Dimension Independence in Unconstrained Private ERM via Adaptive Preconditioning

Peter Kairouz † Mónica Ribero†‡ Keith Rush † Abhradeep Thakurta§

August 18, 2020

1 Introduction

In this paper we revisit the problem of private empirical risk minimization (ERM) with differential privacy [CMS11, BST14, SCS13, ACG+16, BFTT19, MRTZ17, WL+17, INS+19, PSY+19, TAM19, FKT20, STT20]. We show that for unconstrained convex empirical risk minimization if the observed gradients of the objective function along the path of private gradient descent [SCS13, BST14, STT20] lie in a low-dimensional subspace (smaller than the ambient dimensionality of $p$), then using noisy adaptive preconditioning (a.k.a., noisy Adaptive Gradient Descent (AdaGrad) [MS10, DHS11, Haz19]) we obtain a regret composed of two terms: a constant multiplicative factor of the original Adagrad regret and an additional regret due to noise. In particular, we show that if the gradients lie in a constant rank subspace, then one can achieve an excess empirical risk of $\tilde{O}(1/\epsilon n)$, as $T \to \infty$, where $\epsilon$ is the privacy budget and $n$ the number of samples, compared to the worst-case achievable bound of $\tilde{\Theta}(\sqrt{p}/\epsilon n)$. (These results are formalized in Corollaries 4.1 and 4.2). This result implies that, by running noisy Adagrad for long enough, we can bypass DP-SGD bound $\tilde{O}(\sqrt{p}/\epsilon n)$. While [JT14, STT20] show dimension independent excess empirical risk bounds for the restrictive setting of convex generalized linear problems optimized over unconstrained subspace, our results operate with general convex functions in unconstrained minimization.

Along the way, we do a perturbation analysis of noisy AdaGrad, which may be of independent interest. Our utility guarantee for the private ERM problem follows as a corollary to the regret guarantee of noisy AdaGrad. The regret guarantee (stated in Theorem 2.1) can be asymptotically interpreted as the original regret from the non-private AdaGrad, and an additional regret term that depends on the noise added.

In a concurrent and independent work, authors in [ZWB20] also achieve a bound that depends only logarithmically in $p$. By assuming access to a public dataset with $m$ records, the gradient is pre-conditioned with a projection learned on the public dataset, obtaining a bound of $O(1/\epsilon n + \log p/\sqrt{m})$.

1.1 Problem Definition

Given a data set $D = \{d_1, \ldots, d_n\}$, and an objective function $L(\theta; D) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; d_i)$, the goal is to design an $(\epsilon, \delta)$-differentially private algorithm $A_{\text{priv}}$ that outputs a model $\theta_{\text{priv}} \in \mathbb{R}^p$ that approximately solves the following optimization problem: $\min_{\theta} L(\theta; D)$. In terms of accuracy we consider the traditional excess empirical risk defined as follows:

$$\text{Risk}(\theta_{\text{priv}}) = L(\theta_{\text{priv}}; D) - \min_{\theta} L(\theta; D).$$

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 $\ell_2$-norm.

---

*This is a preliminary note. A complete paper will soon follow.
†Google Research. {kairouz, mribero, krush}@google.com
‡Department of Electrical and Computer Engineering, The University of Texas at Austin. Work done while author was interning at Google.
§Google Research - Brain. {athakurta}@google.com
Online convex optimization: To solve the private ERM problem, we will model it along the lines of online convex optimization [Haz19, SS+11]. 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. It is a well-known standard idea called online to batch conversion [Haz19] 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 set of traditional online learning [Haz19]. Formally, given a sequence of loss functions $F = \{f_1, \ldots, f_T\}$ (with each $f_t : \mathbb{R}^p \to \mathbb{R}$) arriving online, the objective is to design an algorithm to output a sequence of models $\{\theta_1, \ldots, \theta_T\}$ s.t. the following is minimized:

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

Throughout this paper, we will call an algorithm $A$ to be a “low-regret” algorithm if it outputs a sequence of models s.t. the regret in (2) is $o(1)$. In principle each of the the loss functions $f_t \in F$ 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 $F$ are assumed to be convex in its first parameter. Furthermore, we will assume that the loss functions are Lipschitz bounded, i.e., $\forall \theta \in \mathbb{R}^p, f \in F : \|\nabla \theta f(\theta)\|_2 \leq L$.

1.2 Notation

We use $\|\cdot\|_2$ to denote the $\ell_2$ 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 $\|\cdot\|_{\text{op}}$ to denote the operator norm of a matrix, defined as $\|A\|_{\text{op}} = \max\{\|\lambda_i\| : \lambda_i$ eigenvalue value of $A \}$. We use $[T]$ to denote the time interval $[T] = \{1, \ldots, T\}$. 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 $\nabla f$ to denote $\nabla f(\theta)$.

Definition 1.1. The Gaussian Orthogonal Ensemble (GOE) is the random matrix model of symmetric matrices $M_n$ where the upper triangular entries have distribution $N(0, 1)$, and the diagonal entries $N(0, 2)$. We use $\mu_{\text{GOE}}$ to denote the distribution of a matrix generated from this model.

2 Analyzing Noisy Adaptive Gradient Descent (Noisy-AdaGrad)

In this section we revisit AdaGrad (Adaptive Gradient Descent) [DHS11, MS10, Haz19] that 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 [HAK07], and online mirror descent [BTN01, SSSS09] is the use of a gradient pre-conditioner. It allows much tighter regret guarantees if the gradients of the loss functions come from a (close to) low-rank subspace.

In this section we study a noisy version of AdaGrad, where adaptive pre-conditioner is perturbed with a Gaussian Orthogonal Ensemble (GOE) (Definition 1.1), and the observed gradients are perturbed with spherical Gaussian noise. Assuming unit variance for the GOE and the noise added to the gradients, and that gradients of the loss function along the trajectory of the models output by noisy AdaGrad lie in a constant rank subspace, we show the following: Asymptotically, the regret of noisy AdaGrad is at most that of AdaGrad with additional terms due to noise, which is bounded by $O\left(\frac{1}{\sqrt{T}}\right)$.

2.1 Algorithm Description

The original Adagrad algorithm proposes the update

$$\theta_{t+1} = \arg \min_{\theta \in \mathbb{R}^p} ||\theta - (\theta_t - \eta \nabla f_t)||^2_{\mathbb{C}^n},$$

(3)
where \( G_t = \sqrt{\sum_{i=0}^{t-1} \nabla_i \nabla_i^T} \) and \( \| \cdot \|_A \) denotes the Mahalanobis seminorm defined as \( \| \cdot \|_A = \sqrt{\langle \cdot , A \cdot \rangle} \) for \( A \) symmetric and positive-semidefinite.

