# Private Adaptive Gradient Methods for Convex Optimization

**Hilal Asi** [*,1,2] **John Duchi** [2,3] **Alireza Fallah** [4,1] **Omid Javidbakht** [5] **Kunal Talwar** [5]

## Abstract

We study adaptive methods for differentially private convex optimization, proposing and analyzing differentially private variants of a Stochastic Gradient Descent (SGD) algorithm with adaptive stepsizes, as well as the AdaGrad algorithm. We provide upper bounds on the regret of both algorithms and show that the bounds are (worst-case) optimal. As a consequence of our development, we show that our private versions of AdaGrad outperform adaptive SGD, which in turn outperforms traditional SGD in scenarios with non-isotropic gradients where (non-private) Adagrad provably outperforms SGD. The major challenge is that the isotropic noise typically added for privacy dominates the signal in gradient geometry for high-dimensional problems; approaches to this that effectively optimize over lower-dimensional subspaces simply ignore the actual problems that varying gradient geometries introduce. In contrast, we study non-isotropic clipping and noise addition, developing a principled theoretical approach; the consequent procedures also enjoy significantly stronger empirical performance than prior approaches.

## 1. Introduction

While the success of stochastic gradient methods for solving empirical risk minimization has motivated their adoption across much of machine learning, increasing privacy risks in data-intensive tasks have made applying them more challenging (Dwork et al., 2006b): gradients can leak users' data, intermediate models can compromise individuals, and even final trained models may be non-private without substantial care. This motivates a growing line of work developing private variants of stochastic gradient descent (SGD), where algorithms guarantee differential privacy by perturbing individual gradients with random noise (Duchi et al., 2013; Smith & Thakurta, 2013a; Abadi et al., 2016; Duchi et al., 2018; Bassily et al., 2019; Feldman et al., 2020). Yet these noise addition procedures typically fail to reflect the geometry underlying the optimization problem, which in non-private cases is essential: for high-dimensional problems with sparse parameters, mirror descent and its variants (Beck & Teboulle, 2003; Nemirovski et al., 2009) are essential, while in the large-scale stochastic settings prevalent in deep learning, AdaGrad and other adaptive variants (Duchi et al., 2011) provide stronger theoretical and practical performance. Even more, methods that do not adapt (or do not leverage geometry) can be provably sub-optimal, in that there exist problems where their convergence is much slower than adaptive variants that reflect appropriate geometry (Levy & Duchi, 2019).

To address these challenges, we introduce PAGAN (Private AdaGrad with Adaptive Noise), a new differentially private variant of stochastic gradient descent and AdaGrad. Our main contributions center on a few ideas. Standard methods for privatizing adaptive algorithms that add isometric (typically Gaussian) noise to gradients necessarily reflect the worst-case behavior of functions to be optimized and eliminate the geometric structure one might leverage for improved convergence. By carefully adapting noise to the actual gradients at hand, we can both achieve convergence rates that reflect the observed magnitude of the gradients—similar to the approach of Bartlett et al. (2007) in the non-private case—which can yield marked improvements over the typical guarantees that depend on worst-case magnitudes. (Think, for example, of a standard normal variable: its second moment is 1, while its maximum value is unbounded.) Moreover, we propose a new private adaptive optimization algorithm that analogizes AdaGrad, showing that under certain natural distributional assumptions for the problems—similar to those that separate AdaGrad from non-adaptive methods (Levy & Duchi, 2019)—our private versions of adaptive methods significantly outperform the standard non-adaptive private algorithms. Additionally, we prove several lower bounds that both highlight the importance of geometry in the problems and demonstrate the tightness of the bounds our algorithms achieve.

[*]The authors are in alphabetical order. [1]Work done while interning at Apple [2]Department of Electrical Engineering, Stanford University [3]Department of Statistics, Stanford University [4]Department of Electrical Engineering & Computer Science, MIT [5]Apple. Correspondence to: Hilal Asi <asi@stanford.edu>, Alireza Fallah <afallah@mit.edu>.

*Proceedings of the 38th International Conference on Machine Learning*, PMLR 139, 2021. 

### 1.1. Related Work

Since the introduction of differential privacy (Dwork et al., 2006b;a), differentially private empirical risk minimization has been a subject of intense interest (Chaudhuri et al., 2011; Bassily et al., 2014; Duchi et al., 2013; Smith & Thakurta, 2013b). The current standard approach to solving this problem is noisy SGD (Bassily et al., 2014; Duchi et al., 2013; Abadi et al., 2016; Bassily et al., 2019; Feldman et al., 2020). Current bounds focus on the standard Euclidean geometries familiar from classical analyses of gradient descent (Zinkevich, 2003; Nemirovski et al., 2009), and the prototypical result (Bassily et al., 2014; 2019) is that, for Lipschitz convex optimization problems on the $\ell_2$-ball in $d$-dimensions, an $\varepsilon$-differentially private version of SGD achieves excess empirical loss $O(\frac{\sqrt{d}}{n\varepsilon})$ given a sample of size $n$; this is minimax optimal. Alternative approaches use the stability of empirical risk minimizers of (strongly) convex functions, and include both output perturbation, where one adds noise to a regularized empirical minimizer, and objective perturbation, where one incorporates random linear noise in the objective function before optimization (Chaudhuri et al., 2011). It is worth noting that, although objective perturbation may provide certain adaptivity properties to the geometry of the problem, its privacy guarantees are limited to the smooth and convex setting, hence it cannot be used in nonconvex applications such as deep neural networks.

Given the success of private SGD for such Euclidean cases and adaptive gradient algorithms for modern large-scale learning, it is unsurprising that recent work attempts to incorporate adaptivity into private empirical risk minimization (ERM) algorithms (Zhou et al., 2020; Kairouz et al., 2020). In this vein, Zhou et al. (2020) propose a private SGD algorithm where the gradients are projected to a low-dimensional subspace—which is learned using public data—and Kairouz et al. (2020) developed an $\varepsilon$-differentially private variant of Adagrad which (similarly) projects the gradient to a low rank subspace. These works show that excess loss $\widetilde{O}(\frac{1}{n\varepsilon}) \ll \frac{\sqrt{d}}{n\varepsilon}$ is possible whenever the rank of the gradients is small. Yet these both work under the assumption that gradient lie in (or nearly in) a low-dimensional subspace; this misses the contexts for which adaptive algorithms (AdaGrad and its relations) are designed (Duchi et al., 2011; McMahan & Streeter, 2010). Indeed, most stochastic optimization algorithms rely on particular dualities between the parameter space and gradients; stochastic gradient descent requires Euclidean spaces, while mirror descent works in an $\ell_1/\ell_\infty$ duality (that is, it is minimax optimal when optimizing over an $\ell_1$-ball while gradients belong to an $\ell_\infty$ ball). AdaGrad and other adaptive algorithms, in contrast, are optimal in an (essentially) dual geometry (Levy & Duchi, 2019), so that for such algorithms, the interesting geometry is when the parameters belong (e.g.) to an $\ell_\infty$ box and the gradients are sparse—but potentially from a very high-rank space. Indeed, as Levy & Duchi (2019) show, adaptive algorithms achieve benefits only when the sets over which one optimizes are quite different from $\ell_2$ balls; the private projection algorithms in the papers by Kairouz et al. (2020) and Zhou et al. (2020) achieve bounds that scale with the $\ell_2$-radius of the underlying space, suggesting that they may not enjoy the performance gains one might hope to achieve using an appropriately constructed and analyzed adaptive algorithm.

