Parameter-Free Locally Differentially Private Stochastic Subgradient Descent

Kwang-Sung Jun
University of Arizona
kjun@cs.arizona.edu

Francesco Orabona
Boston University
francesco@orabona.com

Abstract

We consider the problem of minimizing a convex risk with stochastic subgradients guaranteeing ϵ-locally differentially private (ϵ-LDP). While it has been shown that stochastic optimization is possible with ϵ-LDP via the standard SGD (Song et al., 2013), its convergence rate largely depends on the learning rate, which must be tuned via repeated runs. Further, tuning is detrimental to privacy loss since it significantly increases the number of gradient requests. In this work, we propose BANCO (Betting Algorithm for Noisy COins), the first ϵ-LDP SGD algorithm that essentially matches the convergence rate of the tuned SGD without any learning rate parameter, reducing privacy loss and saving privacy budget.

1 Introduction

In this paper, we consider the problem of minimizing the convex risk of a machine learning predictor, guaranteeing local differential privacy. Instead of going through the empirical risk minimization route, we directly optimize the stochastic objective of the risk via stochastic subgradients appropriately sanitized to guarantee the local differential privacy. Our proposed algorithm is a variant of stochastic subgradient descent, with the important feature of *not requiring the tuning of learning rates*. Yet, it guarantees the same convergence rate and local differential privacy of a stochastic subgradient descent procedure with the oracle tuning of its learning rate.

More in details, consider the risk $R(\boldsymbol{w}) : \mathbb{R}^d \to \mathbb{R}$ defined as $R(\boldsymbol{w}) := \mathbb{E}_{\boldsymbol{x} \sim \rho_X}[\ell(\boldsymbol{w}, \boldsymbol{x})]$, where $\ell(\boldsymbol{w}, \boldsymbol{x})$ is convex in the first argument and $\boldsymbol{x}$ represents sensitive data about an individual drawn i.i.d. from ρ_X. In machine learning terms, $R(\boldsymbol{w})$ represents the test loss that we are interested in directly minimizing, while $\ell(\boldsymbol{w}, \boldsymbol{x})$ is the loss on a single sample $\boldsymbol{x}$. It is well-known that, by linearity of the expectation, we can get an unbiased estimate of a subgradient of $R(\boldsymbol{w})$ just using a subgradient w.r.t. the first argument of $\ell(\cdot, \boldsymbol{x})$ where $\boldsymbol{x}$ is drawn i.i.d. from ρ_X. Hence, we can construct sanitized subgradients simply returning a noisy version $\mathcal{G}(\boldsymbol{w}) \in \partial\ell(\boldsymbol{w}, \boldsymbol{x}) + \boldsymbol{\xi}$ where the noise $\boldsymbol{\xi}$ guarantees the ϵ-Local Differential Privacy (ϵ-LDP), which prevents the private information from being identified. We elaborate more on the problem definition and the assumptions in Section 2.

One immediately obvious algorithm to try for ϵ-LDP is the standard stochastic subgradient descent (SGD) algorithm together with the popular choice of density $p(\boldsymbol{\xi}) \propto \exp(-\frac{\epsilon}{2}\|\boldsymbol{\xi}\|)$ (Song et al., 2013; 2015; Wu et al., 2017). Let us consider the SGD with a constant learning rate η for simplicity.[1] The convergence rate of SGD w.r.t. the optimum $\boldsymbol{w}^* = \min_{\boldsymbol{w}} R(\boldsymbol{w})$ after one pass over T samples is $\mathbb{E}[R(\boldsymbol{w}_T)] - R(\boldsymbol{w}^*) = O(\frac{\|\boldsymbol{w}^*\|^2}{\eta T} + \frac{d^2}{\epsilon^2}\eta)$. Notice that the convergence rate largely depends on the learning rate, and setting $\eta = \frac{\epsilon/d}{\sqrt{T}}$ results in the rate of $O(\frac{d}{\epsilon}\|\boldsymbol{w}^*\|^2/\sqrt{T})$. While the optimal tuning $\|\boldsymbol{w}^*\|\frac{\epsilon/d}{\sqrt{T}}$ would result in a much better rate of $O(\frac{d}{\epsilon}\|\boldsymbol{w}^*\|/\sqrt{T})$, the value $\|\boldsymbol{w}^*\|$ cannot be known beforehand. Therefore, practitioners need to tune the learning rate by trying out various values over multiple runs. However, this significantly increases the number of subgradient requests, which can be

---

[1]The same argument holds for adaptive learning rates such as Duchi et al. (2010).

interpreted as the privacy cost. This also implies that, when there is a fixed budget of privacy loss, one would have to use a smaller ϵ to allow multiple runs, which slows down the convergence.

