Average Stability is Invariant to Data Preconditioning. Implications to Exp-concave Empirical Risk Minimization

Alon Gonen　　　　　　　　　　　　　　　　　　　　　　ALONGNN@CS.HUJI.AC.IL
School of Computer Science and Engineering
The Hebrew University
Jerusalem, Israel

Shai Shalev-Shwartz　　　　　　　　　　　　　　　　　　　SHAIS@CS.HUJI.AC.IL
School of Computer Science and Engineering
The Hebrew University
Jerusalem, Israel

Editor: Ingo Steinwart

Abstract

We show that the average stability notion introduced by Kearns and Ron (1999); Bousquet and Elisseeff (2002) is invariant to data preconditioning, for a wide class of generalized linear models that includes most of the known exp-concave losses. In other words, when analyzing the stability rate of a given algorithm, we may assume the optimal preconditioning of the data. This implies that, at least from a statistical perspective, explicit regularization is not required in order to compensate for ill-conditioned data, which stands in contrast to a widely common approach that includes a regularization for analyzing the sample complexity of generalized linear models. Several important implications of our findings include: a) We demonstrate that the excess risk of empirical risk minimization (ERM) is controlled by the preconditioned stability rate. This immediately yields a relatively short and elegant proof for the fast rates attained by ERM in our context. b) We complement the recent bounds of Hardt et al. (2015) on the stability rate of the Stochastic Gradient Descent algorithm.

1. Introduction

Central to statistical learning theory is the notion of (algorithmic) *stability*. Since being introduced by Bousquet and Elisseeff (2002), deep connections between the *generalization* ability and the algorithmic stability of a learning algorithm have been established. It was shown by Shalev-Shwartz et al. (2010); Mukherjee et al. (2006) that stability characterizes learnability. Furthermore, in expectation, some notion of stability exactly equals the generalization error of an algorithm (namely, to the gap between true loss and train loss).

For generalized linear learning problems, a prominent geometric property which upper bounds the stability rate is the condition number of the loss function. While uniform convergence bounds Shalev-Shwartz and Ben-David (2014)[Chapter 4] mostly yield bounds that scale with $1/\sqrt{n}$, where n is the size of the sample, *well-conditioned* problems admit faster (stability) rates that scale linearly with $1/n$. The caveat is that typical (large-scale) machine learning problems are ill-conditioned. While we defer the precise definition of the *condition number* to the next part, let us mention that the condition number is controlled

by two related quantities corresponding to both the choice of the loss function and the choice of the coordinate system. In a nutshell, our paper establishes the following result:

> The average stability of ERM is invariant to the choice of the coordinate system.

While this observation admits a one-line proof, it has far-reaching implications. In particular, in this paper we use this observation to establish fast rates for empirical risk minimization.

The rest of the paper is organized as follows. In Section 2 we define the setting and proceed to provide basic definitions and results in stability analysis. In Section 3 we state and prove our main result. Section 4 discusses the implications to linear regression as well as improved bounds on the stability of SGD. Related work is discussed in Section 5.

2. Preliminaries

2.1 Setup

We consider the problem of minimizing the *risk* associated with *generalized linear model*:

$$\min_{w \in \mathcal{W}} L(w) := \mathbb{E}_{(x,y) \sim \mathcal{D}}[\phi_y(w^\top x)] \ . \tag{1}$$

Here, both the *domain* $\mathcal{W}$ and the *instance space* $\mathcal{X}$ are assumed to be compact and convex subsets of $\mathbb{R}^d$. We denote by $\mathcal{D}$ an arbitrary probability distribution defined over $\mathcal{X} \times \mathcal{Y}$. Each element y in the *label set* $\mathcal{Y}$ induces a twice differentiable[1] *loss function* of the form $\phi_y : \{w^\top x : w \in \mathcal{W}, x \in \mathcal{X}\} \to \mathbb{R}_+$. We make the following assumptions on the loss function:

(A1) For each $y \in \mathcal{Y}$, ϕ_y is ρ-Lipschitz, i.e., $|\phi_y'(z)| \le \rho$ for all z.
(A2) For each $y \in \mathcal{Y}$, ϕ_y is α-strongly convex, i.e., $\phi_y''(z) \ge \alpha$ for all z.

Our main example is the following formulation of *linear regression* Orabona et al. (2012).

