(Nearly) Dimension Independent Private ERM with AdaGrad Rates via Publicly Estimated Subspaces

Peter Kairouz KAIROUZ@GOOGLE.COM
Google Research

Mónica Ribero MRIBERO@UTEXAS.EDU
The University of Texas at Austin[*]

Keith Rush KRUSH@GOOGLE.COM
Google Research

Abhradeep Thakurta ATHAKURTA@GOOGLE.COM
Google Research

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We revisit the problem of empirical risk minimziation (ERM) with differential privacy. We show that noisy AdaGrad, given appropriate knowledge and conditions on the subspace from which gradients can be drawn, achieves a regret comparable to traditional AdaGrad plus a well-controlled term due to noise. We show a convergence rate of $O(\mathbf{Tr}\left(G_T\right)/T)$, where G_T captures the geometry of the gradient subspace. Since $\mathbf{Tr}\left(G_T\right) = O(\sqrt{T})$ we can obtain faster rates for convex and Lipschitz functions, compared to the $O(1/\sqrt{T})$ rate achieved by known versions of noisy (stochastic) gradient descent with comparable noise variance. In particular, we show that if the gradients lie in a known constant rank subspace, and assuming algorithmic access to an envelope which bounds decaying sensitivity, one can achieve faster convergence to an excess empirical risk of $\widetilde{O}(1/\varepsilon n)$, where ε is the privacy budget and n the number of samples. Letting p be the problem dimension, this result implies that, by running noisy Adagrad, we can bypass the DP-SGD bound $\widetilde{O}(\sqrt{p}/\varepsilon n)$ in $T = (\varepsilon n)^{2/(1+2\alpha)}$ iterations, where $\alpha \geq 0$ is a parameter controlling gradient norm decay, instead of the rate achieved by SGD of $T = \varepsilon^2 n^2$. Our results operate with general convex functions in both constrained and unconstrained minimization.

Along the way, we do a perturbation analysis of noisy AdaGrad, which is of independent interest. Our utility guarantee for the private ERM problem follows as a corollary to the regret guarantee of noisy AdaGrad.

Keywords: Differential privacy, convex optimization, adaptive gradient descent

1. Introduction

Differentially private convex optimization is a fundamental problem for machine learning practitioners. Empirical Risk Minimization (ERM) in particular is foundational in most learning tasks, many of which are posed over datasets with sensitive information that can be leaked through model parameters (Fredrikson et al., 2015; Wu et al., 2016; Shokri et al., 2017). Differential privacy (Dwork et al., 2006b,a) has therefore been adopted in optimization when training machine learning models to limit user data exposure.

In current applications, models are usually many times over-parametrized. This is a major problem for private settings, where the the optimal model θ^* cannot be released, but we must release

[*] Work done while author was interning at Google.

rather a private model θ_{priv} . For a model dimensionality of p, a naive privatization incurs an excess empirical risk with *lower bound* linear in $\sqrt{p}$ (Bassily et al., 2014).

In this paper we propose noisy-AdaGrad, a novel optimization algorithm that leverages gradient pre-conditioning and knowledge of the subspace in which gradients lie to recover AdaGrad regret rates (i.e., $O\left(\frac{\mathbf{Tr}\,(G_T)}{T}\right)$, where G_T is the adaptive pre-conditioner defined in Equation 15 in Algorithm 1), and dimension independent excess risk bounds. We propose a general framework to study noisy versions of Adaptive pre-conditioning (a.k.a. AdaGrad (McMahan and Streeter, 2010; Duchi et al., 2011; Hazan, 2019)). Further, our analysis identifies a simple condition under which AdaGrad-style rates can be achieved in the differentially-private ERM problem: that of oracle access to a constant-factor envelope of the maximum gradient norm across data samples as training progresses (See Definition 2).

Each of the assumed pieces of input data is well-justified in practice. While it is possible to privately estimate the exact gradient subspace under a strong assumption on the linear dependency of gradients (Singhal and Steinke, 2021); this is hard to verify in practice for the general convex optimization setting. Yet, knowledge about the gradient subspace is often available through public data that is easily accessible, for instance through "opt-in" users (Beimel et al., 2013; Xin and Jaakkola, 2014; Alon et al., 2019; Zhou et al., 2020c). For example, for Generalized Linear Models (GLMs) this subspace corresponds to the feature space determined by the column space of the data matrix (see for example Song et al. (2020)). Knowledge of the maximum gradient norm can be had by observing the training procedure, and gradient norms for many classes of well-studied problems decay uniformly for all data samples, e.g. those studied in Bassily et al. (2018); Ma et al. (2018). We leave it as an open problem to design a differentially private algorithm for computing this envelope.

1.1. Notation

We use $\|\cdot\|_2$ to denote the ℓ_2 norm of a vector. We denote by $\lambda_i(A)$ the i-th largest eigenvalue of matrix A, $\lambda_{\text{min}>0}(A)$ the smallest positive eigenvalue of A, and $\|\cdot\|_{op}$ to denote the operator norm of a matrix, defined as $\|A\|_{op} = \max\{|\lambda_i| : \lambda_i$ eigenvalue value of A $\}$. We will occasionally use restricted inner products with respect to a subspace U, denoted $\langle\cdot,\cdot\rangle|_U$, as a shorthand for $\langle\cdot, P_U\cdot\rangle$ where P_U denotes orthogonal projection in ℓ_2 onto the subspace U.

$\|\cdot\|_A$ denotes the Mahalanobis seminorm, defined as $\|\cdot\|_A = \sqrt{\langle\cdot, A\cdot\rangle}$ for A symmetric and positive-semidefinite. The dual norm to a norm $\|\cdot\|$ is defined as $\|x\|^* = \sup_{y:\|y\|\leq 1}\langle x, y\rangle$. The dual norm of the above matrix norm is given by $\|x\|_A^* = \|x\|_{A^{-1}}$. We use $[T]$ to denote the time interval $[T] = \{1, ..., T\}$. Finally, in the considered setting, f_t will be constant over time, so we will denote $f_t = f$, and to simplify notation we use ∇_t to denote $\nabla f(\theta_t)$.

1.2. Problem Definition

Let $D = \{d_1, \ldots, d_n\}$ be a given data set drawn from a distribution $\mathcal{P}$, $\ell(\cdot, d)$ a map defining the loss on data point d, and an objective function $\mathcal{L}(\theta; D) = \frac{1}{n}\sum_{i=1}^{n}\ell(\theta; d_i)$. The goal is to design an (ε, δ)-differentially private algorithm $\mathcal{A}_{\mathsf{priv}}$ that outputs a model $\theta_{\mathsf{priv}} \in \mathcal{C} \subseteq \mathbf{R}^p$ that approximately solves the following optimization problem:

$$\min_{\theta\in\mathcal{C}} \mathcal{L}(\theta; D). \tag{1}$$

In terms of accuracy we consider the traditional excess empirical risk defined as follows:

$$\mathsf{Risk}(\theta_{\mathsf{priv}}) = \mathcal{L}(\theta_{\mathsf{priv}}; D) - \min_{\theta} \mathcal{L}(\theta; D). \tag{2}$$

Being consistent with the literature on private convex ERM, we will assume each of the loss functions $\ell(\theta; d)$ is convex and L-Lipschitz in its first parameter w.r.t. the ℓ_2-norm.

Online convex optimization: To solve the private ERM problem, we will model it along the lines of online convex optimization (Hazan, 2019; Shalev-Shwartz et al., 2011). First, we will propose a noise-tolerant algorithm for the traditional online convex optimization, and then use that algorithm and its analysis to design a differentially private ERM algorithm with a bound on the excess empirical risk. We use the well-known *online to batch conversion* (Hazan, 2019) to translate the regret guarantee for an online algorithm to that of excess empirical risk of a convex optimization problem.

We adhere to the standard regret minimization setting of traditional online learning (Hazan, 2019). Formally, given a sequence of loss functions $\mathcal{F} = \{f_1, \ldots, f_T\}$ (with each $f_t : \mathbf{R}^p \to \mathbf{R}$) arriving online, the objective is to design an algorithm to ouput a sequence of models $\{\theta_1, \ldots, \theta_T\}$ such that the following is minimized:

$$\mathsf{Regret}_T(\mathcal{F}; \mathcal{A}) = \frac{1}{T}\sum_{t=1}^{T} f_t(\theta_t) - \min_{\theta} \frac{1}{T}\sum_{t=1}^{T} f_t(\theta) \tag{3}$$

Throughout this paper, we will call an algorithm $\mathcal{A}$ to be a "low-regret" algorithm if it outputs a sequence of models such that the regret in (3) is $o(1)$. In principle each of the the loss functions $f_t \in \mathcal{F}$ can be chosen adaptively (and adversarially) based on the models output so far, i.e., $\theta_1, \ldots, \theta_{t-1}$. In this paper we will primarily focus on the *convex setting*, where the loss functions in $\mathcal{F}$ are assumed to be convex in its first parameter. Furthermore, we will assume that the loss functions are Lipschitz bounded, i.e., $\forall \theta \in \mathbf{R}^p, f \in \mathcal{F} : \|\partial_\theta f(\theta)\|_2 \leq L$.

1.3. Our Contributions

Our main contribution is to obtain dimension independent excess risk bounds for differentially private ERM through adaptive pre-conditioning. Our contributions can be stated as follows.

A noise tolerant AdaGrad-style algorithm: We design Noisy-AdaGrad ($\mathcal{A}_{\mathsf{noisy-AdaGrad}}$, Algorithm 1), a novel noise tolerant optimization algorithm with adaptive preconditioning that, under appropriate parameter selection, satisfies $(\varepsilon, \delta)-$ differential privacy. The algorithm differs from AdaGrad in three main respects: (1) it uses a gradient perturbed with Gaussian noise; (2) the preconditioner is updated with clean gradients and then perturbed with a noise matrix drawn from the Gaussian Orthogonal Ensemble; (3) we introduce a projection step that is intended to maintain the trajectory of the descent algorithm in the gradients' subspace. We assume (noisy) oracle access $\widetilde{V}_t$ to V_t, the orthogonal matrix whose columns span the gradient subspace at iteration t, and before taking a gradient step we project the update step using $\widetilde{V}_t\widetilde{V}_t^T$. In over-parameterized regimes, this step allows us to significantly decrease the effect of noise, when gradients lie in a low rank subspace, a common characteristic in high-dimensional problems (Agarwal et al., 2019; Gur-Ari et al., 2018). In practice, this subspace can be computed from public data (Beimel et al., 2013; Alon et al., 2019; Zhou et al., 2020c).

Dimension independent and AdaGrad-style regret rates with noisy gradient subspace: We provide a dimension independent low regret bound for $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$ in Theorem 3.1 that recovers

AdaGrad rates given access to a simple sensitivity oracle, improving over previous gradient descent rates. This is, to the best of our knowledge, the first work to analyse a noisy version of full matrix AdaGrad where both the gradient and pre-conditioner are independently noised. Our main regret bound is the following.