Here we present the noisy AdaGrad algorithm\(^1\). It differs from the traditional AdaGrad on the following three aspects: i) The preconditioner matrix at each stage is a noisy perturbation \( H_t \) of the traditional pre-conditioner, i.e., \( G_t \). ii) The state updates (\( \theta_t \rightarrow \theta_{t+1} \)) are dependent on noisy gradients, i.e., \( \nabla_t + b_t \), and iii) Before applying on the gradients, the pre-conditioners (\( H_t \)’s) are projected onto the rank \( k_t \) SVD subspace defined by the eigenvalues s.t. \( \lambda_i(H_t) > \alpha(t) \) for some hyperparameter \( \alpha(t) \). Here \( \alpha \) is a predefined hyperparameter, that controls the overall accuracy of the noisy AdaGrad algorithm.

\[ \begin{align*}
\text{Algorithm 1: Noisy Adagrad } & (A_{\text{noisy-AdaGrad}}) \\
\text{Input:} & \text{ Learning rate } \eta, \theta_0 \in \mathbb{R}^p, \text{ Gradient noise standard deviation } \sigma_b, \text{ GOE scaling } \sigma_B, \text{ projection threshold } \alpha_t, \text{ for } t \in [T] \\
\text{Result:} & \{ \theta_t \} \\
S_0 & \leftarrow 0; \\
G_0 & \leftarrow 0; \\
\text{for } t=1 \text{ to } T \text{ do} \\
& \text{Predict } \theta_t, \text{ suffer loss } f(\theta_t) ; \\
& \text{Update:} \\
& S_t = S_{t-1} + \nabla_t \nabla_t^T, \quad G_t = S_t^{1/2}, \quad B_t = \sigma_B M_p \quad \text{and} \quad M_p \sim \mu_{\text{GOE}} \\
& H_t = \Pi_{\alpha_t}(G_t + B_t) \quad \text{(4)} \\
\text{end} \\
\end{align*} \]

where \( \Pi_{\alpha} = \Pi_{\alpha}(t) \) denotes the projection that zeroes out eigenvalues s.t. \( \lambda_i(H_t) < \alpha \). Denote \( k_t = \text{rank}(\Pi_{\alpha_t}) \).

2.2 Regret Analysis

In this section we provide the regret analysis of noisy AdaGrad in Theorem 2.1. One can interpret the regret as a composition of two terms: i) \( O(\text{Tr}(G_T)/T) \) which is the same as in the original AdaGrad algorithm, and ii) a term that depends on the noisy perturbation, which as we mentioned earlier can be upper bounded by \( O(1/\sqrt{T}) \) for loss functions whose gradients come from constant rank sub-spaces.

The proof of Theorem 2.1 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 \( t \in [T] \).

Before stating the theorem, we define the set of all strongly convex regularization functions with a fixed and bounded Hessian as

\[ \mathcal{H} = \{ X \in \mathbb{R}^{p \times p} : \text{Tr}(X) \leq 1, X \succeq 0 \} \quad \text{(7)} \]

**Theorem 2.1.** Assume \( \| \nabla_t \|_2 \leq L \), gradient noise \( b_t \sim N(0, \sigma_b I_p) \), \( B_t \) is a scaled Wigner matrix, namely \( B_t = \sigma_B M_{p,t} \) for some \( M_{p,t} \sim \mu_{\text{GOE}} \). Let \( C_{\text{ada}} = \max_{t} \| \theta_t - \theta^* \|_2 \) and assume the learning rate is set as \( \eta = C_{\text{ada}} \). Let \( c_1 \) be an absolute constant, and let \( \alpha_t = 3c_1 \sqrt{\text{rank}(G_t) \sigma_B \log t} \) be the hyperparameter in Algorithm 1 (Algorithm \( A_{\text{noisy-AdaGrad}} \)) determining the rank \( k_t \) of the pre-conditioner. Assume that for some \( t_0, \sigma_B \geq 0 \) the following holds: \( \lambda_{\min > 0}(G_t) \geq \alpha \) for all \( t \geq t_0 \), where \( \lambda_{\min > 0}(G_t) \) denotes the smallest positive eigenvalue of \( G_t \). Then the following is true:
\[
E[\text{Regret}_T(\mathcal{F}; A_{\text{noisy-AdaGrad}})] = E \left[ \frac{C_{\text{ada}}}{2T} O \left( \sqrt{\min_{H \in \mathcal{H}} \sum_{t=0}^{T} \| \nabla_t \|_{H}^2 + \sigma_b^2 \sum_{t=0}^{T} \text{rank}(G_t) + t_0 \sigma_B L + \sum_{t=0}^{T} \frac{c_t \sqrt{\text{rank}(G_t) \sigma_B \log t}}{\lambda_{\text{min} > 0}(G_t)} } \right) \right].
\]

**Interpretation of Theorem 2.1 for low-rank pre-conditioners:** For the ease of exposition, let us assume \(C_{\text{ada}} \leq 1\), and \(L = 1\). We can observe that the expected regret can be decomposed into two terms: one that depends on \(O \left( \frac{1}{\sqrt{T}} \right)\), meaning we are at a multiplicative factor of the best pre-conditioner’s regret rate (analogous to original AdaGrad algorithm [Haz19]), and additional terms introduced due to the injection of noise. To understand these terms asymptotically, we further assume constant rank and noise variances, \(k_t = \sigma_b = \sigma_B = O(1)\).

The additional terms simplify to \(O \left( \frac{1}{\sqrt{T}} E \left[ \sum_{t=1}^{T} \sigma(G_t) + t_0 + \sum_{t=0}^{T} \log t \sigma \lambda_{\text{min} > 0}(G_t) \right] \right)\). Since we assumed \(k_t = O(1)\), we will assume \(\lambda_{k_t}(G_t) = \Theta(\sqrt{T})\) and \(t_0 = O(\sqrt{T} \sigma_B \log t)\). Therefore, the additional regret due to the noise simplifies to \(O \left( \frac{1}{\sqrt{T}} \right)\).

**Interpretation of Theorem 2.1 without low-rank assumption:** Let \(\sigma_b, \sigma_B\) as above, but now assume \(\lambda_{b_t}(G_t) = \Theta(\sqrt{T})\), meaning the pre-conditioner could have a spectrum uniformly spread across all dimensions, then the additional regret increases its dependence on dimension to \(O \left( \frac{1}{\sqrt{T}} \right)\).