Departing from the cumbersome and privacy-costing parameter tuning, we propose a new algorithm BANCO (Betting Algorithm for Noisy COins) that enjoys the convergence rate $\tilde{O}(\frac{d}{\epsilon}\|\boldsymbol{w}^*\|/\sqrt{T})$ without having to tune any parameters, where $\tilde{O}$ hides logarithmic factors. At the heart of the algorithm is a decomposition that isolates the work of adapting to the norm of the optimum $\|\boldsymbol{w}^*\|$ and delegates it to a one-dimensional stochastic optimization algorithm. We describe our algorithm and highlight the proof sketch in Section 3 and conclude with future directions in Section 4.

Related work. While there is a large body of work on differentially private (DP) empirical risk minimization, there is less research on studying the generalization properties of the obtained solution. Moreover, most of the work is based on the notion of differential privacy rather than on the stronger notion of local differential privacy (LDP). We remind the readers that the DP is useful when the data provider trusts the learning agents to be DP and allows them to access sensitive data (or data provider may possess the learners internally) whereas LDP is useful when the data provider does not trust the learning agents and thus have to release sanitized version of individual data. Due to space constraints, we focus on discussing the most relevant works. The notion of LDP was introduced in Duchi et al. (2014) and Song et al. (2013; 2015). Duchi et al. (2014) study the generalization of locally differentially private SGD, focusing on the optimal dependency on the number of dimensions, but assuming bounded domains to optimally tune the learning rates. Song et al. (2013; 2015) use locally differentially private SGD to minimize the regularized ERM problems. Wu et al. (2017) study the convergence of SGD to minimize the ERM problem and guarantee differential privacy through the output perturbation method, and they also require the knowledge of the norm of the optimal solution to optimally tune the learning rates. Bassily et al. (2014) studies the generalization properties of locally differentially private SGD, but it requires a bounded domain to set the learning rates or the presence of a regularizer, that in turn must be tuned. Instead, our work gives a generalization guarantee for a locally differentially private SGD with one pass over the data and no learning rates to tune.

The problem of parameter-free ϵ-LDP SGD can be seen as a noisy version of the parameter-free SGD problem (Streeter and McMahan, 2012; Orabona and Pal, 2016; Orabona and Tommasi, 2017; Cutkosky and Orabona, 2018). However, the extension of existing (near-optimal) parameter-free SGDs to noisy gradients is not immediately clear since the noise makes the gradients unbounded while the parameter-free SGDs require bounded gradients. Specifically, naively feeding in the unbounded gradients can even make the algorithm undefined, let alone the lack of convergence guarantees; see (Jun and Orabona, 2019, Section 3) for a detailed argument. Our solution in this paper goes back to the more general version of Online Convex Optimization (OCO) (Zinkevich, 2003) and reconsider its formulation to allow unbounded noise, which is more general than the ϵ-LDP setting. In this paper, however, we focus on the LDP setting; we refer the interested readers to Jun and Orabona (2019).

2 Problem Definition and Preliminaries

In this section, we describe our notations and provide relevant background.

Locally Differentially Private (LDP) SGD. An ϵ-differentially private algorithm must guarantee that the log-likelihood ratio of the outputs of the algorithm under two databases differing in a single individual's data is smaller than ϵ (Dwork et al., 2006). In the stricter definition of local differential privacy (LDP) (Wasserman and Zhou, 2010; Kasiviswanathan et al., 2011; Duchi et al., 2014; Song et al., 2013; 2015) an untrusted algorithm is allowed to access a perturbed version of a sensitive data point only through a sanitization interface. In particular, the sanitization mechanism must guarantee that the log-likelihood ratio of the data of two individuals x and x' is smaller than ϵ.