In more recent work, Yu et al. (2021) use PCA to decompose gradients into two orthogonal subspaces, allowing separate learning rate treatments in the subspaces, and achieve promising empirical results, but they provide no provable convergence bounds. Also related to the current paper is Pichapati et al.'s AdaClip algorithm (2020); they obtain parallels to the non-private convergence guarantees of Bartlett et al. (2007) for private SGD. In contrast to our analysis here, their analysis applies to smooth non-convex functions, while our focus on convex optimization allows more complete convergence guarantees and associated optimality results.

## 2. Preliminaries and notation

Before proceeding to the paper proper, we give notation. Let $\mathcal{Z}$ be a sample space and $P$ a distribution on $\mathcal{Z}$. Given a function $F : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$, convex in its first argument, and a dataset $\mathcal{S} = (z_1, \ldots, z_n) \in \mathcal{Z}^n$ of $n$ points drawn i.i.d. $P$, we wish to privately find the minimizer of the empirical loss

$$\operatorname*{argmin}_{x \in \mathcal{X}} f(x; \mathcal{S}) \coloneqq \frac{1}{n} \sum_{i=1}^{n} F(x; z_i). \tag{1}$$

We suppress dependence on $\mathcal{S}$ and simply write $f(x)$ when the dataset is clear from context. We use the standard definitions of differential privacy (Dwork et al., 2006b;a):

**Definition 2.1.** *A randomized algorithm $M$ is $(\varepsilon, \delta)$-*differentially private *if for all neighboring datasets $\mathcal{S}, \mathcal{S}' \in \mathcal{Z}^n$ and all measurable $O$ in the output space of $M$,*

$$\mathbb{P}(M(\mathcal{S}) \in O) \le e^{\varepsilon} \mathbb{P}(M(\mathcal{S}') \in O) + \delta.$$

*If $\delta = 0$, then $M$ is $\varepsilon$-*differentially private.

It will also be useful to discuss the tail properties of random variables and vectors:

**Definition 2.2.** *A random variable $X$ is $\sigma^2$* sub-Gaussian *if $\mathbb{E}[\exp(s(X - \mathbb{E}[X]))] \le \exp((\sigma^2 s^2)/2)$ for all $s \in \mathbb{R}$. A zero-mean random vector $X \in \mathbb{R}^d$ is $\Sigma$-sub-Gaussian if for any vector $a \in \mathbb{R}^d$, $a^\top X$ is $a^\top \Sigma a$ sub-Gaussian.*

We also frequently use different norms and geometries, so it is useful to recall the definition of Lipschitz functions:

**Definition 2.3.** *A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is $G$-*Lipschitz with respect to norm $\|\cdot\|$ over $\mathcal{W}$ *if for every $w_1, w_2 \in \mathcal{W}$,*

$$|\Phi(w_1) - \Phi(w_2)| \le G \|w_1 - w_2\|.$$

A convex function $\Phi$ is $G$-Lipschitz over an open set $\mathcal{W}$ if and only if $\|\Phi'(w)\|_* \le G$ for any $w \in \mathcal{W}$ and $\Phi'(w) \in \partial\Phi(w)$, where $\|y\|_* = \sup\{x^\top y \mid \|x\| \le 1\}$ is the dual norm of $\|\cdot\|$ (Hiriart-Urruty & Lemaréchal, 1993).

**Notation** We define $\mathrm{diag}(a_1, \ldots, a_d)$ as a diagonal matrix with diagonal entries $a_1, \ldots, a_d$. To state matrix $A$ is positive (semi)definite, we use the notation $A \succ 0_{d\times d}$ ($A \succcurlyeq 0_{d\times d}$). For $A \succcurlyeq 0_{d\times d}$, let $E_A$ denote the ellipsoid $\{x : \|x\|_A \le 1\}$ where $\|x\|_A = \sqrt{x^\top A x}$ is the Mahalanobis norm, and $\pi_A(x) = \mathrm{argmin}_y\{\|y - x\|_2 \mid y \in E_A\}$ is the projection of $x$ onto $E_A$. For a set $\mathcal{X}$, $\mathrm{diam}_{\|\cdot\|}(\mathcal{X}) = \sup_{x,y\in\mathcal{X}} \|x - y\|$ denotes the diameter of $\mathcal{X}$ with respect to the norm $\|\cdot\|$. For the special case of $\|\cdot\|_p$, we write $\mathrm{diam}_p(\mathcal{X})$ for simplicity. For an integer $n \in \mathbb{N}$, we let $[n] = \{1, \ldots, n\}$.

## 3. Private Adaptive Gradient Methods

In this section, we study and develop PASAN and PAGAN, differentially private versions of Stochastic Gradient Descent (SGD) with adaptive stepsize (Algorithm 1) and Adagrad (Duchi et al., 2011) (Algorithm 2). The challenge in making these algorithms private is that adding isometric Gaussian noise—as is standard in the differentially private optimization literature—completely eliminates the geometrical properties that are crucial for the performance of adaptive gradient methods. We thus add noise that adapts to gradient geometry while maintaining privacy.

More precisely, our private version of adaptive optimization algorithms proceed as follows: to privatize the gradients, we first project them to an ellipsoid capturing their geometry, then adding non-isometric Gaussian noise whose covariance corresponds to the positive definite matrix $A$ that defines the ellipsoid. Finally, we apply the adaptive algorithm's step with the private gradients. We present our private versions of SGD with adaptive stepsizes and Adagrad in Algorithms 1 and 2, respectively.

Before analyzing the utility of these algorithms, we provide their privacy guarantees in the following lemma (see Appendix B.1 for its proof).

**Lemma 3.1.** *There exist constants $\bar{\varepsilon}$ and $c$ such that, for any $\varepsilon \le \bar{\varepsilon}$, and with $T = cn^2/b^2$, Algorithm 1 and Algorithm 2 are $(\varepsilon, \delta)$-differentially private.*

Having established the privacy guarantees of our algorithms, we now proceed to demonstrate their performance. To do so, we introduce an assumption to that, as we shall see presently, will allow us to work in gradient geometries different than the classical Euclidean ($\ell_2$) one common to current private optimization analyses.