## 3 Proof of Theorem 2.1

### 3.1 Proof sketch

Following 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 well 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 in the original Adagrad proof and can be bounded by \(\text{Tr}(G_t)\), up to a multiplicative factor due to the noisy, projected pre-conditioner. The connection is attained thanks to Lemma 3.5 that uses Woodbury identity to calculate the inverse of a sum of matrices \((G_t + B_t)\) 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 3.5. Finally, we track the error introduced by the projection using Davis-Kahan theorem [Dav63].

### 3.2 Preliminaries

Below we interpret \(H^{-1}\) as the Moore-Pensore pseudoinverse and for \(t = 1,...,T\). Let \(A_t = \text{rowspace}(H_t) = \text{rowspace}(\Pi_0(\Pi_0(G_t + B_t)))\), so that \(A_t^\perp = \ker(H_t)\). We will decompose \(\mathbb{R}^p\) into the following mutually orthogonal subspaces:

\[
B_t = \text{rowspace}(G_t)|_{A_t},
C_t = \text{rowspace}(G_t)|_{A_t^\perp},
D_t = \ker(G_t).
\]

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 \(A\). To ease notation we drop the subscript \(\sigma_B\) in \(t_0, \sigma_B\).

Following the update rule in Eq (6)
\[ \|\theta_{t+1} - \theta^*\|_{H_t}^2 = \|\theta_t - \eta H_t^{-1}(\nabla_t + b_t) - \theta^*\|_{H_t}^2 \]
\[ = \|\theta_t - \theta^*\|_{H_t}^2 - 2\eta \langle H_t^{-1}(\nabla_t + b_t), \theta_t - \theta^* \rangle_{H_t} + \eta^2 \|\nabla_t + b_t\|_{H_t^{-1}}^2. \]

We have that
\[ \langle H_t^{-1}(\nabla_t + b_t), \theta_t - \theta^* \rangle_{H_t} = \langle \nabla_t + b_t, \theta_t - \theta^* \rangle|_{A_t}. \]

Rearranging,
\[ \langle \nabla_t + b_t, \theta_t - \theta^* \rangle|_{A_t} = \frac{1}{2\eta} (\|\theta_t - \theta^*\|_{H_t}^2 - \|\theta_{t+1} - \theta^*\|_{H_t}^2) + \frac{\eta}{2} \|\nabla_t + b_t\|_{H_t^{-1}}. \]

Taking conditional expectation over \(b_t\), conditioned on \(b_1, ... b_{t-1}, B_1, ..., B_t\) the left hand side becomes
\[ \mathbb{E}_{b_t}[\langle \nabla_t + b_t, \theta_t - \theta^* \rangle|_{A_t}|b_1, ..., b_{t-1}, B_1, ..., B_t] = \langle \nabla_t, \theta_t - \theta^* \rangle|_{A_t}. \]

Traditionally, we could now use convexity to bound the regret by using the identity \(h(\theta_t) - h(\theta^*) \leq \langle \nabla_t, \theta_t - \theta^* \rangle\).

Notice though that we could have lost some signal after the projection step, and \(\langle \nabla_t, \theta_t - \theta^* \rangle \neq \langle \nabla_t, \theta_t - \theta^* \rangle|_{A_t}\) may not hold.

However, we know that
\[ \langle \nabla_t, \theta_t - \theta^* \rangle = \langle \nabla_t, \theta_t - \theta^* \rangle_{B_t} + \langle \nabla_t, \theta_t - \theta^* \rangle|_{C_t} + \langle \nabla_t, \theta_t - \theta^* \rangle|_{D_t}. \]

Furthermore, by construction \(\nabla_t \in \text{rowspace}(G_t)\) and thus: i) its product will be zero on \(D_t\) and ii) we can interchange \(B_t\) and \(A_t\), since \(B_t \subseteq A_t\), then
\[ \langle \nabla_t, \theta_t - \theta^* \rangle \leq \langle \nabla_t, \theta_t - \theta^* \rangle|_{A_t} + \langle \nabla_t, \theta_t - \theta^* \rangle|_{C_t}. \]

Completing this in Equation 9 and using the fact that \(b_t\)-s are independent, we obtain
\[ \langle \nabla_t, \theta_t - \theta^* \rangle \leq \langle \nabla_t, \theta_t - \theta^* \rangle|_{C_t} \]
\[ + \frac{1}{2\eta} (\|\theta_t - \theta^*\|_{H_t}^2 - \|\theta_{t+1} - \theta^*\|_{H_t}^2) \]
\[ + \frac{\eta}{2} \left( \|\nabla_t\|_{H_t^{-1}}^2 + \mathbb{E}_{b_t}[\|b_t\|_{H_t^{-1}}^2] \right). \]

Now we can invoke convexity, \(h(\theta_t) - h(\theta^*) \leq \langle \nabla_t, \theta_t - \theta^* \rangle\) and \(h(\sum_t \theta_t) \leq \sum_t h(\theta_t)\).

Combining these facts and taking the sum over \(t\),
\[ h\left(\frac{1}{T} \sum_t \theta_t\right) - h(\theta^*) \leq \sum_t \frac{1}{T} \langle \nabla_t, \theta_t - \theta^* \rangle|_{C_t} \]
\[ + \frac{1}{2\eta T} (\|\theta_t - \theta^*\|_{H_t}^2 - \|\theta_{t+1} - \theta^*\|_{H_t}^2) \]
\[ + \frac{\eta}{2T} \left( \mathbb{E}_{b_t}[\|b_t\|_{H_t^{-1}}^2|B_t] \right) \]
\[ + \frac{\eta}{2T} \left( \|\nabla_t\|_{H_t^{-1}}^2 \right). \]

Using law of total expectation,
\[ \mathbb{E}[h\left(\frac{1}{T} \sum_{t} \theta_t\right) - h(\theta^*)] \leq \mathbb{E}_{\theta_0,\ldots,\theta_{T-1},B_1,\ldots,B_T} \left[ \frac{1}{2T} \sum_{t=0}^{T} \left( \|\theta_t - \theta^*\|_{H_t}^2 - \|\theta_{t+1} - \theta^*\|_{H_{t+1}}^2 \right) \right] \]

\[ + \frac{\eta}{2T} \left( \mathbb{E}_{b_t} [\|b_t\|_{H_{t-1}}^2 | B_t ] \right) \]

\[ + \frac{\eta}{2T} \left( \|\nabla_{\theta_t}\|_{H_{t-1}}^2 \right) \]

\[ + \frac{1}{T} \langle \nabla_{\theta_t}, \theta_t - \theta^* \rangle |_{C_t} \] (13)

The following lemma is used to ensure the spectrum of \( B_t \) is bounded with high probability.