Definition 1 (Local Differential Privacy). *Let $D = (X_1, \ldots, X_n)$ be a sensitive dataset where each $X_i \sim \rho_X$ corresponds to data about individual i. A randomized sanitization mechanism M which outputs a disguised version $(U_1, \ldots U_n)$ of D is said to provide ϵ-local differential privacy to individual i, if*

$$\sup_{S} \sup_{x,x' \in \mathcal{D}} \frac{\mathbb{P}[U_i \in S | X_i = x]}{\mathbb{P}[U_i \in S | X_i = x']} \leq \exp(\epsilon),$$

where the probability is w.r.t. the randomization in the sanitization mechanism.

Algorithm 1 Betting Algorithm for Noisy COins (BANCO) for Differentially Private SGD

1: Set $\boldsymbol{w}_1 = \boldsymbol{q}_1 = \boldsymbol{0} \in \mathbb{R}^d$
2: **for** $t = 1$ **to** T **do**
3:    Receive a noisy negative subgradient $\hat{\boldsymbol{g}}_t$ such that $\mathbb{E}[\hat{\boldsymbol{g}}_t] \in -\partial\ell(\boldsymbol{w}_t, \boldsymbol{x})$ where $\boldsymbol{x} \sim \rho_X$
4:    Update magnitude: $m_{t+1} = \frac{1}{2a} \int_{-a}^{a} \beta \exp\left(\beta \sum_{s=1}^{t} \langle \hat{\boldsymbol{g}}_s, \boldsymbol{q}_s \rangle - \beta^2 t \left(\frac{\sigma^2}{2} + G^2\right)\right) \mathrm{d}\beta$
      where $a = \min\left(\frac{k_1}{G}, \frac{1}{b}\right)$, and $k_1 = 0.6838$
5:    Update direction: $\boldsymbol{q}_{t+\frac{1}{2}} = \boldsymbol{q}_t - \frac{\hat{\boldsymbol{g}}_t}{\sqrt{\sum_{s=1}^{t} \|\hat{\boldsymbol{g}}_s\|_2^2}}$
6:    Project direction onto $L2$ ball: $\boldsymbol{q}_{t+1} = \boldsymbol{q}_{t+\frac{1}{2}} \cdot \min\left(1, \left\|\boldsymbol{q}_{t+\frac{1}{2}}\right\|_2^{-1}\right)$
7:    Update the weight vector: $\boldsymbol{w}_{t+1} = m_{t+1} \boldsymbol{q}_{t+1} \in \mathbb{R}^d$
8: **end for**
9: Return $\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{w}_t$

Denote by $\partial\ell(\boldsymbol{w}, \boldsymbol{x})$ the subdifferential set of ℓ w.r.t. its first argument evaluated at $(\boldsymbol{w}, \boldsymbol{x})$. The local differential setting can be specialized to SGD, as described in the introduction, by a mechanism that, upon receiving a subgradient in $\partial\ell(\boldsymbol{w}, \boldsymbol{x})$ with $\boldsymbol{x} \overset{i.i.d.}{\sim} \rho_X$, returns a noisy version $\mathcal{G}(\boldsymbol{w}) \in \partial\ell(\boldsymbol{w}, \boldsymbol{x}) + \boldsymbol{\xi}$, where the noise $\boldsymbol{\xi}$ guarantees the ϵ-local differential privacy.

Assumptions. Throughout, we work with *negative* subgradients for notational simplicity. Let $\hat{\boldsymbol{g}}_t$ be the negative subgradient received at time step t. Let $\boldsymbol{g}_t = \mathbb{E}_t[\hat{\boldsymbol{g}}_t]$ be the true negative subgradient and $\boldsymbol{\xi}_t = \hat{\boldsymbol{g}}_t - \mathbb{E}_t[\hat{\boldsymbol{g}}_t]$ where we use the notation $\mathbb{E}_t$ to denote $\mathbb{E}[\cdot \mid \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{t-1}]$.