**Assumption 1.** *There exists a function $G : \mathcal{Z} \times \mathbb{R}^{d\times d} \to \mathbb{R}_+$ such that for any diagonal $C \succ 0$, the function $F(\cdot; z)$ is $G(z; C)$-Lipschitz with respect to the Mahalanobis norm $\|.\|_{C^{-1}}$ over $\mathcal{X}$, i.e., $\|\nabla f(x; z)\|_C \le G(z; C)$ for all $x \in \mathcal{X}$.*

**Algorithm 1** Private Adaptive SGD with Adaptive Noise (PASAN)

**Require:** Dataset $\mathcal{S} = (z_1, \ldots, z_n) \in \mathcal{Z}^n$, convex set $\mathcal{X}$, mini-batch size $b$, number of iterations $T$, privacy parameters $\varepsilon, \delta$;

Choose arbitrary initial point $x^0 \in \mathcal{X}$;
**for** $k = 0$ to $T - 1$ **do**
  Sample a batch $\mathcal{D}_k := \{z_i^k\}_{i=1}^b$ from $\mathcal{S}$ uniformly with replacement;
  Choose ellipsoid $A_k$;
  Set $\tilde{g}^k := \frac{1}{b}\sum_{i=1}^b \pi_{A_k}(g^{k,i})$ where $g^{k,i} \in \partial F(x^k; z_i^k)$;
  Set $\hat{g}^k = \tilde{g}^k + \sqrt{\log(1/\delta)}/(b\varepsilon)\xi^k$ where $\xi^k \overset{\text{iid}}{\sim} \mathsf{N}(0, A_k^{-1})$;
  Set $\alpha_k = \alpha/\sqrt{\sum_{i=0}^k \|\hat{g}^i\|_2^2}$;
  $x^{k+1} := \mathrm{proj}_{\mathcal{X}}(x^k - \alpha_k \hat{g}^k)$;
**end for**
**Return:** $\overline{x}^T := \frac{1}{T}\sum_{k=1}^T x^k$

The moments of the this Lipschitz constant $G$ will be central to our convergence analyses, and to that end, for $p \ge 1$ we define the shorthand

$$G_p(C) := \mathbb{E}_{z\sim P}\left[G(z; C)^p\right]^{1/p}. \tag{2}$$

The quantity $G_p(C)$ are the $p$th moments of the gradients in the Mahalanobis norm $\|\cdot\|_C$; they are the key to our stronger convergence guarantees and govern the error in projecting our gradients. In most standard analyses of private optimization (and stochastic optimization more broadly), one takes $C = I$ and $p = \infty$, corresponding to the assumption that $F(\cdot, z)$ is $G$-Lipschitz for all $z$ and that subgradients $F'(x, z)$ are uniformly bounded in both $x$ and $z$. Even when this is the case—which may be unrealistic—we always have $G_p(C) \le G_\infty(C)$, and in realistic settings there is often a significant gap; by depending instead on appropriate moments $p$, we shall see it is often possible to achieve far better convergence guarantees than would be possible by relying on uniformly bounded moments. (See also the discussion of Barber & Duchi (2014) on these issues in the context of mean estimation.)

An example may be clarifying:

**Example 1:** Let $g : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function, let $F(x; Z) = g(x) + \langle x, Z\rangle$ where $Z \in \mathbb{R}^d$ and the coordinates $Z_j$ are independent $\sigma_j^2$-subgaussian, and $C \succ 0$ be diagonal. Then by standard moment bounds, if

**Algorithm 2** Private Adagrad with Adaptive Noise (PAGAN)

**Require:** Dataset $\mathcal{S} = (z_1, \ldots, z_n) \in \mathcal{Z}^n$, convex set $\mathcal{X}$, mini-batch size $b$, number of iterations $T$, privacy parameters $\varepsilon, \delta$;

1: Choose arbitrary initial point $x^0 \in \mathcal{X}$;
2: **for** $k = 0$ to $T-1$ **do**
3: Sample a batch $\mathcal{D}_k := \{z_i^k\}_{i=1}^b$ from $\mathcal{S}$ uniformly with replacement;
4: Choose ellipsoid $A_k$;
5: Set $\tilde{g}^k := \frac{1}{b}\sum_{i=1}^b \pi_{A_k}(g^{k,i})$ where $g^{k,i} \in \partial F(x^k; z_i^k)$;
6: Set $\hat{g}^k = \tilde{g}^k + \sqrt{\log(1/\delta)}/(b\varepsilon)\xi^k$ where $\xi^k \overset{\text{iid}}{\sim} \mathsf{N}(0, A_k^{-1})$;
7: Set $H_k = \operatorname{diag}\left(\sum_{i=0}^k \hat{g}^i \hat{g}^{i^T}\right)^{\frac{1}{2}} / \alpha$;
8: $x^{k+1} = \operatorname{proj}_{\mathcal{X}}(x^k - H_k^{-1}\hat{g}^k)$ where the projection is with respect to $\|\cdot\|_{H_k}$;
9: **end for**
10: **Return:** $\overline{x}^T := \frac{1}{T}\sum_{k=1}^T x^k$

$\|\nabla g(x)\|_C \le \mu$ we have

$$G_p(C) \le \mu + O(1)\sqrt{p}\sqrt{\sum_{j=1}^d C_{jj}\sigma_j^2}. \tag{3}$$

(See Appendix B.2 for the proof.) As this bound shows, while $G_\infty$ is infinite in this example, $G_p$ is finite. As a result, our analysis extends to settings in which the stochastic gradients are not uniformly bounded. $\diamond$

While we defined $G_p(C)$ by taking expectation with respect to the original distribution $P$, we mainly focus on empirical risk minimization and thus require the empirical Lipschitz constant for a given dataset $\mathcal{S}$:

$$\hat{G}_p(\mathcal{S}; C) := \left(\frac{1}{n}\sum_{i=1}^n G(z_i; C)^p\right)^{1/p}. \tag{4}$$

A calculation using Chebyshev's inequality and that $p$-norms are increasing immediately gives the next lemma:

**Lemma 3.2.** *Let $\mathcal{S}$ be a dataset with $n$ points sampled from distribution $P$. Then with probability at least $1 - 1/n$, we have*

$$\hat{G}_p(\mathcal{S}; C) \le G_p(C) + G_{2p}(C) \le 2G_{2p}(C),$$

It is possible to get bounds of the form $\hat{G}_p(\mathcal{S}; C) \lesssim G_{kp}(C)$ with probability at least $1 - 1/n^k$ using Khintchine's inequalities, but this is secondary for us.

Given these moment bounds, we can characterize the convergence of both algorithms, defering proofs to Appendix B.

### 3.1. Convergence of PASAN

We first start with PASAN (Algorithm 1). Similarly to the non-private setting where SGD (and its adaptive variant) are most appropriate for domains $\mathcal{X}$ with small $\ell_2$-diameter $\operatorname{diam}_2(\mathcal{X})$, our bounds in this section mostly depend on $\operatorname{diam}_2(\mathcal{X})$.

