\mathbf{x} \in \mathcal{X}} \mu_t(\mathbf{x}) + \sqrt{\beta_{t+1}} \sigma_t(\mathbf{x}), \tag{2}$$

where $\mu_t(\mathbf{x})$ and $\sigma_t(\mathbf{x})$ are the GP posterior mean and standard deviation after t samples (Rasmussen & Williams, 2006). Additionally β_{t+1} is a parameter that trades off the *exploitation* of maximizing $\mu_t(\mathbf{x})$ and the *exploration* of maximizing $\sigma_t(\mathbf{x})$. Srinivas et al. (2010) proved that given certain assumptions on $f_{\mathcal{D}}$ and fixed, non-zero observation noise ($\sigma^2 > 0$), selecting hyper-parameters $\mathbf{x}$ to maximize eq. (2) is a no-regret Bayesian optimization procedure: $\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} f_{\mathcal{D}}(\mathbf{x}^*) - f_{\mathcal{D}}(\mathbf{x}_t) = 0$, where $f_{\mathcal{D}}(\mathbf{x}^*)$ is the maximizer of eq. (1). For the no-noise setting, de Freitas et al. (2012) give a no-regret algorithm that maximizes eq. (2) as a subroutine.

The primary question this work aims to answer is: given $\hat{\mathbf{x}} \triangleq \arg\max_{t \leq T} f_{\mathcal{D}}(\mathbf{x}_t)$, how can we release private versions of $\hat{\mathbf{x}}$ and $f_{\mathcal{D}}(\hat{\mathbf{x}})$ that are close to $\mathbf{x}^*$ and $f_{\mathcal{D}}(\mathbf{x}^*)$?

Setting. To answer this question, let us define a GP over hyper-parameters $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ *and* datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{S}$ as follows: $\mathcal{GP}\big(0, k_1(\mathcal{D}, \mathcal{D}') \otimes k_2(\mathbf{x}, \mathbf{x}')\big)$. A prior of this form is a multi-task GP (Bonilla et al., 2008; Swersky et al., 2013). The function $k_1(\mathcal{D}, \mathcal{D}')$ defines a set kernel (e.g., the number of records that differ between $\mathcal{D}$ and $\mathcal{D}'$). For k_2, we focus on the squared exponential: $k_2(\mathbf{x}, \mathbf{x}') = \exp\big(-\|\mathbf{x} - \mathbf{x}'\|_2^2/(2\ell^2)\big)$ or Matérn kernels: (*e.g.,* $k_2(\mathbf{x}, \mathbf{x}') = (1 + \sqrt{5}r/\ell + (5r^2)/(3\ell^2)) \exp(-\sqrt{5}r/\ell)$, for $r = \|\mathbf{x} - \mathbf{x}'\|_2$), for a fixed ℓ, as they have known GP information gain bounds (Srinivas et al., 2010). As defined, the kernels k_2 are normalized (*i.e.*, $k_2(\mathbf{x}, \mathbf{x}) = 1$).

Assumption 1. *We have a maximization problem of type eq. (1), where all possible dataset functions* $[f_1, \ldots, f_{2^{|\mathcal{S}|}}]$ *are Gaussian process distributed* $\mathcal{GP}\big(0, k_1(\mathcal{D}, \mathcal{D}') \otimes k_2(\mathbf{x}, \mathbf{x}')\big)$ *for all* $\mathcal{D}, \mathcal{D}' \in \mathcal{S}$ *and* $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$*, where* $|\mathcal{X}| < \infty$.

Similar Gaussian process assumptions have been made in previous work (Srinivas et al., 2010). Differently, the dataset kernel k_1 additionally ensures that functions drawn from the GP change smoothly as the dataset is changed (similar to the stability assumptions of Chaudhuri & Vinterbo (2013)). 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.