We assume that the true subgradients are bounded by G: $\|\boldsymbol{g}_t\|_2 \le G$. Furthermore, the noise $\boldsymbol{\xi}_t$ is conditionally zero-mean and has conditional finite variance:

$$\mathbb{E}_t\left[\|\boldsymbol{\xi}_t\|_2^2\right] \le \sigma^2, \forall t, \tag{1}$$

for some $\sigma > 0$. We also assume a tail condition such that $\boldsymbol{\xi}_t$ is conditionally sub-exponential with parameters $(\sigma_{\mathrm{ID}}^2, b)$:

$$\max_{\boldsymbol{a}:\|\boldsymbol{a}\|_2 \le 1} \mathbb{E}_t\left[\exp(\beta\langle\boldsymbol{\xi}_t, \boldsymbol{a}\rangle)\right] \le \exp\left(\beta^2 \sigma_{\mathrm{ID}}^2/2\right), \ \forall|\beta| \le 1/b. \tag{2}$$

One can show that, when (2) is achieved with equality, we have $\sigma_{\mathrm{ID}}^2 \le \sigma^2$. The intuition of the condition above is that the tail of the noise $\boldsymbol{\xi}_t$ behaves well in any direction. This noise definition covers a wide range of distributions, including Gaussian and Laplace. If $d = 1$, we have $\sigma^2 = \sigma_{\mathrm{ID}}^2$. This is not true in general and the relationship depends on the noise distribution. If $\boldsymbol{\xi}_t \sim \mathcal{N}(0, s^2\mathbf{I})$, then one can see that $\sigma_{\mathrm{ID}}^2 = s^2$ and $\sigma^2 = ds^2$. As another example, the Laplace mechanism noise used in differentially-private learning satisfies the tail condition above; see Jun and Orabona (2019, Lemma 7).

3 Parameter-free Stochastic Optimization with Noise

We now describe BANCO that achieves essentially the convergence rate of optimally-tuned SGD without repeated runs for tuning the learning rate, thus being truly one-pass. BANCO decomposes the optimization process into two parts: one for optimizing the direction and the other for the magnitude of the solution vector. For the direction, we use projected stochastic gradient descent with adaptive learning rates (Orabona and Pál, 2018). While, for the magnitude we use a parameter-free algorithm based on coin-betting (Orabona and Pal, 2016; Jun and Orabona, 2019). We then construct the solution vector as the multiplication of the two solutions. The full pseudocode can be found in Algorithm 1. Note that m_{t+1} has a closed form solution as follows: with shorthands $x = \sum_{s=1}^{t} \langle \hat{\boldsymbol{g}}_s, \boldsymbol{q}_s \rangle$ and $y = t(\sigma^2/2 + G^2)$,

$$m_{t+1} = \frac{e^{-a(ay+x)}\left(\sqrt{\pi}x \exp\left(\frac{(2ay+x)^2}{4y}\right)\left(\mathrm{erf}(\frac{2ay+x}{2\sqrt{y}}) + \mathrm{erf}(\frac{2ay-x}{2\sqrt{y}})\right) + 2\sqrt{y}(1 - e^{2ax})\right)}{8ay^{3/2}}.$$

For improving numerical precision for computing m_{t+1}, we refer to (Koolen, 2015).

We now show that BANCO is a parameter-free ϵ-LDP SGD algorithm. Consider the Laplace sanitization mechanism that adds noise with probability density function $\rho_{\boldsymbol{\xi}}(\boldsymbol{z}) \propto \exp(-\frac{\epsilon}{2}\|\boldsymbol{z}\|_2)$, which makes the subgradients very similar to one another. Song et al. (2015) proved that this

mechanism is ϵ-locally differentially private and ensures $\mathbb{E}\left[\|\boldsymbol{\xi}_t\|_2^2\right] \le \frac{4(d^2+d)}{\epsilon^2}$, satisfying (1). Further, Jun and Orabona (2019, Lemma 7) show that this mechanism satisfies (2) with $\sigma_{\text{ID}}^2 = 18d^2/\epsilon^2$ and $b = \epsilon/4$.