**Lemma 3.1** (Corollary 2.3.5 in [Tao12]). Suppose that the coefficients of matrix \( M \in \mathbb{R}^{p \times p} \) are independent, have zero mean and uniformly bounded by 1. Then there exists absolute constants \( C, c > 0 \) such that for all \( A \geq C \)

\[ \mathbb{P}(\|M\|_{op} > A\sqrt{p}) \leq C \exp(-cAp) \]

**Remark:** Notice that using Lemma 3.1 and union bound, we can ensure with high probability that \( \|B_t\|_{op} = O(\sqrt{p \log(T)} \sigma_B) \) for all times \( T \). Furthermore, as shown in Lemma 3.3 if the eigengap of \( G_t \) is larger than \( \alpha(t) \), we have that, with high probability, \( H_t \) captures a noisy version of the subspace spanned by \( G_t \). Since the overall regret is bounded by \( O(C_{ada} L) \) in the worst-case, we will condition on the set defined in Lemma 3.3 for the remainder of the proof.

We use linearity of expectation and bound the four main terms in this expression independently.

### 3.3 First term: \( \langle \nabla_{\theta_t}, \theta_t - \theta^* \rangle |_{C_t} \)

This term corresponds to the component of \( \nabla_{\theta_t} \) we could have lost in the projection. It is introduced by the instability of eigenvectors to small perturbations. The amount of perturbation largely depends on the eigengap of the perturbed matrix, and this is the reason why we introduced a threshold: we rather pay a constant regret in the first iterations \( t < t_0 \) rather than advancing in a random direction; once \( G_t \) has a sufficiently large eigengap, this interaction can be controlled using Davis-Kahan [Davis63] theorem below.

Recall \( t_0 \) is defined as the iteration such that \( G_t \) has an eigengap of at least \( \alpha_t \), meaning \( \lambda_t(G_t) > \alpha(t) \) for all non-zero eigenvalue of \( G_t \), and for all \( t > t_0 \). While \( t < t_0 \), we will assume the worst-case, where \( \lambda_t(G_t) < \alpha(t) \), meaning we are zeroing out the whole rowspace of \( G_t \), and then \( \langle \nabla_{\theta_t}, \theta_t - \theta^* \rangle |_{C_t} = \langle \nabla_{\theta_t}, \theta_t - \theta^* \rangle \). In this case,

\[ \left| \sum_{t=0}^{t_0} \langle \nabla_{\theta_t}, \theta_t - \theta^* \rangle \right| \leq \left| \sum_{t=0}^{t_0} \langle \nabla_{\theta_t}, \theta_t - \theta^* \rangle \right| \]

\[ \leq t_0 C_{ada} L \] (15)

Now we consider the iterations \( t > t_0 \), where the eigengap of \( G_t \) is larger than \( \alpha_t \). For a matrix \( A \), let \( P_A^k \) denote the projection onto its top-k eigenvectors. Note that \( (I - P_A^k) \) projects onto the span of the bottom \( p - k \) eigenvectors.

The inner product can be controlled using a variant of Davis and Kahan’s classic \( \sin \theta \) theorem, providing bounds on angles between perturbed subspaces:

**Theorem 3.2** (Davis-Kahan Theorem). For any matrices \( A \) and \( B \) of like dimensions, for which \( \lambda_i(A) > \lambda_j(B) \),

\[ \|P_A(I - P_B^{i-1})\|_{op} \leq \frac{\|A - B\|_{op}}{\lambda_i(A) - \lambda_j(B)} \] (16)

We will rely on a slight extension of this version of Davis-Kahan which keeps track of the projections on the right-hand side of the above inequality; the result we need is shown in Lemma 3.4.
Lemma 3.3. For $H_t$, $G_t$, $B_t$ as in the statement of the theorem, assuming $\lambda_{\min > 0}(G_t) > \alpha(t)$ for $t > t_0$, and $\text{rank}(G_t) \geq 1$, for any $\eta > 0$ there is some universal $c$ such that the event

$$E = \bigcap_{t=t_0}^{T} E_t = \bigcap_{t=t_0}^{T} \left\{ \|P_{G_t}^{(j_t)} B_t\|_{op} \leq c \sqrt{\text{rank}(G_t) \log(T)} \right\}$$

satisfies

$$\Pr(E) > 1 - \eta$$

Proof. We will consider the complement of these sets $E_t$, and show that their probabilities sum up to some small constant. Let $\eta > 0$.

Begin by noting that Lemma 3.1 immediately implies

$$\Pr(\|B_t\|_{op} \geq A \sqrt{p \log(t)}) \leq C \exp\left(-cA \log(t)p\right)$$

and therefore an appropriate choice of $A$ can be made such that

$$\sum_{t=0}^{T} \Pr(\|B_t\|_{op} \geq A \sqrt{p \log(t)}) \leq 1 - \eta$$

And we will show that the probability of the events $E_t$ can sum up to some small value under the assumption $\|B_t\|_{op} \leq c_1 \sqrt{p \log(t)}$ for some universal $c_1$.

Note that

$$\|P_{G_t}^{(j_t)} B_t (I - P_{G_t}^{(k_t)})\|_{op} \leq \|P_{G_t}^{(j_t)} B_t\|_{op}.$$ 

Now, $P_{G_t}^{(j_t)}$ can be written as multiplication by $V \Sigma V^T$ for $V$ orthogonal, $\Sigma$ diagonal matrix of 1s and 0s associated to the appropriate eigenvalues. Since the GOE is invariant under orthogonal conjugations, this implies that the distribution of $P_{G_t}^{(j_t)} B_t$ is identical to the GOE distribution on matrices of $\text{rank}(C_t) \leq \text{rank}(G_t)$. That is, for $\simeq$ denoting distributional equality,

$$P_{G_t}^{(j_t)} B_t = V \Sigma V^T B_t \simeq V \Sigma V^T V B_t V^T = V \Sigma B_t V^T \simeq VMV^T$$

where $M$ is the GOE over $\text{rank}(\Sigma) \times \text{rank}(\Sigma)$ matrices.

Therefore applying Lemma 3.1 again to this lower-dimensional GOE, we obtain

$$\Pr(\|P_{G_t}^{(j_t)} B_t\|_{op} \geq A \sqrt{\text{rank}(G_t) \log(t)}) \leq C \exp\left(-cA \log(t)\text{rank}(G_t)\right) \leq C \exp\left(-cA \log(t)\right)$$

and again an appropriate choice of $A$ implies

$$\Pr(\|B_t\|_{C_t} \geq A \sqrt{\text{rank}(G_t) \log(t)}) \leq 1 - \eta$$

and the conclusion follows a fortiori.