Example 1 (Linear Regression:) *Let $\mathcal{X}$ be any compact and convex subset of $\mathbb{R}^d$ and $\mathcal{Y}$ be an interval of the form $[-Y, Y]$. The domain $\mathcal{W}$ is given by*

$$\mathcal{W} = \{w \in \mathbb{R}^d : (\forall x \in \mathcal{X}) \ |w^\top x| \le Y\} \ .$$

For all $y \in \mathcal{Y}$, let ϕ_y be the square loss, $\phi_y(z) = \frac{1}{2}(z - y)^2$. Note that for any $y \in \mathcal{Y}$ and $z \in \{w^\top x : w \in \mathcal{W}, x \in \mathcal{X}\}$,

$$|\phi_y'(z)| = \frac{1}{2}|2(z - y)| \le |z| + |y| \le 2Y, \quad \|\phi_y''(z)\| = 1 \ .$$

Hence, the assumptions (A1-2) are satisfied with $\rho = 2Y$ and $\alpha = 1$.

More generally, our setting captures most of the known *exp-concave* functions Kivinen and Warmuth (1999). A twice-continuously differentiable function $f : \mathcal{W} \to \mathbb{R}$ is said to be $\bar{\alpha}$-exp-concave if $\nabla^2 f(w) \succeq \bar{\alpha} \nabla f(w) \nabla f(w)^\top$ for all $w \in \mathcal{W}$.

---

1. As we do not require smoothness of the loss function, our results can easily be extended to continuous but non-differentiable functions.

Lemma 1 *Consider a risk of the form (1) that satisfies the assumptions* **(A1-2)**. *Then, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the function $w \in \mathcal{W} \mapsto \phi_y(w^\top x)$ is α/ρ^2-exp concave.*

Proof Fix a pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The gradient and the Hessian of the map $\ell(w) = \phi_y(w^\top x)$ are given by

$$\nabla \ell(w) = \phi'(w^\top x)x, \quad \nabla^2 \ell(w) = \phi''(w^\top x)xx^\top . \tag{2}$$

By assumption $|\phi'(w^\top x)| \le \rho$ and $\phi''(w^\top x) \ge \alpha$, hence ℓ is α/ρ^2-exp concave. ∎

A learning algorithm $\mathcal{A}$ receives as an input a training sequence (a.k.a. sample) of n i.i.d. pairs, $S = ((x_i, y_i))_{i=1}^n \sim \mathcal{D}^n$, and outputs a predictor, $\mathcal{A}(S) \in \mathcal{W}$. The *empirical risk function*, $\hat{L} : \mathcal{W} \to \mathbb{R}$, is defined as

$$\hat{L}_S(w) = \hat{L}(w) = \frac{1}{n} \sum_{i=1}^n \underbrace{\phi_{y_i}(w^\top x_i)}_{:=\hat{\ell}_i(w)} . \tag{3}$$

In this paper we focus on the ERM algorithm, whose output is a minimizer of the empirical risk.[2] We denote the output of the ERM by $\hat{w}(S)$, or simply $\hat{w}$ when S is understood from the context. The generalization error and the excess risk of $\hat{w}$ are defined by $L(\hat{w}) - \hat{L}(w)$ and $L(\hat{w}) - L(w^\star)$, respectively. For ERM, it is immediate that any upper bound on the generalization error translates into the same bound on the excess risk.

Remark 2 *While we mostly focus on exact ERM, it should be emphasized that our results are easily extended to any algorithm that approximately minimizes the empirical risk. The formulation of Lemma 3 below highlights this idea.*

2.2 Stability

In this section we review basic definitions and results on stability. For completeness, we also provide proofs of the stated results.

Let $S = ((x_i, y_i))_{i=1}^n$ be a training sequence. For every $i \in [n]$, let $\hat{w}_i$ be a minimizer of the risk w.r.t. $S \setminus \{(x_i, y_i)\}$, namely,

$$\hat{w}_i \in \underset{w \in \mathcal{W}}{\operatorname{argmin}} \frac{1}{n-1} \sum_{j \ne i} \hat{\ell}_j(w) .$$

The *average stability* of ERM is defined as

$$\Delta(S, \mathcal{W}) = \frac{1}{n} \sum_{i=1}^n (\hat{\ell}_i(\hat{w}_i) - \hat{\ell}_i(\hat{w})) . \tag{4}$$

We omit the dependency on $\mathcal{W}$ when it is clear from the context. The next lemma shows that the expected generalization error of the ERM is equal to the expected average stability.