Theorem 1.1 (Informal version of Theorem 3.1) *Let V_t be the orthogonal matrix whose column space is the tracked gradient subspace up to time t, and $\widetilde{V}_t$ an approximation returned by an oracle. Let γ be a bound on the subspaces' principal angle difference, i.e., $\|V_tV_t^T - \widetilde{V}_t\widetilde{V}_t^T\|_{op} \leq \gamma$. Let L be the gradient ℓ_2−norm bound, C the diameter of the constraint set $\mathcal{C}$, and assume $L = C = O(1)$. Letting $\sigma_b^2(t)$ be the gradient noise variance, then running $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$ on $\mathcal{L}(\theta; D)$ for T iterations we get*

$$\mathbb{E}[\mathsf{Regret}_T(\mathcal{F}; \mathcal{A}_{\mathsf{noisy-AdaGrad}})]$$
$$\leq O\left(\mathbb{E}\left[\sqrt{\mathsf{Regret}_T(\mathcal{F}; \mathcal{A}_{\mathsf{AdaGrad}})^2 + \frac{\mathbf{Tr}\left(G_T\right)\sum_{t=1}^{T}\sigma_b^2(t)\mathbf{Tr}\left(G_t^{-1}\right)}{T^2}} + \gamma\right]\right) \tag{4}$$

Our result can be interpreted as follows.

- The AdaGrad regret term in our bound reduces to $O(\mathbf{Tr}\left(G_T\right)/T)$, improving over SGD which achieves regret $O(1/\sqrt{T})$.
- The second term only depends on the gradient space dimension, dictated by the clean pre-conditioner G_t, unlike DP-SGD where this term linear in $\sqrt{p}$. By incorporating a projection to gradient subspace, we obtain dimension independence. Furthermore, we show in Corollary 3 that by adapting the gradient noise at each iteration to be similar in scale to the gradient, we obtain faster rates: again, $O(\mathbf{Tr}\left(G_T\right)/T)$.
- An additive factor γ accounting for subspace estimation mismatch. We use Davis-Kahan $\sin(\theta)$ theorem to bound errors due to rotation of the problem space.

This analysis can be of independent interest and is crucial for any differential privacy guarantee, since the pre-conditioner used for AdaGrad contains the full history of gradients. In practice, the alternative to private AdaGrad has been to update the pre-conditioner with noisy gradients and rely on the post-processing property of differential privacy.

Dimension independent excess empirical risk bounds for private AdaGrad with public data: Our third result is to derive an excess risk bound that addresses the case where noise parameters are set to provide differential privacy. Our algorithm uses public data to compute the projection matrix $\widetilde{V}_t\widetilde{V}_t^T$ that forces the descent algorithm to stay in the gradient subspace, and the analysis derives a dimension independent excess risk bound for differentially private AdaGrad. Setting gradient and pre-conditioner noise variances appropriately, $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$ is differentially private and we obtain an excess risk of $\frac{1}{\varepsilon n}$ in $T = (\varepsilon n)^{2/(1+2\alpha)}$, where α controls the decay rate of gradients norm. This means that if $\alpha > 0$ we reach the excess risk faster than (P)DP-SGD, that has running time $T = \varepsilon^2 n^2$. Additionally, we include in Lemma 5 the non-trivial computation of the pre-conditioner's sensitivity.

Corollary 1.2 (Informal version of Corollary 6) *Given a minimization problem where the subspace spanned by gradients has bounded rank $k < p$, running $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$ with appropriate noise parameters is (ε, δ)- differentially private and the expected excess risk of $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$ is* $O\left(\frac{\sqrt{k \log(1/\delta)}}{\varepsilon n}\right)$.

Dimension independent excess empirical risk for DP-SGD without public data: Finally, we extend previous results from Song et al. (2020) and show that for unconstrained minimization, DP-SGD, without any gradient subspace knowledge, is enough to obtain an excess risk bound of $O\left(\frac{\sqrt{k \log(1/\delta)}}{\varepsilon n}\right)$ independent of dimension.

Theorem 1.3 (Informal version of Theorem 5.1) *Let $\theta_0 = \mathbf{0}$ be the initial point of $\mathcal{A}_{\mathsf{DP-GD}}$. Let $\theta^* = \arg\min_{\theta \in \mathbb{R}^p}$ and $M = VV^T$ be the projector to the gradient eigenspace. Letting L be the gradient ℓ_2−norm bound, setting the constraint set $\mathcal{C} = \mathbb{R}^p$, and running $\mathcal{A}_{\mathsf{DP-GD}}$ on $\mathcal{L}(\theta; D)$ for $T = \varepsilon^2 n^2$ and appropriate choice of learning rate,*

$$\mathbb{E}[\mathcal{L}(\theta_{\mathsf{priv}}; D)] - \mathcal{L}(\theta^*; D) \leq \frac{L\|\theta^*\|_M \sqrt{1 + 2rank(M) \log(1/\delta)}}{\varepsilon n} \tag{5}$$

The concurrent work of Zhou et al. (2020c) studies a similar problem, incorporating Gaussian noise to privatize the gradient and a publicly available projection to gradient subspace to achieve dimension independence for differentially private SGD. The methods of proof, however, are significantly different, and the results presented here conditionally achieve faster convergence.

1.4. Techniques

In this section we describe the main techniques leveraged to obtain the above results. Our contributions are structured as follows: we first analyse the regret of noisy-Adagrad, introduced in Algorithm 1. Second, we use this analysis to provide excess risk bounds. Third, and finally, we translate these results for the case when the noise in $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$ is intended to provide privacy.

Noisy-Adagrad. The first part of the proof of Theorem 3.1 bounds the regret of our noisy-AdaGrad algorithm, relying on matrix perturbation analysis. The proof follows standard convexity arguments to bound the regret with a linear approximation, resulting in the four terms in Equation 6. Although this expression is analogous to the original AdaGrad regret bound, the analysis in our case is much more involved due to gradient noise b_t and pre-conditioner noise B_t. Given that we need our bound in terms of the original pre-conditioner G_t, we introduce several findings and key lemmas that allow us to achieve this. We summarize them below.

Equation 6 is composed of four terms that can be independently bounded: a potential drop term that captures closeness to the optimum, a gradient noise norm term, a gradient norm term, and a projection error term.

$$\begin{aligned} f\left(\frac{1}{T}\sum_t \theta_t\right) - f(\theta^*) \leq & \sum_t \frac{1}{2\eta T}\left(\|\theta_t - \theta^*\|^2_{H_t} - \|\theta_{t+1} - \theta^*\|^2_{H_t}\right) \\ & + \frac{\eta}{2T}\left(\mathbb{E}_{b_t}[\|b_t\|^2_{H_t^{-1}} | B_t]\right) + \frac{\eta}{2T}\left(\|\nabla_t\|^2_{H_t^{-1}}\right) + \frac{1}{T}\langle \nabla_t, (\theta_t - \theta^*)\rangle|_{C_t} \end{aligned} \tag{6}$$

for C_t the kernel of H_t intersected with the rowspace of G_t.

We briefly describe the additional difficulties of analyzing our algorithm compared to traditional AdaGrad or DP-SGD.

The first term, the potential drop involving matrix norms $\|\cdot\|_{H_t}$, is traditionally bounded using a telescoping argument, resulting in a $\mathbf{Tr}\,(G_t)$ term. Here we first need to manipulate this expression and rely on trace definition and properties, and the fact that B_t is zero mean to obtain a similar result in terms of G_t and not H_t.

To bound the second and third terms involving matrix norm $\|\cdot\|_{H_t^{-1}}$, we prove structural Lemma 12. This Lemma uses an analog of the Woodbury identity to calculate the restricted pseudoinverse of a sum of matrices (in this case the pre-conditioner matrix G_t, and the pre-conditioner noise B_t).

The last term makes use of the Davis-Kahan theorem (Theorem C.3) to bound the principal angle difference between two subspaces: this allows to measure how much signal is lost by projecting onto a perturbed subspace.

Achieving excess empirical risk guarantee. It is a well-known standard idea called *online to batch conversion* (Cesa-Bianchi et al., 2004; Shalev-Shwartz et al., 2009) to translate the regret guarantee for an online algorithm to that of excess empirical risk of a convex optimization problem.

Providing privacy. To set noise values, we compute the ℓ_2 sensitivity of gradients and pre-conditioner. Since individual data point loss functions ℓ are assumed $L-$Lipschitz, the sensitivity of the overall loss function's gradients can be bounded by $\frac{L}{n}$. The pre-conditioner sensitivity is more involved, since at each iteration t, it utilizes the full history of gradients. We show in Lemma 5 that it can be bounded by $L\sqrt{\frac{t}{n}}$. To the best of our knowledge, this is the first time the ℓ_2 sensitivity of the pre-conditioner is explicitly computed; previous private Adagrad results relied on the post-processing property of differential privacy, and used private gradients to update the pre-conditioner. This easy fix turns out to be inefficient since it adds bias to the pre-conditioner, slowing down the exploration advantage (large learning rates in unexplored directions) of the original AdaGrad algorithm.

Finally, relying on the Gaussian mechanism and strong composition of differential privacy, we show $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$ can be adapted for privacy and achieve an excess risk of $\frac{1}{\varepsilon n}$.

Using standard techniques, one can work with an ℓ_2-norm regularized loss and derive excess population risk guarantees (see Theorem 2 in Shalev-Shwartz et al. (2009)) .

1.5. Related Work

Differentially private ERM has been widely studied theoretically and empirically (Chaudhuri et al., 2011; Bassily et al., 2014; Song et al., 2013; Abadi et al., 2016; Bassily et al., 2019; McMahan et al., 2017; Wu et al., 2017; Iyengar et al., 2019; Pichapati et al., 2019; Thakkar et al., 2019; Feldman et al., 2020; Song et al., 2020). It was established by Bassily et al. (2014) that the excess risk in the constrained setting for any differentially private optimization algorithm over convex functions is lower bounded by $\Omega\left(\frac{\sqrt{p}}{\varepsilon n}\right)$.

Noisy versions of AdaGrad where the pre-conditioner is updated with information from noisy gradients have also been studied (Xie et al., 2020; Zhou et al., 2020a,b). In each of these cases the excess risk bounds depend on ambient dimension. When using pre-conditioning we aim at working in the intrinsic subspace of the problem; therefore our work is also tangential to differentially private and noisy subspace estimation, which have been a broad area of study (Dwork et al., 2014; Upadhyay and Upadhyay, 2020).

Closer to our contributions, recent work has attempted to eliminate the dependence in dimension by relaxing assumptions in the optimization setting (Song et al., 2020; Zhou et al., 2020c; Yu et al., 2021). Below we point out how our contributions differ from these contemporary works.

Comparison to Song et al. (2020): Song et al. (2020) relax the constrained optimization assumption and show that in the unconstrained setting it is possible to obtain a dimension independent bound for Generalized Linear Models (GLMs). The excess risk depends on the rank of the feature matrix, rather than the ambient dimension. First, our results extend the class of problems and show that this is true for general convex Lipschitz functions in theorem 5.1: Unconstrained DP-SGD excess risk depends only on rank (VV^T) where the columns of V span the gradient subspace. Although this bound depends on V, the algorithm is oblivious of V. Second, the advantage of Noisy-AdaGrad over these results is that the proof in Song et al. (2020) requires an apriori bound on the per-sample gradient norm ($L(t)$) at each time step $t \in [T]$ to set the optimal learning rate, and achieve the $O\left(\frac{1}{\sqrt{T}}\right)$ excess empirical risk, while Noisy-Ada only needs online access to $L(t)$ at iteration t to set the noise, and achieves an excess empirical risk of $O\left(\frac{\mathbf{Tr}\,(G_T)}{T}\right)$, for a constant learning rate. This is because the optimal learning rate depends only on the constraint set diameter. Compared to apriori knowledge of all $L(t)$ values for $t \in [T]$, online access is a more achievable assumption in practice, for example through adaptive clipping (Pichapati et al., 2019; Thakkar et al., 2019).