Differential Privacy. One of the most theoretically sound 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 (Ganta et al., 2008; Sweeney, 1997; Narayanan & Shmatikov, 2008). Given an algorithm $\mathcal{A}$ that outputs a value $\mathbf{x}$ when run on dataset $\mathcal{D}$, the goal of differential privacy is to 'hide' the effect of a small change in $\mathcal{D}$ on the output of $\mathcal{A}$. Note that any non-trivial private algorithm must include some amount of randomness to guarantee such a change in $\mathcal{D}$ is unobservable in the output $\mathbf{x}$ of $\mathcal{A}$ (Dwork & Roth, 2013). 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* $\mathbf{x} \in \text{Range}(\mathcal{A})$ *and for all neighboring datasets* $\mathcal{D}, \mathcal{D}'$ *(i.e.,* $\mathcal{D}$ *and* $\mathcal{D}'$ *differ in one record) we have that*

$$\Pr\big[\mathcal{A}(\mathcal{D}) = \mathbf{x}\big] \leq e^{\epsilon} \Pr\big[\mathcal{A}(\mathcal{D}') = \mathbf{x}\big] + \delta. \tag{3}$$

The parameters ϵ, δ guarantee how private $\mathcal{A}$ is; the smaller, the more private. If $\delta = 0$, we say the algorithm is simply ϵ-differentially private. For a survey on differential privacy see Dwork & Roth (2013). Two popular methods for making an algorithm ϵ-differentially private are: (a) the Laplace mechanism (Dwork et al., 2006b), in which we add random noise to $\mathbf{x}$ and (b) the exponential mechanism (McSherry & Talwar, 2007), which draws a random output $\tilde{\mathbf{x}}$ such that $\tilde{\mathbf{x}} \approx \mathbf{x}$. For each mechanism we define the *global sensitivity*, describing how much $\mathcal{A}$ changes when $\mathcal{D}$ changes.

Definition 2. *(Laplace mechanism) The **global sensitivity** of an algorithm $\mathcal{A}$ over all neighboring datasets $\mathcal{D}, \mathcal{D}'$ (i.e., $\mathcal{D}, \mathcal{D}'$ differ in one record) is: $\Delta_{\mathcal{A}} \triangleq \max_{\mathcal{D},\mathcal{D}' \in \mathcal{S}} \|\mathcal{A}(\mathcal{D}) - \mathcal{A}(\mathcal{D}')\|_1$.*
*(Exponential mechanism) The **global sensitivity** of a function $q\colon \mathcal{S} \times \mathcal{X} \to \mathbb{R}$ over all possible neighboring datasets $\mathcal{D}, \mathcal{D}'$ is: $\Delta_q \triangleq \max_{\mathcal{D},\mathcal{D}' \in \mathcal{S}; \mathbf{x} \in \mathcal{X}} \|q(\mathcal{D}, \mathbf{x}) - q(\mathcal{D}', \mathbf{x})\|_1$.*

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

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

We now describe our method for privately releasing the best hyper-parameters and validation accuracies from Bayesian optimization. Due to space limits we defer all proofs to the full paper version.

3 With observation noise

Observation noise occurs in many real-world modeling settings such as sensor measurement prediction (Krause et al., 2008). Our algorithm makes use of (Srinivas et al., 2010) and reports private Bayesian optimization quantities. In all sections that follow, for notational simplicity we will occasionally omit the subscript $\mathcal{D}$ for quantities: y, f, μ, σ^2 (similarly, for $\mathcal{D}'$: y', f', μ', σ'^2).

3.1 Private near-maximum $\mathbf{x}$

We guarantee that releasing $\tilde{\mathbf{x}}$ is private (Corollary 1) and that it is near-optimal (Theorem 2). We first derive the global sensitivity of $\mu_T(\mathbf{x})$ with probability at least $1-\delta$. Then we will show show that releasing $\tilde{\mathbf{x}}$ via the exponential mechanism is (ϵ, δ)-differentially private. Finally, we prove that $\mu_T(\tilde{\mathbf{x}})$ is close to $f(\mathbf{x}^*)$, the true maximizer of eq. (1). The global sensitivity of $\mu_T(\mathbf{x})$ is bounded:

Theorem 1. *Given assumption 1, for any two neighboring datasets $\mathcal{D}, \mathcal{D}'$ and for all $\mathbf{x} \in \mathcal{X}$ with probability at least $1-\delta$ we have the upper bound on the global sensitivity of μ_T:*

$$|\mu'_T(\mathbf{x}) - \mu_T(\mathbf{x})| \leq 2\sqrt{\beta_{T+1}} + \sigma_1 \sqrt{2 \log\left(3|\mathcal{X}|/\delta\right)},$$