---

2. The compactness of $\mathcal{W}$ implies that both the true and the empirical risks admit minimizers.

Lemma 3 *Let $\mathcal{A}$ be a possibly randomized algorithm and denote by $\hat{w}$ its output. The generalization error of $\mathcal{A}$ satisfies*

$$\mathbb{E}_{S\sim\mathcal{D}^{n-1}}[L(\hat{w}) - \hat{L}(\hat{w})] = \mathbb{E}_{S\sim\mathcal{D}^n}[\Delta(S)] . \tag{5}$$

Furthermore, if $\mathcal{A}$ satisfies, for every sample S, $\mathbb{E}[\hat{L}(\hat{w})] \leq \min_{w\in\mathcal{W}} \hat{L}(w) + \epsilon$, where the expectation is with respect to $\mathcal{A}$'s own randomization, then the excess risk of $\mathcal{A}$ is bounded by

$$\mathbb{E}_{S\sim\mathcal{D}^{n-1}}[L(\hat{w}) - L(w^\star)] \leq \mathbb{E}_{S\sim\mathcal{D}^n}[\Delta(S)] + \epsilon . \tag{6}$$

Proof Since $\hat{w}_i$ does not depend on the i.i.d. pair (x_i, y_i),

$$\mathbb{E}_{S\sim\mathcal{D}^n}[\hat{\ell}_i(\hat{w}_i)] = \mathbb{E}_{S\sim\mathcal{D}^{n-1}}[L(\hat{w}(S))] , i = 1, \ldots, n .$$

By linearity of expectation, we obtain

$$\mathbb{E}_{S\sim\mathcal{D}^n}\left[\frac{1}{n}\sum_{i=1}^n \hat{\ell}_i(\hat{w}_i)\right] = \mathbb{E}_{S\sim\mathcal{D}^{n-1}}[L(\hat{w}(S))] .$$

Therefore,

$$\mathbb{E}[\Delta(S)] = \mathbb{E}_{S\sim\mathcal{D}^n}\left[\frac{1}{n}\sum_{i=1}^n \hat{\ell}_i(\hat{w}_i)\right] - \mathbb{E}_{S\sim\mathcal{D}^n}\left[\frac{1}{n}\sum_{i=1}^n \hat{\ell}_i(\hat{w})\right] = \mathbb{E}_{S\sim\mathcal{D}^{n-1}}[L(\hat{w})] - \mathbb{E}[\hat{L}(\hat{w})] .$$

This establishes the first claim.

Next, by assumption, for every S, $\mathbb{E}[\hat{L}(\hat{w})] \leq \hat{L}(w^\star) + \epsilon$. Hence,

$$\mathbb{E}_{S\sim\mathcal{D}^n}[\hat{L}(\hat{w})] \leq \mathbb{E}_{S\sim\mathcal{D}^n}[\hat{L}(w^\star)] + \epsilon = L(w^\star) + \epsilon .$$

Combining this inequality with the first claim, concludes the proof. ∎

2.2.1 STABILITY OF WELL-CONDITIONED OBJECTIVES

Lemma 3 motivates us to derive an upper bound on the average stability. A key quantity that governs $\Delta(S)$ is the condition number of the objective. We next provide exact definitions and discuss this relation.

Fix a training sequence S. We denote the empirical covariance matrix by

$$\hat{C} := \hat{C}(S) = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top .$$

The (average) empirical condition number of $\hat{C}$ is defined as

$$\kappa(\hat{C}) = \frac{\text{tr}(\hat{C})}{\lambda_{\min}(\hat{C})} ,$$

where $\operatorname{tr}(\hat{C})$ is the trace of $\hat{C}$ and $\lambda_{\min}(\hat{C})$ is the smallest nonzero eigenvalue of $\hat{C}$. We define the functional condition number as the ratio between the squared Lipschitz parameter and the strong convexity parameter:

$$\kappa(\phi) = \frac{\rho^2}{\alpha} \ .$$

Finally, we define the condition number of the objective as the product between the empirical and the functional condition number:

$$\kappa = \kappa(\hat{C})\kappa(\phi) \ .$$

Lemma 4 *For every training sequence S,*

$$\Delta(S) \le \frac{2\kappa}{n} = \frac{2\kappa(\hat{C})\kappa(\phi)}{n} = \frac{2\rho^2}{\alpha\,n}\kappa(\hat{C}) \ . \tag{7}$$

To the best of our knowledge, this result has only been proved in the context of regularized loss minimization (e.g., the bound on the uniform stability in Shalev-Shwartz and Ben-David (2014)[Corollary 13.6]). Inspecting the proofs, one can notice that the role of regularization is merely to ensure the strong convexity of the objective. This simple observation is crucial for our development.