Lemma 3.4. For $G_t$, $B_t$ as in the theorem statement, we have the following for any $j_t, k_t \in \mathbb{Z}^+$:

$$\|P_{G_t}^{(j_t)} (I - P_{G_t}^{(k_t)})\|_{op} \leq \frac{\|P_{G_t}^{(j_t)} B_t\|_{op}}{\lambda_{j_t}(G_t) - \lambda_{k_t+1}(G_t + B_t)}$$

Proof. Our proof of this statement will essentially follow the body of the proof of Theorem 7 in [McS04]. Notice that we need only consider the symmetric case, as $G_t$ and $B_t$ are both symmetric matrices.

Let $A = G_t$, $B = G_t + B_t$. Algebraic manipulation shows

$$P_{A}^{(j_t)} B_t (I - P_{B}^{(k_t)}) = AP_{A}^{(j_t)} (I - P_{B}^{(k_t)}) - P_{A}^{(j_t)} (I - P_{B}^{(k_t)}) B$$

7
Let $x$ be an extremizer of the operator norm of $P_{G_t}^{(j_t)}(I - P_{G_t+B_t}^{(k_t)})$. Then
\[
|P_A^{(j_t)}B_t(I - P_B^{(k_t)})x| \geq |AP_A^{(j_t)}(I - P_B^{(k_t)})x| - |P_A^{(j_t)}(I - P_B^{(k_t)})Bx|
\]
Therefore
\[
\|P_A^{(j_t)}B_t\|_{op} \geq \|P_A^{(j_t)}B_t(I - P_B^{(k_t)})\|_{op} \geq \lambda_{j_t}(G_t)\|P_A^{(j_t)}(I - P_B^{(k_t)})\|_{op} - \lambda_{k_t}(G_t + B_t)\|P_A^{(j_t)}(I - P_B^{(k_t)})\|_{op}
\]
and rearranging yields the statement.

Recall that $C_t$ corresponds to the intersection of the least $p - k_t$ eigenvectors of $H_t$ and the rowspace($G_t$). That is, for $j_t$ the index of the smallest nonzero eigenvalue of $G_t$,
\[
\Pi_{C_t} = P_{G_t}^{j_t}(I - P_{G_t+B_t}^{k_t})
\]
where we replaced $P_{H_t}^{k_t}$ with $P_{G_t+B_t}^{k_t}$, as they share top $k_t$ eigenvectors.

Due to the eigengap assumption on $G_t$ for large $t$,
\[
\lambda_{k_t+1}(G_t + B_t) = O(\sqrt{p}\sigma_B \log(t)) \leq c\lambda_{\min>0}(G_t)
\]
for $c < 1$. Therefore we may apply Lemma 3.4

\[
\langle \nabla_t, \theta_t - \theta^* \rangle_{C_t} = \nabla_t^T P_{G_t}^{j_t}(I - P_{H_t}^{k_t})(\theta_t - \theta^*)
\leq \|\nabla_t\|_2\|P_{G_t}^{j_t}(I - P_{H_t}^{k_t})(\theta_t - \theta^*)\|_2
\leq C_{\text{ada}}L\|P_{G_t}^{j_t}(I - P_{H_t}^{k_t})\|_{op}
\leq C_{\text{ada}}L\|P_{G_t}^{j_t}B_t\|_{op} \lambda_{j_t}(G_t) - \lambda_{k_t+1}(G_t + B_t)
\leq C_{\text{ada}}L\|P_{G_t}^{j_t}B_t\|_{op} \lambda_{j_t}(G_t) - c\sqrt{p}\log(t)
\leq C_{\text{ada}}L\sqrt{\text{rank}(G_t)}\sigma_B \log(t) \lambda_{\min>0}(G_t)
\]

Where 17 uses Lemma 3.4, 18 follows due to $t > t_0$, and 19 uses the conclusions of Lemma 3.3

3.4 Second Term: $\|\theta_t - \theta^*\|_{H_t}^2 - \|\theta_{t+1} - \theta^*\|_{H_t}^2$

\[
\sum_{t=0}^{T}(\|\theta_t - \theta^*\|_{H_t}^2 - \|\theta_{t+1} - \theta^*\|_{H_t}^2)
\leq \|\theta_0 - \theta^*\|_{H_0}^2 - \|\theta_{T+1} - \theta^*\|_{H_T}^2 + \sum_{t=1}^{T-1}(\theta_t - \theta^*)(H_t - H_{t-1})(\theta_t - \theta^*)
\]

The first term on the right hand side is 0, since $H_t = 0$ and the second one is non-positive thanks to the projection step, so we can bound this entire term as
\[
\leq \sum_{t=1}^{T} C_{\text{ada}}^2 \sigma_{\max}(H_t - H_{t-1})
\]

By linearity of the trace and projections,

\[
= C_{\text{ada}}^2 \sum_{t=1}^{T-1} \text{Tr}(\Pi_k(G_t)) - \text{Tr}(\Pi_k(G_{t-1})) + \text{Tr}(\Pi_k(B_t)) - \text{Tr}(\Pi_k(B_{t-1}))
\]

Since \(G_t - G_{t-1}\) is positive semi-definite,

\[
\leq C_{\text{ada}}^2 \sum_{t=1}^{T-1} \text{Tr}(G_t) - \text{Tr}(G_{t-1}) + \text{Tr}(\Pi_k(B_t)) - \text{Tr}(\Pi_k(B_{t-1}))
\]

Now, \(E[\Pi_k(B_t)] = 0\), 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

\[
E \left[ \sum_t \left( \| \theta_t - \theta^* \|_{H_t^{-1}}^2 - \| \theta_{t+1} - \theta^* \|_{H_t^{-1}}^2 \right) \right]
\]

\[
\leq E \left[ C_{\text{ada}}^2 \sum_{t=1}^{T-1} \text{Tr}(G_t) - \text{Tr}(G_{t-1}) + \text{Tr}(\Pi_k(B_t)) - \text{Tr}(\Pi_k(B_{t-1})) \right]
\]

\[
\leq E \left[ C_{\text{ada}}^2 \sum_{t=1}^{T-1} \text{Tr}(G_t) - \text{Tr}(G_{t-1}) \right]
\]

\[
= E[C_{\text{ada}}^2 \text{Tr}(G_T)]
\]

(20)

\section{3.5 \(E[\| b_t \|_{H_t^{-1}}^2]\)}

For the third and fourth term we will used the following lemma.