**Theorem 1.** *Let $\mathcal{S} \in \mathcal{Z}^n$ and $C \succ 0$ be diagonal, $p \ge 1$, and assume that $\hat{G}_p(\mathcal{S}; C) \le G_{2p}(C)$. Consider running* PASAN *(Algorithm 1) with $\alpha = \operatorname{diam}_2(\mathcal{X})$, $T = cn^2/b^2$, $A_k = \frac{1}{B^2}C$, where*[1]

$$B = 2G_{2p}(C)\left(\frac{\operatorname{diam}_{\|\cdot\|_{C^{-1}}}(\mathcal{X})n\varepsilon}{\operatorname{diam}_2(\mathcal{X})\sqrt{\operatorname{tr}(C^{-1})}\sqrt{\log(1/\delta)}}\right)^{1/p}$$

*and $c$ is the constant in Lemma 3.1. Then*

$$\mathbb{E}[f(\overline{x}^T; \mathcal{S}) - \min_{x \in \mathcal{X}} f(x; \mathcal{S})] \le \mathcal{O}\left(\frac{\operatorname{diam}_2(\mathcal{X})}{T}\sqrt{\sum_{k=1}^T \mathbb{E}[\|g^k\|_2^2]} + \operatorname{diam}_2(\mathcal{X})G_{2p}(C) \times \left(\frac{\sqrt{\operatorname{tr}(C^{-1})\ln\frac{1}{\delta}}}{n\varepsilon}\right)^{\frac{p-1}{p}}\left(\frac{\operatorname{diam}_{\|\cdot\|_{C^{-1}}}(\mathcal{X})}{\operatorname{diam}_2(\mathcal{X})}\right)^{\frac{1}{p}}\right),$$

*where the expectation is taken over the internal randomness of the algorithm.*

To gain intuition for these bounds, note that for large enough $p$, the bound from Theorem 1 is approximately

$$\operatorname{diam}_2(\mathcal{X})\Bigg(\underbrace{\frac{1}{T}\sqrt{\sum_{k=1}^T \mathbb{E}[\|g^k\|_2^2]}}_{=:R_{\text{std}}(T)} + G_{2p}(C)\cdot\frac{\sqrt{\operatorname{tr}(C^{-1})\log(1/\delta)}}{n\varepsilon}\Bigg). \tag{5}$$

The term $R_{\text{std}}(T)$ in (5) is the standard non-private convergence rate for SGD with adaptive stepsizes (Bartlett et al., 2007; Duchi, 2018) and (in a minimax sense) is unimprovable even without privacy; the second term is the cost of privacy. In the standard setting of gradients uniformly bounded in $\ell_2$-norm, where $C = I$ and $p = \infty$, this bound recovers the standard rate $\operatorname{diam}_2(\mathcal{X})G_\infty(I)\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon}$. However, as we show in our examples, this bound can offer significant improvements whenever $C \ne I$ such that $\operatorname{tr}(C^{-1}) \ll d$ or $G_{2p}(C) \ll G_\infty$ for some $p < \infty$.

[1] We provide the general statement of this theorem for positive $B$ in Appendix B.4

### 3.2. Convergence of PAGAN

Having established our bounds for PASAN, we now proceed to present our results for PAGAN (Algorithm 2). In the non-private setting, adaptive gradient methods such as Adagrad are superior to SGD for constraint sets such as $\mathcal{X} = [-1, 1]^d$ where $\text{diam}_\infty(\mathcal{X}) \ll \text{diam}_2(\mathcal{X})$. Following this, our bounds in this section will depend on $\text{diam}_\infty(\mathcal{X})$.

**Theorem 2.** *Let $\mathcal{S} \in \mathcal{Z}^n$ and $C \succ 0$ be diagonal, $p \geq 1$, and assume that $\hat{G}_p(\mathcal{S}; C) \leq G_{2p}(C)$. Consider running* PAGAN *(Algorithm 2) with $\alpha = \text{diam}_\infty(\mathcal{X})$, $T = cn^2/b^2$, $A_k = \frac{1}{B^2}C$, where*

$$B = 2G_{2p}(C)\left(\frac{\text{diam}_{\|\cdot\|_{C^{-1}}}(\mathcal{X}) n\varepsilon}{\text{diam}_\infty(\mathcal{X})\sqrt{\log(1/\delta)}\,\text{tr}(C^{-\frac{1}{2}})}\right)^{1/p}$$

*and $c$ is the constant in Lemma 3.1. Then*

$$\begin{aligned}&\mathbb{E}[f(\overline{x}^T; \mathcal{S}) - \min_{x\in\mathcal{X}} f(x; \mathcal{S})]\\ &\leq \mathcal{O}\Bigg(\frac{\text{diam}_\infty(\mathcal{X})}{T}\sum_{j=1}^d \mathbb{E}\left[\sqrt{\sum_{k=1}^T (g_j^k)^2}\right] + \text{diam}_\infty(\mathcal{X}) \times \\ &\quad G_{2p}(C)\left(\frac{\sqrt{\ln\frac{1}{\delta}}\,\text{tr}(C^{-\frac{1}{2}})}{n\varepsilon}\right)^{\frac{p-1}{p}}\left(\frac{\text{diam}_{\|\cdot\|_{C^{-1}}}(\mathcal{X})}{\text{diam}_\infty(\mathcal{X})}\right)^{\frac{1}{p}}\Bigg),\end{aligned}$$

*where the expectation is taken over the internal randomness of the algorithm.*

To gain intuition, we again consider the large $p$ case, where Theorem 2 simplifies to roughly

$$\text{diam}_\infty(\mathcal{X})\Bigg(\underbrace{\frac{1}{T}\sum_{j=1}^d \mathbb{E}\left[\sqrt{\sum_{k=1}^T (g_j^k)^2}\right]}_{=:R_{\text{ada}}(T)} + G_{2p}(C)\left(\frac{\sqrt{\log(1/\delta)}\,\text{tr}(C^{-\frac{1}{2}})}{n\varepsilon}\right)\Bigg).$$

In analogy with Theorem 1, the first term $R_{\text{ada}}(T)$ is the standard error for non-private Adagrad after $T$ iterations (Duchi et al., 2011)—and hence unimprovable (Levy & Duchi, 2019)—while the second is the privacy cost. In some cases, we may have $\text{diam}_\infty(\mathcal{X}) = \text{diam}_2(\mathcal{X})/\sqrt{d}$, so private Adagrad can offer significant improvements over SGD whenever the matrix $C$ has polynomially decaying diagonal.