Comparison to Zhou et al. (2020c) and Yu et al. (2021): In a concurrent and independent work, Zhou et al. (2020c) provided dimension independent bounds via a variant of projected version of stochastic gradient descent (PDP-SGD), using public data to compute the projector to the gradient subspace. In another concurrent work, Yu et al. (2021) too relies on access to auxiliary data to learn the gradient subspace principal components. Their algorithm, Gradient Embedding Perturbation (GEP) uses the auxiliary data to express the gradient as a sum of the projection onto the principal components and a residual gradient. This allows them to privatize independently large and small components. Although improving over DP-SGD, their bound still has a dependence on ambient dimension since both components are used, but the magnitude of noise is significantly smaller in the residual subspace. In section 5 we show that public data is not necessary for DP-SGD to achieve dimension independence in the unconstrained setting, relaxing the public data assumption of Zhou et al. (2020c) and Yu et al. (2021), and obtaining a better result in this setting by removing the subspace mismatch term obtained by projections.

The main distinction of our line of work with this concurrent work is that we take full advantage of pre-conditioning and develop a careful analysis of it in the private/noisy setting. This is of utmost importance to bridge the gap between private and non-private optimization, and take a first step in this direction. While concurrent work also exploits the gradient subspace, they do not consider pre-conditioning the gradient.

In Corollary 3 we show pre-conditioning can translate to faster convergence under knowledge of $L(t)$, an envelope for individual data point gradient norms; while algorithms proposed by Zhou et al. (2020c), and Yu et al. (2021) converge as $\frac{1}{\sqrt{T}}$, Noisy-Ada converges as $\frac{\mathbf{Tr}\,(G_T)}{T}$ and $\frac{\mathbf{Tr}\,(G_T)}{T} = o\left(\frac{1}{\sqrt{T}}\right)$ in this setting (see Remark 4 for details). Setting constant noise $\sigma_b(t) = O(L)$, PDP-SGD requires $T_1 = (\varepsilon n)^2$, while Noisy-Ada requires $T_2 = (\varepsilon n)^{2/(1+2\alpha)}$ where measures the decay of gradients norm over the optimization process, thus $T_2 \leq T_1$.

Finally, the matrix perturbation analysis and techniques we develop can be of independent interest to other areas in learning theory like private principal component analysis.

2. Background

In this section we introduce the necessary tools for the analysis of our Noisy-AdaGrad algorithm. We start by introducing the traditional AdaGrad algorithm, followed by standard differential privacy definitions.

AdaGrad. AdaGrad (Adaptive Gradient Descent) (Duchi et al., 2011; McMahan and Streeter, 2010; Hazan, 2019) achieves low-regret for convex loss functions. One of the main features that separates AdaGrad from other online convex optimization algorithms like follow-the-regularized-leader, online gradient descent (Hazan et al., 2007), and online mirror descent (Ben-Tal and Nemirovski, 2001; Shalev-Shwartz et al., 2009) is the use of a gradient pre-conditioner. It allows much tighter regret guarantees if the gradients of the loss functions come from a constant (close to) low-rank subspace.

The original AdaGrad algorithm (Appendix A.1) proposes the following update with a convex constraint set $\mathcal{C}$

$$\theta_{t+1} = \arg\min_{\theta \in \mathcal{C}} \|\theta - (\theta_t - \eta G_t^{-1} \nabla_t)\|_{G_t}^2, \tag{7}$$

AdaGrad is derived by analyzing the optimal (strongly convex) regularization function to use in hindsight, that would minimize the regret of an online convex optimization algorithm. Concretely, consider the set of all strongly convex regularization functions with a fixed and bounded Hessian in the set

$$\mathcal{H} = \{X \in \mathbb{R}^{p \times p} : \mathbf{Tr}(X) \leq 1, X \succeq 0\} \tag{8}$$

AdaGrad achieves a regret bound that is within a constant factor of $2C$ of the regret achieved by the best, fixed pre-conditioner in hindsight. We formalize this in theorem 2.1

Theorem 2.1 *(Theorem 5.11. in Hazan (2019), originally Theorem 6 in McMahan and Streeter (2010) and Theorem 8 in Duchi et al. (2011)) Let $\{\theta_t\}$ be defined by Algorithm 2 with parameters $\eta = C$, where $C = \max_{\boldsymbol{u} \in \mathcal{K}} \|\boldsymbol{u} - \theta_0\|_2$, θ_0 is the initial value, and $\mathcal{K} \subseteq \mathbb{R}^p$ is a bounded convex compact set.*

Then for any $\theta^* \in \mathcal{K}$,

$$\mathsf{Regret}_T(\mathcal{F}; \mathcal{A}_{\mathsf{Adagrad}}) \leq 2C \sqrt{\min_{H \in \mathcal{H}} \sum_t \|\nabla_t\|_H^{*2}} \tag{9}$$

Differential Privacy. Originally proposed by Dwork et al. (2006a,b), differential privacy is a framework protecting single records in a database by bounding the probability of re-identifying any record from a query output. In this paper we limit ourselves to approximate differential privacy, and we rely on the Gaussian mechanism and Renyi composition theorem to provide these privacy guarantees (see Appendix A.2). Formally,

Definition 1 ((approximate) Differential Privacy (Dwork et al., 2006a,b)) *A randomized algorithm $\mathcal{A}$ that receives as input a dataset D is $(\varepsilon, \delta)-$ differentially private if, for any pair of neighboring datasets D and D'(Definition 9), and any set of events $\mathcal{S}$ in the range of $\mathcal{A}$,*

$$\mathbf{Pr}(\mathcal{A}(D) \in \mathcal{S}) \leq e^{\varepsilon}\mathbf{Pr}(\mathcal{A}(D') \in \mathcal{S}) + \delta,$$

where the probability is taken over the random coins of $\mathcal{A}$.

3. Analysis of Noisy-Adagrad

In this section we present and study a noisy version of AdaGrad (see Algorithm 1), where the adaptive pre-conditioner is perturbed with a matrix sampled from the Gaussian Orthonormal Ensemble (GOE) (Definition 8), and the observed gradients are perturbed with spherical Gaussian noise. Assuming that gradients of the loss function along the trajectory of the models output by noisy AdaGrad lie in an accessible constant rank subspace, and an oracle providing an asymptotically correct estimate of the maximum gradient across data samples, we show: *Asymptotically, the regret of noisy AdaGrad is within a constant factor of traditional AdaGrad.*

Formally, let V be an orthonormal matrix whose columns span the gradient subspace. It is shown in Song et al. (2020) that for generalized linear models (GLMs), unconstrained DP-(S)GD achieves a dimension independent bound. The proof relies on restricting the analysis to the feature subspace, which corresponds for these problems to the gradient subspace spanned by the columns of V. Even though the algorithm is oblivious to V, by tracking the error only in this region, in expectation the error is dimension independent. We extend this result to constrained optimization, by introducing Algorithm 1 that utilizes V_t, the matrix whose columns span the gradient subspace up to time t, to achieve dimension-independent bounds for this constrained setting. In practice it is highly unlikely we can compute the true subspace, but it is often the case that we have (noisy) oracle access to the subspace. For example, when there is public data available, it is possible to compute a noisy version $\widetilde{V}_t$ of V_t. In 3.1 we prove we can still obtain dimension independence with an extra factor of γ that accounts for the distribution difference between the real subspace and the one obtained from the oracle.

3.1. Algorithm Description

Here we describe the noisy AdaGrad algorithm $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$ presented in Algorithm 1. It differs from the traditional AdaGrad in three ways: i) The pre-conditioner matrix at each stage is a noisy perturbation H_t of the traditional pre-conditioner; ii) The state updates ($\theta_t \to \theta_{t+1}$) are dependent on noisy gradients, i.e., $\widetilde{\nabla}_t = \nabla_t + b_t$ where b_t represents the noise; iii) before applying on the gradients, the pre-conditioners (H_t's) are projected onto the rank k_t subspace defined by $\widetilde{V}_t$, the matrix returned by the subspace oracle.

3.2. Regret Analysis

In this section we provide the regret analysis of noisy AdaGrad in Theorem 3.1. One can interpret the regret as a composition of three terms: i) $O(\mathbf{Tr}(G_T)/T)$ which is the same as in the original AdaGrad algorithm; ii) a term that depends on the gradient noise, which as we mentioned earlier can be upper bounded by $O(\mathbf{Tr}\,(G_T)/T)$ given a sensitivity oracle; and iii) a term γ that bounds the error from a noisy projection obtained from the subspace oracle.

Algorithm 1: Noisy Adagrad ($\mathcal{A}_{\mathsf{noisy-AdaGrad}}$) with gradient subspace oracle

```
Input: Learning rate $\eta$, $\theta_0 \in \mathbf{R}^p$, Gradient noise standard deviation $\sigma_b(t)$, GOE scaling $\sigma_B(t)$,
       oracle access to $\widetilde{V}_t$ estimate of $V_t$ for $t \in [T]$, $S_0 \leftarrow \mathbf{0}$
for t=1 to T do
    Predict $\theta_t$, suffer loss $f(\theta_t)$ ;
    Update
        $S_t = S_{t-1} + \nabla_t \nabla_t^T, \quad G_t = S_t^{1/2}, \quad B_t = \sigma_B(t) M_p \quad \text{and} \quad M_p \sim \mu_{GOE}$   (10)
        $\widetilde{V}_t \leftarrow$ Gradient subspace returned by the oracle
        $H_t = \Pi_t(G_t + B_t) \quad \text{where } \Pi_t = \widetilde{V}_t \widetilde{V}_t^T$   (11)
        $\widetilde{\nabla}_t = \nabla_t + b_t \quad \text{where } b_t \sim \mathcal{N}(0, \sigma_b^2(t) I_p)$
    Denote $k_t = \text{rank}(H_t)$.
        $y_{t+1} = \theta_t - \eta H_t^{-1} \widetilde{\nabla}_t$
        $\theta_{t+1} \in P_{\mathcal{C}}^{H_t}(y_{t+1})$   (12)
    where $P_{\mathcal{C}}^H(y) = \arg\min_{\theta \in \mathcal{C}} \|\theta - y\|_H$ denotes the projection over the convex set $\mathcal{C}$ using
    the semi-norm determined by $H$.
end
Result: $\{\theta_t\}_{t=1}^T$
```

The proof of Theorem 3.1 goes through a careful matrix perturbation analysis, that controls the perturbation of the subspace spanned by the non-noisy pre-conditioner G_t at each time step $t \in [T]$. Recall that $\lambda_{\min>0}(G_t)$ denotes the smallest positive eigenvalue of G_t.