Proof **(of Lemma 4)** We first assume that $\hat{C}$ is of full rank. Note that for all w, the Hessian of $\hat{L}$ at w is given by

$$\nabla^2 \hat{L}(w) = \frac{1}{n}\sum_{i=1}^{n}\phi''(w^\top x_i)x_i x_i^\top \succeq \frac{1}{n}\sum_{i=1}^{n}\alpha\, x_i x_i^\top = \alpha\,\hat{C} \ . \tag{8}$$

In particular, $\hat{L}$ is strongly convex and $\hat{w}$ is uniquely defined. Denote the strong convexity parameter of $\hat{L}$ by $\hat{\mu}$. We also denote the Lipschitz parameter of each $\hat{\ell}_i$ by $\hat{\rho}_i$ and define $\hat{\rho}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\rho}_i^2$. We will shortly derive upper and lower bounds on these parameters, but first let us relate them to the average stability.

Fix some $i \in [n]$ and let $\hat{\Delta}_i = \hat{\ell}_i(\hat{w}_i) - \hat{\ell}_i(\hat{w})$ (we do not assume that $\hat{w}_i$ is uniquely defined). The $\hat{\rho}_i$-Lipschitzness of $\hat{\ell}_i$ yields the bound

$$\hat{\Delta}_i \le \hat{\rho}_i\|\hat{w}_i - \hat{w}\| \ .$$

The $\hat{\mu}$-strong convexity of $\hat{L}$ implies (e.g. using Shalev-Shwartz (2011)[Lemma 2.8]) that

$$\frac{\hat{\mu}}{2}\|\hat{w}_i - \hat{w}\|^2 \le \hat{L}(\hat{w}_i) - \hat{L}(\hat{w}) \ .$$

On the other hand, since $\hat{w}_i$ minimizes the risk over $S \setminus \{(x_i, y_i)\}$, we have that

$$\hat{L}(\hat{w}_i) - \hat{L}(\hat{w}) = \frac{\sum_{j\ne i}(\hat{\ell}_j(\hat{w}_i) - \hat{\ell}_j(\hat{w}))}{n} + \frac{\hat{\ell}_i(\hat{w}_i) - \hat{\ell}_i(\hat{w})}{n} \le 0 + \frac{\hat{\Delta}_i}{n} \ .$$

Combining the bounds, we conclude the following bound for every $i \in [n]$:

$$\hat{\Delta}_i^2 \leq \hat{\rho}_i^2 \|\hat{w}_i - \hat{w}\|^2 \leq \frac{2\hat{\rho}_i^2}{\hat{\mu}}(\hat{L}(\hat{w}_i) - \hat{L}(\hat{w})) \leq \frac{2\hat{\rho}_i^2}{n\hat{\mu}}\hat{\Delta}_i \ .$$

Dividing by $\hat{\Delta}_i$ (we may assume w.l.o.g. that $\hat{\Delta}_i > 0$), we obtain

$$\hat{\Delta}_i \leq \frac{2\hat{\rho}_i^2}{n\hat{\mu}} \ . \tag{9}$$

Let us remark that at this point, we can deduce a bound of $\max_{i \in [n]} \frac{2\hat{\rho}_i^2}{n\hat{\mu}}$ on the uniform stability. This matches the bound in Shalev-Shwartz and Ben-David (2014)[Corollary 13.6]. We next proceed to establish the claimed bound on the average stability.

By averaging (9) over $i = 1, \ldots, n$, we obtain

$$\hat{\Delta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\Delta}_i \leq \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\rho}_i^2\right)\frac{2}{n\hat{\mu}} = \frac{2\hat{\rho}^2}{n\hat{\mu}} \ . \tag{10}$$

It remains to derive bounds on $\hat{\rho}$ and $\hat{\mu}$. Note that

$$\|\nabla\hat{\ell}_i(w)\|^2 = \|\phi'(w^\top x_i)x_i\|^2 \leq \rho^2\|x_i\|^2 = \rho^2\mathrm{tr}(x_i x_i^\top) \ .$$

Hence, $\hat{\rho}_i^2 \leq \rho^2\mathrm{tr}(x_i x_i^\top)$. By averaging, we obtain that $\hat{\rho}^2 \leq \rho^2\mathrm{tr}(\hat{C})$. Next, using (8) we obtain that $\hat{\mu} \geq \alpha\lambda_d(\hat{C})$. By substituting the bounds on $\hat{\rho}^2$ and $\hat{\mu}$ in (10), we conclude the desired bound.