\textbf{Lemma 3.5.} Define \(H = G + B\), for \(H, G, B \in \mathbb{R}^{p \times p}\) invertible matrices such that \(2\|B\|_{op} \leq \lambda_p(G)\) where \(\lambda_p(G)\) denotes the eigengap for matrix \(G\). Then for \(v \in \mathbb{R}^p\),

\[
\| v \|_{H^{-1}}^2 \leq \frac{4}{3} v^T G^{-1} v
\]

\textbf{Proof.} Since \(G\) and \(B\) can be inverted, we can use Woodbury identity to calculate \(H^{-1} = (G+B)^{-1}\).

\[
\| v \|_{H^{-1}}^2 = v^T H^{-1} v
\]

\[
= v^T (G^{-1} - G^{-1}(GB^{-1} + I)^{-1}) v
\]

\[
= v^T G^{-1} v - v^T G^{-1}(GB^{-1} + I)^{-1} v
\]

\[
\leq v^T G^{-1} v + |\text{Tr}(v^T G^{-1}(GB^{-1} + I)^{-1} v)|
\]

(21)
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,

\[ \|v^T G^{-1} v + |\text{Tr}(v u^T G^{-1}(GB^{-1} + I)^{-1})| \]  

Using the trace duality property,

\[ \leq v^T G^{-1} v + \|v^T G^{-1} v\|_1 \|(GB^{-1} + I)^{-1}\|_\infty \]  

\[ \leq v^T G^{-1} v \left( 1 + \max_{i,j} \frac{\lambda_i(B)}{\lambda_i(B) + \lambda_j(G)} \right) \]  

(24)

(25)

We have that \( \lambda_j(G) \geq \lambda_p(G_t) \). Since 2\(\|B\|_{op} \leq \lambda_j(G) \) for all \( j \), the term on the right is maximized when \( \lambda_j(G) = \lambda_p(G) \), and \( \lambda_i(B) = \frac{\lambda_p(G)}{2} \)

\[ \leq v^T G^{-1} v \left( 1 + \frac{\lambda_p(G)/2}{\lambda_k(G)/2 + \lambda_p(G)} \right) \]  

\[ \leq v^T G^{-1} v (1 + 1/3) \]  

\[ \leq \frac{4}{3} v^T G^{-1} v \]  

(26)

\[ \square \]

Now, to bound the norm of \( b_t \) under \( H_t^{-1} \). We proceed to decompose \( \mathbb{R}^p \) into mutually orthogonal subspaces, concretely, into the mutually orthogonal row and null spaces of \( G_t \). Abusing notation we use \( G_t \) and \( B_t \) for the projected operators.

\[ \mathbb{R}^p = \ker(G_t) \perp \bigoplus \ker(G_t) \]

Call these spaces respectively \( A \) and \( B \). These cover all of \( \mathbb{R}^p \), and are mutually orthogonal. Therefore we may write:

\[ \|b_t\|_{H_t^{-1}}^2 = (b_t, H_t^{-1}b_t) \]  

\[ = (b_t, H_t^{-1}A b_t) + (b_t, H_t^{-1}B b_t) \]

On \( A \), both \( B_t \) and \( G_t \) can be inverted as both linear transformations are full-rank on this subspace. Further, every nonzero eigenvalue of \( H_t \) is at least \( \alpha(t) \), whereas by the restriction assumed by Lemma 3.3, the eigenvalues of \( B_t \) are at most \( \frac{\alpha(t)}{3} \). Consequently by Weyl’s inequality the minimum nonzero eigenvalue of \( \Pi_{\ker(H_t)^\perp} G_t \) must be at least \( \frac{2 \alpha(t)}{3} \). Therefore we can use Lemma 3.3 and that \( b_t \) is zero mean spherical noise with variance \( \sigma_b^2 \) to obtain

\[ \mathbb{E}_{b_t}[\|b_t\|_{H_t^{-1}}^2] \leq \frac{4}{3} b_t^T G_t^{-1} b_t \]

\[ = \frac{4 \sigma_b^2}{3} \text{Tr}(G_t^{-1}) \]

\[ \leq \frac{4 \sigma_b^2}{3} \cdot \text{rank}(G_t) \]

\[ \leq \frac{4 \sigma_b^2}{3} \cdot \frac{\text{rank}(G_t)}{3 \lambda_{\min > 0}(G_t)} \]  

(27)

We claim that the composition of projections onto the top-\( k_t \) eigenspace of \( G_t + B_t \) and the kernel of \( G_t \) is 0. This conclusion can alternately be stated as: no vector in the kernel of \( G_t \) can be in the top-\( k_t \) eigenspace of \( B_t + G_t \).

As \( H_t = B_t \) on \( \ker(G_t) \), this conclusion is implied by showing that

\[ \|B_t\|_{op} < \lambda_{k_t}(H_t) \]
but this is immediate by assumption. Therefore $H_t$ is in fact the zero operator on $B$, and does not contribute to the bound on $\|b_t\|_{H_t^{-1}}^2$.

Taking the sum over $t$,

$$
\sum_t \mathbb{E}_{b_t} \left[ \|b_t\|_{H_t^{-1}}^2 | b_1, ..., b_{t-1}, B_1, ..., B_T \right] \leq \sigma_B^2 \left( \frac{4}{3} \sum_{t=1}^{T} \frac{\text{rank}(G_t)}{\lambda_{\min>0}(G_t)} \right)
$$

(28)

Taking expectation over the remaining terms,

$$
\mathbb{E} \left[ \sum_t \|b_t\|_{H_t^{-1}}^2 \right] \leq \sigma_B^2 \mathbb{E} \left[ \frac{4}{3} \sum_{t=1}^{T} \frac{\text{rank}(G_t)}{\lambda_{\min>0}(G_t)} \right]
$$

(29)

### 3.6 Fourth term: $\|\nabla t\|_{H_t^{-1}}$

Paralleling the proof in the previous section, Section 3.5, using the space decomposition and Lemma 3.5, we have that

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

(30)

Below we will bound this term using the following lemma.

**Lemma 3.6** (Lemma 5.15 in [Haz19]),

$$
\sum_t \|\nabla t\|_{G_t^{-1}}^2 \leq 2 \text{Tr}(G_T)
$$

Taking the sum over $t$, applying Lemma 3.6 and taking expectation over the conditioned terms,

$$
\mathbb{E} \left[ \sum_t \|\nabla t\|_{H_t^{-1}}^2 \right] \leq \mathbb{E} \left[ \sum_t \|\nabla t\|_{H_t^{-1}}^2 \right] \leq \mathbb{E} \left[ \frac{4}{3} \sum_t \nabla_t^T G_t^{-1} \nabla_t \right] \leq \mathbb{E} \left[ \frac{4}{3} \cdot 2 \cdot \text{Tr}(G_T) \right]
$$

(31)