To clarify the advantages and scalings we expect, we may consider an extremely stylized example with sub-Gaussian distributions. Assume now—in the context of Example 1—that we are optimizing the random linear function $F(x; Z) = \langle x, Z\rangle$, where $Z$ has independent $\sigma_j^2$-sub-Gaussian compoments. In this case, by assuming that $p = \log d$ and taking $C_{jj} = \sigma_j^{-4/3}$ and $b = 1$, Theorem 2 guarantees that PAGAN (Algorithm 2) has convergence

$$\mathbb{E}[f(\overline{x}^T; \mathcal{S}) - \min_{x\in\mathcal{X}} f(x; \mathcal{S})] \leq \tag{6}$$

$$\mathcal{O}(1)\text{diam}_\infty(\mathcal{X})\left[R_{\text{ada}}(T) + \frac{(\sum_{j=1}^d \sigma_j^{2/3})^{3/2}}{n\varepsilon}\log\frac{d}{\delta}\right]. \tag{7}$$

On the other hand, for PASAN (Algorithm 1), with $p = \log d$, $b = 1$, the choice $C_{jj} = \sigma_j^{-1}$ optimizes the bound of Theorem 1 and yields

$$\mathbb{E}[f(\overline{x}^T; \mathcal{S}) - \min_{x\in\mathcal{X}} f(x; \mathcal{S})] \leq \tag{8}$$

$$\mathcal{O}(1)\text{diam}_2(\mathcal{X})\left[R_{\text{std}}(T) + \frac{\sum_{j=1}^d \sigma_j}{n\varepsilon}\log\frac{d}{\delta}\right]. \tag{9}$$

Comparing these results, two differences are salient: $\text{diam}_\infty(\mathcal{X})$ replaces $\text{diam}_2(\mathcal{X})$ in Eq. (**??**), which can be an improvement by as much as $\sqrt{d}$, while $(\sum_{j=1}^d \sigma_j^{2/3})^{3/2}$ replaces $\sum_{j=1}^d \sigma_j$, and Hölder's inequality gives

$$\sqrt{d}\sum_{j=1}^d \sigma_j \geq \left(\sum_{j=1}^d \sigma_j^{2/3}\right)^{3/2} \geq \sum_{j=1}^d \sigma_j.$$

Depending on gradient moments, there are situations in which PAGAN offers significant improvements; these evidently depend on the expected magnitudes of the gradients and noise, as the $\sigma_j$ terms evidence. As a special case, consider $\mathcal{X} = [-1, +1]^d$ and assume $\{\sigma_j\}_{j=1}^d$ decrease quickly, e.g. $\sigma_j = 1/j^{3/2}$. In such a setting, the upper bound of PAGAN is roughly $\frac{\text{poly}(\log d)}{n\varepsilon}$ while PASAN achieves $\frac{\sqrt{d}}{n\varepsilon}$.

# 4. Some approaches to unknown moments

As the results of the previous section demonstrate, bounding the gradient moments allows us to establish tighter convergence guarantees; it behooves us to estimate them with accuracy sufficient to achieve (minimax) optimal bounds.

### 4.1. Unknown moments for generalized linear models

Motivated by the standard practice of training the last layer of a pre-trained neural network (Abadi et al., 2016), in this section we consider algorithms for generalized linear models, where we have losses of the form $F(x; z) = \ell(z^T x)$ for $z, x \in \mathbb{R}^d$ and $\ell : \mathbb{R} \to \mathbb{R}_+$ is a convex and 1-Lipschitz loss. As $\nabla F(x; z) = \ell'(z^T x)z$, bounds on the Lipschitzian moments (2) follow from moment bounds on $z$ itself, as $\|\nabla F(x; z)\| \leq \|z\|$.

The results of Section 3 suggest optimal choices for $C$ under sub-Gaussian assumptions on the vectors $z$, where in our stylized cases of $\sigma_j$-sub-Gaussian entries, $C_j = \sigma_j^{-4/3}$ minimizes our bounds. Unfortunately, it is hard in general to estimate $\sigma_j$ even without privacy (Duchi, 2019). Therefore, we make the following bounded moments ratio assumption, which relates higher moments to lower moments to allow estimation of moment-based parameters (even with privacy).

**Definition 4.1.** *A random vector $z \in \mathbb{R}^d$ has* moment ratio $r < \infty$ *if for all* $1 \le p \le 2\log d$ *and* $1 \le j \le d$

$$\mathbb{E}[z_j^p]^{2/p} \le r^2 p \cdot \mathbb{E}[z_j^2].$$

When $z$ satisfies Def. 4.1, we can provide a private procedure (Algorithm 3) that provides good approximation to the second moment of coordinates of $z_j$—and hence higher-order moments—allowing the application of a minimax optimal PAGAN algorithm. We defer the proof to Appendix C.

**Theorem 3.** *Let $z$ have moment ratio $r$ (Def. 4.1) and let $\sigma_j^2 = \mathbb{E}[z_j^2]$. Let $\beta > 0$, $T = \frac{3}{2}\log d$,*

$$n \ge 1000 r^2 \log \frac{8d}{\beta} \max\left\{ \frac{T\sqrt{d}\log^2 r \log\frac{T}{\delta}}{\varepsilon}, r^2 \right\},$$

*and* $\max_{1\le j\le d} \sigma_j = 1$*. Then Algorithm 3 is $(\varepsilon,\delta)$-DP and outputs $\hat{\sigma}$ such that with probability $1-\beta$,*

$$\frac{1}{2}\max\{\sigma_j, d^{-3/2}\} \le \hat{\sigma}_j \le 2\sigma_j \quad \textit{for all } j \in [d]. \qquad (10)$$

*Moreover, when condition (10) holds,* PAGAN *(Alg. 2) with $\hat{C}_j = (r\hat{\sigma}_j)^{-4/3}/4$, $p = \log d$ and $b = 1$ has convergence*

$$\mathbb{E}[f(\overline{x}^T; \mathcal{S}) - \min_{x\in\mathcal{X}} f(x;\mathcal{S})] \le R_{\text{ada}}(T) + \mathcal{O}(1)\mathrm{diam}_\infty(\mathcal{X}) r \frac{\left(\sum_{j=1}^d \sigma_j^{2/3}\right)^{3/2}}{n\varepsilon}\log\frac{d}{\delta}.$$

**Algorithm 3** Private Second Moment Estimation

**Require:** Dataset $\mathcal{S} = (z_1,\ldots,z_n) \in \mathcal{Z}^n$, number of iterations $T$, privacy parameters $\varepsilon, \delta$;
Set $\Delta = 1$ and $S = [d]$
**for** $t = 1$ to $T$ **do**
  **for** $j \in S$ **do**
    $\rho_t \leftarrow 4r\Delta\log r$
    $\hat{\sigma}_{t,j}^2 \leftarrow \frac{1}{n}\sum_{i=1}^n \min(z_{i,j}^2, \rho_t^2) + \xi_{t,j}$ where $\xi_{t,j} \sim \mathsf{N}(0, \rho_t^4 T^2 d\log(T/\delta)/n^2\varepsilon^2)$
    **if** $\hat{\sigma}_{t,j} \ge 2^{-t-1}$ **then**
      $\hat{\sigma}_j \leftarrow 2^{-t}$
      $S \leftarrow S \setminus \{j\}$
    **end if**
  **end for**
  $\Delta \leftarrow \Delta/2$
**end for**
**for** $j \in S$ **do**
  $\hat{\sigma}_j \leftarrow 2^{-T}$
**end for**
**Return:** $(\hat{\sigma}_1,\ldots,\hat{\sigma}_d)$

## 5. Lower bounds for private optimization

To give a more complete picture of the complexity of private stochastic optimization, we now establish (nearly) sharp lower bounds, which in turn establish the minimax optimality of PAGAN and PASAN. We establish this in two parts, reflecting the necessary dependence on geometry in the problems (Levy & Duchi, 2019): in Section 5.1, we show that PAGAN achieves optimal complexity for minimization over $\mathcal{X}_\infty = \{x \in \mathbb{R}^d : \|x\|_\infty \le 1\}$. Moreover, in Section 5.2 we show that PASAN achieves optimal rates in the Euclidean case, that is, for domain $\mathcal{X}_2 = \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$.