Note that if $\hat{C}$ is not of full rank, we can replace each vector $x \in \mathbb{R}^d$ with $U^\top x$, where the columns of U form an orthonormal basis for span($\{x_1, \ldots, x_n\}$), without affecting $\hat{\Delta}, \hat{\Delta}_1, \ldots, \hat{\Delta}_n$ (this modification is only for the sake of the analysis). As a result, the new covariance matrix is of full rank and its eigenvalues are $\lambda_1(\hat{C}), \ldots, \lambda_{\min}(\hat{C})$. Repeating the above arguments, we conclude the proof. ∎

Let us specify the bound to linear regression as formulated in Example (1). As $\alpha = 1$ and $\rho = 2Y$, the functional condition number is $4Y^2$. Hence, the average stability is bounded by

$$\Delta(S) \leq \frac{4Y^2}{n}\hat{\kappa}(\hat{C}) \ . \tag{11}$$

Using Lemma 3 we deduce the same bound on the excess risk. The weakness of this bound stems from the fact that empirically, the empirical condition number tends to be huge (e.g., see the empirical study in Gonen et al. (2016)).

In the next section we show that the (dependence on the) empirical condition number associated with our arbitrary choice of coordinate system can be replaced by the empirical condition number obtained by an optimal preconditioning.

3. Preconditioned Stability

We are now in position to describe our main result. Let P be a (symmetric) positive definite matrix, S_P be the training set obtained by replacing every x_j with $\tilde{x}_j = P^{-1/2}x_j$, and $\mathcal{W}_P = P^{1/2}\mathcal{W}$. We call $P^{-1/2}$ a *preconditioner*. Recall the definition of average stability from Equation (4). Our main theorem is:

Theorem 5 *For any training sequence S and positive definite matrix P,*

$$\Delta(S_P, \mathcal{W}_P) = \Delta(S, \mathcal{W}) \ .$$

In words, the average stability is invariant to the choice of the coordinate system.

Proof The crucial observation is that the empirical risk minimization with respect to S_P over the domain $\mathcal{W}_P$ is equivalent to the ERM w.r.t. S over the domain $\mathcal{W}$ in the following sense. For any pair $(w, \tilde{w} = P^{1/2}w) \in \mathcal{W} \times \mathcal{W}_P$ and any $j \in [n]$, the prediction $(\tilde{w})^\top \tilde{x}_j$ is equal to the prediction $w^\top x_j$. Therefore, the empirical risks $\hat{L}_{S_p}(\tilde{w})$ and $\hat{L}_S(w)$ are equal. By associating the corresponding minimizers of the empirical risk (i.e., $\hat{w}$ is associated with $P^{1/2}\hat{w}$ and for any $i \in [n]$, $\hat{w}_i$ is associated with $P^{1/2}\hat{w}_i$), we conclude our proof. ∎

Theorem 5 tells us that we can analyze the stability of S_P instead of the stability of S. Crucially, this is true for every P, even one that is chosen based on S. Therefore, the expected suboptimality is upper bounded by the expected value of the quantity, $\inf_{P \succ 0} \Delta(S_P, \mathcal{W}_P)$, which we refer to as the *preconditioned average stability*. Equipped with this observation, we next choose P that leads to a minimal condition number, and consequently obtain a tighter bound on the excess risk.

Note that for every $P \succ 0$, the empirical covariance matrix that corresponds to the preconditioned training sequence, S_P, is

$$\frac{1}{n} \sum_{i=1}^n (P^{-1/2}x_i)(P^{-1/2}x_i)^\top = P^{-1/2} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) P^{-1/2} = P^{-1/2}\hat{C}P^{-1/2} \ .$$

When $\hat{C}$ is of full rank, by choosing $P = \hat{C}$, we obtain that

$$\kappa(\underbrace{P^{-1/2}\hat{C}P^{-1/2}}_{I}) = \frac{\mathrm{tr}(I)}{\lambda_{\min}(I)} = d \ .$$

If $\hat{C}$ is not of full rank, we can add arbitrary "noise" in directions that do not lie in the column space of $\hat{C}$. For example, by choosing $P = \hat{C} + \delta(I - \hat{C}\hat{C}^\dagger)$, (where δ can be any positive scalar), we obtain that $\kappa(P^{-1/2}\hat{C}P^{-1/2}) = \mathrm{rank}(\hat{C}) \leq d$. It is easy to see that in both cases, we obtain the minimal value of $\kappa(P^{-1/2}\hat{C}P^{-1/2})$ over all matrices $P \succ 0$. Combining this bound with Lemma 3 and Lemma 4, we arrive at the following conclusion.