Finally, putting together the four expressions,

$$
\begin{align*}
h \left( \frac{1}{T} \sum_t \theta_t \right) - h(\theta^*) & \leq \mathbb{E} \left[ \left( \frac{C_{\text{ada}}^2}{2\eta T} + \frac{8\eta}{3 \cdot 2T} \right) \text{Tr}(G_T) \right] \\
& \quad + \frac{\eta \sigma_B^2}{2T} \mathbb{E} \left[ \sum_t \frac{\text{rank}(G_t)}{\lambda_{\min>0}(G_t)} \right] \\
& \quad + \frac{t_0 L C_{\text{ada}}}{T} + \frac{C_{\text{ada}} L}{T} \sum_{t=t_0}^{T} c_1 \sqrt{\frac{\text{rank}(G_t) \sigma_B}{\lambda_{\min>0}(G_t)}}
\end{align*}
$$

(32)

and replacing $\eta = C_{\text{ada}}$
\[ h\left(\frac{1}{T} \sum_{t} \theta_t\right) - h(\theta^*) \leq \mathbb{E} \left[ \frac{C_{\text{ada}}}{2T} \left( \frac{11}{3} \text{Tr}(G_T) + \sigma_b^2 \sum_{t} \frac{\text{rank}(G_t)}{\lambda_{\min>0}(G_t)} + 2t_0LC_{\text{ada}} + 2 \sum_{t=t_0}^{T} c_1 \sqrt{\text{rank}(G_t)} \sigma_B \right) \right]. \] (33)

The result follows by replacing \(\text{Tr}(G_T)\) above using the following lemma, that relates \(G_t\) and the best pre-conditioner at hindsight.

Lemma 3.7.
\[
\sqrt{\min_{H \in \mathcal{H}} \sum_{t} \|\nabla_t\|_H^2} = \text{Tr}(G_T)
\] (34)

Remark: Notice that in step 23 the \(L_1\) and \(L_\infty\) norms can be replaced by any \(p\) and \(q\) such that \(\frac{1}{p} + \frac{1}{q} = 1\) to obtain a better bound.

4 Private Pre-conditioned Gradient Descent for ERM

In this section we will use Noisy-AdaGrad algorithm to define an \((\varepsilon, \delta)\)-differentially private algorithm \(A_{\text{priv}}\) that approximately minimizes the excess empirical risk defined in (1). 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 \(\theta_{\text{priv}} = \frac{1}{T} \sum_{t=1}^{T} \theta_t\) output by Algorithm 1 (Algorithm \(A_{\text{noisy-AdaGrad}}\)), then
  \[ \mathbb{E} \left[ \text{Risk}(\theta_{\text{priv}}) \right] \leq \mathbb{E} \left[ \text{Regret_T}(F; A_{\text{noisy-AdaGrad}}) \right]. \] (This follows from standard use of Jensen’s inequality.)

- **Computing \((\varepsilon/2, \delta/2)\)- private pre-conditioner:** With the above formulation, standard use of Renyi composition theorem [Mir17] implies that ensuring \(\sigma_B = O\left(\frac{L^2 \sqrt{T \log(1/\delta)}}{\varepsilon n}\right)\) in Algorithm \(A_{\text{noisy-AdaGrad}}\) ensures \((\varepsilon/2, \delta/2)\)-differential privacy to the computation of all the \(H_i\)’s in Algorithm \(A_{\text{noisy-AdaGrad}}\).

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

With these observations, and composition for \((\varepsilon, \delta)\)-differential privacy [DR14] we can ensure the above variant of noisy AdaGrad is \((\varepsilon, \delta)\)-differentially private. In the following, will use the online to batch conversion mentioned above to bound the excess empirical risk. In particular, we obtain a bound of \(O\left(\frac{1}{\varepsilon n}\right)\) that does not depend 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 \(\Theta\left(\sqrt{p}/(\varepsilon n)\right)\) for private ERM via differentially private gradient descent [BST14].

**Corollary 4.1.** Assume \(C_{\text{ada}} = L = O(1)\) and learning rate \(\eta = O\left(\frac{1}{\varepsilon n}\right)\). If the rank is constant, \(k = O(1), \lambda_{\min>0}(G_t) = \Theta\left(\sqrt{T}\right)\) and accordingly \(t_0 = O(k \sigma_B^2)\), then for \(T = \varepsilon^2 n^2\) the excess empirical risk is \(\bar{O}\left(\frac{1}{\varepsilon n}\right)\), where \(\bar{O}\) hides poly-logarithmic factors in \(T, n\) and \(1/\delta\).

**Proof.** Replacing the corresponding values in Theorem 2.1, the excess empirical risk is
\[ \mathbb{E} \left[ \text{Risk}(\theta^{\text{priv}}) \right] \leq \mathbb{E} \left[ \text{Regret}_T(\mathcal{F}; \mathcal{A}_\text{noisy-AdaGrad}) \right] \]

\[ = \tilde{O} \left( \frac{1}{T} \left[ \left( \sigma_b + \frac{1}{\sqrt{T}} \right) \sqrt{\min_{H \in \mathcal{H}} \sum_t \|\nabla_t \|^2_H + \sigma_b \sum_t \frac{1}{\sqrt{T}} + \frac{T}{\varepsilon_n} + \frac{\sqrt{T}}{\varepsilon_n \sqrt{T}} } \right) \right) \]

\[ = \tilde{O} \left( \frac{1}{T} \left[ \left( \frac{\sqrt{T}}{\varepsilon_n} + \frac{\varepsilon_n}{\sqrt{T}} \right) \sqrt{T} + \frac{\sqrt{T}}{\varepsilon_n} \sqrt{T} + \frac{T}{\varepsilon_n} + \frac{\varepsilon_n}{\varepsilon_n} \right) \right) \]

\[ = \tilde{O} \left( \frac{\varepsilon_n}{T} + \frac{1}{\varepsilon_n} \right) \]

\[ = \tilde{O} \left( \frac{1}{\varepsilon_n} \right). \]

Alternatively, assuming the worst-case scenario, \( G_t \), we get a similar bound with higher dependence on dimension.

**Corollary 4.2.** Assume \( \mathcal{C}_\text{ada} = L = O(1) \). If \( \eta = O(\frac{C_{\text{ada}}}{\sqrt{p} \sigma_b}) \), \( \lambda_{\min}(G_t) = \Theta(\frac{\sqrt{T}}{p}) \), and \( t_0 = \Theta(k \sigma_b^2 \log t) \) then for \( T = \varepsilon^2 n^2 \) the excess empirical risk is \( \tilde{O} \left( \frac{\sqrt{T}}{\varepsilon n} \right) \), where \( \tilde{O} \) hides poly-logarithmic factors in \( T, n, \) and \( (1/\delta) \).