As one of our foci here is for data with varying norms, we prove lower bounds for sub-Gaussian data—the strongest setting for our upper bounds. In particular, we shall consider linear functionals $F(x; z) = z^T x$, where the entries $z_j$ of $z$ satisfy $|z_j| \le \sigma_j$ for a prescribed $\sigma_j$; this is sufficient for the data $Z$ to be $\frac{\sigma_j^2}{4}$-sub-Gaussian (Vershynin, 2019). Moreover, our upper bounds are conditional on the observed sample $\mathcal{S}$, and so we focus on this setting in our lower bounds, where $|g_j| \le \sigma_j$ for all subgradients $g \in \partial F(x; z)$ and $j \in [d]$.

### 5.1. Lower bounds for $\ell_\infty$-box

The starting point for our lower bounds for stochastic optimization over $\mathcal{X}_\infty$ is the following lower bound for the problem of estimating the sign of the mean of a dataset. This will then imply our main lower bound for private optimization. We defer the proof of this result to Appendix D.1.

**Proposition 1.** *Let $M$ be $(\varepsilon,\delta)$-DP and $\mathcal{S} = (z_1,\ldots,z_n)$ where $z_i \in \mathcal{Z} = \{z \in \mathbb{R}^d : |z_j| \le \sigma_j\}$. Let $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$ be the mean of the dataset $\mathcal{S}$. If $\sqrt{d}\log d \le n\varepsilon$, then*

$$\sup_{\mathcal{S}\in\mathcal{Z}^n} \mathbb{E}\left[\sum_{j=1}^d |\bar{z}_j| 1\{\mathrm{sign}(M_j(\mathcal{S})) \ne \mathrm{sign}(\bar{z}_j)\}\right] \ge \frac{(\sum_{j=1}^d \sigma_j^{2/3})^{3/2}}{n\varepsilon\log^{5/2} d}.$$

We can now use this lower bound to establish a lower bound for private optimization over the $\ell_\infty$-box by an essentially straightforward reduction. Consider the problem

$$\operatorname*{minimize}_{x\in\mathcal{X}_\infty} f(x;\mathcal{S}) := -\frac{1}{n}\sum_{i=1}^n x^T z_i = -x^T\bar{z},$$

where $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$ is the mean of the dataset. Letting $x_{\mathcal{S}}^\star \in \mathrm{argmin}_{x\in\mathcal{X}_\infty} f(x;\mathcal{S})$, we have the following result.

**Theorem 4.** *Let $M$ be $(\varepsilon,\delta)$-DP and $\mathcal{S} \in \mathcal{Z}^n$, where $\mathcal{Z} = \{z \in \mathbb{R}^d : |z_j| \le \sigma_j\}$. If $\sqrt{d}\log d \le n\varepsilon$, then*

$$\sup_{\mathcal{S}\in\mathcal{Z}^n} \mathbb{E}\left[f(M(\mathcal{S});\mathcal{S}) - f(x^\star_{\mathcal{S}};\mathcal{S})\right] \ge \frac{(\sum_{j=1}^d \sigma_j^{2/3})^{3/2}}{n\varepsilon \log^{5/2} d}.$$

**Proof** For a given dataset $\mathcal{S}$, the minimizer $x^\star_j = \text{sign}(\bar{z}_j)$. Therefore for every $x$ we have

$$\begin{aligned} f(x;\mathcal{S}) - f(x^\star;\mathcal{S}) &= \|\bar{z}\|_1 - x^T\bar{z} \\ &\ge \sum_{j=1}^d |\bar{z}_j| 1\{\text{sign}(x_j) \neq \text{sign}(\bar{z}_j)\}. \end{aligned}$$

As $\text{sign}(M(\mathcal{S}))$ is $(\varepsilon,\delta)$-DP by post-processing, the claim follows from Proposition 1 by taking expectations. □

Recalling the upper bounds that PAGAN achieves in Section 3.2, Theorem 4 establishes the tightness of these bounds to within logarithmic factors.

### 5.2. Lower bounds for $\ell_2$-ball

Having established PAGAN's optimality for $\ell_\infty$-box constraints, in this section we turn to proving lower bounds for optimization over the $\ell_2$-ball, which demostrate the optimality of PASAN. The lower bound builds on the lower bounds of Bassily et al. (2014). Following their arguments, let $\mathcal{X}_2 = \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$ and consider the problem

$$\underset{x\in\mathcal{X}_2}{\text{minimize}}\, f(x;\mathcal{S}) := -\frac{1}{n}\sum_{i=1}^n x^T z_i = -x^T\bar{z}.$$

The following bound follows by appropriate re-scaling of the data points in Theorem 5.3 in (Bassily et al., 2014)

**Proposition 2.** *Let $M$ be $(\varepsilon,\delta)$-DP and $\mathcal{S} = (z_1,\ldots,z_n)$ where $z_i \in \mathcal{Z} = \{z \in \mathbb{R}^d : \|z\|_\infty \le \sigma_j\}$. Then*

$$\sup_{\mathcal{S}\in\mathcal{Z}^n} \mathbb{E}\left[f(M(\mathcal{S});\mathcal{S}) - f(x^\star_{\mathcal{S}};\mathcal{S})\right] \ge \min\left(\sigma\sqrt{d}, \frac{d\sigma}{n\varepsilon}\right).$$

Using Proposition 2, we can establish the tight lower bounds—to within logarithmic factors—for PASAN (Section 3). We defer the proof to Appendix D.3.

**Theorem 5.** *Let $M$ be $(\varepsilon,\delta)$-DP and $\mathcal{S} = (z_1,\ldots,z_n)$ where $z_i \in \mathcal{Z} = \{z \in \mathbb{R}^d : |z_j| \le \sigma_j\}$. If $\sqrt{d} \le n\varepsilon$, then*

$$\sup_{\mathcal{S}\in\mathcal{Z}^n} \mathbb{E}\left[f(M(\mathcal{S});\mathcal{S}) - f(x^\star_{\mathcal{S}};\mathcal{S})\right] \ge \frac{\sum_{j=1}^d \sigma_j}{n\varepsilon\log d}.$$

## 6. Experiments

We conclude the paper with several experiments to demonstrate the performance of PAGAN and PASAN algorithms. We perform experiments both on synthetic data, where we may control all aspects of the experiment, and a real-world example training large-scale private language models.

