Can Implicit Bias Explain Generalization? Stochastic Convex Optimization as a Case Study

June 3, 2020

Abstract

The notion of implicit bias, or implicit regularization, has been suggested as a means to explain the surprising generalization ability of modern-days overparameterized learning algorithms. This notion refers to the tendency of the optimization algorithm towards a certain structured solution that often generalizes well. Recently, several papers have studied implicit regularization and were able to identify this phenomenon in various scenarios.

We revisit this paradigm in arguably the simplest non-trivial setup, and study the implicit bias of Stochastic Gradient Descent (SGD) in the context of Stochastic Convex Optimization. As a first step, we provide a simple construction that rules out the existence of a *distribution-independent* implicit regularizer that governs the generalization ability of SGD. We then demonstrate a learning problem that rules out a very general class of *distribution-dependent* implicit regularizers from explaining generalization, which includes strongly convex regularizers as well as non-degenerate norm-based regularizations. Certain aspects of our constructions point out to significant difficulties in providing a comprehensive explanation of an algorithm's generalization performance by solely arguing about its implicit regularization properties.

1 Introduction

One of the great mysteries of contemporary machine learning is the impressive success of *unregularized* and *overparameterized* learning algorithms. In detail, current machine learning practice is to train models with far more parameters than samples and let the algorithm *fit* the data, oftentimes without any type of regularization. In fact, these algorithms are so overcapacitated that they can even memorize and fit random data (Neyshabur et al., 2015; Zhang et al., 2017). Yet, when trained on real-life data, these algorithms show remarkable performance in generalizing to unseen samples.

This phenomenon is often attributed to what is described as the *implicit-regularization* of an algorithm (Neyshabur et al., 2015). Implicit regularization roughly refers to the learner's preference to implicitly choosing certain structured solutions *as if* some explicit regularization term appeared in its objective. As a canonical example, in linear optimization one can show that various forms of gradient descent, an apriori unregularized algorithm, behaves identically as regularized risk minimization penalized with the squared Euclidean norm on the parameters (Cesa-Bianchi and Lugosi, 2006).

Understanding implicit regularization poses several interesting challenges. For example: how can we find the implicit bias of a given learning algorithm? what is the rate of convergence towards the biased solution? how (and if) does it govern the generalization of an algorithm? and, when and what types of regularizations can account for and explain the generalization in modern-days machine learning?

Towards answering these questions we revisit a fundamental setting that was extensively studied in recent years: Stochastic Convex Optimization (SCO), focusing on the SGD optimization algorithm. In contrast to most previous work, we do not attempt to identify the implicit bias in specific problems. Instead, we study these questions in the general case, and we construct examples which rule out the existence of potential regularizers in general. To some extent, these constructions demonstrate a behavior that might seem counter-intuitive or contradictory to the implicit-bias point of view.

Besides being a well-studied and well-understood model for learning, an important trait of SCO which makes it suitable for our investigation is that learning cannot in general be performed by naive *Empirical Risk Minimization* (ERM). In detail, the work of Shalev-Shwartz et al. (2009) showed the existence of SCO instances where naive-ERM fails but *regularized*-ERM succeeds. Thus, we view SCO as a natural test-bed for exploring the role of regularization and its relation to generalization. Compellingly, the generalization of SGD in SCO is well-established, and we are left with the question of how well can we account for generalization through an investigation of its bias.

1.1 Our Contributions

Implicit distribution-independent bias. We begin with a simple construction which demonstrates that SGD does not have any *distribution-independent* implicit bias. To show that, we construct a case where SGD *does not* converge to a Pareto-efficient (not even approximately) solution with respect to the empirical loss and a given regularization penalty. In fact, this result is also true for Gradient Descent over smooth functions. In other words, our construction here involves a distribution supported on a single smooth convex function.

Our result is general and rules out any (reasonable) regularizer from being the implicit bias of SGD in this distribution-independent setting. Since the Euclidean-norm distance is the immediate suspect for the implicit regularization of SGD, the first step towards achieving the result is to rule out that Euclidean norm is the implicit bias of SGD. We thus construct an example of a function with a plateau of minimizers where SGD does not converge to the closest point in Euclidean-norm sense. While the result might not seem surprising, it is the technical engine behind the further constructions we provide. Previous to this work, Suggala et al. (2018) showed that gradient descent with an infinitely small step size (that is, gradient flow), might diverge from the closest point, and we provide a complementary construction combined with a full rigorous analysis for fixed step-size gradient descent.

Implicit distribution-dependent bias. Having ruled out the possibility of a problem-independent regularizer, we proceed to study the more compelling *distribution-dependent* implicit regularization. The question here is whether for every distribution over convex functions, we can associate a regularizer r such that SGD tries to (approximately) find a Pareto-efficient solution with respect to r and the empirical loss (notice that we allow the regularizer to depend on the distribution, but *not* on the specific sample received by SGD.)

We first show that we can rule out the effect of strongly-convex regularizers in the relevant regime of learning (where the dimension and the number of training examples are of roughly the same order). In fact, we rule out a more general class of regularizers that have large range on sets with large diameters. Namely, in any ball with large diameter the regularizer shows preference towards a certain point.

We then continue and demonstrate a distribution where, given an input sample, there is a very large set of possible solutions that share the same empirical loss and the same regularization penalty, and yet, SGD chooses its solution arbitrarily within this set. Here, by "very large" we mean from a learning-theoretical point of view; namely, this set is large enough so that, in general, empirical risk minimization restricted to the set will fail (and yet, it appears that this is exactly what SGD does). In other words, no regularizer r is sufficient for narrowing down the set of possible SGD solutions to the point where non-trivial generalization can be deduced without appealing to other properties of the specific problem.

Implicit bias in constant dimension. Several of our constructions are given in high dimension, namely the number of parameters is larger than the number of examples. One could argue that this is the interesting regime, nevertheless it is still worthy to understand the role of implicit bias when the dimension of the problem is smaller than number of examples. Here we cannot rule out the role of implicit bias in a similar fashion to before - namely, due to uniform convergence, any algorithm that is constrained to the unit ball will generalize and this implicit bias is indeed the explanation to that. It is interesting though to understand the existence of specific regularizers (such as, e.g., strongly convex regularizers).

While we do not provide an answer to this question, we make an intermediate step. Our final construction is in a slightly relaxed model, where the instances are non-convex, but the expected loss function is convex. While this result may be limited, because of the non-convexity, we stress that the learning guarantees of SGD are completely applicable to this setting: namely, SGD does learn the problem (as it is convex in expectation). We show that for any *strictly quasi-convex* regularizer, namely a regularizer that has preference for a single point in any convex regime, the algorithm will not converge to the optimal solution with optimal regularization penalty (even though it converges to a convex domain where seemingly it can improve its parameter choice towards the regularized solution).

1.2 Related Work

Implicit regularization, or implicit bias, has received considerable attention in the past few years. Starting with Neyshabur et al. (2015); Zhang et al. (2017), it was suggested that implicit regularization might explain the success of networks to improve test error by increasing network size beyond what is needed to achieve zero training error. Subsequently, a line of work has focused on identifying implicit regularization in various problems and domains, e.g., matrix factorization (Gunasekar et al., 2017; Arora et al., 2019), linearly separable data (Soudry et al., 2018; Gunasekar et al., 2018b), as well as deep networks (Neyshabur, 2017; Neyshabur et al., 2017) and others (Nacson et al., 2019; Nakajima and Sugiyama, 2010; Lin et al., 2016; Gunasekar et al., 2018a). Our work here can be seen as an attempt to investigate the limitations of implicit regularization. Most similarly to this work, Suggala et al. (2018) provides an example of a problem where gradient flow does not converge to closest Euclidean solution. Here we focus on the more concrete SGD algorithm with a fixed step size, and give finite-time analysis. We are also able to harness our example to construct further new constructions that rule out a richer class of implicit-type regularization schemes.

This work can also be seen as an attempt towards separation between *learnability* and *regularization*. Besides regularization, several other useful notions have been suggested as surrogates of learnability. Most classically, uniform convergence (Blumer et al., 1989) has been shown to be equivalent to learnability in the binary, distribution-independent model of PAC learning (Valiant, 1984). As discussed, Shalev-Shwartz et al. (2009) showed that in the stochastic convex setting naive-ERM fails (but not regularized-ERM), hence learnability and uniform convergence are no longer equivalent. The constructions of Shalev-Shwartz et al. (2009) were later substantially strengthened by (Feldman, 2016). More recently, Nagarajan and Kolter (2019) also provided an example that rules out uniform convergence, perhaps in the strictest sense. Their construction, though, does exhibit tangible implicit regularization, which account to the generalization of the algorithm.

Another useful notion is the *stability* of a learning algorithm. Stability is very much related to regularization: e.g., regularizing empirical risk minimization with a strongly convex function induces stability (Bousquet and Elisseeff, 2002). As such, constructing unstable convex problems could also serve as a means to rule out certain types of implicit regularizers. Our examples are in fact stable, and as such, could also be interpreted as a certain weak separation between stability and regularization.

2 Preliminaries

2.1 The Setup: Stochastic Convex Optimization

We consider the following standard setting of stochastic convex optimization. A learning problem consists of a fixed domain $\mathcal{W}$, which for concreteness we will assume it to be a closed and bounded set in $\mathbb{R}^d$ for some finite d, a class of functions $f(\mathbf{w}; z)$ that are convex over $\mathbf{w}$, and an unknown distribution D over a random variable z. The objective of the learner is to minimize:

$$F(\mathbf{w}) := \mathop{\mathbb{E}}_{z\sim D}[f(\mathbf{w}; z)].$$

The goal of the learner, given a sample $S = \{z_1, \dots, z_T\}$ of T i.i.d. examples from the distribution D, is to return a parameter vector $\mathbf{w}_S$ such that

$$\mathop{\mathbb{E}}_{S}[F(\mathbf{w}_S)] < \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) + \epsilon, \tag{1}$$

for a desired target accuracy $\epsilon > 0$. (The sample size T may be determined based on ϵ.)

We make the following assumptions throughout. We will generally assume that the functions f are also 1-Lipschitz, namely $\|\nabla_{\mathbf{w}} f(w, z)\| \le 1$ for all values of z and $\mathbf{w} \in \mathcal{W}$, and that the domain $\mathcal{W}$ has diameter at most 1, that is $\max_{\mathbf{w}_1,\mathbf{w}_2\in\mathcal{W}} \|\mathbf{w}_1 - \mathbf{w}_2\| \le 1$.[1] We will also discuss strongly-convex functions (or regularizers): We say that a convex function is λ-*strongly convex* if for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$ and we have: $f(\mathbf{w}_1) \ge f(\mathbf{w}_2) + \nabla f(\mathbf{w}_2)^\top(\mathbf{w}_1 - \mathbf{w}_2) + \lambda\|\mathbf{w}_1 - \mathbf{w}_2\|^2$.

[1]Since our main focus in this paper is on impossibility results, fixing the Lipschitz constant and the diameter both to 1 does not harm the generality of the setup.

2.2 Gradient Descent and Stochastic Gradient Descent

The main focus of this paper is the well-known Stochastic Gradient Descent (SGD) algorithm. Given a sample $S = \{z_1, \ldots, z_T\}$ and a step-size parameter $\eta > 0$, SGD initializes at $\mathbf{w}^{(1)} = \mathbf{0}$ and performs iterations:

$$\forall\ t = 1, \ldots, T: \quad \mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla f(\mathbf{w}^{(t)}; z_t)\ , \quad \text{and outputs:} \quad \mathbf{w}_S = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}. \tag{2}$$

The standard SGD analysis guarantees the following (see, e.g., Shalev-Shwartz and Ben-David (2014)):

Theorem. *Let $B, \rho > 0$. Further assume that $F(\cdot)$ is convex and $\|\nabla f(w, z)\| \le \rho$ for all w, z. Suppose that SGD is run for T iterations on the sample $S = \{z_1, \ldots, z_T\}$ with step size $\eta = \sqrt{B^2/(\rho^2 T)}$. Then,*

$$\mathop{\mathbb{E}}_{S}[F(\mathbf{w}_S)] - F(\mathbf{w}^\star) \le \frac{B\rho}{\sqrt{T}}, \tag{3}$$

where here $\mathbf{w}^\star \in \arg\min_{\mathbf{w}:\|\mathbf{w}\|\le B} F(\mathbf{w})$.