Corollary 6 *Consider the optimization problem (1), where for all $y \in \mathcal{Y}$, ϕ_y is ρ-Lipschitz and α-strongly convex. The expected excess risk of empirical risk minimization is bounded by*

$$\mathbb{E}_{S \sim \mathcal{D}^{n-1}}[L(\hat{w}) - L(w^\star)] \leq \mathbb{E}_{S \sim \mathcal{D}^n}[\Delta(S)] = \mathbb{E}_{S \sim \mathcal{D}^n}[\inf_{P \succ 0} \Delta(S_P)] \leq \frac{2\rho^2 d}{\alpha n} \ .$$

4. Some Implications

4.1 Linear Regression

We start by specifying our bounds to linear regression (Example (1)).

Corollary 7 (Linear Regression) *Consider linear regression as formulated in Example (1). The expected excess risk of ERM is bounded by*

$$\mathbb{E}_{S \sim \mathcal{D}^{n-1}}[L(\hat{w}) - L(w^\star)] \leq \Delta(S) \leq \frac{4Y^2 d}{n} \ .$$

Comparing the bounds in (11) and Corollary 7, we see that the dependence on $\hat{\kappa}(\hat{C})$ is replaced by the optimal empirical condition number, $\hat{\kappa}(I) = d$. As we mentioned above, this gap tends to be huge in practice.

As we discuss in Section 5, standard bounds for this setting depend on the geometry of $\mathcal{X}$ and $\mathcal{W}$. On the contrary, it follows from the generalized Cauchy-Schwarz inequality that for any choice of a norm $\|\cdot\|$ on $\mathbb{R}^d$, our bound applies to the sets[3]

$$\mathcal{X} = \mathcal{B}_{\|\cdot\|} = \{x \in \mathbb{R}^d : \|x\| \leq 1\}, \quad \mathcal{W} = Y\mathcal{B}_{\|\cdot\|^\star} := \{w \in \mathbb{R}^d : \|w\|^\star \leq Y\} \tag{12}$$

4.2 The Average Stability of Stochastic Gradient Descent

One of the most widely used algorithms in machine learning is Stochastic Gradient Descent (SGD). Besides its computational simplicity, its popularity stems also from its generalization abilities Shalev-Shwartz and Ben-David (2014)[Section 14.5]. Recently, Hardt et al. (2015) studied the (uniform) stability of SGD in various settings. Following our notation, theorem 3.9 of their paper implies a bound of $\max_i \frac{2\hat{\rho}_i^2}{\gamma n}$ on the uniform stability, where γ is the strong convexity of the entire objective, and for any $i \in [n]$, $\hat{\rho}_i$ is the Lipschitz parameter of $\hat{\ell}_i$. As the proof of Lemma 4 reveals, γ can be bounded by $\alpha\hat{\kappa}(\hat{C})$ and $\hat{\rho}_i$ is at most $\hat{\rho}^2\|x_i\|^2$. In particular, the bound depends on the choice of the coordinate system.

As implied by Bubeck (2015)[Theorem 3.2], SGD can be viewed in our context as an (approximate) ERM. Hence, the average stability of SGD is invariant to the choice of the coordinate system and the stability rate of SGD is bounded as in Corollary 6. It should be noted that our bound holds in the regime where SGD converges to a minimizer of the empirical risk, whereas the result of Hardt et al. holds for the entire trajectory of SGD.

5. Related Work

5.1 Slower rates

One of the most direct techniques for establishing bounds on the excess risk is via analyzing the Rademacher complexity Bartlett and Mendelson (2003) of the associated class of predictors. In our setting, these techniques have been employed by Kakade et al. (2009) to

---

3. In fact, under mild additional assumptions on $\mathcal{X}$, any two sets $\mathcal{X}$ and $\mathcal{W}$ that satisfy our assumptions can be presented in this way. Assume that $\mathcal{X}$ is symmetric (i.e., $x \in \mathcal{X}$ iff $-x \in \mathcal{X}$) and $0 \in \text{int}(\mathcal{X})$. Then it is known Conway (2013) that $\mathcal{X}$ induces a norm on $\mathbb{R}^d$ through the Minkowsky functional

$$\|x\| := p(x) = \inf\{t \in \mathbb{R} : x \in tB\} \ .$$

It is immediate that the closed unit ball $\{x : \|x\| \leq 1\}$ is $\mathcal{X}$ itself. Therefore, the dual norm is simply the support function of $\mathcal{X}$

$$\|w\|^\star = \max_{x \in \mathcal{X}} w^\top x \ .$$

It follows that $\mathcal{X}$ and $\mathcal{W}$ can be described as in (12).

establish bounds of order $1/\sqrt{n}$ on the generalization error of ERM. We refer to these rates as slower due to the inferior dependence on the sample size n.