### 6.1. Regression on Synthetic Datasets

If our PAGAN algorithm indeed captures the aspects of AdaGrad and other adaptive methods, we expect it to outperform other private stochastic optimization methods at the least in those scenarios where AdaGrad improves upon stochastic gradient methods—as basic sanity check. To that end, in our first collection of experiments, we compare PAGAN against standard implementations of private AdaGrad and SGD methods. We also compare our method against Projected DP-SGD (PDP-SGD), a recent method proposed by Zhou et al., which projects the noisy gradients into a low-dimensional subspace, given by the top $k$ eigenvectors of the second moment of gradients.

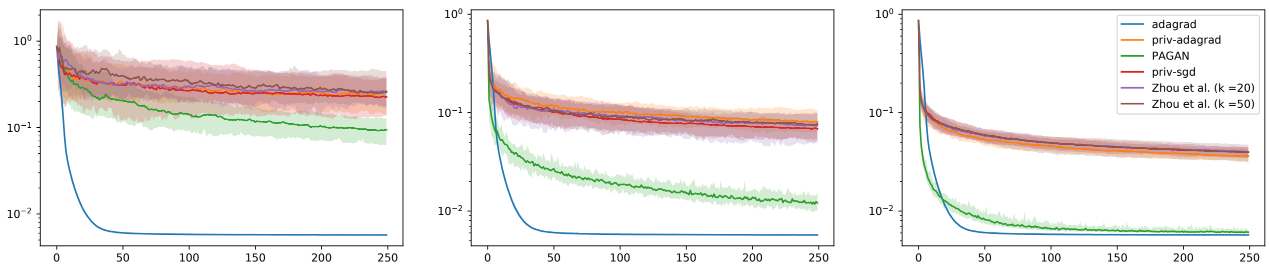
We consider an absolute regression problem with data $z_i = (a_i, b_i) \in \mathbb{R}^d \times \mathbb{R}$ and loss $F(x; a_i, b_i) = |\langle a_i, x\rangle - b_i|$. Given $n$ datapoints $(a_1,b_1),\ldots,(a_n,b_n)$, we wish to

$$\text{minimize}\;\; f(x) = \frac{1}{n}\sum_{i=1}^n |\langle a_i, x\rangle - b_i|. \tag{11}$$

We construct our dataset by first drawing a (population) optimal $x^\star \sim \mathsf{Uni}\{-1,1\}^d$, then sampling $a_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \text{diag}(\sigma)^2)$ for a vector $\sigma \in \mathbb{R}^d_+$. We set $b_i = \langle a_i, x^\star\rangle + \xi_i$ for noise $\xi_i \overset{\text{iid}}{\sim} \mathsf{Lap}(0,\tau)$, where $\tau \ge 0$.

We compare six (divided into three pairs) algorithms in this experiment. The first are non-private SGD and AdaGrad. The second are the naive implementations of private SGD (PASAN) and AdaGrad (PAGAN), Algorithm 1 and Algorithm 2, respectively, with $A_k = I$. Finally, we run PAGAN with the optimal diagonal matrix scaling $A_k$ as we derive in Section 3.2. In our experiments, we use the parameters $n = 5000$, $d = 100$, $\sigma_j = j^{-3/2}$, $\tau = 0.01$, and the batch size for all methods is $b = 70$. As optimization methods are sensitive to stepsize choice in general (even non-privately (Asi & Duchi, 2019)), we run each method with different values of initial stepsize in $\{0.005, 0.01, 0.05, 0.1, 0.15, 0.2, 0.4, 0.5, 1.0\}$ to find the best stepsize value. Then we run each method $T = 30$ times and report the median of the loss as a function of the iterate with 95% confidence intervals.

Figure 1 demonstrates the results of this experiment. Each plots the loss of the methods against iteration count in various privacy regimes. In the high-privacy setting (Figure 1(a)), we see that the performance of all private methods is worse than the non-private algorithms, though PAGAN (Alg. 2) seems to be outperforming other algorithms. As we increase the privacy parameter—reducing privacy preserved—we see that PAGAN quickly starts to enjoy faster convergence, resembling non-private AdaGrad. In contrast, the standard implementation of private AdaGrad—even in the moderate privacy regime with $\varepsilon = 4$—appears to obtain the slower convergence of SGD rather than the Adaptive




(a) (b) (c)

**Figure 1.** Sample loss as a function of the iterate for various optimization methods for synthetic absolute regression problem (11) with varying privacy parameters $\varepsilon$. (a) $\varepsilon = 0.1$. (b) $\varepsilon = 1$. (c) $\varepsilon = 4$.

methods. This is consistent with the predictions our theory makes: the isometric Gaussian noise addition that standard private stochastic gradient methods (e.g. PASAN and variants) employ eliminates the geometric properties of gradients (e.g., sparsity) that adaptive methods can leverage—indeed, must leverage (Levy & Duchi, 2019)—for improved convergence.

### 6.2. Training Private Language Models on WikiText-2

Following our simulation results, we study the performance of PAGAN for fitting a next word prediction model. Here, we train a variant of a recurrent neural network with Long-Short-Term-Memory (LSTM) (Hochreiter & Schmidhuber, 1997) on the WikiText-2 dataset (Merity et al., 2017), which is split into train, validation, and test sets. We further split the train set to 59,674 data points, where each data point has 35 tokens. The input data to the model consists of a one-hot-vector $x \in \{0, 1\}^d$, where $d = 8{,}000$. The first 7,999 coordinates correspond to the most frequent tokens in the training set, and the model reserves the last coordinate for unknown/unseen tokens. We train a full network, which consists of a fully connected embedding layer $x \mapsto Wx$ mapping to 120 dimensions, where $W \in \mathbb{R}^{120\times 8000}$; two layers of LSTM units with 120 hidden units, which then output a vector $h \in \mathbb{R}^{120}$; followed by a fully connected layer $h \mapsto \Theta h$ expanding the representation via $\Theta \in \mathbb{R}^{8000\times 120}$; and then into a softmax layer to emit next-word probabilities via a logistic regression model. The entire model contains 2,160,320 trainable parameters.

We use Abadi et al.'s moments accountant analysis to track the privacy losses of each of the methods. In each experiment, for PAGAN we use gradients trained for one epoch on a held-out dataset (a subset of the WikiText 103 dataset (Merity et al., 2017) which does not intersect with WikiText-2) to estimate moment bounds and gradient norms, as in Section 4; these choices—while not private—reflect the common practice that we may have access to public data that provides a reasonable proxy for the actual moments on our data. Moreover, our convergence guarantees in Section 3 are robust in the typical sense of stochastic gradient methods (Nemirovski et al., 2009), in that mis-specifying the moments by a multiplicative constant factor yields only constant factor degradation in convergence rate guarantees, so we view this as an acceptable tradeoff in practice. It is worth noting that we ignore the model trained over the public data and use that one epoch solely for estimating the second moment of gradients.