Theorem 3.1 *Let V_t be the orthogonal matrix whose column space is the tracked gradient subspace up to time t, and $\widetilde{V}_t$ an approximation returned by an oracle. Let γ be a bound on the subspaces' principal angle difference, i.e., $\|V_t V_t^T - \widetilde{V}_t \widetilde{V}_t^T\|_{op} \leq \gamma$, and assume $\widetilde{V}_t \widetilde{V}_t^T (I - V_t V_t^T) = 0$. Let L be the gradient ℓ_2-norm bound, C the diameter of the constraint set $\mathcal{C}$, and assume $L = C = O(1)$. Letting η be the learning rate, and $\sigma_b^2(t)$ be the gradient noise variance, then running $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$ on $\mathcal{L}(\theta; D)$ for T iterations we get*

$$\mathbb{E}[\mathsf{Regret}_T(\mathcal{F}; \mathcal{A}_{\mathsf{noisy-AdaGrad}})] \leq O\left(\mathbb{E}\left[\left(\frac{1}{\eta T} + \frac{\eta}{T}\right)\mathbf{Tr}\left(G_T\right) + \frac{\eta}{T}\sum_t \sigma_b^2(t)\mathbf{Tr}\left(G_t^{-1}\right) + \gamma\right]\right) \tag{13}$$

Comparison with traditional Adagrad: We first introduce a definition that will allow us to determine conditions under which it is possible to achieve AdaGrad rates.

Definition 2 *Let $\mathcal{A}$ be an optimization algorithm for solving Problem 1, that at time t outputs result θ_t. We define $L_{\mathcal{L},\mathcal{A}}(t)$ as the function that asymptotically bounds from above and below the gradient norm $\max_i \|\nabla \ell_i(\theta_t)\|_2$ at iteration t of algorithm $\mathcal{A}$, i.e., $L_{f,\mathcal{A}}(t) = \Theta(\max_i \|\nabla \ell_i(\theta_t)\|_2)$.*

We will drop the subindices $f, \mathcal{A}$, since it will refer to our loss function $\mathcal{L}$ and algorithm $\mathcal{A}_{\textsf{noisy-AdaGrad}}$ in our paper.

AdaGrad achieves a regret bound that is within a constant factor of $2C$ of the regret achieved by the best, fixed pre-conditioner in hindsight. We formalize this in Theorem 2.1. Selecting the learning rate that minimizes the expression on the right of Equation 13 we obtain the result in the informal Theorem 1.1.

Assume constant rank $k_t = O(1)$ smaller than the problem dimension p. In the worst case, when $\sigma_b(t) = \Theta(1)$, these terms balance to $O(1/\sqrt{T})$ and we obtain the same rates achieved by PDP-SGD. Assuming $\sigma_b(t) = L(t)$, these additional terms simplify to $O\left(\mathbf{Tr}\left(G_T\right)/T + \gamma\right)$. That is, in this setting we recover AdaGrad rates:

Corollary 3 *(Appendix C.5) Let $\sigma_b(t) = O(\|\nabla_t\|_2)$ in Algorithm 1. With an appropriate learning rate, the overall regret of $\mathcal{A}_{\textsf{noisy-AdaGrad}}$ is $O(\mathbf{Tr}\left(G_T\right)/T + \gamma)$. Further, if $\frac{1}{T}\sum_t \|\nabla_t\|_2 = o(1)$ then $O(\mathbf{Tr}\left(G_T\right)/T) = o(1/\sqrt{T})$.*

Remark 4 *How to access $\|\nabla_t\|_2$ to design the noise is an open direction that we leave for future work. In practice we can find a bound on the expected norm schedule of the gradient, or rely on estimating it from public data, add adaptive gradient clipping to our algorithm according to this schedule (see for example Pichapati et al. (2019)), and design the noise according to these clipping values to obtain the desired rates. For example, if $\|\nabla_t\|_2$ is decreasing as $O\left(\frac{1}{\sqrt[4]{T}}\right)$ then $\mathbf{Tr}\left(G_T\right) = O\left(\sqrt{\sum_t 1/\sqrt[4]{t}}\right) = O(T^{3/8})$, and the regret decreases as $O(T^{-5/8})$, improving over SGD whose rates are in the order of $O(1/\sqrt{T})$.*

3.3. Proof sketch

Dimension independence is obtained thanks to the following observations: (1) the projection step given by $V_t V_t^T$ in Eq 11 allows us to work in a k dimensional subspace instead of a p dimensional one; (2) even if this projection "erases" part of the real update, this error also lies in a k dimensional subspace by assumption.

More concretely, the proof is structured as follows: paralleling traditional convergence proofs for descent algorithms, we will expand the expression $\|\theta_{t+1} - \theta^*\|_{H_t}$ to obtain an expression involving $\langle \nabla_t, \theta_t - \theta^* \rangle$, and bound the regret using convexity.

Four terms are introduced that we will bound independently: two of them, one that depends on $\|\theta_{t+1} - \theta^*\|_{H_t} - \|\theta_t - \theta^*\|_{H_t}$, and the norm of the gradients under H^{-1}, are analogous to the original AdaGrad proof; relying on an orthogonal decomposition of $\mathbf{R}^p$ we work only with the noisy, projected pre-conditioner, and these terms can be finally bounded by $\mathbf{Tr}(G_t)$, up to a multiplicative factor. The connection is attained first by decomposing $\mathbf{R}^p$, isolating the subspace where H_t is invertible. Thanks to the above observation (1) the relevant subspaces are k-dimensional. Then, restricted to this space, we rely on Lemma 12 that uses Woodbury identity to calculate the inverse of a sum of matrices (restricted G_t and B_t in this case), and Holder's inequality. The third term is the norm of b_t, the gradient noise that is similarly bounded using Lemma 12. Finally, we track the error introduced by the projection using Davis-Kahan theorem (Davis, 1963) which again also lies in a k-dimensional subspace (observation (2)).

4. Private Pre-conditioned Gradient Descent for ERM

4.1. Estimating subspace with public data

In this section we will use Noisy-AdaGrad algorithm to define an (ε, δ)-differentially private algorithm $\mathcal{A}_{\mathsf{priv}}$ that approximately minimizes the excess empirical risk defined in (2). Our main contribution is in the low-rank unconstrained setting where, compared with original AdaGrad, we only pay an additional price of scale $\widetilde{O}\left(\frac{1}{\varepsilon n}\right)$, independent of dimension. To do so, we make the following observations:

- **Online to batch conversion:** If we set each of the loss function to be identical to $f_t(\theta) = \mathcal{L}(\theta; D)$, and set $\theta^{\texttt{priv}} = \frac{1}{T}\sum\limits_{t=1}^{T}\theta_t$ output by Algorithm 1 (Algorithm $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$), then $\mathbb{E}\left[\mathsf{Risk}(\theta^{\texttt{priv}})\right] \leq \mathbb{E}\left[\mathsf{Regret}_T(\mathcal{F}; \mathcal{A}_{\mathsf{noisy-AdaGrad}})\right]$. (This follows from standard use of Jensen's inequality.)

- **Computing $(\varepsilon/2, \delta/2)$-private pre-conditioner :** Lemma 5 below and standard use of Renyi composition theorem (Mironov, 2017) imply that ensuring $\sigma_B(t) = O\left(\frac{L\sqrt{Tt\log(1/\delta)}}{\varepsilon\sqrt{n}}\right)$ in Algorithm $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$ ensures $\left(\frac{\varepsilon}{2}, \frac{\delta}{2}\right)$-differential privacy to the computation of all the $\boldsymbol{H}_t$'s in Algorithm $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$.

- **Ensuring all noisy gradients preserve $(\varepsilon/2, \delta/2)$-differential privacy:** By the same argument as above, ensuring $\sigma_b(t) = O\left(\frac{L\sqrt{T\log(1/\delta)}}{\varepsilon n}\right)$ in Algorithm $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$ ensures $\left(\frac{\varepsilon}{2}, \frac{\delta}{2}\right)$-differential privacy to the computation of all the $(\nabla_t + b_t)$'s in Algorithm $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$.

With these observations, and composition for (ε, δ)-differential privacy (Dwork and Roth, 2014) we can ensure the above variant of noisy AdaGrad is (ε, δ)-differentially private. In the following, we will use the online to batch conversation mentioned above to bound the excess empirical risk. In particular, we obtain a bound of $\widetilde{O}(\frac{1}{\varepsilon n})$ that does not depends on the dimensionality p. We formalize this result in the following corollary. In the setting where the pre-conditioner does not satisfy low-rank assumption, we will recover the traditional upper bound of $\widetilde{\Theta}(\sqrt{p}/(\varepsilon n))$ for private ERM via differentially private gradient descent (Bassily et al., 2014), since $\mathbf{Tr}\left(G_t\right)$ will be growing with the dimension.

Lemma 5 *(Appendix C.6) Let $G_t = \sqrt{\sum_t \nabla_t \nabla_t^T}$ be the preconditioner formed at iteration t. Let ℓ_i be an $L-$Lipschitz loss function on datapoint d_i for $i = 1, ..., n$, and n the total number of records. Then the preconditioner's ℓ_2-sensitivity is given by $\Delta_2(G_t) = O\left(L\sqrt{\frac{t}{n}}\right)$*

Corollary 6 *(Appendix C.7) Assume the subspace spanned by accumulated gradients is bounded by a constant $k < p$. Let α be a non-negative real number such that $\|\nabla_t\|_2 = O(\frac{1}{t^\alpha})$. With appropriate choice of η, and for $\gamma = O(\frac{1}{\varepsilon n})$, after $T = (\varepsilon n)^{2/(1+2\alpha)}$, the excess risk of noisy-subspace $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$ is $O\left(\frac{\sqrt{k\log(1/\delta)}}{\varepsilon n}\right)$.*

Remark 7 *Connecting with the discussion in Remark 4, Corollary 6 does not assume noise $\sigma_b(t) = O(\|\nabla_t\|)$ since doing so would violate privacy. However, access to this quantity would give us enough information to find the sensitivity of the gradient ∇ in certain settings and further improve rates. For example, under an interpolation assumption (see Bassily et al. (2018); Ma et al. (2018)), a bound on the norm of the average gradient $\|\nabla\mathcal{L}(\theta_t; D)\|_2$ implies a bound on the envelope $L(t)$ of individual gradients $\|\nabla\ell_i(\theta_t)\|_2$ (Definition 2).*

For now, we rather assume constant noise, leaving us with suboptimal rates. However, we still reach the excess risk bound in fewer iterations than DP-SGD and PDP-SGD, in the case where gradient norm is decreasing ($\alpha > 0$ in Corollary 6): we require $T = (\varepsilon n)^{2/(1+2\alpha)}$, compared to $T = \varepsilon^2 n^2$ in (P)DP - SGD.

4.2. Discussion: Privately Estimating the Subspace may not Help

- A natural way to avoid using public data is to privately estimate the subspace with differential privacy. Under the strong assumption that the span of accumulated gradients are in "general position" (no trivial linear dependencies among the data points), Singhal and Steinke (2021) show that exact recovery of the subspace is possible with probability 1. However, this is a very strong assumption difficult to verify in our setting.

- In the general setting, even if we estimate the subspace with 1/2 the data, there will be a dependence on $\gamma = \sqrt{p}/n$ by the best known upper bound. (See Theorem 2 in Dwork et al. (2014).) This is fundamental in the constrained optimization setting, where there exists a lower bound of $\Omega(\sqrt{p}/\varepsilon n)$.