We will also discuss in this paper the procedure of *Gradient Descent* (GD). In our context, the gradient descent algorithm takes steps using the full gradient with respect to a sample $S = \{z_1, \ldots, z_T\}$:

$$\forall\ t = 1, \ldots, T: \quad \mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla F_S(\mathbf{w}^{(t)})\ , \quad \text{and outputs:} \quad \mathbf{w}_S = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}. \tag{4}$$

where here $F_S(\mathbf{w}) = \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}; z_t)$ is the *empirical loss* of $\mathbf{w}$.

Other variants of SGD. While the above version of SGD is perhaps the most standard one, there are other variants that can be considered. For example, it is common to consider, instead of a fixed step-size, a decaying step-size (where η may depend on t), as well as taking the last SGD iterate rather than the average iterate. We focus on the version in Eq. (2) for several reasons. First, taking the last iterate is not always justified and attains suboptimal rates (see **?**). Second, the algorithm in Eq. (2) is also the more challenging variant to argue about, in the sense that averaging and taking small fixed step size induces bias towards initialization, and as such, is more strongly regularized (and indeed, the constructions we provide here can be readily modified to address a decaying step-size or the last iterate.[2]) Nevertheless, it could be an interesting future work to derive a natural variant of SGD whose implicit regularization properties induce the desired generalization guarantees.

2.3 Regularized (Structural) Risk Minimization

Another well studied approach to perform learning is through *regularization*, Regularized Empirical Risk Minimization (ERM) solves the following minimization problem:

$$\widehat{w}_\lambda = \arg\min_{w \in \mathcal{W}} \big\{F_S(\mathbf{w}) + \lambda r(w)\big\}, \tag{5}$$

where $\lambda \in \mathbb{R}^+$, and $r(w) : \mathbb{R} \mapsto \mathbb{R}^+$ is a regularization function. When $f(\mathbf{w}; z)$ is Lipschitz-bounded and $r(\mathbf{w}), \lambda$ are properly chosen this method leads to a principled learning algorithm. For example, in the case $r(\mathbf{w}) = \|\mathbf{w}\|^2$, Bousquet and Elisseeff (2002) showed that with the correct choice of λ, Regularized ERM is guaranteed to generalize.

3 Regularization

We next discuss the different classes of regularizers we will consider in this paper. While some of the results we provide make little to no assumptions on the regularizers, sometimes we would like to add further structure and rule out specific classes as the implicit bias of SGD, in other cases we would like to formally explain in what sense we might assume that the regularizer does not allow a comprehensive explanation of the implicit bias.

Most generally, a regularizer is any function $r : \mathcal{W} \to \mathbb{R}_+$. We will however make the following basic assumptions on the regularizers, to avoid degenerate cases:

[2]In fact, the proofs will be significantly simpler; for example, in the proof overview we actually consider the last iterate for simplicity.

- $r(0) = 0$, and r is non-constant over $\mathcal{W} \backslash \{0\}$;
- r is upper semi-continuous; namely, for every point $\mathbf{w} \in \mathcal{W}$ and every $\epsilon_0 > 0$ there exists a neighborhood $B_{\delta_0}(\mathbf{w}) = \{\mathbf{u} : \|\mathbf{w} - \mathbf{u}\| < \delta_0\}$ for which $r(\mathbf{u}) > r(\mathbf{w}) - \epsilon_0$ if $\mathbf{u} \in B_{\delta_0}(\mathbf{w})$.

Any regularizer that satisfies these properties will be said to be an *admissible regularizer* (or shortly, a regularizer).

The first assumption above is only for normalization: the algorithms we will consider are all initialized at zero and will prefer the zero solution if it is a minimizer of the empirical error. Hence, this assumption is made almost w.l.o.g. The second assumption is perhaps somewhat stronger, but it is intended to rule out pathological examples. For example, one could consider a regularizer r which is 0 on almost all points, but is 1 on the negligible, dense set of real numbers that SGD would never reach. One could argue that r is an implicit bias of SGD, however, it does not capture our intuition of a regularizer. Thus, we add an assumption that a point penalized by the regularizer should also be penalized under small perturbations.

3.1 Strongly-convex Regularizers

While some of the results we will present are given for general (admissible) regularizers, it is natural and expected to study more structured classes of regularizers and ask if they induce the generalization properties of a certain algorithm. One natural family of such regularizers is the class of λ-*strongly-convex* functions, which we will also assume are 1-Lipschitz. As discussed in length, many of the prominent generalization results are provided in the context of strongly convex regularizers (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2009).

Strongly-convex regularizers come with a very natural property which allows us to rule out such regularizers on certain problems: a strongly convex function always attains a *unique* minimizer on any convex set. As such we can always identify if the output of an algorithm minimizes (approximately) the strongly-convex regularizer, by comparing the output to the minimizer of the regularizer over the given empirical risk.

3.2 General (Admissible) Regularizers

Studying implicit bias that does not stem from a strongly convex regularizer is no less important; however, it becomes much more subtle to rule out the latter. Once the regularizer is allowed to have non-unique minima we should be more careful in stating what we mean when we say it *does not explain generalization*. In fact, almost any plausible algorithm can be said to be implicitly biased on any given distribution. For example, the fact that the regularizer is constrained to the unit ball is a form of algorithmic bias—but as was shown by Shalev-Shwartz et al. (2009), it cannot explain generalization in the SCO setting.

Towards clarifying what we mean by "explain generalization", let us consider the following: given a regularizer r and an algorithm $\mathcal{A}$ that outputs a solution $\mathcal{A}(S)$ on a sample S, define the set of "competitive" solutions

$$K_{S,r}(\mathcal{A}(S)) = \{\mathbf{w} \in \mathcal{W} \ : \ F_S(\mathbf{w}) \le F_S(\mathcal{A}(S)) \ \text{ and } \ r(\mathbf{w}) \le r(\mathcal{A}(S))\}. \tag{6}$$

For shorthand, we will also use the notation $K_{S,r}(\mathcal{A})$ instead of $K_{S,r}(\mathcal{A}(S))$.

In words, $K_{S,r}(\mathcal{A})$ is the set of solutions that are comparable with (or better than) the output of $\mathcal{A}$, with respect to both the empirical loss and the regularization penalty. For example, consider a regularized ERM, as in Eq. (5), then $K_{S,r}(\mathcal{A})$ depicts *all* minimizers of Eq. (5) with comparable regularization penalty. For example, with a strongly-convex regularizer r one can observe that the set $K_{S,r}(\mathcal{A})$ is in fact a set of a single *unique* solution.

More generally, if a regularizer r is said to be the implicit bias of an algorithm $\mathcal{A}$, and as such it explains the generalization of the algorithm, it is expected that the set $K_{S,r}(\mathcal{A})$ would be "small" in the sense that choosing an arbitrary solution from it should provide principled guarantees. If we cannot attain such guarantees without further investigation of the problem and algorithm, we argue that the regularizer does not provide a comprehensive explanation of generalization. This motivates the following definition for studying more general regularizers than, say, strongly convex ones:

Definition 1. *Let us say that a set K is (T, ϵ_0)*-statistically complex *if for some distribution D over 1-Lipschitz convex functions, given T i.i.d. samples we have that with probability at least $1/10$ that for some $\mathbf{w} \in K$ it holds that $\frac{1}{T}\sum_{i=1}^{T} f(\mathbf{w}, z_i) = 0$, yet $\mathbb{E}_z[f(\mathbf{w}, z)] > \epsilon_0$.*

Note that the statistical complexity of the set K is measured with respect to an *arbitrary* distribution D over convex functions: this captures our requirement that the set $K_{S,r}(\mathcal{A})$ should explain generalization, *without further investigation of the problem.* In other words, it could be that for a correct choice of a regularizer, on a specific problem, all the models in $K_{S,r}(\mathcal{A})$ will generalize. But here we want to ensure that this generalization does not stem from any further structure in the problem that is not captured by the regularizer. Thus, we require that this set will be "simple" on any arbitrary distribution over convex functions.

4 Results

4.1 Distribution Independent Implicit Regularization

We start with the natural question, whether there is some distribution independent implicit regularization being promoted by SGD. As a warm-up we begin by ruling out the existence of a distribution-independent strongly convex regularizer that plays the role of the implicit bias of SGD. This family of regularizers is already very interesting, and has been studied extensively in the literature of stochastic convex optimization (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2009).

Theorem 1. *For every λ-strongly convex r, there is a distribution D_r and $\mathbf{w}_r \in \mathcal{W}$ such that, with probability 1, SGD with step size $0 < \eta < \frac{1}{3}$ over an input sample S of size $T = \Omega(1/(\lambda\eta))$ outputs $\mathbf{w}_S$ such that:*

$$F_S(\mathbf{w}_r) \leq F_S(\mathbf{w}_S), \qquad \textit{and} \quad r(\mathbf{w}_r) \leq r(\mathbf{w}_S) - \Theta(\lambda) \ .$$

In words, for any strongly convex regularizer there exists an instance problem where SGD chooses a solution that is sub-optimal in terms of both empirical error, and regularization penalty.
The last result can be extended to general (admissible) regularizers. Here, the rate of divergence from a Pareto optimal solution depends on the structure of the regularizer r. This dependence of the divergence-rate on the regularizer r is unavoidable. Indeed if we consider a regularizer r such that $r \approx 0$, it is not hard to be convinced that it would take SGD longer to become r-suboptimal.

Theorem 2. *For every admissible regularizer r, there are constants $c_r > 0$, a distribution D_r, and $\mathbf{w}_r \in \mathcal{W}$ such that, with probability 1 over the input sample S, SGD with step size $0 < \eta < \frac{1}{3}$ and sample size $T_r = \Omega_r(1/\eta)$ outputs $\mathbf{w}_S$ such that:*

$$F_S(\mathbf{w}_r) \leq F_S(\mathbf{w}_S), \qquad \textit{and} \qquad r(\mathbf{w}_r) \leq r(\mathbf{w}_S) - c_r \ .$$

The $\Omega_r(\cdot)$ notation hides constant that may depend on the regularizer r. The dependence on the regularizer is expected here, as we would need a very strong level of accuracy if we want to rule out a nearly-constant regularizer, for example.

4.2 Distribution-Dependent Implicit Regularization

Having ruled out a class of implicit regularizers in the distribution-independent model, we next move on to discuss the possibility of distribution dependent regularizers.

Theorem 3. *Suppose we run SGD with $\eta = O(1/\sqrt{T})$ learning rate. There exists a distribution D over $\mathbb{R}^d$, where $d = \Theta(T)$ such that for any λ-strongly convex regularizer r, with probability $\Theta(1)$, SGD outputs $\mathbf{w}_S$ for which there is a $\mathbf{w}^\star$, such that*

$$F_S(\mathbf{w}^\star) \leq F_S(\mathbf{w}_S), \qquad \textit{and} \quad r(\mathbf{w}^*) \leq r(\mathbf{w}_S) - \Theta(\lambda) \ .$$

Utilizing a construction of a statistically complex set due to Feldman (2016), we can also obtain the following result:

Theorem 4. *There exists a distribution D over $\mathbb{R}^d$, such that if we run SGD on a sample of size T with learning rate $\eta = O(1/\sqrt{T})$, then for any regularizer r we have that with probability at least $1/10$ (over the sample) the set $K_{S,r}(\mathbf{w}_S)$ is $(2T, \Theta(1))$-statistically complex.*

In words, Theorem 4 asserts that for a certain given distribution D the output of SGD cannot be interpreted as coming from a "small" structured family of solutions, that would generalize regardless of other specialized properties of the particular learning problem.