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

This global sensitivity bound implies that the exponential mechanism guarantees privacy:

Corollary 1. *Let $\mathcal{A}(\mathcal{D})$ denote Algorithm 1 applied on dataset $\mathcal{D}$. Given assumption 1, the quantity $\tilde{\mathbf{x}}$ is (ϵ, δ)-differentially private,* i.e.,

$$\Pr\big[\mathcal{A}(\mathcal{D}) = \tilde{\mathbf{x}}\big] \leq e^{\epsilon} \Pr\big[\mathcal{A}(\mathcal{D}') = \tilde{\mathbf{x}}\big] + \delta$$

for any pair of neighboring datasets $\mathcal{D}, \mathcal{D}'$.

We show that even though we release a noisy hyper-parameter setting $\tilde{\mathbf{x}}$, it is near-optimal.

Theorem 2. *Given assumption 1 the near-optimal guarantee for releasing $\tilde{\mathbf{x}}$ holds:*

$$\mu_T(\tilde{\mathbf{x}}) \geq f(\mathbf{x}^*) - 2\sqrt{\beta_T} - q - \frac{2\Delta}{\epsilon}(\log|\mathcal{X}| + a)$$

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

Algorithm 1 Private BO (noisy)

Input: $\mathcal{D}$; $\mathcal{X}$; T; (ϵ, δ); $\sigma^2_{\mathcal{D},0}$; γ_T; $\mu_{\mathcal{D},0} = 0$
for $t = 1$ **to** T **do**
  $\beta_t = 2\log(|\mathcal{X}| t^2 \pi^2 / (3\delta))$
  Select $\mathbf{x}_t$ using eq. (2)
  Observe $y_{\mathcal{D},t}$, given $\mathbf{x}_t$
  Update $\mu_{\mathcal{D},t}$ and $\sigma^2_{\mathcal{D},t}$
end for
$c = 2\sqrt{\big(1 - k(\mathcal{D}, \mathcal{D}')\big) \log\big(3|\mathcal{X}|/\delta\big)}$
$q = \sigma\sqrt{4\log(3/\delta)}$
$C_1 = 8/\log(1 + \sigma^{-2})$
Draw $\tilde{\mathbf{x}} \in \mathcal{X}$ w.p. $\Pr[\mathbf{x}] \propto \exp\Big(\frac{\epsilon \mu_{\mathcal{D},T}(\mathbf{x})}{2(2\sqrt{\beta_{T+1}} + c)}\Big)$
$y^* = \max_{t \leq T} y_{\mathcal{D},t}$
Draw $\theta \sim \mathrm{Lap}\Big[\frac{\sqrt{C_1 \beta_T \gamma_T}}{\epsilon\sqrt{T}} + \frac{c}{\epsilon} + \frac{q}{\epsilon}\Big]$
$\tilde{y} = y^* + \theta$
return $\tilde{\mathbf{x}}, \tilde{y}$

3.2 Private near-maximum y

We show releasing $\tilde{y}$ via the Laplace mechanism is (ϵ, δ)-differentially private (Theorem 3) and that $\tilde{y}$ is close to $f(\mathbf{x}^*)$ (Theorem 4). We bound global sensitivity of the maximum y as such:

Theorem 3. *Given assumption 1, the sensitivity bound for the maximum y holds w.p. at least $1-\delta$:*

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

where the maximum Gaussian process information gain γ_T is bounded above for the squared exponential and Matern kernels (Srinivas et al., 2010).

Given bound on the sensitivity of the maximum y, the Laplace mechanism yields a private algorithm:

Corollary 2. *Let $\mathcal{A}(\mathcal{D})$ denote Algorithm 1 run on dataset $\mathcal{D}$. Given assumption 1, releasing $\tilde{y}$ is (ϵ,δ)-differentially private,* i.e.,

$$\Pr[\mathcal{A}(\mathcal{D}) = \tilde{y}] \leq e^{\epsilon}\Pr[\mathcal{A}(\mathcal{D}') = \tilde{y}] + \delta.$$

Further, $\tilde{y}$, despite being a noisy maximum is close to $f(\mathbf{x}^*)$:

Theorem 4. *Given the assumptions of Theorem 1, we have the following bound,*

$$|\tilde{y} - f(\mathbf{x}^*)| \leq \sqrt{2\log(2T/\delta)} + \frac{\Omega}{T} + a\left(\frac{\Omega}{\epsilon T} + \frac{c}{\epsilon} + \frac{q}{\epsilon}\right),$$

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

Because releasing either $\tilde{\mathbf{x}}$ or $\tilde{y}$ is (ϵ,δ)-differentially private, releasing both private quantities in Algorithm 1 guarantees $(2\epsilon, 2\delta)$-privacy for validation dataset $\mathcal{D}$ (due to the composition properties of (ϵ,δ)-differential privacy (Dwork et al., 2006a)).

4 Without observation noise

If we can observe function evaluations exactly: $y_{\mathcal{D},t} = f_{\mathcal{D}}(\mathbf{x}_t)$ note that we can use the same algorithm to report a private maximum $\mathbf{x}$ as above. Theorems 1 and 2 still hold (note $q=0$ in Theorem 2). However, as γ_T approaches infinity as $\sigma^2 \to 0$ we extend results from the previous section to the exact observation case via the regret bounds of de Freitas et al. (2012).

Algorithm 2 Private BO (noise free)

Input: $\mathcal{D}$; $\mathcal{X} \subseteq \mathbb{R}^d$; T; (ϵ,δ); A,τ; assumptions on $f_{\mathcal{D}}$ in de Freitas et al. (2012)
Run method of de Freitas et al. (2012), resulting in noise free observations: $f_{\mathcal{D}}(\mathbf{x}_1),\ldots,f_{\mathcal{D}}(\mathbf{x}_T)$
$c = 2\sqrt{\big(1-k(\mathcal{D},\mathcal{D}')\big)\log(2|\mathcal{X}|/\delta)}$
Draw $\theta \sim \mathrm{Lap}\left[\frac{A}{\epsilon}e^{-\frac{2\tau}{(\log 2)^{d/4}}} + \frac{c}{\epsilon}\right]$
return $\tilde{f} = \max_{2\leq t\leq T} f_{\mathcal{D}}(\mathbf{x}_t) + \theta$

4.1 Private near-maximum f

We show that releasing $\tilde{f}$ is private (Corollary 3) and that $\tilde{f}$ is nearly $f(\mathbf{x}^*)$ (Theorem 6). The global sensitivity of the maximum f is:

Theorem 5. *Given assumption 1 and the assumptions de Freitas et al. (2012), Theorem 2,*

$$|\max_{2\leq t\leq T} f'(\mathbf{x}_t) - \max_{2\leq t\leq T} f(\mathbf{x}_t)| \leq Ae^{-\frac{2\tau}{(\log 2)^{d/4}}} + c$$

w.p. at least $1-\delta$ for $c = 2\sqrt{\big(1-k(\mathcal{D},\mathcal{D}')\big)\log(2|\mathcal{X}|/\delta)}$.

Now we apply the Laplace mechanism to release $\tilde{f}$.

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

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

Theorem 6. *Given the assumptions of Theorem 5, we have the utility guarantee for Algorithm 2:*

$$|\tilde{f} - f(\mathbf{x}^*)| \leq \Omega + a\Big(\frac{\Omega}{\epsilon} + \frac{c}{\epsilon}\Big)$$

w.p. at least $1-(\delta+e^{-a})$ for $\Omega = Ae^{-\frac{2\tau}{(\log 2)^{d/4}}}$.

Thus, our private maximum $\tilde{f}$ is near-optimal. We have introduced methods for privately releasing the best hyper-parameters and validation accuracies in the case of exact and noisy observations. 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.

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.

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.

Chaudhuri, Kamalika and Vinterbo, Staal A. A stability-based validation procedure for differentially private machine learning. In *NIPS*, pp. 2652–2660, 2013.

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

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 *SIGMOD-SIGACT-SIGART symposium on principles of database systems*, pp. 202–210. ACM, 2003.

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.

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.

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.

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.

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, C. E. and Williams, C. K. I. Gaussian processes for machine learning. 2006.

Shah, Amar, Wilson, Andrew Gordon, and Ghahramani, Zoubin. Student-t processes as alternatives to gaussian processes. In *AISTATS*, 2014.

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

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.

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.