  An open question that remains is if there exists a more direct analysis of private AdaGrad in the unconstrained setting that could achieve dimension independence without oracle access to the gradient subspace. In Section 5 we prove that this is possible for general convex functions in the unconstrained setting with only DP-SGD.

- Subspace estimation from public data: This problem has been widely explored in the literature, where a covariance matrix is to be estimated from m (public) records sampled from distribution $\mathcal{P}$. More concretely, it is shown for example in Zhou et al. (2020c) that under natural assumptions $\gamma < O\left(\sqrt{\frac{\log p}{m}}\right)$ in the worst case scenario.

5. Interlude: Dimension independence in Unconstrained DP-SGD

Below we extend the results in Song et al. (2020), and show that unconstrained DP-SGD achieves dimension independence for general convex functions, without access to public data.

Theorem 5.1 (Dimension independence in unconstrained optimization (Appendix C.8)) *Let $\theta_0 = \mathbf{0}$ be the initial point of $\mathcal{A}_{\mathsf{DP-GD}}$. Let $\theta^* = \arg\min_{\theta\in\mathbb{R}^p} f(\theta)$ and $M = VV^T$ be the projector to the gradients eigenspace. Letting L be the gradient ℓ_2-norm bound, setting the constraint set $\mathcal{C} = \mathbb{R}^p$, and running $\mathcal{A}_{\mathsf{DP-GD}}$ on $\mathcal{L}(\theta; D)$ for $T = \varepsilon^2 n^2$ and an appropriate choice of learning rate,*

$$\mathbb{E}[\mathcal{L}(\theta_{\mathsf{priv}}; D)] - \mathcal{L}(\theta^*; D) \leq \frac{L\|\theta^*\|_M\sqrt{1 + 2rank(M)\log(1/\delta)}}{\varepsilon n} \qquad (14)$$

6. Discussion

We provide several insights that widen the understanding differentially private constrained and unconstrained optimization. First, with knowledge of the subspace where the gradients lie, it is possible to obtain bounds in terms of the trace of the pre-conditioner. This last one in turn, encodes the intrinsic dimension of the data, a smoother definition of the rank. Formally, the intrinsic dimension is defined for a positive-semidefinite matrix A as the quantity

$$\text{intdim}(A) = \frac{\mathbf{Tr}\left(A\right)}{\|A\|_{op}}$$

It measures the number of dimensions where A has spectral content (see Tropp (2015)). We can interpret our bound $\frac{\mathbf{Tr}\left(G_T\right)}{T} = \frac{\text{intdim}(A)\|G_T\|_{op}}{T}$ as being dependent on the intrinsic dimension, rather than p, and the rate at which gradients are decreasing, captured by $\frac{\|G_T\|_{op}}{T}$

Second, we introduce the importance of a gradient norm schedule during the optimization is necessary to guarantee differential privacy without sacrificing running time. We leave it as a future direction the exploration of differentially private algorithms that provide access to this envelope.

Acknowledgments

The authors would like to thank H. Brendan McMahan, Jalaj Upadhyay, Naman Agarwal, and Thomas Steinke for their helpful comments towards improving the paper.

References

Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS'16)*, pages 308–318, 2016.

Naman Agarwal, Brian Bullins, Xinyi Chen, Elad Hazan, Karan Singh, Cyril Zhang, and Yi Zhang. Efficient full-matrix adaptive regularization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 102–110. PMLR, 09–15 Jun 2019.

Noga Alon, Raef Bassily, and Shay Moran. Limits of private learning with access to public data. *arXiv preprint arXiv:1910.11519*, 2019.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 2014 IEEE 55th Annual Symp. on Foundations of Computer Science (FOCS)*, pages 464–473, 2014.

Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of SGD in non-convex over-parametrized learning. *CoRR*, abs/1811.02564, 2018.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 11279–11288, 2019.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 363–378. Springer, 2013.

Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Stephen Campbell and Carl Meyer. *Generalized Inverses of Linear Transformations*. SIAM, 2009.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

Chandler Davis. The rotation of eigenvectors by a perturbation. *Journal of Mathematical Analysis and Applications*, 6(2):159–173, 1963.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

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

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

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proc. of the Third Conf. on Theory of Cryptography (TCC)*, pages 265–284, 2006b.

Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *STOC*, pages 11–20, 2014.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in linear time. In *Proc. of the Fifty-Second ACM Symp. on Theory of Computing (STOC'20)*, 2020.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, page 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery.

Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, 2019.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.

H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, 2010.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X Yu, Sashank J Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

Vikrant Singhal and Thomas Steinke. Privately learning subspaces. *arXiv preprint arXiv:2106.00001*, 2021.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248. IEEE, 2013.

Shuang Song, Om Thakkar, and Abhradeep Thakurta. Characterizing private clipped gradient descent on convex generalized linear problems. *arXiv preprint arXiv:2006.06783*, 2020.

Matthew Streeter and H Brendan McMahan. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

Om Thakkar, Galen Andrew, and H. Brendan McMahan. Differentially private learning with adaptive clipping. *CoRR*, abs/1905.03871, 2019. URL `http://arxiv.org/abs/1905.03871`.

Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

Jalaj Upadhyay and Sarvagya Upadhyay. A framework for private matrix analysis. *arXiv preprint arXiv:2009.02668*, 2020.

X. Wu, M. Fredrikson, S. Jha, and J. F. Naughton. A methodology for formalizing model-inversion attacks. In *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, pages 355–370, 2016. doi: 10.1109/CSF.2016.32.

Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey F. Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD*, 2017.

Yuege Xie, Xiaoxia Wu, and Rachel Ward. Linear convergence of adaptive stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 1475–1485. PMLR, 2020.

Yu Xin and Tommi Jaakkola. Controlling privacy in recommender systems. Neural Information Processing Systems, 2014.

Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. *arXiv preprint arXiv:2102.12677*, 2021.

Yingxue Zhou, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Arindam Banerjee. Private stochastic non-convex optimization: Adaptive algorithms and tighter generalization bounds. *arXiv preprint arXiv:2006.13501*, 2020a.

Yingxue Zhou, Belhal Karimi, Jinxing Yu, Zhiqiang Xu, and Ping Li. Towards better generalization of adaptive gradient methods. *Advances in Neural Information Processing Systems*, 33, 2020b.

Yingxue Zhou, Zhiwei Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private sgd with gradient subspace identification. *arXiv preprint arXiv:2007.03813*, 2020c.

Appendix A. Additional background details.

A.1. AdaGrad algorithm

Below we present an adaptation to our notation of the original AdaGrad algorithm from Duchi et al. (2011).

Algorithm 2: Adagrad ($\mathcal{A}_{\mathsf{AdaGrad}}$)

```
Input: Learning rate $\eta > 0$, initial accumulator $\delta \geq 0$, bounded convex constraint set $\mathcal{X}$
$\theta_1 \leftarrow \mathbf{0}$;
$S_0 \leftarrow \mathbf{0}$;
$G_0 \leftarrow \mathbf{0}$;
$H_0 \leftarrow \mathbf{0}$;
for $t$=1 to $T$ do
    Predict $\theta_t$, suffer loss $f(\theta_t)$ ;
    Update:
        $$S_t = S_{t-1} + \nabla_t \nabla_t^T, \quad G_t = S_t^{1/2}, \tag{15}$$
        $$H_t = \delta \boldsymbol{I} + \boldsymbol{G}_t \tag{16}$$
        $$\theta_{t+1} = \arg\min_{\theta \in \mathcal{X}} \|\theta - (\theta_t - \eta \nabla_t)\|_{H_t}^2, \tag{17}$$
end
Result: $\{\theta_t\}$
```

A.2. Differential Privacy

Definition 8 *The Gaussian Orthogonal Ensamble (GOE) is the random matrix model of symmetric matrices M_n where the upper triangular entries have distribution $\mathcal{N}(0,1)$, and the diagonal entries $\mathcal{N}(0,2)$. We use μ_{GOE} to denote the distribution of a matrix generated from this model. .*

Definition 9 *We say that two datasets D and D' are neighbors, and use the notation $D \sim D'$, if they differ in exactly one record, meaning there is exactly one record that is present in one and not in the other.*

Theorem 10 (Theorem A.1. in Dwork and Roth (2014)) *Let f be an arbitrary function with range in $\mathbf{R}^p$. Define its ℓ_2-sensitivity as $\Delta_2(f) = \max_{D \sim D'} \|f(D) - f(D')\|_2$. Let $\varepsilon \in (0,1)$, $c^2 > 2\ln(1.25\delta)$, and $\sigma \geq \frac{c\Delta_2(f)}{\varepsilon}$. The Gaussian mechanism with parameter σ that adds noise $\mathcal{N}(0,\sigma^2)$ to all p components is $(\varepsilon,\delta)-$differentially private.*

Gaussian mechanism and strong composition: It follows from Renyi composition (Proposition 1 and 3 in Mironov (2017)) that ensuring $\sigma = O\left(\frac{\Delta_2(f)\sqrt{T\log(1/\delta)}}{\varepsilon}\right)$ preserves $(\varepsilon,\delta)-$differential privacy when composing T times the Gaussian mechanism with parameter σ. See Appendix A.2 for details and notation.

Appendix B. Proof of Theorem 3.1

B.1. Proof of Theorem 3.1

Below we present the detailed proof of Theorem 3.1, and defer the proofs of structural Lemmas to Section C. We first split the regret in four terms in Section B.1.1 and bound each of these independently.

B.1.1. PRELIMINARIES

Notice that H_t may not be full rank, so we interpret H_t^{-1} as the Moore-Penrose pseudoinverse for $t = 1, ..., T$. Indeed all inverses below on potentially non-full-rank matrices should be interpreted as pseudoinverses, and their corresponding norms as semi-norms.

To aid our analysis we will occasionally decompose into subspaces. Let $D_t = \text{rowspace}(H_t)$, so that $D_t^\perp = ker(H_t)$. Let:

$$\begin{aligned} E_t &= \text{rowspace}(G_t) \cap D_t \\ C_t &= \text{rowspace}(G_t) \cap D_t^\perp \\ F_t &= \ker(G_t) \end{aligned}$$

Lemma 11 *(Appendix C.1) Under the same assumptions of Theorem 3.1*

$$\begin{aligned} \mathbb{E}\left[f\left(\frac{1}{T}\sum_t \theta_t\right) - f(\theta^*)\right] \leq \mathbb{E}_{b_1,\ldots,b_{T-1},B_1,\ldots,B_T}\Bigg[&\sum_{t=0}^{T} \frac{1}{2\eta T}\left(\|\theta_t - \theta^*\|_{H_t}^2 - \|\theta_{t+1} - \theta^*\|_{H_t}^2\right) \\ &+ \frac{\eta}{2T}\left(\mathbb{E}_{b_t}[\|b_t\|_{H_t^{-1}}^2|B_t]\,]\right) \\ &+ \frac{\eta}{2T}\left(\|\nabla_t\|_{H_t^{-1}}^2\right) \\ &+ \frac{1}{T}\langle\nabla_t, \theta_t - \theta^*\rangle|_{C_t}\Bigg] \end{aligned} \tag{18}$$

We bound the four terms in this expression independently.