4.3 Implicit Bias in Constant Dimension

In the results above we provided constructions in spaces with more parameters than samples. We next discuss the case $d \ll T$, which is interesting for certain contexts.

Regarding Theorem 4, we again point out that such a result cannot hold in the aforementioned regime. Indeed, in this case uniform convergence over the unit-ball applies. In that sense, restricting an algorithm to choose a solution in the unit ball provides an inductive bias that provides generalization guarantees. But what about Theorem 3? It is interesting to know if we can rule out regularizers that are not benign like the unit ball.

We do not know the answer to this question and we leave it as an open problem. Nevertheless we can provide the following intermediate result in a slightly more relaxed setting, where the instances may be non-convex, but the expected loss function is indeed convex. Thus, SGD's learning guarantee still implies.

We will state the next result for a slightly larger class of regularizers than merely convex regularizers. Recall that a function f is called *quasi-convex* if $f(\lambda x + (1-\lambda)y) \le \max\{f(x), f(y)\}$ for every $0 \le \lambda \le 1$ and $x, y \in \mathcal{K}$. and *strictly* quasi-convex, if $f(\lambda x + (1-\lambda)y) < \max\{f(x), f(y)\}$

Theorem 5. *There exists a distribution D over $\mathbb{R}^2$, on not necessarily convex functions, such that $\mathbb{E}_z[f(\mathbf{w}; z)] = 0$ for every $\mathbf{w} \in \mathcal{W}$, for every strictly quasi-convex regularizer r, and for large enough T, if $\eta = \Theta(1/\sqrt{T})$ then with some positive probability, $\Theta(1)$, there exists $\mathbf{w}^\star$ such that:*

$$F_S(\mathbf{w}^\star) \le F_S(\mathbf{w}_S); \qquad r(\mathbf{w}^\star) < r(\mathbf{w}_S); \qquad \|\mathbf{w}_S - \mathbf{w}^\star\| = \Theta(1).$$

5 Constructions

Here we give a high level description of the constructions as well as the proofs of the main results. We remark that our full proofs address the average of the SGD iterates (as presented in Eq. (2)); for simplicity of exposition, though, here we will mainly focus on the last iterate.

5.1 Distribution Independent Regularization

Our constructions build upon the following class of functions in $\mathbb{R}^2$. Let A be a set of the form $\{(\alpha, \theta) : b_1 \le \alpha \le b_2\}$, where θ, b_1, b_2 are parameters of the set and Σ is a PSD matrix. We then consider the function $f_{A,\Sigma}$ defined as follows:

$$f_{A,\Sigma}(\mathbf{w}) = \frac{1}{2} \min_{\mathbf{v} \in A} \{(\mathbf{w} - \mathbf{v})^\top \Sigma (\mathbf{w} - \mathbf{v})\}. \tag{7}$$

One can observe that these functions are convex, and further the gradient of $f_{A,\Sigma}$ at point $\mathbf{w}$ will equal

$$\nabla f_{A,\Sigma}(\mathbf{w}) = \Sigma(\mathbf{w} - \mathbf{v}(\mathbf{w})), \tag{8}$$

where

$$\mathbf{v}(\mathbf{w}) = \arg\min_{\mathbf{v} \in A} \{(\mathbf{w} - \mathbf{v})^\top \Sigma (\mathbf{w} - \mathbf{v})\}.$$

Warm-up: GD need not converge to a minimal-norm solution. We start by showing how we can construct a function (of the type in Eq. (7)) that does not converge to minimal norm solution. Let us take a concrete case where

$$A = \{(\alpha, 1) : 0 \le \alpha \le \infty\}; \qquad \Sigma = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}.$$

We will suppress dependence on A and Σ, and simply write f. The main observation is that the trajectory of f is characterized by two phases.

At the first phase the closest point to $\mathbf{w}^{(t)}$ (w.r.t. Σ-norm) is at the boundary of A (i.e $\alpha = 0$). At this phase, $\mathbf{w}^{(t)}$ can be seen to move "towards" the center of the interval, namely $w_1^{(t)}$ is increasing (see Eq. (8)). At the end of this phase, $w_1^{(t)}$, is sufficiently large irrespective of the step size $\eta > 0$.

The second phase, starts when $\mathbf{e}_2 \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ stops being the closest point, and the closest point to $\mathbf{w}^{(t)}$ is at the interior of the interval. One can show that at this phase, the gradient moves upward hence $w_1^{(t)}$

does not decrease and overall the trajectory will converge to a point away from $\mathbf{e}_2$: the Euclidean closest minimizer to 0.

To see that when $\mathbf{v}(\mathbf{w})$ is at the interior of A then $\nabla f(\mathbf{w}) \propto \mathbf{e}_1$, consider the following scalar function:

$$g(a) = (\mathbf{w} - (a,1))^T \Sigma (\mathbf{w} - (a,1)).$$

Our assumption is that g attains its minimum at $0 < v_1$. Taking the derivative at v_1 and equating to 0 (because the minimum is attained at the interior), we can see that $g'(v_1) = (\mathbf{w} - (v_1, 1))^T \Sigma \mathbf{e}_1 = 0$. Hence, $\nabla f(\mathbf{w}) = (\mathbf{w} - v(\mathbf{w}))\Sigma \perp \mathbf{e}_1$.

The trajectory of $\mathbf{w}^{(t)}$ is illustrated in Fig. 1 (green line).

No strongly-convex implicit bias. The construction above is the heart of most of our results. Let us illustrate how it rules out a strongly convex regularizer (in the distribution-independent setting) and attain Theorem 1.

The key property of strongly-convex regularizers is that in any convex set they have a unique minimum. Moreover, two far away points cannot simultaneously attain close-to-minimal value. This is in fact the only property we will use. Thus, our result can in fact be extended to any regularizer that is a “tie-breaker"- namely it always prefers a single unique solution amongst a class of possible solutions with large diameter.

The construction above will allow us to generate two instances of convex learning problems, where SGD converges to two far away points. The first instance is the standard Euclidean distance. Namely we take a function f_1 of the form in Eq. (7), with Σ the identity and A with boundaries $(-\infty, \infty)$. In this case SGD is biased towards $\mathbf{e}_2$ (gray dashed line in Fig. 1). The second instance, f_2, is the construction above where SGD is biased towards another point on the interval (green line in Fig. 1).

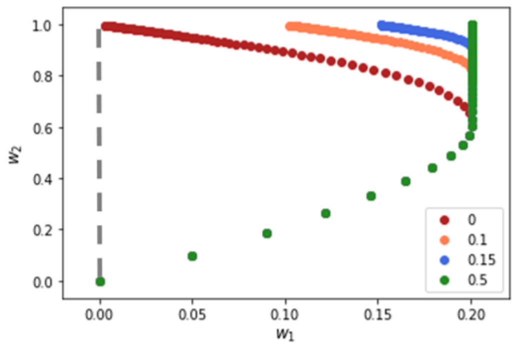

Figure 1: Simulation of objective functions where GD does not converge to closest point. At the first phase, the gradient points toward right: specifically the vector $\Sigma\mathbf{e}_2$ (see Eq. (7)), which causes it to diverge from $\mathbf{e}_2$. The red, orange, blue and green trajectories are simulations of the same objective function when we replace A from *Eq.* (7) with $A' = \{(\alpha, 1), 0 \le \alpha \le b\}$ for values $b = \{0, 0.1, 0.15, \infty\}$ respectively.

Now both points are global minima, for both f_1 and f_2, hence if SGD is implicitly biased towards solutions with minimum regularization penalty r, we must have that $r(\mathbf{e}_2) = r(\mathbf{v})$, where $\mathbf{v}$ is the choice of SGD when it observes f_2. However, if r is strongly convex, because $\|\mathbf{e}_2 - \mathbf{v}\| = O(1)$, there has to be a point on the interval between them that attain a strictly lesser regularization penalty, moreover it also attains minimal loss value. This contradicts the existence of such an r.

The general case. Our second result (Theorem 2) rules out the existence of any distribution-independent regularizer. In contrast with the strongly-convex case we can not give uniform bounds that depend on parameters of strong convexity. As such, the rates depend on the regularizer.

But the construction here is similar. We basically start with the assumption that there are two points $\mathbf{w}_1$ and $\mathbf{w}_2$ with different regularization penalty, and we want to construct two functions f_1, f_2 that maps $\mathbf{w}_1, \mathbf{w}_2$ to the same empirical loss. It might seem that through a simple linear transformation that maps, say, $\mathbf{w}_1$ to $\mathbf{e}_2$ and $\mathbf{w}_2$ to $\mathbf{v}$ we can reduce this case to the case above. However, there is some subtlety since gradient descent is not invariant to linear transformations.[3]

Towards this, we extend the construction above by constructing a more general example, where we can tune the point of convergence of SGD to *any* point on the interval between $\mathbf{v}$ and $\mathbf{e}_2$. This allows us to avoid scaling, and use only rotations (which GD is invariant to) in order to reduce the problem to the former case. This is done by changing the set A from allowing $0 \le \alpha < \infty$, to adding a second boundary condition on the right. In Fig. 1 we illustrate how this changes the trajectory (red, orange and blue lines).

5.2 Distribution-Dependent Implicit Bias

We next discuss our second sets of results that argue about *distribution-dependent regularization.* Here we want to study if, for a given distribution, the set of solutions on which SGD converges has some

[3]We note though that it can be turned to an affine invariant optimization algorithm (Koren and Livni, 2017).

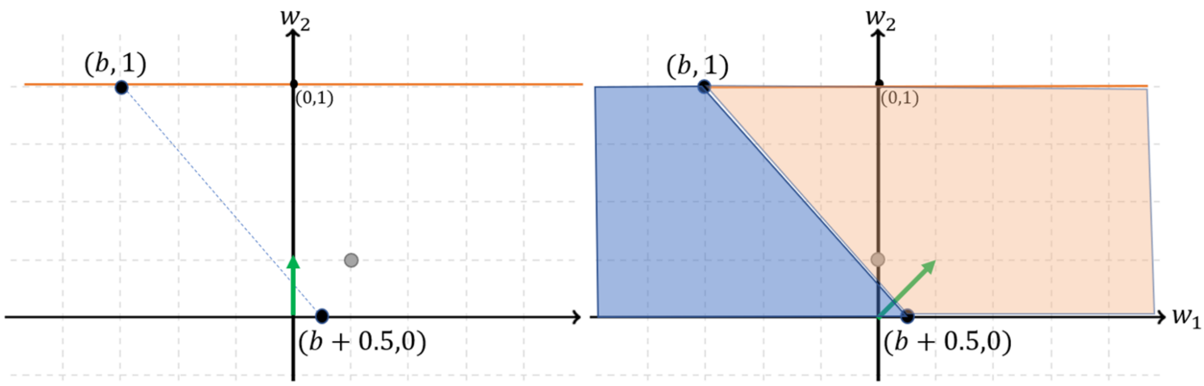

Figure 2: Depiction of the auxiliary construction in Lemma 1. The left sketch illustrate the first function where the gradient of the loss always points upward, hence on a x-axis parallel line it is constant. The right figure illustrate the second loss function. Here at a regime that include 0 the gradient points sideways, and at the second regime, the gradient points vertically upwards

meaningful structure on which we can argue why it generalizes.