**Proof.** Notice that if all eigenvalues of \( G_t \) are growing, then for \( t > t_0 \) rowspace(\( G_t ) = \mathbb{R}^p \), and \( C_t = \ker(G_t + B_t) = \emptyset \), since \( \lambda_p(H_t) > \alpha(t) \), so we do not need to account for the last term.

Replacing the corresponding values in Theorem 2.1, we obtain

\[ \mathbb{E} \left[ \text{Risk}(\theta^{\text{priv}}) \right] \leq \mathbb{E} \left[ \text{Regret}_T(\mathcal{F}; \mathcal{A}_\text{noisy-AdaGrad}) \right] \]

\[ = \tilde{O} \left( \frac{1}{T} \left[ \left( \sigma_b \frac{1}{\sqrt{p}} + \frac{1}{\sqrt{p} \sigma_b} \right) \sqrt{\min_{H \in \mathcal{H}} \sum_t \|\nabla_t \|^2_H + \sigma_b \sum_t \frac{p}{\sqrt{T}} } \right) \right) \]

\[ = \tilde{O} \left( \frac{1}{T} \left[ \left( \frac{\sqrt{T}}{\varepsilon n} \frac{1}{\varepsilon_n \sqrt{T}} \right) \sqrt{T} + \frac{\sqrt{T}}{\varepsilon_n \sqrt{T}} p \sqrt{T} \right) \right) \]

\[ = \tilde{O} \left( \frac{\sqrt{p}}{\varepsilon n} + \frac{\varepsilon_n}{\sqrt{p} T} + \frac{\sqrt{T}}{\varepsilon_n} \right) \]

\[ = \tilde{O} \left( \frac{\sqrt{p}}{\varepsilon n} \right). \]

Our main contribution is in the low-rank unconstrained setting where, compared with original Adagrad, we only pay an additional factor \( \tilde{O} \left( \frac{1}{\varepsilon n} \right) \), independent of dimension.

Our results do not contradict the lower bound of \( \tilde{\Omega}(\frac{\sqrt{T}}{\varepsilon n}) \) for \( \varepsilon, \delta \)-Differentially Private minimization algorithms since we are working in the unconstrained setting. In the full rank setting, we obtain an excess risk \( \tilde{O}(\frac{\sqrt{T}}{\varepsilon n}) \) as DP-SGD.

### 4.1 Infeasibility of Extending Our Results to Constrained ERM

Lower bounds on constrained private ERM show that error rate scaling with dimension \( p \) is inevitable [BST14]. Our results show however dimension independent bounds, and this is due to the unconstrained setting assumption: concretely, by avoiding the projection to a constraint set in equation 8. Under constrained optimization this step would become
\[ \| z_{t+1} - \theta^* \|_{H_t} = 0 \]

Figure 1: Infeasibility of extending results in Theorem 2.1 to constrained ERM. When projecting \( \theta_t \). Before projection, \( \| \theta_t - \eta H_t^{-1} \nabla_t - \theta^* \|_{H_t}^2 = 0 \), but after projecting \( \| \Pi_C(\theta_t - \eta H_t^{-1} \nabla_t) - \theta^* \|_{H_t}^2 = \frac{\sqrt{2} - 1}{2} \) making impossible the following step, since it is not true in general that

\[ \| \Pi_C(\theta_t - \eta H_t^{-1} \nabla_t) - \theta^* \|_{H_t}^2 \leq \| \theta_t - \eta H_t^{-1} \nabla_t - \theta^* \|_{H_t}^2. \]

To illustrate this, let consider the example in Figure 1. In this case, \( C \) is the \( \ell_2 \) ball centered at the origin in \( \mathbb{R}^2 \), \( \theta^* = [0, \frac{1}{2}] \), \( z_{t+1} = \theta_t - \eta H_t^{-1} \nabla_t = [1, 1] \), and a positive semidefinite matrix defining the norm \( H = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \), meaning it takes only coordinate \( y \) into account. Consequently,

\[ \| \Pi_C(z_{t+1}) - \theta^* \|_{H_t}^2 = \frac{\sqrt{2} - 1}{2} > \| z_{t+1} - \theta^* \|_{H_t}^2 = 0. \]

References

[ACG+16] Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security (CCS’16), pages 308–318, 2016.

[BFTT19] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alchê-Buc, Emily B. Fox, and Roman Garnett, editors, 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.

[BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In Proc. of the 2014 IEEE 55th Annual Symp. on Foundations of Computer Science (FOCS), pages 464–473, 2014.
[BTN01] Aharon Ben-Tal and Arkadi Nemirovski. Lectures on modern convex optimization: analysis, algorithms, and engineering applications. SIAM, 2001.

[CMS11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. Journal of Machine Learning Research, 12(Mar):1069–1109, 2011.

[Dav63] Chandler Davis. The rotation of eigenvectors by a perturbation. Journal of Mathematical Analysis and Applications, 6(2):159–173, 1963.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research, 12(7), 2011.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4):211–407, 2014.

[FKT20] 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.

[HAK07] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. Machine Learning, 69(2-3):169–192, 2007.

[Haz19] Elad Hazan. Introduction to online convex optimization. arXiv preprint arXiv:1909.05207, 2019.

[INS+19] 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.

[JT14] Prateek Jain and Abhradeep Guha Thakurta. (near) dimension independent risk bounds for differentially private learning. In International Conference on Machine Learning, pages 476–484, 2014.

[McS04] Frank McSherry. Spectral methods for data analysis. PhD thesis, Citeseer, 2004.

[Mir17] Ilya Mironov. Rényi differential privacy. In 2017 IEEE 30th Computer Security Foundations Symposium (CSF), pages 263–275. IEEE, 2017.

[MRTZ17] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. arXiv preprint arXiv:1710.06963, 2017.

[MS10] 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.

[PSY+19] 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.

[SCS13] 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.

[SS+11] Shai Shalev-Shwartz et al. Online learning and online convex optimization. Foundations and trends in Machine Learning, 4(2):107–194, 2011.

[SSSS09] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009, 2009.

[STT20] Shuang Song, Om Thakkar, and Abhradeep Thakurta. Characterizing private clipped gradient descent on convex generalized linear problems. arXiv preprint arXiv:2006.06783, 2020.
[TAM19] Om Thakkar, Galen Andrew, and H. Brendan McMahan. Differentially private learning with adaptive clipping. *CoRR*, abs/1905.03871, 2019.

[Tao12] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[WLK+17] 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 Semih Salihoglu, Wenchao Zhou, Rada Chirkova, Jun Yang, and Dan Suciu, editors, *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD*, 2017.

[ZWB20] Yingxue Zhou, Zhiwei Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private sgd with gradient subspace identification. *arXiv preprint arXiv:2007.03813*, 2020.