B.2. First Term: $\|\theta_t - \theta^*\|_{H_t}^2 - \|\theta_{t+1} - \theta^*\|_{H_t}^2$

$$\begin{aligned} &\sum_{t=0}^{T}(\|\theta_t - \theta^*\|_{H_t}^2 - \|\theta_{t+1} - \theta^*\|_{H_t}^2) \\ &\qquad\qquad \leq \|\theta_0 - \theta^*\|_{H_0}^2 - \|\theta_{T+1} - \theta^*\|_{H_T} + \sum_{t=1}^{T-1}(\theta_t - \theta^*)(H_t - H_{t-1})(\theta_t - \theta^*) \end{aligned}$$

The first term on the right hand side is 0, since $H_0 = 0$ and the second one is non-positive thanks to the projection step, so we can bound this entire term as

$$\leq \sum_{t=1}^{T} C_{\mathsf{ada}}^2 \sigma_{max}(H_t - H_{t-1})$$
$$\leq \sum_{t}^{T} C_{\mathsf{ada}}^2 \mathbf{Tr}(H_t - H_{t-1})$$

By linearity of the trace and projections,

$$= C_{\mathsf{ada}}^2 \sum_{t}^{T-1} \mathbf{Tr}(\Pi_t(G_t)) - \mathbf{Tr}(\Pi_t(G_{t-1})) + \mathbf{Tr}(\Pi_t(B_t)) - \mathbf{Tr}(\Pi_t(B_{t-1}))$$

Since $G_t - G_{t-1}$ is positive semi-definite,

$$\leq C_{\mathsf{ada}}^2 \sum_{t=1}^{T-1} \mathbf{Tr}(G_t) - \mathbf{Tr}(G_{t-1}) + \mathbf{Tr}(\Pi_t(B_t)) - \mathbf{Tr}(\Pi_t(B_{t-1}))$$

Now, $\mathbb{E}[\Pi_t(B_t)] = \mathbf{0}$, so taking expected value on both sides respect to $B_1, ..., B_T$, conditioned on $b_1, ..., b_{t-1}$, using linearity of expectation and independence of b_t-s and B_t-s

$$\begin{aligned}
&\mathbb{E}\left[\sum_t \left(\|\theta_t - \theta^*\|_{H_t}^2 - \|\theta_{t+1} - \theta^*\|_{H_t}^2\right)\right] \\
&\quad \leq \mathbb{E}\left[C_{\mathsf{ada}}^2 \sum_{t=1}^{T-1} \mathbf{Tr}(G_t) - \mathbf{Tr}(G_{t-1}) + \mathbf{Tr}(\Pi_t(B_t)) - \mathbf{Tr}(\Pi_t(B_{t-1}))\right] \\
&\quad \leq \mathbb{E}\left[C_{\mathsf{ada}}^2 \sum_{t=1}^{T-1} \mathbf{Tr}(G_t) - \mathbf{Tr}(G_{t-1})\right] \\
&\quad = \mathbb{E}[C_{\mathsf{ada}}^2 \mathbf{Tr}(G_T)] \qquad (19)
\end{aligned}$$

B.3. Second term: $\mathbb{E}[\|b_t\|_{H_t^{-1}}^2]$

For the third and fourth term we will restrict attention to the range of G_t, utilizing the following Lemma.

Lemma 12 *(Appendix C.4) Define $C = A + B$, for A, B, C linear operators on $\mathbf{R}^p$ such that B and $B^{-1}A + I$ are invertible. Then for $v \in im(A)$, $u \in \mathbf{R}^p$,*

$$\left|\langle u, C^{-1}v\rangle\right| \leq \frac{4}{3}\left|\langle u, A^{-1}v\rangle\right|$$

Now, to bound the norm of b_t under H_t^{-1}. By sign considerations we may assume H_t to be positive semidefinite, and similarly we may drop the projection Π_t from this portion of the analysis. We proceed to decompose $\mathbf{R}^p$ into the row and null spaces of G_t:

$$\mathbf{R}^p = \ker(G_t)^\perp \bigoplus \ker(G_t)$$

Call these spaces respectively A, and B, we can write:

$$\|b_t\|^2_{H_t^{-1}} = \langle b_t, H_t^{-1} b_t \rangle = \langle b_t, H_t^{-1}|_A b_t \rangle + \langle b_t, H_t^{-1}|_B b_t \rangle$$

B_t is full-rank with probability 1. As G_t is the sum of projectors, its rowspace is contained in its image, so $A \subseteq \mathrm{im}(G_t)$. Finally, B_t's continuous eigenvalue distribution implies, for $v \in \mathbf{R}^n$, $v \in \mathrm{im}(B_t^{-1}G_t + I)$ with probability 1. Since $\mathbf{R}^n$ is finite-dimensional this yields the invertibility with probability 1.

Therefore we can apply Lemma 12 to $C = H_t, A = G_t$ and $B = B_t$. Using that b_t is zero mean spherical noise with variance $\sigma_b^2(t)$ we obtain

$$\begin{aligned}\mathbb{E}_{b_t}[\|b_t\|^2_{H_t^{-1}}|b_1,\ldots,b_{t-1},B_1,\ldots,B_T] &\leq \frac{4}{3}\mathbb{E}_{b_t}\left[b_t^T G_t^{-1} b_t\right] \\ &= \frac{4\sigma_b^2(t)}{3}\mathbf{Tr}(G_t^{-1}) \end{aligned} \tag{20}$$

Where the inequality in the second step follows from the use independence and linearity of expectation:

$$\mathbb{E}_{b_t}\left[b_t^T G_t^{-1} b_t\right] = \mathbb{E}_{b_t}\left[\sum_{i,j} G_{ij}^{-1} b_{t_i} b_{t_j}\right] = \sum_{i,j}\mathbb{E}_{b_t}\left[G_{ij}^{-1} b_{t_i} b_{t_j}\right] = \sum_i \mathbb{E}_{b_t}\left[G_{ii}^{-1} b_{t_i}^2\right] = \sigma_b^2(t)\mathbf{Tr}(G_t^{-1})$$

We claim that the composition of Π_t and projection onto the kernel of G_t is 0. This is immediately implied by the assumption $\widetilde{V}_t \widetilde{V}_t^T (I - V_t V_t^T) = 0$. Therefore H_t is in fact the zero operator on subspace B, and does not contribute to the bound on $\|b_t\|^2_{H_t^{-1}}$.

Taking the sum over t, and expectation over remaining terms,

$$\mathbb{E}\left[\sum_t \|b_t\|^2_{H_t^{-1}}\right] \leq \mathbb{E}\left[\frac{4}{3}\sum_{t=1}^T \sigma_b^2(t)\mathbf{Tr}\left(G_t^{-1}\right)\right] \tag{21}$$

B.4. Third term: $\|\nabla_t\|_{H_t^{-1}}$

Paralleling the proof in the previous section, Section B.3, using the space decomposition and Lemma 12, we have that

$$\|\nabla_t\|^2_{H_t^{-1}} \leq \frac{4}{3}\nabla_t^T G_t^{-1} \nabla_t \tag{22}$$

Below we will bound this term using the following lemma.

Lemma 13 (Lemma 5.15 in Hazan (2019))

$$\sum_t \|\nabla_t\|^2_{G_t^{-1}} \leq 2\mathbf{Tr}(G_T)$$

Taking the sum over t, applying Lemma 13, and taking expectation over the conditioned terms,

$$\mathbb{E}\left[\sum_t \|\nabla_t\|^2_{H_t^{-1}}\right] \leq \mathbb{E}\left[\frac{4}{3} \cdot 2 \cdot \mathbf{Tr}(G_T)\right] \tag{23}$$

B.5. Fourth term: $\langle \nabla_t, \theta_t - \theta^* \rangle|_{C_t}$

This term corresponds to the component of ∇_t we could have lost in the projection due to an innacurate gradient subspace estimation.

Recall that C_t corresponds to the intersection of the accumulated gradient subspace (row space of G_t) with the kernel of H_t. So this term can be expanded as follows, using Cauchy-Schwartz for the first inequality and Davis-Kahan theorem for the second one (see Appendix C.3).

$$\begin{aligned}
\langle \nabla_t, \theta_t - \theta^* \rangle|_{C_t} &= \nabla_t^T P_{G_t}^{j_t}(I - P_{H_t}^{k_t})(\theta_t - \theta^*) \\
&\leq \|\nabla_t\|_2 \|P_{G_T}^{j_t}(I - P_{H_t})(\theta_t - \theta^*)\|_2 \\
&\leq CL\|V_t V_t^T - \widetilde{V}_t \widetilde{V}_t^T\|_{op} \\
&\leq CL\gamma
\end{aligned} \tag{24}$$

B.6. Putting these estimates together

Finally, putting together the four expressions,

$$\begin{aligned}
f(\frac{1}{T}\sum_t \theta_t) - f(\theta^*) \leq \mathbb{E}&\left[\left(\frac{C_{\mathsf{ada}}^2}{2\eta T} + \frac{8\eta}{3 \cdot 2T}\right)\mathbf{Tr}(G_T)\right] \\
&+ \frac{\eta}{2T}\mathbb{E}\left[\sum_t \sigma_b^2(t)\mathbf{Tr}\left(G_t^{-1}\right)\right] \\
&+ \gamma
\end{aligned} \tag{25}$$

We obtain the informal version in Theorem 1.1 by picking the minimizing learning rate.

Appendix C. Proof of structural lemmas and Theorem 5.1

C.1. Lemma 11

Under the same assumptions of Theorem 3.1

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_t \theta_t\right) - f(\theta^*)\right] \le \mathbb{E}_{b_1,\dots,b_{T-1},B_1,\dots,B_T}\left[\sum_{t=0}^{T}\frac{1}{2\eta T}\left(\|\theta_t-\theta^*\|_{H_t}^2 - \|\theta_{t+1}-\theta^*\|_{H_t}^2\right)\right. \tag{26}$$

$$+\frac{\eta}{2T}\left(\mathbb{E}_{b_t}[\|b_t\|_{H_t^{-1}}^2|B_t]\,]\right)$$
$$+\frac{\eta}{2T}\left(\|\nabla_t\|_{H_t^{-1}}^2\right)$$
$$\left.+\frac{1}{T}\langle\nabla_t,\theta_t-\theta^*\rangle|_{C_t}\right] \tag{27}$$

Proof Recall that given H, we define the scalar product $\langle x,y\rangle_H := \langle x, Hy\rangle$, and we use the notation $\cdot|_A$ to denote the output of a transformation restricted to subspace A.