Note that so far, our problem instances considered only a single function and the results were applicable to GD also. Here, though, in the distribution dependent setting such an example cannot work. Indeed, given a single function as an instance problem, SGD behaves deterministically and the solution it chooses is a unique solution which trivially generalizes.

No strongly-convex distribution-dependent bias. We next discuss our argument that rules out a strongly convex regularizer, even if it may depend on the distribution at hand. We again utilize the property that a strongly convex regularizer obtains approximately minimum solutions only on a small diameter around the unique minimum.

Our strategy is as follows: assume that there are two samples S_1 and S_2 such that, when SGD observes S_1 it converges to $\mathbf{w}_1$ and when it observes S_2 it converges to $\mathbf{w}_2$. However, assume also that $\|\mathbf{w}_1 - \mathbf{w}_2\| = \Theta(1)$, and that the empirical loss of $\mathbf{w}_1$ and $\mathbf{w}_2$ is comparable, on both samples: namely $F_{S_1}(\mathbf{w}_1) = F_{S_2}(\mathbf{w}_1)$, and similarly with $\mathbf{w}_2$.

In the case above, as we argued in the distribution-independent case, clearly the algorithm failed to choose the minimizer of the regularization penalty, in at least one of the realization S_1 or S_2. So if S_1 and S_2 are equally likely, we obtain that with probability half (conditioned on the event that we saw S_1 or S_2) the algorithm failed to minimize r. Now, if the probability to observe one of such couple of samples S_1, S_2 is positive, then we obtain the desired result.

To generate this setting, we rely on the following auxiliary construction in $\mathbb{R}^2$. We construct two functions such that, if SGD observes the first function, at the first iteration, then the gradient points upward and right. But if SGD observes the second function, at the first iteration, then the gradient points upward. This ensures that in each case SGD will move towards a different solution. If the size of the gradient is constant then the gap between the two iterations will be $\Theta(\eta)$.

We will also construct the examples in such a way that both points enter a regime where all points obtain the same empirical loss on both functions. This construction can in fact be done using piece-wise linear functions and it is illustrated in Fig. 2. We also give the formal statement here:

Lemma 1. *For every constant $0 < c < 1$, there are two 1-Lipschitz functions $f(\mathbf{w}; \pm 1)$ over $\mathbb{R}^2$ such that if $\mathbf{v}_1 = -\nabla f(0;1)$ and $\mathbf{v}_{-1} = -\nabla f(0;-1)$ and $c < \frac{1}{2}\eta < 1$ then $\|\mathbf{v}_1 - \mathbf{v}_{-1}\| \geq 1/4$ and $f(\eta \mathbf{v}_1; z) = f(\eta \mathbf{v}_{-1}; z)$ for any $z \in \{-1, 1\}$.*

We next utilize the above construction to generate the problem in $\mathbb{R}^d$. Note that the construction above generates a problem where SGD will converge to two different solutions with distance η but same empirical loss (after one step). Indeed, we just need to randomly pick one of these functions.

We next want to amplify the distance. To do that, we consider $d = \Omega(T)$ Cartesian copies of $\mathbb{R}^2$. Then at each example, we show one of the functions above, at one of the products. Assuming enough coordinates were seen only once (which is going to happen w.h.p.), the variance on each sub-plane will be η^2: if we have $\Theta(T)$ such coordinates, the overall variance is going to be $\Theta(T\eta^2)$ which ensures that we will converge to far away solutions on different realizations of the problem, if $\eta = \Theta(1/\sqrt{T})$.

SGD might be biased towards statistically-complex sets. Next, we derive Theorem 4 which addresses implicit regularization in a much broader setting. As discussed, here we cannot rule out the existence of an implicit bias; indeed, some form of an implicit bias always exists. We attempt, though, to understand how the implicit bias can explain generalization.

The result shows that for any regularizer: the set $K_{S,r}(\mathbf{w}_S)$ which is the set of comparable solutions to the one outputted by SGD, given the empirical loss and regularization penalty, can be large up to the fact that choosing an arbitrary solution from this set can, in principle, lead to over-fitting (over general convex problems). Thus, to argue that the algorithm did generalize, further structure in the problem needs to be taken into account. And this is true for *any* regularizer.

Our construction is similar to the previous case in Theorem 3 up to some modification. Therefore, let us show that in the construction above $K_{S,r}$ will be $(T/6, \Theta(1))$-statistically complex. This is less than what we actually desire. We, in fact, observed T examples and not $T/6$. Indeed, in the construction above, we showed that if we project the output of SGD to the observed coordinates, we obtain a solution of the form $(\mathbf{v}_{\pm 1}, \mathbf{v}_{\pm 1}, \cdots, \mathbf{v}_{\pm 1}) \in (\mathbb{R}^2)^T$, where $\mathbf{v}_1, \mathbf{v}_{-1}$ are as in Lemma 1. By projecting this set, it can be seen to be a copy of (up to some rescaling) the normalized unit cube $\mathcal{M} = \{\pm\eta, \pm\eta, \cdots, \pm\eta\} \in \mathbb{R}^T$. This is true since $\|\eta \cdot \mathbf{v}_1 - \eta \cdot \mathbf{v}_{-1}\| = \Theta(\eta)$.

Here, we rely on a construction by Feldman (2016). In order to show that uniform convergence is not equivalent to learnability in the convex optimization setting, Feldman showed (in our terminology) that the set $\mathcal{M} \in \mathbb{R}^T$ is $(T/6, 1/4)$ statistically complex, if $\eta = \Theta(1/\sqrt{T})$.

As discussed, this is less than what we want, as we actually want a set that is at least $(T, \Theta(1))$ statistically complex. To tackle this, on each iteration we show the learner a loss function over multiple pairs of coordinates. Namely, if in the example above we drew at each iteration $f(\mathbf{w}; z)$ where $z \sim D$, now in each iteration we show the algorithm $\frac{1}{k}\sum_{i=1}^{k} f(\mathbf{w}; z_i)$, where z_i are i.i.d. This will reduce the step-size on each coordinate a little bit but if k is constant we will still present a constant loss. On the other hand, now projecting on observed coordinates, SGD will converge to a solution in $(\mathbf{v}_{\pm 1}, \mathbf{v}_{\pm 1}, \ldots, \mathbf{v}_{\pm 1}) \in (\mathbb{R}^2)^{\Theta(kT)}$. Thus we only need a constant $k > 6$ so that the algorithm will converge to a $(T, \Theta(1))$-statistically complex set.

5.3 Implicit Bias in Constant Dimension

We next provide a construction in $\mathbb{R}^2$ that again rules out a class of regularizers, in particular strongly convex regularizers (and more generally, strictly quasi-convex regularizers).

In a similar fashion to previous constructions, we make SGD choose from a set of solutions, that exhibit comparable empirical loss. While the dimension of previous constructions depended on T, this construction does not. However, for the construction we relax the assumption that $f(\mathbf{w}; z)$ are convex, but F remains convex. Note that the learning guarantees of SGD are completely applicable to this setting.

Our construction relies on a 2-dimensional square, centered at the origin. Inside the square, SGD makes a simple 2-dimensional random walk, while when it exits from the square, it continues to perform a random walk in just one dimension (denoted as y), while the other coordinate (denoted as x) remains the same. As a result, the optimizer of F_S is independent of w_x.

We study the event that $\mathbf{w}$ will stay inside the square for enough iterations to ensure that the variance of w_x will be larger than some constant, but eventually $\mathbf{w}$ exit from the square to make F_S independent of w_x. This will result with a set of solutions that share the same empirical error and also SGD can converge to each one of them.

References

S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7411–7422, 2019. 3

A. C. Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941. 13

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989. 3

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar): 499–526, 2002. 3, 4, 5, 6

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 12

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006. 1

C.-G. Esseen. On the liapunov limit error in the theory of probability. *Ark. Mat. Astr. Fys.*, 28:1–19, 1942. 13

V. Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. In *Advances in Neural Information Processing Systems*, pages 3576–3584, 2016. 3, 6, 10, 12

S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017. 3

S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018a. 3

S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018b. 3

T. Koren and R. Livni. Affine-invariant online optimization and the low-rank experts problem. In *Advances in Neural Information Processing Systems*, pages 4747–4755, 2017. 8

J. Lin, R. Camoriano, and L. Rosasco. Generalization properties and implicit regularization for multiple passes sgm. In *International Conference on Machine Learning*, pages 2340–2348, 2016. 3

M. S. Nacson, J. D. Lee, S. Gunasekar, P. H. P. Savarese, N. Srebro, and D. Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 3420–3428, 2019. 3

V. Nagarajan and J. Z. Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 11611–11622, 2019. 3

S. Nakajima and M. Sugiyama. Implicit regularization in variational bayesian matrix factorization. In *ICML*, pages 815–822, 2010. 3

B. Neyshabur. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017. 3

B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. 1, 3

B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017. 3

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning:From Theory to Algorithms*. Cambridge University, 2014. ISBN 978-1-107-05713-5. 4

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *COLT*, 2009. 1, 3, 5, 6, 12

D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018. 3

A. Suggala, A. Prasad, and P. K. Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, pages 10608–10619, 2018. 2, 3

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. 3

P. van Beek. An application of fourier methods to the problem of sharpening the berry-esseen inequality. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 23(3):187–196, 1972. 13

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. 1, 3

A Technical Preliminaries

A.1 Distance Functions

Throughout, we will repeatedly use functions of the following form in our constructions:

$$f_{A,\Sigma}(\mathbf{w}) = \frac{1}{2}\min_{v\in A}\Big\{(\mathbf{w}-\mathbf{v})^\top \Sigma(\mathbf{w}-\mathbf{v})\Big\}, \tag{9}$$

where A is a set of the form $A = \{(\alpha,\theta) : -b_1 < \alpha < b_2\}$, $\theta \in \mathbb{R}$, $b_1, b_2 \in \mathbb{R}_+$ and Σ is a PSD matrix of the following form

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma/2 \\ \sigma/2 & 1 \end{bmatrix}.$$

Because A is convex, it is known that a function f of the aforementioned form, depicted in Eq. (9), is indeed a convex function (see Boyd and Vandenberghe (2004) example 3.16). If there is no reason for confusion we will omit dependence in A and Σ and simply write $f(\mathbf{w})$.

It can be seen that for a function $f_{A,\Sigma}$ of the form in Eq. (9), the gradient is given by

$$\nabla f(\mathbf{w}) = \Sigma(\mathbf{w}-\mathbf{v}(\mathbf{w})),$$

where we denote $\mathbf{v}(\mathbf{w}) = \arg\min_{v\in A}(\mathbf{w}-\mathbf{v})^\top\Sigma(\mathbf{w}-\mathbf{v})$. As a corollary one can obtain the following expressions for the gradient

$$\nabla f(\mathbf{w}) = \begin{cases} \Sigma(\mathbf{w}-(b_1,\theta)) & w_1 + \frac{1}{2\sigma}(w_2-\theta) < b_1 \\ \begin{pmatrix} 0 \\ \frac{3(w_2-\theta)}{4}\end{pmatrix} & b_1 < w_1 + \frac{1}{2\sigma}(w_2-\theta) < b_2 \\ \Sigma(\mathbf{w}-(b_2,\theta)) & b_2 \le w_1 + \frac{1}{2\sigma}(w_2-\theta) \end{cases} \tag{10}$$

A.2 Feldman's Statistically Complex Set

A key technical tool in the proof of Theorem 4 is a construction by Feldman, Feldman (2016), of a statistically complex set in $\mathbb{R}^d$. While Feldman's construction is not the first to show that the sample complexity of an ERM algorithm may scale with the dimension, it greatly improved over previous construction Shalev-Shwartz et al. (2009), and showed that the dependence may be *linear* in the dimension.