5.2 Dependence on Norm

Note that since both the uniform and the average stability of ERM are bounded above by its generalization error Shalev-Shwartz et al. (2010), the bounds of Kakade et al. (2009) translate into bounds on the average stability.

Unlike our fast rates, the exact bounds depend on the geometry of the set $\mathcal{X}$ and $\mathcal{W}$. For example: a) If both $\mathcal{X}$ and $\mathcal{W}$ are the Euclidean unit ball in $\mathbb{R}^d$, then the obtained bound scales with $1/\sqrt{n}$. b) If $\mathcal{X}$ is the unit ℓ_∞-ball and $\mathcal{W}$ is the ℓ_1-ball, then the obtained bound scales with $\sqrt{\log(d)/n}$.

5.3 Lower bounds on the excess risk

Lower bounds for stochastic minimization of exp-concave functions have been studied in Mahdavi et al. (2015). For our setting, theorem 2 in this paper implies a bound of $\Omega(d/n)$ on the excess risk of any algorithm.

For the special case of linear regression with

$$\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}, \ \mathcal{W} = \{w \in \mathbb{R}^d : \|w\|_2 \leq B\}, \ \mathcal{Y} = [-Y, Y] \tag{13}$$

Shamir (2014) proved the lower bound $\Omega\left(\min\{Y^2, \frac{B^2+dY^2}{n}, \frac{BY}{\sqrt{n}}\}\right)$ on the generalization error of ERM. The left-most term is trivially attained by the predictor $w = 0$. The middle term is attained by combining the Vovk-Azoury-Warmuth forecaster Azoury and Warmuth (2001); Vovk (2001) with standard online-to-batch conversions (Cesa-Bianchi et al. (2004)). Last, the right term is attained by ERM, as implied by the aforementioned upper bound of Kakade et al. (2009).

It is left open whether the middle term in the lower bound is attained by ERM. Note that if $B = \omega(\sqrt{d}Y)$, then the middle term in the above lower bound is asymptotically larger than our upper bound. However, since in the setting of Shamir (2014), Equation (13)) the magnitude of the predictions is not uniformly upper bounded by Y, no contradiction arises.

5.4 Stability and Regularization

Previous work Bousquet and Elisseeff (2002); Shalev-Shwartz et al. (2010) studied the rate of uniform stability in various settings. For our setting, their bounds on the expected risk are identical to the bound in Lemma 4. As we explained above, these fast rates are often worse than the so-called slower rates due to the dependence on the empirical condition number. The standard approach for tackling this problem is add a regularization term. By adding the regularization term $\lambda\|w\|^2$ to the objective, one effectively increases the eigenvalues of $\hat{C}$ by λ, and consequently, the overall condition number is decreased. However, as explained in Shalev-Shwartz and Ben-David (2014)[Section 13.4], this modification usually does not preserve the fast rates.[4]

---

4. Namely, when tuning λ, we need to ensure that any $\epsilon/2$-approximate minimizer with respect to the regularized objective is also an ϵ-approximate minimizer with respect to the unregularized objective. As

5.5 Stability and Exp-concavity

Informally, exp-concavity can be seen as a local and weaker form of strong convexity. Indeed, the Online Newton Step (ONS) of Hazan et al. (2007), which has been designed for online minimization of exp-concave functions, achieves improved (logarithmic) regret bounds that resemble the regret bounds for strongly convex functions Hazan et al. (2007). The online-to-batch analysis of Mahdavi et al. (2015) yields a bound on the excess risk that coincides with our bounds up to logarithmic factors. The main shortcoming of the ONS algorithm is that it employs expensive iterations (the runtime per iteration scales at least quadratically with d). Hence, it is natural to ask whether there exist simpler algorithms that achieve fast rates.

This question was answered affirmatively by Koren and Levy (2015). This work, which is most closely related to our work, considers the minimization of a risk of the form $F(w) = \mathbb{E}[f(w, Z)]$, where for any z, $f(\cdot, z)$ is $\bar{\beta}$-smooth[5] and $\bar{\alpha}$-exp-concave function. They established fast rates for any algorithm that minimizes the *regularized* risk $\hat{L}(w) + \frac{1}{n}R(w)$, where $R(w)$ is assumed to be a 1-strongly convex function (e.g., one can set $R(w) = \frac{1}{2}\|w\|^2$). While exp-concavity is weaker than strong convexity, Koren and Levy (2015)[section 4.2] interprets exp-concavity as strong convexity in the (local) norm induced by the outer products of the gradients and the regularization term. In other words, the problem is well-conditioned with respect to this local norm. Note that their formulation is more general in the sense that they do not assume a GLM structure. However, it should be emphasized that all the known exp-concave functions in machine learning are of the form (1)).