Theorem 2. *Assume that $\ell(\boldsymbol{w}, \boldsymbol{x})$ is convex in the first argument $\boldsymbol{w} \in \mathbb{R}^d$ and has its subgradients' L2-norm bounded by 1, where the subgradient is with respect to the first argument. Let the noise $\hat{\boldsymbol{g}}_t - \mathbb{E}[\hat{\boldsymbol{g}}_t]$ follow the density $\rho(\boldsymbol{z}) \propto \exp(-\frac{\epsilon}{2}\|\boldsymbol{z}\|_2)$. Then, for any $\boldsymbol{w}^\star \in \mathbb{R}^d$, after one pass over T samples Algorithm 1 guarantees*

$$\mathbb{E}\left[R\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_t\right)\right] - R(\boldsymbol{w}^\star) \le O\left(\frac{d\|\boldsymbol{w}^\star\|_2}{\epsilon\sqrt{T}}\sqrt{\ln\left(1 + \frac{d^2\|\boldsymbol{w}^\star\|_2 T}{\epsilon^2}\right)} + \frac{1}{T}\right).$$

This convergence rate matches the private SGD of Wu et al. (2017) up to polylogarithmic terms, with the important difference that we do not need to assume the knowledge of the norm of the optimal solution $\boldsymbol{w}^\star$ to tune the learning rates. Furthermore, the rate in Theorem 2 is unimprovable up to logarithmic factors as shown in Jun and Orabona (2019). Finally, we remark that Algorithm 1 is in fact an instance of Jun and Orabona (2019), and more generic algorithms are available via a reduction therein, which can be of independent interest. Also, the Gaussian noise can be used as well, resulting a better dependency in the dimension of the space, but in the weaker (ϵ, δ)-LDP.

3.1 Proof Sketch

The proof is mainly concerned with regret, a standard quantity in online learning that measures the suboptimality with an arbitrary comparator $\boldsymbol{u}$. Theorem 2 is a direct consequence of the expected regret bound we describe below using so-called the online-to-batch conversion (Cesa-Bianchi et al., 2004). Theorem 3 below shows that the expected regret of Algorithm 1 is decomposed into two expected regrets, each for the magnitude learner m_t and the direction learner $\boldsymbol{q}_t$. The fact that we require the expected regret of the direction learner w.r.t. the *unit norm comparator* frees us from tuning its learning rate, delegating the burden of adaptation to the magnitude learner. In turn, the magnitude algorithm is parameter-free, giving the best guarantee without having to tune any learning rates. The proof is simple and immediate from Cutkosky and Orabona (2018). Hereafter, we consider linear functions that can be written as $\ell(\boldsymbol{w}, \boldsymbol{x}) = \langle -\boldsymbol{g}_t, \boldsymbol{w}\rangle$ for some $\boldsymbol{g}_t$ without loss of generality (due to the convexity).

Theorem 3. *Suppose the direction learner obtains expected regret $\text{Regret}_T^{\text{D}}(\boldsymbol{u}) \coloneqq \sum_{t=1}^{T}\langle \hat{\boldsymbol{g}}_t, \boldsymbol{u} - \boldsymbol{q}_t\rangle$ for any competitor $\boldsymbol{u}$ in the unit ball $S \subset \mathbb{R}^d$ and the magnitude learner obtains expected regret $\text{Regret}_T^{\text{M}}(v) \coloneqq \sum_{t=1}^{T} s_t \cdot (v - m_t)$ for any competitor $v \in \mathbb{R}$ where $s_t = \langle \hat{\boldsymbol{g}}_t, \boldsymbol{q}_t\rangle$. Then, the iterates $m_t\boldsymbol{q}_t$ guarantee*

$$\mathbb{E}\,\text{Regret}_T(\boldsymbol{u}) \coloneqq \mathbb{E}\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{u} - m_t\boldsymbol{q}_t\rangle = \mathbb{E}\sum_{t=1}^{T}\langle \hat{\boldsymbol{g}}_t, \boldsymbol{u} - m_t\boldsymbol{q}_t\rangle \le \text{Regret}_T^{\text{M}}(\|\boldsymbol{u}\|) + \|\boldsymbol{u}\|\,\text{Regret}_T^{\text{D}}\left(\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right),$$

where we define $\boldsymbol{u}/\|\boldsymbol{u}\| = \boldsymbol{0}$ when $\boldsymbol{u} = \boldsymbol{0}$.