We will exploit here Feldman's set in order to construct an example where SGD essentially picks arbitrarily an element from a statistically complex set, akin to ERM, and we will need the following statement due to Feldman

Theorem 6 (Essentially Theorem 3.3 in Feldman, 2016)**.** *Let $\mathcal{W}_d = \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$. There exists a distribution D over 1-Lipschitz convex functions such that given a sample $|S| < d/6$ drawn i.i.d from D then w.p. $1/2$ (over the sample S) there exists $\mathbf{w} \in \mathcal{W}_d$ such that*

$$\frac{1}{|S|}\sum_{t=1}^{|S|} f(\mathbf{w}, z_t) = 0, \tag{11}$$

but

$$\mathop{\mathbb{E}}_{z\sim D}[f(\mathbf{w}, z)] = 1/4. \tag{12}$$

We will need a slightly stronger version of the theorem which is an immediate corollary

Corollary 6.1. *Let $A \subseteq \mathcal{W}_d$, such that $|A| \ge 2^{d-1}$, then A is $(d/6, 1/4)$-statistically complex.*

Proof. For two vectors $\mathbf{v} \in \{-1,1\}^d$ and an element $\mathbf{w} \in \mathcal{W}_d$ let $\mathbf{v} * \mathbf{w} \in \mathcal{W}_d$ be the pointwise product between $\mathbf{w}$ and $\mathbf{v}$, i.e.

$$(\mathbf{v} * \mathbf{w})_i = \mathbf{v}_i \cdot \mathbf{w}_i.$$

Let D be the distribution from Theorem 6 and consider a distribution where we draw uniformly an elements $\mathbf{v} \in \{-1,1\}^d$ and a sample S of size d/6 i.i.d from D. One can show that with probability $|A|/(2^{d+1})$ we have that there exists an elements $\mathbf{w} \in A$ such that

$$\frac{1}{|S|}\sum_{t=1}^{|S|} f(\mathbf{v} * \mathbf{w}; z_t) = 0 \tag{13}$$

but

$$\mathop{\mathbb{E}}_{z\sim D} f(\mathbf{v} * \mathbf{w}; z) = 1/4. \tag{14}$$

In particular, there exists a $\mathbf{v}$ such that with probability $\frac{|A|}{2^{d+1}}$, Eqs. (13) and (14) holds for some $\mathbf{w} \in A$ over the random sample S. Thus, we can define a convex Lipschitz mapping parameterized by $\mathbf{z}$ such that

$$f_{\mathbf{v}}(\mathbf{w}; z) = f(\mathbf{v} * \mathbf{w}; z).$$

From the above discussion if we draw $z \sim D$ we can see that this distribution demonstrates that A is (d/6,1/4)-statistically complex

□

A.3 Berry-Esseen Theorem

A very important and valuable tool for analysing the behavior of random walks that we will use is the well-known Berry-Esseen Theorem, discovered independently in Berry (1941); Esseen (1942).

Theorem 7 (Berry Esseen Theorem)**.** *Let $X_1, X_2, \ldots, X_T$ be zero mean and independent random variables, with $\mathbb{E}(X_i^2) = \sigma_i^2$ and $\mathbb{E}(|X_i^3|) = \rho_i$. Let $S_T = \frac{1}{\sqrt{\sum_{i=1}^T \sigma_i^2}} \sum_{i=1}^T X_i$, then we have*

$$|P(S_T \le a) - \Phi(a)| \le C_{BE}(\sum_{i=1}^T \sigma_i^2)^{-3/2} \sum_{i=1}^T \rho_i,$$

where $C_{BE} < 1$ is an absolute constant, and $\Phi(a)$ is the CDF of a unit variate zero-mean Gaussian random variable.

For a bound $C_{\text{BE}} < 1$ of the absolute constant see, for example, van Beek (1972). We will need the following technical Lemma which is derived via Theorem 7:

Lemma 2. *Let $k \ge 0$, and assume $T > 2 \cdot k$. If X_t is a random variables such that*

$$X_t = \begin{cases} c\frac{T-t}{T} & w.p.\ 1/4 \\ -c\frac{T-t}{T} & w.p.\ 1/4 \\ 0 & w.p.\ 1/2 \end{cases},$$

and $I = \{1, 2, \ldots, T/k\}$, then

$$P\left(\left|\frac{1}{\sqrt{T}}\sum_{i\in I} X_i\right| < a\frac{c}{\sqrt{50k}}\right) \le \mathit{erf}(a) + \sqrt{\frac{50^3 k}{T}},$$

where $\mathit{erf}(a) = \Phi(a) - \Phi(-a)$ is the error function.

Proof. First, we lower bound $\sum_{i\in I}\sigma_i^2$, and obtain that:

$$\begin{aligned}
\sum_{i\in I} E[|X_i|^2] &= \frac{c^2}{2}\sum_{i\in I}\left(\frac{T-t}{T}\right)^2 \\
&\ge \frac{c^2}{2T^2}\sum_{t=1}^{T/k}(T-t)^2 \\
&= \frac{c^2}{2T^2}\sum_{t=0}^{T/k-1}\left(\left(\frac{k-1}{k}\right)T + t\right)^2 \\
&\ge \frac{c^2}{2T^2}\max\left\{\frac{(k-1)^2}{k^2}T^2\frac{T}{k}, \sum_{t=0}^{T/k-1} t^2\right\}
\end{aligned}$$

We also have that for $T > 2 \cdot k$:

$$\sum_{t=0}^{T/k-1} t^2 = \frac{T/k\,(T/k-1)\,(2T/k-1)}{6} \geq \frac{T^3}{12k^3}$$

Taken together we obtain that

$$\begin{aligned}\sum_{i\in I} \mathbb{E}[|X_i|^2] &\geq \frac{c^2 T}{2} \max\left\{\frac{(k-1)^2}{k^2}, \frac{1}{12k^2}\right\}\\ &\geq \frac{c^2 T}{50k}\end{aligned}$$

Next, we lower bound $\sum \rho_i$:

$$\sum_{i\in I} \mathbb{E}[|X_i|^3] \leq \frac{1}{2}\sum_{t=1}^{T/k} \mathbb{E}[|c\frac{(T-t)}{T}|^3] = \frac{c^3}{2T^3}\sum_{t=1}^{T/k}(T-t)^3 \leq \frac{c^3 T}{2T^3 k}T^3 \leq \frac{c^3 T}{2k}$$

Taken together we obtain that

$$\begin{aligned}P\left(\left|\frac{1}{\sqrt{T}}\sum X_i\right| < a\frac{c}{\sqrt{50k}}\right) &\leq P\left(\left|\frac{1}{\sqrt{\sum_{i=1}^T \sigma_i^2}}\sum X_i\right| < a\right)\\ &\leq \Phi(a) - \Phi(-a) + 2\left(\sum_{i=1}^T \sigma_i^2\right)^{-3/2}\sum_{i=1}^T \rho_i\\ &\leq \Phi(a) - \Phi(-a) + \frac{(50k)^{3/2}c^3 T}{c^3 T^{3/2} k}\\ &\leq \Phi(a) - \Phi(-a) + \sqrt{\frac{50^3 k}{T}}\end{aligned}$$

□

B Proofs: Distribution Independent Regularizers

As discussed, the main technical difficulty behind the distribution-independent-regularization results is the existence of a problem instance where GD does not converge to the minimal norm solution:

Theorem 8 (GD doesn't converge to closest Euclidean norm solution)**.** *For every $0 < \theta_2 \leq 1$, and $0 < \theta_1 \leq 0.025 \cdot \theta_2$, there exists a positive, convex and smooth function f_{θ_1,θ_2} such that for $0 < \eta < \frac{1}{3}$, GD outputs $\mathbf{w}_S$ such that*

$$\|\mathbf{w}_S - (\theta_1, \theta_2)\| < \frac{120}{\eta T}$$

but

$$f_{\theta_1,\theta_2}((0,\theta_2)) = f_{\theta_1,\theta_2}(\theta_1,\theta_2) = 0$$

Proof. For $0 \leq \theta_2 \leq 1$ and $0 \leq \theta_1 \leq 0.025 \cdot \theta_2$ let us define the set: $A_{\theta_1,\theta_2} = \{(\alpha, \theta_2) : 0 \leq \alpha \leq \theta_1\}$. In turn, we define the function $f_{\theta_1,\theta_2}(\mathbf{w})$ to be:

$$f_{\theta_1,\theta_2}(\mathbf{w}) = \arg\min_{v\in A_{\theta_1,\theta_2}} \frac{1}{2}(\mathbf{w}-\mathbf{v})^T \Sigma(\mathbf{w}-\mathbf{v}), \tag{15}$$

where we let,

$$\Sigma = \begin{pmatrix} 1 & 0.5\\ 0.5 & 1\end{pmatrix}.$$

We begin with a proposition that describes the trajectory of GD over the function f_{θ_1,θ_2}.

Lemma 3. *Let $\mathbf{w}^{(1)}, ..., \mathbf{w}^{(T)}$ be the sequence defined by running GD over f_{θ_1,θ_2} with step size $\eta \leq \frac{1}{3}$, and T iteration. Then there exist $\frac{1}{2\cdot\eta} \leq t_0 \leq \frac{3}{\eta}, t_0 \leq t_1 \leq t_0 + \frac{7}{\eta}$ s.t.*

$$\text{For } 1 \leq t \leq t_0\text{:} \quad \boldsymbol{w}^{(t)} = \left(I - (I - \eta\Sigma)^{t-1}\right)\xi_0 \qquad \text{where } \xi_0 = (0, \theta_2) \tag{16}$$

next,

$$\text{For } t_0 < t \leq t_1\text{:} \quad \mathbf{w}^{(t)} = \begin{pmatrix} w_1^{(t_0)} \\ (1 - \frac{3\eta}{4})^{t-t_0}\left[w_2^{(t_0)} - \theta_2\right] + \theta_2 \end{pmatrix} \tag{17}$$

and,

$$\text{For } t_1 < t \leq T\text{:} \quad \boldsymbol{w}^{(t)} = \left(I - (I - \eta\Sigma)^{t-t_1}\right) \cdot \xi_1 + (I - \eta\Sigma)^{t-t_1} \cdot \mathbf{w}^{(t_1)} \qquad \text{where } \xi_1 = (\theta_1, \theta_2) \tag{18}$$

This trajectory is depicted in Fig. 1 for different values of θ_1. We now show how to derive Theorem 8 from Lemma 3 and deter the proof of Lemma 3 to the end of the section. Note that, by notation above, to prove Theorem 8, we only need to show that $\|w_S - \xi_1\| \leq \frac{120}{\eta T}$– the fact that $f_{\theta_1,\theta_2}(0, \theta_2) = f_{\theta_1,\theta_2}(\theta_1, \theta_2)$ is immediate from the definition of f.

We will first need to bound the sizes $\|\xi_0\|, \|\xi_1\|, \|\mathbf{w}^{(t)}\|$. One can easily observe that $\|\xi_1\|, \|\xi_0\| < 1.5$. Following the trajectory path of $\mathbf{w}^{(t)}$, provided in Lemma 3 we can also provide a bound on $\mathbf{w}^{(t)}$:

Namely,

- If $t \leq t_0$ we have that $\|\mathbf{w}^{(t)}\| < \|\xi_0\| \leq 1$
- If $t_0 \leq t \leq t_1$, then $\|\mathbf{w}^{(t)}\| \leq \|\mathbf{w}^{(t_0)}\| + \theta_2 \leq 2$.
- And if $t \geq t_1$ we have that $\|\mathbf{w}^{(t)}\| \leq \|\mathbf{w}^{(t_1)}\| + \|\xi_1\| < 5$

Taken together we have that $\|\mathbf{w}^{(t)}\| < 5$.