In our experiments, we evaluate the performance of the trained models with validation- and test-set perplexity. While we propose adaptive algorithms, we still require hyperparameter tuning, and thus perform a hyper-parameter search over three algorithm-specific constants: a multiplier $\alpha \in \{0.1, 0.2, 0.4, 0.8, 1.0, 10.0, 50.0\}$ for step-size, mini-batch size $b \in \{50, 100, 150, 200, 250\}$, and projection threshold $B \in \{0.05, 0.1, 0.5, 1.0\}$. Each run of these algorithms takes $< 4$ hours on a standard workstation without any accelerators. We trained the LSTM model above with PAGAN, comparing its performance to DP-SGD (Abadi et al., 2016) and PDP-SGD proposed by Zhou et al.. To do a fair comparison, and similar to PAGAN, we use the second moment estimate cmpouted over the public data for PDP-SGD as well, and do not update this estimate over iterates. Finally, we also include completely non-private SGD and AdaGrad for refrence. For each of the privacy levels $\varepsilon \in \{1, 3\}$ we consider, we present the performance of each algorithm in terms of best validation set and test-set perplexity in Figure 2 and Table 1. In our results, LargeAux and SmallAux correspond to two different sizes for the public dataset, where in the former the size of public dataset is equal to the training set, and in the latter, it is equal to one fifth of the size of training set. The code is available online[2].

We highlight a few messages present in Figure 2. First, PA-

[2] `https://github.com/apple/ml-private-adaptive-gradient-methods`

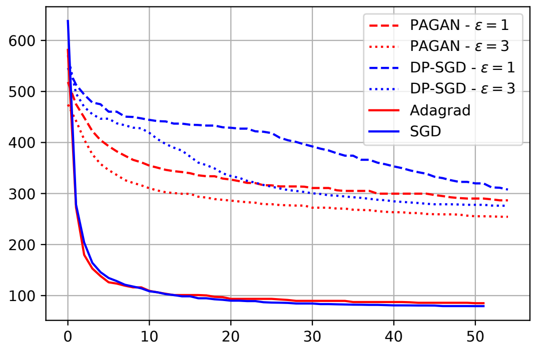



**Figure 2.** Minimum validation perplexity versus training epoch for PAGAN and the standard differentially private stochastic gradient method (DP-SGD) (Abadi et al., 2016), varying privacy levels $\varepsilon \in \{1, 3\}$.

**Table 1.** Test perplexity error of different methods. For reference, non-private SGD and AdaGrad (without clipping) achieve 75.45 and 79.74, respectively.

| Algorithm | $\varepsilon = 3$ | $\varepsilon = 1$ |
|---|---|---|
| DP-SGD (Abadi et al., 2016) | 240.32 | 287.81 |
| **SmallAux** | | |
| Zhou et al. with $k = 50$ | 280.19 | – |
| PAGAN | **236.03** | **265.35** |
| **LargeAux** | | |
| PAGAN | **228.72** | **251.01** |

GAN consistently outperforms the non-adaptive methods—though all allow the same hyperparameter tuning—at all privacy levels, excepting the non-private $\varepsilon = +\infty$, where PAGAN without clipping is just AdaGrad and its performance is comparable to the non-private stochastic gradient method. Certainly, there remain non-negligible gaps between the performance of the private methods and non-private methods, but we hope that this is a step at least toward effective large-scale private optimization and modeling.

## 7. Acknowledgement

The authors like to thank Vitaly Feldman and Jalaj Upadhyay for helpful discussions in the process of preparing this paper, as well as Daniel Levy for comments on an earlier draft. Part of this work was done while HA and AF were interning at Apple.

## References

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

Asi, H. and Duchi, J. C. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019. URL https://arXiv.org/abs/1810.05633.

Asi, H., Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: Optimal rates in $\ell_1$ geometry. *arXiv:2103.01516 [cs.LG]*, 2021.

Barber, R. F. and Duchi, J. C. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv:1412.4451 [math.ST]*, 2014.

Bartlett, P. L., Hazan, E., and Rakhlin, A. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems 20*, 2007.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th Annual Symposium on Foundations of Computer Science*, pp. 464–473, 2014.

Bassily, R., Feldman, V., Talwar, K., and Thakurta, A. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems 32*, 2019.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

Duchi, J. C. Introductory lectures on stochastic convex optimization. In *The Mathematics of Data*, IAS/Park City Mathematics Series. American Mathematical Society, 2018.

Duchi, J. C. Information theory and statistics. Lecture Notes for Statistics 311/EE 377, Stanford University, 2019. URL http://web.stanford.edu/class/stats311/lecture-notes.pdf. Accessed May 2019.

Duchi, J. C., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

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

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

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

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology (EUROCRYPT 2006)*, 2006a.

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

Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: Optimal rates in linear time. In *Proceedings of the Fifty-Second Annual ACM Symposium on the Theory of Computing*, 2020.

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

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Kairouz, P., Ribero, M., Rush, K., and Thakurta, A. Dimension independence in unconstrained private ERM via adaptive preconditioning. *arXiv:2008.06570 [cs.LG]*, 2020.

Levy, D. and Duchi, J. C. Necessary and sufficient geometries for gradient methods. In *Advances in Neural Information Processing Systems 32*, 2019. URL https://arxiv.org/abs/1909.10455.

McMahan, B. and Streeter, M. Adaptive bound optimization for online convex optimization. In *Proceedings of the Twenty Third Annual Conference on Computational Learning Theory*, 2010.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. In *Proceedings of the Fifth International Conference on Learning Representations*, 2017.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Pichapati, V., Suresh, A. T., Yu, F. X., Reddi, S. J., and Kumar, S. AdaCliP: Adaptive clipping for private SGD, 2020.

Smith, A. and Thakurta, A. (Nearly) optimal algorithms for private online learning in full-information and bandit settings. In *Advances in Neural Information Processing Systems 26*, 2013a.

Smith, A. and Thakurta, A. Differentially private feature selection via stability arguments, and the robustness of the Lasso. In *Proceedings of the Twenty Sixth Annual Conference on Computational Learning Theory*, pp. 819–850, 2013b. URL http://proceedings.mlr.press/v30/Guha13.html.

Talwar, K., Thakurta, A., and Zhang, L. Nearly optimal private lasso. In *Advances in Neural Information Processing Systems 28*, pp. 3025–3033, 2015.

Vershynin, R. *High Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2019.

Yu, D., Zhang, H., Chen, W., and Liu, T.-Y. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=7aogOj_VYO0.

Zhou, Y., Wu, Z. S., and Banerjee, A. Bypassing the ambient dimension: Private SGD with gradient subspace identification. *arXiv:2007.03813 [cs.LG]*, 2020.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.