For the direction learner, one can invoke any algorithm that guarantees the optimal regret with respect to competitors in the unit ball. Here, we have used projected online gradient descent with the scale-free learning rates in Orabona and Pál (2018). One can then immediately obtain the expected regret bound with noisy subgradients:

$$\mathbb{E}\left[R_T^{\text{D}}\left(\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2}\right)\right] = O\left(\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\hat{\boldsymbol{g}}_t\|_2^2}\right]\right) \overset{(a)}{=} O\left(\sqrt{\sum_{t=1}^{T}\left(\mathbb{E}\|\boldsymbol{g}_t\|_2^2 + \sigma^2\right)}\right),$$

where (a) uses Jensen's inequality and the fact that $\mathbb{E}[\|\hat{\boldsymbol{g}}_t\|^2] = \mathbb{E}[\|\boldsymbol{g}_t\|_2^2] + \sigma^2$.

Finally, for the magnitude learner, we have instantiated the coin betting algorithm of Jun and Orabona (2019) that has the following regret bound:

$$\mathbb{E}[R_T^D(u)] = O\left(|u|\max\left\{(1+b)\ln\left(|u|(1+b)\right), \sqrt{(1+\sigma_{\text{ID}}^2)T\ln\left(|u|(1+\sigma_{\text{ID}}^2)T + 1\right)}\right\} + 1\right).$$

4

4 Future Work

Our study opens up numerous research directions. First, one immediate difference in our upper bound from the standard SGD algorithm with adaptive step size is that we do not have a data-dependent regret bound; we have $(G^2 + \sigma^2)T$ rather than $\mathbb{E}\left[\sum_{t=1}^{T} \|\hat{\boldsymbol{g}}_t\|^2\right]$. It would be interesting to investigate whether data-dependent bounds are possible. Second, it would be desirable not to require the knowledge of the noise through (σ^2, b) so we can have a single algorithm that can adapt to various types of noise. Finally, high probability regret bounds are a straightforward research direction.

References

R. Bassily, A. Smith, and A. Thakurta. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds. *arXiv preprint arXiv:1405.7085*, 2014.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Trans. Inf. Theory*, 50(9):2050–2057, 2004.

A. Cutkosky and F. Orabona. Black-box reductions for parameter-free online learning in Banach spaces. In *Proc. of the Conference on Learning Theory (COLT)*, 2018.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *COLT*, 2010.

J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Privacy aware learning. *Journal of the ACM*, 61(6): 38, 2014.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

K.-S. Jun and F. Orabona. Parameter-free online convex optimization with sub-exponential noise. In *Proc. of the Conference on Learning Theory (COLT)*, 2019.

S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

W. M. Koolen. Implementing squint, 2015. http://blog.wouterkoolen.info/Squint_implementation/post.html.

F. Orabona and D. Pal. Coin betting and parameter-free online learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 577–585. Curran Associates, Inc., 2016.

F. Orabona and D. Pál. Scale-free online learning. *Theoretical Computer Science*, 716:50–69, 2018. Special Issue on ALT 2015.

F. Orabona and T. Tommasi. Training deep networks without learning rates through coin betting. In *Advances in Neural Information Processing Systems*, pages 2160–2170, 2017.

S. Song, K. Chaudhuri, and A. D. Sarwate. Stochastic gradient descent with differentially private updates. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 245–248. IEEE, 2013.

S. Song, K. Chaudhuri, and A. Sarwate. Learning from data with heterogeneous noise using SGD. In *Proc. of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 894–902, 2015.

M. Streeter and B. McMahan. No-regret algorithms for unconstrained online convex optimization. In *Advances in Neural Information Processing Systems 25*, pages 2402–2410. Curran Associates, Inc., 2012.

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

X. Wu, F. Li, A. Kumar, K. Chaudhuri, S. Jha, and J. Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proc. of the 2017 ACM International Conference on Management of Data*, pages 1307–1322. ACM, 2017.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proc. of ICML*, pages 928–936, 2003.