Finally, by simple calculation we can show that the singular values of Σ are 3/2 and 1/2. Hence,

$$\|\,(I - \eta\Sigma)\,\|_\infty \leq (1 - \frac{\eta}{2}). \tag{19}$$

where $\|\cdot\|_\infty$ denotes the spectral norm of a matrix. We are now ready to show that $\mathbf{w}_S$ converges to ξ_1:

$$\begin{aligned}
\|\mathbf{w}_S - \xi_1\|_2 \leq \frac{1}{T}\sum_{t=1}^{T}\|(\mathbf{w}^{(t)} - \xi_1)\|_2 &= \frac{1}{T}\sum_{t=1}^{t_1}\|(\mathbf{w}^{(t)} - \xi_1)\|_2 + \frac{1}{T}\sum_{t=t_1+1}^{T}\|(\mathbf{w}^{(t)} - \xi_1)\|_2 \\
&\leq \frac{10t_1}{T} + \frac{1}{T}\sum_{t=t_1+1}^{T}\|\mathbf{w}^{(t)} - \xi_1\|_2 && \|\mathbf{w}^{(t)}\|, \|\xi_1\| < 5 \\
&= \frac{100}{\eta T} + \frac{1}{T}\sum_{t=t_1+1}^{T}\|(1 - \eta\Sigma)^{t-t_1} \cdot (\mathbf{w}^{(t_1)} - \xi_1)\| && t_1 < 10/\eta,\ \textit{Eq. } (18) \\
&\leq \frac{100}{\eta T} + \frac{1}{T}\sum_{t=t_1+1}^{T}\|(I - \eta\Sigma)^{t-t_1}\|_\infty \cdot \|\mathbf{w}^{(t_1)} - \xi_1\|_2 \\
&\leq \frac{100}{\eta T} + \frac{10}{T}\sum_{t=1}^{T-t_1}\|(I - \eta\Sigma)\|_\infty^t && \|\mathbf{w}^{(t_1)} - \xi_1\| < 10 \\
&\leq \frac{100}{\eta T} + \frac{10}{T}\sum_{t=0}^{\infty}(1 - \frac{\eta}{2})^t && \textit{Eq. } (19) \\
&\leq \frac{100}{\eta T} + \frac{10}{T} \cdot \frac{2}{\eta} \leq \frac{120}{\eta T}
\end{aligned}$$

□

Proof of Lemma 3 First note that f_{θ_1,θ_2} is a function of the form depicted in Eq. (9), with parameter $\sigma = 1$. We thus obtain two boundary conditions that governs the behavior of the trajectory:

$$w_1 + \frac{1}{2}w_2 < \frac{1}{2} \cdot \theta_2 \tag{20}$$

$$w_1 + \frac{1}{2}w_2 \geq \frac{1}{2}\cdot\theta_2 + \theta_1, \tag{21}$$

Given η, we claim that Lemma 3 holds if we let t_0 denote the first iterate such that $\mathbf{w}^{(t_0)}$ violates Eq. (20), when running GD, and if t_1 denotes the first iterate for which $\mathbf{w}^{(t)}$ satisfies Eq. (21). We will split the proof into 3 parts, according to GD's trajectory, i.e. $t \leq t_0, t_0 < t \leq t_1, t > t_1$

Claim 9. *There exists $\frac{1}{2\cdot\eta} \leq t_0 \leq \frac{3}{\eta}$ such that $\mathbf{w}^{(t_0)}$ is the first iterate that violates Eq. (20). Further, for any $t \leq t_0$, $\boldsymbol{w}^{(t)}$ can be calculated by Eq. (16). And finally, $0.03\theta_2 \leq w_1^{(t_0)}$.*

Proof. First note that $\mathbf{w}^{(1)}$ satisfies Eq. (20), hence $t_0 \geq 1$. Now, following the calculation of the derivative provided in, Eq. (10) we obtain the following update step: $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta\Sigma\left(\mathbf{w}^{(t)} - \xi_0\right)$. which we can rewrite as:

$$\mathbf{w}^{(t+1)} = (I - \eta\Sigma)\mathbf{w}^{(t)} + \eta\Sigma\xi_0. \tag{22}$$

By induction one can show that for $2 \leq t \leq t_0$:

$$\begin{aligned}\mathbf{w}^{(t)} &= \sum_{i=0}^{t-2}(I-\eta\Sigma)^i\cdot(\eta\Sigma\xi_0)\\ &= \left(I - (I-\eta\Sigma)^{t-1}\right)\xi_0.\end{aligned} \tag{23}$$

This shows that for any $t \leq t_0$, $\mathbf{w}^{(t)}$ can be calculated by Eq. (16). We proceed with the proof to show that $\frac{1}{2\eta} \leq t_0 \leq \frac{3}{\eta}$,

Considering the singular value decomposition of Σ one can show that:

$$(I-\eta\Sigma)^{t-1} = \frac{1}{2}\begin{pmatrix}[(1-3\eta/2)^{t-1} + (1-\eta/2)^{t-1}] & [(1-3\eta/2)^{t-1} - (1-\eta/2)^{t-1}]\\ [(1-3\eta/2)^{t-1} - (1-\eta/2)^{t-1}] & [(1-3\eta/2)^{t-1} + (1-\eta/2)^{t-1}]\end{pmatrix}. \tag{24}$$

Plugging this in Eq. (23), we obtain that for any $t \leq t_0$:

$$\mathbf{w}^{(t)} = \frac{\theta_2}{2}\begin{pmatrix} & [(1-\eta/2)^{t-1} - (1-3\eta/2)^{t-1}]\\ 2- & [(1-\eta/2)^{t-1} + (1-3\eta/2)^{t-1}]\end{pmatrix} \tag{25}$$

To obtain the lower bound on t_0 observe that t_0 satisfies:

$$w_1^{(t_0)} + \frac{1}{2}(w_2^{(t_0)} - \theta_2) \geq 0,$$

Plugging Eq. (25) and dividing by $\theta_2/2$ we obtain that:

$$\left(1-\frac{\eta}{2}\right)^{t_0-1} - \left(1-\frac{3\eta}{2}\right)^{t_0-1} + \frac{1}{2}\left(2 - \left(1-\frac{\eta}{2}\right)^{t_0-1} - \left(1-\frac{3\eta}{2}\right)^{t_0-1} - 2\right) \geq 0.$$

Rearranging terms we get:

$$\frac{1}{2}\left(1-\frac{\eta}{2}\right)^{t_0-1} - \frac{3}{2}\left(1-\frac{3\eta}{2}\right)^{t_0-1} \geq 0$$

Which for $\eta < 1/3$, can be rewritten as:

$$\left(1+\frac{2\eta}{2-3\eta}\right)^{t_0-1} = \left(\frac{2-\eta}{2-3\eta}\right)^{t_0-1} \geq 3. \tag{26}$$

This leads to

$$\begin{aligned} t_0 &\geq \frac{1}{\ln(1+\frac{2\eta}{2-3\eta})} && \ln(3) \geq 1\\ &\geq \frac{2-3\eta}{2\eta} && \ln(x+1) \leq x\\ &= \frac{1}{\eta} - \frac{3}{2}\\ &\geq \frac{1}{2\eta} && \eta \leq \frac{1}{3}\end{aligned}$$

Next we provide an upper bound for t_0. Again, for every $t < t_0$ Eq. (20) is satisfied, which, as we already saw (in Eq. (26)) means that for every $t < t_0$:

$$\forall t < t_0, \quad \left(1 + \frac{2\eta}{2-3\eta}\right)^{t-1} \leq 3. \tag{27}$$

Using the inequality $(1+2/n)^n \geq 3$, we obtain

$$\left(1 + \frac{2\eta}{2-3\eta}\right)^{t-1} \geq (1+\eta)^{t-1} \geq 3^{\frac{\eta}{2}(t-1)}$$

In particular for $t \geq \frac{2}{\eta} + 1$ Eq. (27) is violated and hence $t_0 \leq \frac{3}{\eta}$.

Finally, we provide a lower bound for $w_1^{(t_0)}$. Namely, we want to show that $w_1^{(t_0)} \geq 0.04\theta_2$.

First, by rearranging terms at Eq. (26) we obtain that t_0 is sufficiently large so that $\left(1 - \frac{\eta}{2}\right)^{t_0 - 1} \geq 3\left(1 - \frac{3\eta}{2}\right)^{t_0-1}$. Again applying the formula for $\mathbf{w}^{(t_0)}$ in Eq. (25) we have that:

$$\begin{aligned} w_1^{(t_0)} = \frac{\theta_2}{2} \cdot [(1 - \frac{1}{2}\cdot\eta)^{t_0-1} - (1 - \frac{3}{2}\cdot\eta)^{t_0-1}] &\geq \frac{\theta_2}{4}\left(1 - \frac{1}{2}\eta\right)^{t_0-1} \\ &\geq \frac{\theta_2}{4}\left(1 - \frac{1}{2}\eta\right)^{3/\eta} && t_0 < \frac{3}{\eta} \\ &\geq 2^{-5}\theta_2 && \left(1 - \frac{1}{2n}\right)^n > \frac{1}{2} \end{aligned} \tag{28}$$

This concludes the analysis of the first phase of the trajectory. We next move on to the case $t_0 \leq t \leq t_1$ □

Claim 10. *Let $t_0 \leq t \leq t_1$. Then $\mathbf{w}^{(t)}$ can be calculated by Eq. (17). Moreover $t_1 \leq t_0 + \frac{7}{\eta}$.*

Proof. We again apply the calculation of the derivative provided in Eq. (10) at $t_0 \leq t \leq t_1$ and obtain :

$$\nabla f(\mathbf{w}^{(t)}) = \begin{pmatrix} 0 \\ \frac{3}{4}(w_2^{(t)} - \theta_2) \end{pmatrix}. \tag{29}$$

Note that this proves that $w_1^{(t)} = w_1^{(t_0)}$. For $w_2^{(t)}$, we have that

$$w_2^{(t)} = w_2^{(t-1)}(1 - \frac{3}{4}\eta) + \frac{3}{4}\cdot\eta\cdot\theta_2,$$

which leads by induction to the following:

$$\begin{aligned} w_2^{(t)} &= \left(1 - \frac{3}{4}\eta\right)^{t-t_0} w_2^{(t_0)} + \sum_{i=0}^{(t-t_0)-1}\left(1 - \frac{3}{4}\eta\right)^i \cdot \frac{3\eta\theta_2}{4} \\ &= \left(1 - \frac{3}{4}\eta\right)^{t-t_0} w_2^{(t_0)} + \left(1 - \left(1 - \frac{3}{4}\eta\right)^{t-t_0}\right)\theta_2 \\ &= \left(1 - \frac{3}{4}\eta\right)^{t-t_0}\left[w_2^{(t_0)} - \theta_2\right] + \theta_2. \end{aligned}$$

This shows that for any $t_0 \leq t \leq t_1$ Eq. (17) holds.