The above interpretation of Koren and Levy (2015) inspired us to make one step forward and directly show that regularization is not required as long as a related (preconditioned) problem is well conditioned. Besides the obvious importance of showing the insignificance of regularization in this context, we believe that the notion of preconditioned stability and its relation to the excess risk make these ideas more transparent and simplify the proofs.

The upper bound of Koren and Levy (2015) on the excess risk scales with $\frac{24\beta d}{\bar{\alpha}n} = \frac{24\beta d\rho^2}{\alpha n}$ (recall that the exp-concavity parameter $\bar{\alpha}$ is equal to α/ρ^2). Note that our analysis does not assume smoothness of the loss. This resolves the question raised by Koren and Levy (2015) regarding the necessity of the smoothness assumption. Note that for linear regression, the smoothness is 1, making our bounds identical to the bounds of Koren and Levy (2015) for this special case.

As discussed in Koren and Levy (2015), it is difficult to translate bounds on the average stability into high-probability bounds (while preserving the fast rate and introducing only logarithmic dependence on $1/\delta$).

5.6 Other Techniques and High-Probability Bounds

The bound on the expected excess risk in Corollary 6 can be obtained by using two additional techniques. Both of these techniques also yield high probability bounds. We next survey the corresponding results.

---

explained in Shalev-Shwartz and Ben-David (2014)[Section 13.3], by optimally controlling this tradeoff, we no longer obtain fast rates (i.e., the stability rate scales with $1/\sqrt{n}$ rather than $1/n$).

5. That is, the maximal eigenvalue of the Hessian of f at any point w is at most β.

A recent follow-up work by Mehta (2016) established a bound of $\tilde{O}(d\log(n)+\log(1/\delta)/n)$ on the excess risk of ERM, where δ is the confidence parameter.[6]. He also showed how to get rid of the $\log(n)$ factor by boosting the confidence of ERM. The proof is centered around a Bernstein condition which holds due to the exp-concavity assumption.

Another alternative, is to bound the excess risk by the local Rademacher complexity (LRC) of the associated class of predictors (e.g., using Corollary 5.3 in Bartlett et al. (2005)). In our setting, one can derive bounds on the LRC (e.g., using Koltchinskii (2008)) which coincide with our bounds.

All of these techniques employ arguably heavy machinery and lack the geometric interpretation, which is nicely captured by our notion of preconditioned stability.

Acknowledgments

---

6. The dependence on the exp-concavity parameter as well as the diameter of the loss function is hidden.

References

Katy S Azoury and Manfred K Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, pages 1497–1537, 2005.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

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

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

John B Conway. *A course in functional analysis*, volume 96. Springer Science & Business Media, 2013.

Alon Gonen, Francesco Orabona, and Shai Shalev-Shwartz. Solving ridge regression using sketched preconditioned svrg. *arXiv preprint arXiv:1602.02350*, 2016.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

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

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in neural information processing systems*, pages 793–800, 2009.

Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.

Jyrki Kivinen and Manfred K Warmuth. Averaging expert predictions. In *Computational Learning Theory*, pages 153–167. Springer, 1999.

V Koltchinskii. Oracle inequalities in empirical risk minimization and sparse recovery problems: Lecture notes. Technical report, Technical report, Ecole dete de Probabilites de Saint-Flour, 2008. 12.6, 2008.

Tomer Koren and Kfir Levy. Fast rates for exp-concave empirical risk minimization. In *Advances in Neural Information Processing Systems*, pages 1477–1485, 2015.

Mehrdad Mahdavi, Lijun Zhang, and Rong Jin. Lower and upper bounds on the generalization of stochastic exponentially concave optimization. In *Proceedings of The 28th Conference on Learning Theory*, pages 1305–1320, 2015.

Nishant A Mehta. From exp-concavity to variance control: O (1/n) rates and online-to-batch conversion with high probability. *CoRR*, 2016.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.

Francesco Orabona, Nicolo Cesa-Bianchi, and Claudio Gentile. Beyond logarithmic bounds in online learning. In *AISTATS*, pages 823–831, 2012.

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

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Ohad Shamir. The sample complexity of learning linear predictors with the squared loss. *arXiv preprint arXiv:1406.5143*, 2014.

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.