Following the update rule in Eq 12,

$$\|\theta_{t+1}-\theta^*\|_{H_t}^2 = \|P_{\mathcal{C}}^{H_t}(\theta_t - \eta H_t^{-1}(\nabla_t + b_t)) - \theta^*\|_{H_t}^2 \tag{28}$$
$$\le \|\theta_t - \eta H_t^{-1}(\nabla_t + b_t) - \theta^*\|_{H_t}^2 \tag{29}$$
$$= \|\theta_t-\theta^*\|_{H_t}^2 - 2\eta\langle H_t^{-1}(\nabla_t+b_t),\theta_t-\theta^*\rangle_{H_t} + \eta^2\|\nabla_t+b_t\|_{H_t^{-1}}^2.$$

Where the first steps follows by the contraction property of projections (see Appendix C.2)
We have that

$$\langle H_t^{-1}(\nabla_t+b_t),\theta_t-\theta^*\rangle_{H_t} = \langle\nabla_t+b_t,\theta_t-\theta^*\rangle|_{D_t}. \tag{30}$$

Rearranging,

$$\langle\nabla_t+b_t,\theta_t-\theta^*\rangle|_{A_t} = \frac{1}{2\eta}\left(\|\theta_t-\theta^*\|_{H_t}^2 - \|\theta_{t+1}-\theta^*\|_{H_t}^2\right) + \frac{\eta}{2}\|\nabla_t+b_t\|_{H_t^{-1}}^2 \tag{31}$$

Taking conditional expectation over b_t, conditioned on $b_1, \dots b_{t-1}, B_1, \dots, B_t$ the left hand side becomes

$$\mathbb{E}_{b_t}[\langle\nabla_t+b_t,\theta_t-\theta^*\rangle|_{A_t}|b_1,\dots,b_{t-1},B_1,\dots,B_t] = \langle\nabla_t,\theta_t-\theta^*\rangle|_{A_t}$$

Traditionally, we could now use convexity to bound the regret by using the identity $h(\theta_t) - h(\theta^*) \le \langle\nabla_t,\theta_t-\theta^*\rangle$. Notice though that we could have lost some signal after the projection step, and $\langle\nabla_t,\theta_t-\theta^*\rangle \ne \langle\nabla_t,\theta_t-\theta^*\rangle|_{D_t}$
However, we know that

$$\langle\nabla_t,\theta_t-\theta^*\rangle = \langle\nabla_t,\theta_t-\theta^*\rangle|_{E_t} + \langle\nabla_t,\theta_t-\theta^*\rangle|_{C_t} + \langle\nabla_t,\theta_t-\theta^*\rangle|_{F_t}.$$

Furthermore, by construction $\nabla_t \in \text{rowspace}(G_t)$ and thus: i) its product will be zero on F_t and ii) we can interchange E_t and D_t since $E_t \subseteq D_t$, then

$$\langle \nabla_t, \theta_t - \theta^* \rangle \leq \langle \nabla_t, \theta_t - \theta^* \rangle|_{D_t} + \langle \nabla_t, \theta_t - \theta^* \rangle|_{C_t}$$

Completing this in Equation 30, and using the fact that b_t-s are independent, we obtain

$$\begin{aligned} \langle \nabla_t, \theta_t - \theta^* \rangle \leq & \langle \nabla_t, \theta_t - \theta^* \rangle|_{C_t} \\ & + \frac{1}{2\eta} \left(\|\theta_t - \theta^*\|_{H_t}^2 - \|\theta_{t+1} - \theta^*\|_{H_t}^2 \right) \\ & + \frac{\eta}{2} \left(\|\nabla_t\|_{H_t^{-1}}^2 + \mathbb{E}_{b_t}[\|b_t\|_{H_t^{-1}}^2] \right) \end{aligned} \tag{32}$$

Now we can invoke convexity, $f(\theta_t) - f(\theta^*) \leq \langle \nabla_t, \theta_t - \theta^* \rangle$ and $f(\sum_t \theta_t) \leq \sum_t f(\theta_t)$. Combining these facts and taking the sum over t,

$$\begin{aligned} f\left(\frac{1}{T} \sum_t \theta_t \right) - f(\theta^*) \leq & \sum_t \frac{1}{T} \langle \nabla_t, \theta_t - \theta^* \rangle|_{C_t} \\ & + \frac{1}{2\eta T} \left(\|\theta_t - \theta^*\|_{H_t}^2 - \|\theta_{t+1} - \theta^*\|_{H_t}^2 \right) \\ & + \frac{\eta}{2T} \left(\mathbb{E}_{b_t}[\|b_t\|_{H_t^{-1}}^2 | B_t] \right) \\ & + \frac{\eta}{2T} \left(\|\nabla_t\|_{H_t^{-1}}^2 \right) \end{aligned} \tag{33}$$

Using law of total expectation,

$$\mathbb{E}\left[f\left(\frac{1}{T} \sum_t \theta_t \right) - f(\theta^*) \right] \leq \mathbb{E}_{b_1,\ldots,b_{T-1},B_1,\ldots,B_T} \left[\sum_{t=0}^{T} \frac{1}{2\eta T} \left(\|\theta_t - \theta^*\|_{H_t}^2 - \|\theta_{t+1} - \theta^*\|_{H_t}^2 \right) \right. \tag{34}$$

$$\begin{aligned} & + \frac{\eta}{2T} \left(\mathbb{E}_{b_t}[\|b_t\|_{H_t^{-1}}^2 | B_t]\ ] \right) \\ & + \frac{\eta}{2T} \left(\|\nabla_t\|_{H_t^{-1}}^2 \right) \\ & \left. + \frac{1}{T} \langle \nabla_t, \theta_t - \theta^* \rangle|_{C_t} \right] \end{aligned} \tag{35}$$

■

C.2. Contraction property of projection for arbitrary norms

Lemma 14

Let $\|\cdot\|$ define a seminorm. Let $\mathcal{C}$ be a convex set and $\Pi_{\mathcal{C}}(x) = \arg\min_{v \in \mathcal{C}} \|v - x\|$ be the projection operator to set $\mathcal{C}$. Then for any $v \in \mathcal{C}$,

$$\|\Pi_{\mathcal{C}}(x) - v\| \leq \|x - v\|$$

Proof Notice the contraction and the projection are measured using the same seminorm.

Let $x^* = \Pi_{\mathcal{C}}(x)$, and $v \in \mathcal{C}$, $v \neq x^*$

We first prove that

$$\langle x - x^*, v - x^* \rangle \leq 0 \tag{36}$$

Let $\alpha \in (0, 1)$, then by convexity of $\mathcal{C}$, $x^* + \alpha(v - x^*) \in \mathcal{C}$, so by optimality of x^*

$$\begin{aligned} \|x - x^*\|_2^2 &\leq \|x - (x^* + \alpha(v - x^*))\|_2^2 \\ &= \|x - x^*\|_2^2 + \alpha^2 \|v - x^*\|_2^2 - 2\alpha \langle x - x^*, v - x^2 \rangle \\ \iff \langle x - x^*, v - x^2 \rangle &\leq \frac{\alpha}{2} \|v - x^*\|_2^2 \end{aligned}$$

Which is true for α arbitrarily small, yielding Equation 36

Now, we start going backwards,

$$\begin{aligned} \|x - v\|_2^2 &\geq \|x^* - v\|_2^2 \\ &= \|x^* - x + x - v\|_2^2 \\ &= \|x^* - x\|_2^2 + \|x - v\|_2^2 + 2\langle x^* - x, x - v \rangle \\ &= \|x^* - x\|_2^2 + \|x - v\|_2^2 + 2\langle x - x^*, v - x^* + x^* - x \rangle \\ &= -\|x^* - x\|_2^2 + \|x - v\|_2^2 + 2\langle x - x^*, v - x^* \rangle \end{aligned}$$

Cancelling terms, and rearranging,

$$\iff \|x^* - x\|_2 \geq 2\langle x - x^*, v - x^* \rangle$$

which is true by non-negativity of semi-norms and equation 36

■

C.3. Davis-Kahan Theorem

Theorem 15 (Davis-Kahan Theorem) *For any matrices A and B of like dimensions, for which $\lambda_i(A) > \lambda_j(B)$,*

$$\|P_A^i (I - P_B^{j-1})\|_{op} \leq \frac{\|A - B\|_{op}}{\lambda_i(A) - \lambda_j(B)} \tag{37}$$

C.4. Lemma 12

Define $C = A + B$, for A, B, C linear operators on $\mathbf{R}^n$ such that A and C are positive semi-definite, and B and $(B^{-1}A + I)$ invertible. Then for $v \in im(A)$, $u \in \mathbf{R}^n$,

$$\left| \langle u, C^{-1} v \rangle \right| \leq \frac{4}{3} \left| u^T A^{-1} v \right|$$

Proof Note that $C = B(B^{-1}A + I)$ is invertible by the stated assumptions. Since A, B, $A + B$ and $B^{-1}A + I$ can be inverted for $v \in \mathrm{im}(A)$, we can use the Woodbury identity (in its special case as Hua's identity, which does not rely on global but rather pointwise invertibility on its intermediate terms), to calculate $C^{-1} = (A+B)^{-1}$. We additionally use invertibility of $B^{-1}A + I$ to invert the order of the Moore-Penrose pseudoinversion on the product $A(I + B^{-1}A)$ (see Corollary 1.4.1 of Campbell and Meyer (2009)).

First we compute an expression for $C^{-1}v = (B+A)^{-1}v$ following Hua's identity:

$$\begin{aligned}
&(B+A)\cdot\left[A^{-1} - \left(A + AB^{-1}A\right)^{-1}\right]v \\
&\qquad = \left[BA^{-1} + I - (B+A)\left(A + AB^{-1}A\right)^{-1}\right]v \\
&\qquad = \left[BA^{-1} + I - (B+A)\left(A\left(I + B^{-1}A\right)\right)^{-1}\right]v \\
&\qquad = \left[I + BA^{-1} - B\left(I + B^{-1}A\right)\left(I + B^{-1}A\right)^{-1}A^{-1}\right]v \\
&\qquad = v
\end{aligned} \tag{38}$$

where we have used $AA^{-1}v = v$ since $v \in \text{im}(A)$, invertibility of B to write $BB^{-1}A = A$, and invertibility of $X = I + B^{-1}A$ to both collapse $XX^{-1} = I$ and rewrite $(AX)^{-1} = X^{-1}A^{-1}$. Since $A + B$ is invertible by assumption, (38) implies $C^{-1}v = \left(A^{-1} - \left(A + AB^{-1}A\right)^{-1}\right)v$. Therefore:

$$\begin{aligned}
\left|\langle u, C^{-1}v\rangle\right| &= \left|u^T C^{-1}v\right| \\
&= \left|u^T(A^{-1} - (AB^{-1}A + A)^{-1})v\right| \\
&= \left|v^T A^{-1}v - u^T(AB^{-1}A + A)^{-1})v\right| \\
&\leq \left|u^T A^{-1}v\right| + |\mathbf{Tr}(u^T A^{-1}(B^{-1}A + I)P_{\text{im}(A)})^{-1}v)|
\end{aligned} \tag{39}$$

Where the last step follows by the triangle inequality, and because the trace of a scalar is just that scalar. Using the cyclic property of the trace,

$$= \left|u^T A^{-1}v\right| + |\mathbf{Tr}(uv^T A^{-1}((AB^{-1} + I)P_{\text{im}(A)})^{-1})| \tag{40}$$

Using the trace duality property,

$$\leq \left|u^T A^{-1}v\right| + \|u^T A^{-1}v\|_1 \|(AB^{-1} + I)^{-1})\|_\infty \tag{41}$$

$$\leq \left|u^T A^{-1}v\right| \left(1 + \max_{i,j} \frac{\lambda_i(B)}{\lambda_i(B) + \lambda_j(A)}\right) \tag{42}$$

Notice that we only care about $\lambda_j(B) \geq -\lambda_i(A)$, otherwise it would mean C would have negative eigenvalues, contradicting the PSD assumption. The maximum on (42) is then bounded by $\frac{1}{1+\lambda_{min>0}(A)} \leq 1$. Thus,

$$\begin{aligned}
&\leq v^T A^{-1}v\,(1+1) \\
&\leq 2v^T A^{-1}v
\end{aligned} \tag{43}$$

Remark: Notice that in step 41, the L_1 and L_∞ norms can be replaced by any p and q such than $\frac{1}{p} + \frac{1}{q} = 1$ to obtain a better bound.