We next bound t_1. Recall that t_1 is defined to be the first iterate for which Eq. (21) is satisfied. Let us show that for any t s.t $t_0 + \frac{7}{\eta} < t$ holds, Eq. (21) is satisfied and hence $t_1 \leq t_0 + 7/\eta$. Equivalently we will show that for $t > t_0 + 7/\eta$, the following equation holds:

$$\theta_2 - w_2^{(t)} \leq 2(w_1^{(t)} - \theta_1). \tag{30}$$

Indeed, let $t < t_1$, then

$$\begin{aligned}
2 \cdot (w_1^{(t)} - \theta_1) &= 2 \cdot (w_1^{(t_0)} - \theta_1) && (w_1^{(t)} = w_1^{(t_0)} \text{ by } \textit{Eq. } (17)) \\
&\geq 2 \cdot (2^{-5} \cdot \theta_2 - \theta_1) && (w_1^{(t_0)} \geq 2^{-5}\theta_2 \text{ by } \textit{Eq. } (28)) \\
&\geq 2 \cdot (2^{-5} \cdot \theta_2 - 0.025\theta_2) && (\theta_1 \leq 0.025 \cdot \theta_2) \\
&\geq 0.01 \cdot \theta_2
\end{aligned}$$

Next assume that $t \geq t_0 + \frac{7}{\eta}$, then

$$\begin{aligned}
0.01 \cdot \theta_2 &\geq e^{-\frac{3\eta \cdot (t-t_0)}{4}} \theta_2 && t \geq t_0 + \frac{20}{3\eta} \\
&\geq \left(1 - \frac{3}{4}\eta\right)^{t-t_0} \theta_2 \\
&\geq \left(1 - \frac{3}{4}\eta\right)^{t-t_0} \left[\theta_2 - w_2^{(t_0)}\right] && (w_2^{(t_0)} \geq 0) \\
&= \theta_2 - \mathbf{w}_2^{(t)}. && \textit{Eq. } (17)
\end{aligned}$$

We now move to the last phase of the trajectory. □

Claim 11. *Let $t \geq t_1$, then $\boldsymbol{w}^{(t)}$ can be calculated by Eq. (18).*

Proof. Let $t \geq t_1$ be such that Eq. (21) holds. Then again, we consider the formula of the derivative $\nabla f(\mathbf{w})$ (see Eq. (10)) and have that

$$\nabla f(\mathbf{w}) = \Sigma(\mathbf{w} - \xi_1).$$

We obtain the following recursive formula for t if Eq. (21) holds for all $t_1 \leq t' \leq t$:

$$\begin{aligned}
\mathbf{w}^{(t)} &= (I - \eta\Sigma)\mathbf{w}^{(t-1)} + \eta\Sigma\xi_1 \\
&= (I - \eta\Sigma)^{t-t_1}\mathbf{w}^{(t_1)} + \sum_{i=0}^{t-t_1-1} (I - \eta\Sigma)^i \eta\Sigma\xi_1 \\
&= (I - \eta\Sigma)^{t-t_1}(\mathbf{w}^{(t_1)} - \xi_1) + \xi_1 \qquad (31)
\end{aligned}$$

This shows that $\mathbf{w}^{(t)}$ can be calculated via Eq. (18). It remains thus to show that for any $t \geq t_1$, Eq. (21) always holds. We prove this by induction. Note that for the base case, this follows from the definition of t_1. We can thus assume by induction hypothesis that $\mathbf{w}^{(t)}$ satisfies Eq. (31), and we want to prove that

$$w_1^{(t)} + 1/2 w_2^{(t)} - \theta_1 - 1/2\theta_2 \geq 0.$$

For succinctness, let us write

$$\alpha_t = \left(1 - \frac{3\eta}{2}\right)^{t-t_1}, \quad \text{and,} \quad \beta_t = \left(1 - \frac{\eta}{2}\right)^{t-t_1}.$$

We will denote also $\mathbf{v} = \begin{pmatrix} 1 \\ 1/2 \end{pmatrix}$ Then using Eq. (24) and Eq. (31) we have that

$$\begin{aligned}
w_1^{(t)} + 1/2 w_2^{(t)} - \theta_1 - 1/2\theta_2 &= \mathbf{v}^\top(\mathbf{w}^{(t)} - \xi_1) \\
&= \mathbf{v}^\top(1 - \eta\Sigma)^{t-t_1}(\mathbf{w}^{(t_1)} - \xi_1) && \textit{Eq. } (31) \\
&= (3/2\alpha_t + 1/2\beta_t)(\mathbf{w}_1^{(t_1)} - \theta_1) + (3/2\alpha_t - 1/2\beta_t)(\mathbf{w}_2^{(t_1)} - \theta_2) && \textit{Eq. } (24) \\
&\geq (3/4\alpha_t - 3/4\beta_t)(\mathbf{w}_2^{(t_1)} - \theta_2)) && 2(\mathbf{w}_1^{(t_1)} - \theta_1) \geq \theta_2 - \mathbf{w}_2^{(t_1)} \\
&\geq 0
\end{aligned}$$

where the last inequality is true since $\alpha_t \leq \beta_t$ for $\eta < 1/3$ and we also have that $\mathbf{w}_2^{(t_1)} < \theta_2$. □

This concludes the proof of Lemma 3.

B.1 Proof of Theorem 1

Theorem 1 is an almost immediate corollary of Theorem 8. Indeed, let r be a strongly convex function. Take $\mathbf{e}_2 = (0,1)$ and let $\mathbf{w}_0 = (0.024, 1)$. Consider the set $a_{(\mathbf{e}_2,\mathbf{w}_0)} = \{\alpha\mathbf{e}_2 + (1-\alpha)\mathbf{w}_0 : 0 \le \alpha \le 1\}$. and let

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in a_{\mathbf{e}_2,\mathbf{w}_0}} r(\mathbf{w}).$$

Now we either have $\|\mathbf{w}^* - \mathbf{e}_2\| \ge 0.012$, or $\|\mathbf{w}^* - \mathbf{w}_0\| \ge 0.012$. We will show that in both cases we can choose a function $f \ge 0$, such that $f(\mathbf{e}_2) = f(\mathbf{w}_0) = f(\mathbf{w}^*) = 0$ and that $\mathbf{w}_S$ will satisfy

$$\|\mathbf{w}^* - \mathbf{w}_S\| > 0.01 \tag{32}$$

and that if $\Pi(\mathbf{w}_S)$ is the projection of $\mathbf{w}_S$ on $a_{\mathbf{e}_2,\mathbf{w}_0}$ then

$$\|\mathbf{w}_S - \Pi(\mathbf{w}_S)\| < \frac{120}{T\eta} \tag{33}$$

This will conclude the proof. Indeed, by strong convexity:

$$\begin{aligned}
r(\mathbf{w}_S) &\ge r(\mathbf{w}^*) + \nabla r(\mathbf{w}^*)^\top(\mathbf{w}_S - \mathbf{w}^*) + \lambda\|\mathbf{w}^* - \mathbf{w}_S\|^2 \\
&\ge r(\mathbf{w}^*) + \nabla r(\mathbf{w}^*)^\top(\Pi(\mathbf{w}_S) - \mathbf{w}^*) + \nabla r(\mathbf{w}^*)^\top(\mathbf{w}_S - \Pi(\mathbf{w}_S)) + \lambda\|\mathbf{w}^* - \mathbf{w}_S\|^2 \\
&\ge r(\mathbf{w}^*) - \frac{120}{T\eta} + \lambda\|\mathbf{w}^* - \mathbf{w}_S\|^2 && \text{Lipschitness of } r \\
&\ge r(\mathbf{w}^*) - \frac{120}{T\eta} + 10^{-6}\lambda
\end{aligned}$$

We are thus left with proving the existence of f in each case:

- First assume that $\|\mathbf{w}^* - \mathbf{e}_2\| > 0.012$. We will construct a distribution s.t. $\mathbf{w}^{(t)}$ will converge to $\mathbf{e}_2$: set

  $$f(\mathbf{w}) = \frac{1}{2} \cdot \min_{\mathbf{v}\in a_{(\mathbf{e}_2,\mathbf{w}_0)}} \|\mathbf{w} - \mathbf{v}\|^2$$

  (i.e. $\Sigma = I$). Our distribution D is defined to choose f w.p. 1.

  A simple analysis of the update step of SGD shows that for $\eta < 1$, we have that $\mathbf{w}^{(t+1)} = \sum_{i=0}^{t-1}(1-\eta)^i\eta\mathbf{e}_2$. Hence,

  $$\begin{aligned}
  \|\mathbf{w}_S - \mathbf{e}_2\| &= \|\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)} - \mathbf{e}_{21}\| \le \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{w}^{(t)} - \mathbf{e}_2\| \\
  &= \frac{1}{T}\sum_{t=1}^{T}\|\sum_{i=1}^{t-1}(1-\eta)^i\eta\mathbf{e}_2 - \mathbf{e}_2\| \\
  &= \frac{1}{T}\sum_{t=1}^{T}\|(1-\eta)^t\mathbf{e}_2\| && \sum_{i=1}^{t}(1-\eta)^t = \frac{1-(1-\eta)^{t+1}}{\eta} \\
  &\le \frac{1}{T}\sum_{t=1}^{T}(1-\eta)^t \\
  &\le \frac{1}{T\eta}
  \end{aligned}$$

  In particular we have that

  $$\|\mathbf{w}_S - \Pi(\mathbf{w}_S)\| < \|\mathbf{w}_S - \mathbf{e}_2\| < \frac{120}{T\eta},$$

  also assuming $\frac{1}{T\eta}$ is sufficiently small we have that $\|\mathbf{w}^* - \mathbf{w}_S\| > 0.01$.

- Next we assume that $\|\mathbf{w}^* - \mathbf{w}_0\| > 0.012$. We now apply Theorem 8. and we let $f = f_{\theta_1,\theta_2}$ be as in Theorem 8 with parameters $\theta_1 = 0.024$ and $\theta_2 = 1$. Then, by Theorem 8 we have that $\|\mathbf{w}_S - \mathbf{w}_0\| < \frac{120}{\eta T}$, and we obtain as before that $\|\mathbf{w}^* - \mathbf{w}_S\| > 0.01$ and that $\|\mathbf{w}_S - \Pi(\mathbf{w}_S)\| < \frac{120}{\eta}$, as required.

B.2 Proof of Theorem 2

For a vector $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^2$ let us denote by $\mathbf{w}^\perp := (w_2, -w_1)$. In particular, we have that $\mathbf{w}^\top \mathbf{w}^\perp = 0$ and $\|\mathbf{w}\| = \|\mathbf{w}^\perp\|$. Our proof relies on the following claim which we prove at the end of this section.

Claim 12. *Let r be an admissible regulerizer over $\mathbb{R}^2$. There are two points $\mathbf{w}_1$ and $\mathbf{w}_2$ in the unit ball such that for some $-0.005\|\mathbf{w}_1\| < \delta < 0.005\|\mathbf{w}_1\|$ we have*

$$\mathbf{w}_2 = \mathbf{w}_1 + \delta \mathbf{w}_1^\perp,$$

and $r(\mathbf{w}_1) \neq r(\mathbf{w}_2)$.

We next proceed with proof of Theorem 2. Let $\mathbf{w}_1$ and $\mathbf{w}_2$ be as in Claim 12. First, because GD is invariant to rotations, we can assume w.l.o.g that $\mathbf{w}_1 = \|\mathbf{w}_1\| \cdot \mathbf{e}_2$, and hence $\mathbf{w}_2 = (1, \delta)\|\mathbf{w}_1\|$.

We now set $c_r = \frac{1}{2}|r(\mathbf{w}_1) - r(\mathbf{w}_2)|$. To choose T_r, D_r and $\mathbf{w}_r$ we now look at two cases: if $r(\mathbf{w}_1) > r(\mathbf{w}_2)$ and if $r(\mathbf{w}_1) < r(\mathbf{w}_2)$.

- First suppose $r(\mathbf{w}_1) > r(\mathbf{w}_2)$. By upper-semicontinuity there exists a neighborhood δ_1 such that for every $\mathbf{w}$ s.t. $\|\mathbf{w} - \mathbf{w}_1\| < \delta_1$, satisfies $r(\mathbf{w}) > r(\mathbf{w}_2) + c_r$. We thus set $T_r = \frac{1}{\eta\delta_1}$, and $\mathbf{w}_r = \mathbf{w}_2$.