■

C.5. Corollary 3

Assume the norm of gradients is decreasing as $L(t) = o(1)$, and constant rank for the gradient subspace. Let $\sigma_b(t) = O(L(t))$, then the overall regret of $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$ is $O(\mathbf{Tr}\,(G_T)/T) = o(1/\sqrt{T})$.

Proof We first introduce the following inequality that has been previously used in optimization, see Lemma 1 in Streeter and McMahan (2010) for a proof.

Lemma 16 *For any non-negative real numbers $a_1, a_2, a_3, \cdots, n$,,*

$$\sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum\limits_{j=1}^{i} a_j}} \leq 2\sqrt{\sum_{i=1}^{n} a_i}$$

Notice that $\mathbf{Tr}\,(G_t) = O\left(\sqrt{\sum\limits_{s=1}^{t}(L(s))^2}\right)$. Then selecting the optimal learning rate in expression 13, we have:

$$\begin{aligned}
\mathbb{E}[\mathsf{Regret}_T(\mathcal{F}; \mathcal{A}_{\mathsf{noisy-AdaGrad}})] &\leq O\left(\mathbb{E}\left[\frac{\sqrt{\mathbf{Tr}\,(G_T)^2 + \mathbf{Tr}\,(G_T)\sum_t \sigma_b^2(t)\mathbf{Tr}\,(G_t^{-1})}}{T} + \gamma\right]\right) \\
&\leq O\left(\mathbb{E}\left[\frac{\sqrt{\mathbf{Tr}\,(G_T)^2 + \mathbf{Tr}\,(G_T)\sum_t L(t)^2/\sqrt{\sum_{s=1}^{t} L(s)^2}}}{T} + \gamma\right]\right) \\
&\leq O\left(\mathbb{E}\left[\frac{\sqrt{\mathbf{Tr}\,(G_T)^2 + \mathbf{Tr}\,(G_T)\sqrt{\sum_t L(t)^2}}}{T} + \gamma\right]\right) \qquad (44) \\
&\leq O\left(\mathbb{E}\left[\frac{\sqrt{\mathbf{Tr}\,(G_T)^2 + \mathbf{Tr}\,(G_T)^2}}{T} + \gamma\right]\right) \\
&\leq O\left(\mathbb{E}\left[\frac{\mathbf{Tr}\,(G_T)}{T} + \gamma\right]\right)
\end{aligned} \qquad (45)$$

Where Equation 44 follows from the non-negativity of $L(t)^2$ and inequality in Lemma 16 ■

C.6. Lemma 5 - pre-conditioner sensitivity

Let $G_t = \sqrt{\sum_t \nabla_t \nabla_t^T}$ be the preconditioner formed at iteration t. Let ℓ_i be an $L-$Lipschitz loss function on datapoint d_i for $i = 1, ..., n$, and n the total number of records. Then the preconditioner's ℓ_2-sensitivity is given by $\Delta_2(G_t) = O\left(L\sqrt{\frac{t}{n}}\right)$

Proof Let G_t^D be the preconditioner computed at iteration t with dataset $D = \{d_1, ..., d_n\}$. Let D' be a neighboring dataset, w.l.o.g. $d_n \notin D'$.

Let $g_{t,j} = \frac{\nabla \ell(\theta_t; d_j)}{n}$. By the $L-$Lipschitz condition, $\|g_{i,j}\| \leq \frac{L}{n}$ Recall that $G_t^D = \sqrt{\sum_{i=1}^t \nabla_i \nabla_i^T}$, and $\nabla_i^D = \sum_{j \in D} g_{i,j}$. Let $K_t = \sum_{i=1}^t (\sum_{j=1}^{n-1} g_{i,j})(\sum_{j=1}^{n-1} g_{i,j})^T$

We have then

$$\begin{align}
\|G_t^D - G_t^{D'}\|_2 &= \left\|\sqrt{K_t + \sum_{i=1}^{t}\sum_{j=1}^{n} g_{i,n} g_{i,j}} - \sqrt{K_t}\right\|_F \nonumber \\
&\leq \left\|\sqrt{\sum_{i=1}^{t}\sum_{j=1}^{n} g_{i,n} g_{i,j}}\right\|_F \nonumber \\
&= \sqrt{\mathbf{Tr}\left(\sqrt{\sum_{i=1}^{t}\sum_{j=1}^{n} g_{i,n} g_{i,j}^T}\sqrt{\sum_{i=1}^{t}\sum_{j=1}^{n} g_{i,n} g_{i,j}^T}\right)} \tag{46} \\
&= \sqrt{\mathbf{Tr}\left(\sum_{i=1}^{t}\sum_{j=1}^{n} g_{i,n} g_{i,j}^T\right)} \tag{47} \\
&\leq \frac{L\sqrt{t}}{\sqrt{n}} \tag{48}
\end{align}$$

■

C.7. Corollary 6

Assume the subspace spanned by accumulated gradients is bounded by a constant $k < p$. With appropriate choice of η, and for $\gamma = O(\frac{1}{\varepsilon n})$, the excess risk of noisy-subspace $\mathcal{A}_{\mathsf{noisy-AdaGrad}}$ is $O\left(\frac{\sqrt{\log(1/\delta)}}{\varepsilon n}\right)$.

Proof Recall that $\sigma_b(t) = O(\frac{L\sqrt{T\log(1/\delta)}}{\varepsilon n})$, assume $L = O(1)$, and assume gradients norms are decreasing as $\|\nabla_t\| = O(\frac{1}{t^\alpha})$, then $\mathbf{Tr}\left(G_T\right) = O(\sqrt{\sum_t 1/t^{2\alpha}}) = O(T^{\frac{1-2\alpha}{2}})$ and $\sum_t \mathbf{Tr}\left(G_t^{-1}\right) = O\left(\sum_t \frac{1}{t^{\frac{1-2\alpha}{2}}}\right) = O(T^{\frac{1+2\alpha}{2}})$.

Replacing these values in Theorem 3.1,

$$\mathbb{E}\left[\mathsf{Risk}(\theta^{\mathtt{priv}})\right] \leq \sqrt{\frac{T^{1-2\alpha}}{T^2} + \frac{T^{\frac{1-2\alpha}{2}} T T^{(1+2\alpha)/2}}{\varepsilon^2 n^2 T}} + \gamma$$

$$\leq \sqrt{\frac{1}{T^{1+2\alpha}} + \frac{1}{\varepsilon^2 n^2}} + \frac{1}{\varepsilon n}$$

Then letting $T = \varepsilon n)^{2/(1+2\alpha)}$ we obtain the desired result. ■

C.8. Unconstrained DP-GD

Algorithm 3: DP-GD: Differentially private gradient descent

```
Input: noise variance $\sigma^2$, number of iterations $T$, learning rate $\eta$, gradient oracle $\nabla_t$
for $t$=1 to $T$ do
    $\widetilde{\nabla}_t \leftarrow \nabla_t + b_t$   where $b_t \sim \mathcal{N}(0, \sigma_b^2 I_p)$;
    $\theta_{t+1} = \theta_t - \eta\widetilde{\nabla}_t$
end
Result: $\frac{1}{T}\sum_{t=1}^{T} \theta_t$
```

Theorem 5.1: *Let $\theta_0 = \mathbf{0}$ be the initial point of $\mathcal{A}_{\mathsf{DP-GD}}$. Let $\theta^* = \arg\min_{\theta \in \mathbb{R}^p} f(\theta)$ and $M = VV^T$ be the projector to the gradients eigenspace. Letting L be the gradient ℓ_2−norm bound, setting the constraint set $\mathcal{C} = \mathbb{R}^p$, and running $\mathcal{A}_{\mathsf{DP-GD}}$ on $\mathcal{L}(\theta; D)$ for $T = \varepsilon^2 n^2$ and appropriate learning rate η,*

$$\mathbb{E}[\mathcal{L}(\theta_{\mathsf{priv}}; D)] - \mathcal{L}(\theta^*; D) \leq \frac{L\|\theta^*\|_M \sqrt{1 + 2rank(M)\log(1/\delta)}}{\varepsilon n}$$

Proof We follow standard arguments for analyzing gradient descent Song et al. (2020); Bubeck (2015) . Recall that M is the projector to gradients eigenspace.

We have that

$$\|\theta_{t+1} - \theta^*\|_M^2 = \|\theta_t - \theta^* - \eta(\nabla_t + b_t)\|_M^2$$
$$\leq \|\theta_t - \theta^*\|_M^2 - 2\eta\langle \nabla_t + b_t, \theta_t - \theta^*\rangle + \eta^2\|\nabla_t + b_t\|_M^2$$

Taking expected value respect to b_t conditioned on $b_1, ..., b_{t-1}$,

$$\leq \|\theta_t - \theta^*\|_M^2 - 2\eta\langle \nabla_t, \theta_t - \theta^*\rangle + \eta^2(L^2 + \text{rank}\,(M)\sigma^2)$$

Here we used that ∇_t lies in the subspace M, and $b_t \sim \mathcal{N}(0, \sigma^2 I_p)$

Rearranging,and taking expectation over $b_1, ..., b_{t-1}$

$$\mathbb{E}\left[\langle \nabla_t, \theta_t - \theta^* \rangle\right] \leq \frac{1}{2\eta}(\mathbb{E}\left[\|\theta_t - \theta^*\|_M^2 - \|\theta_{t+1} - \theta^*\|_M^2\right]) + \frac{\eta}{2}(L^2 + \text{rank}\,(M)\sigma^2) \tag{49}$$

By convexity,

$$\mathcal{L}(\theta^{\texttt{priv}}; D) - \mathcal{L}(\theta^*; D) \leq \frac{1}{T}\sum_{t=1}^{T}\langle \nabla_t, \theta_t - \theta^* \rangle$$

So taking the sum over t in Equation 49 and using linearity of expectation we get

$$\mathbb{E}\left[\mathcal{L}(\theta^{\texttt{priv}}; D)\right] - \mathcal{L}(\theta^*; D) \leq \frac{1}{2\eta T}\|\theta_0 - \theta^*\|_M^2 + \frac{\eta}{2}(L^2 + \text{rank}\,(M)\sigma^2)$$

Taking the optimal learning rate η,

$$\leq \|\theta^*\|_M \sqrt{\frac{L^2 + rank(M)\sigma^2}{T}}$$

Setting $\sigma = O(\frac{L\sqrt{T\log(1/\delta)}}{\varepsilon n})$, and $T = \varepsilon^2 n^2$ we obtain the desired result:

$$\mathbb{E}[\mathcal{L}(\theta_{\mathsf{priv}}; D)] - \mathcal{L}(\theta^*; D) \leq \frac{L\|\theta^*\|_M \sqrt{1 + 2\text{rank}(M)\log(1/\delta)}}{\varepsilon n}$$

■