  We are left with choosing D_r. Note that in this case, the regularizer prefers a point with large Euclidean norm over a point with smaller Euclidean norm. Thus, to show it is not the implicit bias of SGD we only need to construct a distribution that is biased towards smaller Euclidean norms:

  Indeed, consider the set $a_{(\mathbf{w}_1,\mathbf{w}_2)} = \{\alpha \mathbf{w}_1 + (1-\alpha)\mathbf{w}_2 : 0 \leq \alpha \leq 1\}$ we set

  $$f(\mathbf{w}) = \frac{1}{2} \cdot \min_{\mathbf{v} \in a_{(\mathbf{w}_1,\mathbf{w}_2)}} \|\mathbf{w} - \mathbf{v}\|^2$$

  Our distribution D_r is defined to choose f w.p. 1. Having defined $T_r, c_r, \mathbf{w}_r$ and D_r we now set out to prove the result.

  A simple analysis of the update step of SGD shows that for $\eta < 1$ we have for every $\mathbf{w}^{(t)}$ that $\mathbf{w}^{(t+1)} = \sum_{i=0}^{t-1} (1-\eta)^i \eta \mathbf{w}_1$. Hence,

  $$\begin{aligned}
  \|\mathbf{w}_S - \mathbf{w}_1\| = \|\frac{1}{T_r}\sum_{t=1}^{T_r} \mathbf{w}^{(t)} - \mathbf{w}_1\| &\leq \frac{1}{T_r}\sum_{t=1}^{T_r} \|\mathbf{w}^{(t)} - \mathbf{w}_1\| \\
  &= \frac{1}{T_r}\sum_{t=1}^{T_r} \|\sum_{i=1}^{t-1}(1-\eta)^i \eta \mathbf{w}_1 - \mathbf{w}_1\| \\
  &= \frac{1}{T_r}\sum_{t=1}^{T_r} \|(1-\eta)^t \mathbf{w}_1\| && \sum_{i=1}^{t}(1-\eta)^t = \frac{1-(1-\eta)^{t+1}}{\eta} \\
  &\leq \frac{1}{T_r}\sum_{t=1}^{T_r}(1-\eta)^t \\
  &\leq \frac{1}{T_r \eta} \\
  &= \delta_1 && T_r = \frac{1}{\delta_1 \eta}
  \end{aligned}$$

  By property of δ_1 we have that $r(\mathbf{w}_S) > r(\mathbf{w}_2) + c_r$.

  But because $\mathbf{w}_2$ is optimal (i.e. attain zero on f), we also have $F_S(\mathbf{w}_S) > F(\mathbf{w}_r)$. This proves the case $r(\mathbf{w}_1) > r(\mathbf{w}_2)$.

- Next, assume that $r(\mathbf{w}_1) < r(\mathbf{w}_2)$. As before we have a neighborhood δ_2 such that if $\|\mathbf{w} - \mathbf{w}_2\| < \delta_2$ then we are guaranteed that $r(\mathbf{w}) > r(\mathbf{w}_1) + c_r$. We choose then $T_r = \frac{1}{\delta_2 \eta}$ and $\mathbf{w}_r = \mathbf{w}_1$.

  To define D_r, we now use the function f_{θ_1,θ_2} from Theorem 8. We assume w.l.o.g that $\delta > 0$, if this is not the case we can use that function $f_{\theta_1,\theta_2}(\mathbf{w}) = f_{\theta_1,\theta_2}(-\mathbf{w})$.

  Let us set $\theta_2 = \|\mathbf{w}_1\|$ and $\theta_1 = |\delta|\|\mathbf{w}_1\| < 0.05\theta_2$. Again, we consider a deterministic distribution D_r that chooses f_{θ_1,θ_2} w.p. 1.

Recall that we assume that $\mathbf{w}_1 = \|\mathbf{w}_1\|\mathbf{e}_2$, hence $\mathbf{w}_1 = (0, \theta_2)$ and $\mathbf{w}_2 = (\theta_1, \theta_2)$. Hence, by Theorem 8, if we run over a sample of size $T_r > \frac{1}{\delta_2 \eta}$, we obtain that

$$\|\mathbf{w}_S - \mathbf{w}_2\| < \delta_2.$$

In particular $r(\mathbf{w}_S) > r(\mathbf{w}_1) + c_r$. But again $F_S(\mathbf{w}_S) \geq F(\mathbf{w}_1)$, because $\mathbf{w}_1$ is optimal.

Proof of Claim 12 First, let us assume that there are $\mathbf{u}, \mathbf{v}$ such that $\|\mathbf{u}\|_2, \|\mathbf{v}\|_2 = a$ and $r(\mathbf{u}) \neq r(\mathbf{v})$ (at the end we will show that for admissible regularizer we always have such two points).

We will also assume that $\|\mathbf{u} - \mathbf{v}\|_2 \leq 10^{-9} \cdot a^2$. If this was not the case we can cover the sphere $\{\mathbf{w} : \|\mathbf{w}\|_2 = a\}$ with balls with radius $10^{-9} \cdot a^2$, and have a constant function at every ball, concluding that r is constant on the sphere (which contradicts our assumption).

Next, we also assume that $|u_2| > \frac{1}{2}a$, (either $|u_1| > \frac{1}{2}a$ or $|u_2| > \frac{1}{2}a$, and the proof is similar in both cases so we will analyse only the later case).

Then, since $u_1^2 + u_2^2 = v_1^2 + v_2^2 = a^2$, one can show that

$$\frac{u_1 - v_1}{v_2 + u_2} = \frac{v_2 - u_2}{u_1 + v_1}.$$

So, by choosing $\delta = \frac{u_1 - v_1}{v_2 + u_2} = \frac{v_2 - u_2}{u_1 + v_1}$ we have that:

$$\begin{aligned}
|\delta| = \frac{|u_1 - v_1|}{|v_2 + u_2|} &= \frac{|u_1 - v_1|}{|v_2| + |u_2|} && (|u_2 - v_2| < 10^{-9}a^2, 0.5a < |u_2|) \\
&\leq 4\frac{|u_1 - v_1|}{a} && (|v_2| + |u_2| > a/4) \\
&\leq 0.0025a && (|u_1 - v_1| < 0.0025/4 \cdot a)
\end{aligned}$$

Using the first equality we can show that $u_1 - \delta u_2 = v_1 + \delta v_2$ Similarly we can show that $u_2 + \delta u_1 = v_2 - \delta v_1$. Taken together we obtain that

$$\mathbf{u} + \delta \mathbf{u}^{\perp} = \mathbf{v} - \delta \mathbf{v}^{\perp}.$$

In particular $r(\mathbf{u} + \delta \mathbf{u}^{\perp}) = r(\mathbf{v} - \delta \mathbf{v}^{\perp})$. Since $r(\mathbf{u}) \neq r(\mathbf{v})$, we either have $r(\mathbf{u}) \neq r(\mathbf{u} + \delta \mathbf{u}^{\perp})$, or $r(\mathbf{v}) \neq r(\mathbf{v} - \delta \mathbf{v}^{\perp})$. In the former case we choose $\mathbf{w}_1 = \mathbf{u}$, whereas in the latter case we choose $\mathbf{w}_1 = \mathbf{v}$.

Finally, so far we assume we can find two points on a sphere with different regularization penalty. Next, we assume that on every sphere r is constant. Assume also to the contrary that for every $-0.0025\|\mathbf{w}\| < \delta < 0.0025\|\mathbf{w}\|$:

$$r(\mathbf{w} + \delta \mathbf{w}^{\perp}) = r(\mathbf{w}).$$

It is not hard to show that in this case r is constant everywhere except maybe 0, making it in-admissible.

C Proofs II: Distribution Dependent Regularization

We start this section by proving the existence of the auxiliary construction in Lemma 1

Lemma 1. *For every constant $0 < c < 1$, there are two 1-Lipschitz functions $f(\mathbf{w}; \pm 1)$ over $\mathbb{R}^2$ such that if $\mathbf{v}_1 = -\nabla f(0; 1)$ and $\mathbf{v}_{-1} = -\nabla f(0; -1)$ and $c < \frac{1}{2}\eta < 1$ then $\|\mathbf{v}_1 - \mathbf{v}_{-1}\| \geq 1/4$ and $f(\eta \mathbf{v}_1; z) = f(\eta \mathbf{v}_{-1}; z)$ for any $z \in \{-1, 1\}$.*

Before we continue with the proof, notice the following immediate corollary of Lemma 1

Corollary 12.1. *For every constant $c > 0$, there is a distribution D over a pair of convex functions $\{f(\mathbf{w}; 1), f(\mathbf{w}; -1)\}$, such that $f(\mathbf{w}; z)$ is a 1-Lipschitz convex function in $\mathbb{R}^2$ and, for every $c < \eta < 1$ denote $\mathbf{v}_{z,\eta} = -\eta \nabla f(0; z)$. Then the following holds:*

- *For every $z \in \{-1, 1\}$ we have that $f(\mathbf{v}_{z,\eta}; z) = f(\mathbf{v}_{-z,\eta}; z)$.*
- $\|\mathbf{v}_{z,\eta} - \mathbf{v}_{-z,\eta}\| > \eta/4$
- $\|\mathbf{v}_{z,\eta}\| \leq \eta$ *for any $z \in \{-1, 1\}$*

Indeed to derive Corollary 12.1 from Lemma 1, take a distribution that w.p. 1/2 picks $f(\mathbf{w}; 1)$ from Lemma 1, and with probability 1/2 picks $f(\mathbf{w}; -1)$. One can observe that the result holds.

Proof of Lemma 1 Let us define $f(\mathbf{w};\pm 1)$ as follows, denote $\mathbf{v}_1 = (1/4, 3/4)$, $\mathbf{v}_{-1} = \frac{3}{4}\cdot \mathbf{e}_2$ and let

$$f(\mathbf{w};z) = \max(0, \mathbf{v}_z^\top \mathbf{w} + c\|\mathbf{v}_z\|^2)$$

It is easy to check that $\nabla f(0;1) = \mathbf{v}_1$ and that $\nabla f(0;-1) = \mathbf{v}_{-1}$, and that $\|\mathbf{v}_1 - \mathbf{v}_{-1}\| \geq 1/4$.

Next, note that if $\eta > c$ then

$$f(\mathbf{v}_{1,\eta};1) = \max\left(0, (-\eta + c)\cdot \|\mathbf{v}_1\|^2\right) = 0 = \max\left(0, -\eta \mathbf{v}_1^\top \mathbf{v}_{-1} + c\|\mathbf{v}_{-1}\|^2\right) = f(\mathbf{v}_{-1,\eta};1)$$

Similarly, $f(\mathbf{v}_{-1,\eta};-1) = 0 = f(\mathbf{v}_{1,\eta};-1)$.

C.1 Proof of Theorem 3

Theorem 3 is an immediate corollary of the following theorem:

Theorem 13. *For every T, there exists a distribution D over $\mathbb{R}^d$ with $d = \Theta(T)$ such that if we run SGD with step size $1/T^2 < \eta < 1$, the following holds: for any regularizer r, w.p. at least $1/10$ over the sample S there is $\mathbf{w}_r \in \mathcal{W}$ such that*

$$\begin{aligned} F_S(\mathbf{w}_r) &\leq F_S(\mathbf{w}_S)\ , \\ r(\mathbf{w}_r) &\leq r(\mathbf{w}_S)\ , \end{aligned}$$

Moreover

$$\|\mathbf{w}_r - \mathbf{w}_S\|_2^2 \ = \ \Theta(\eta^2 T)\ .$$

To see how Theorem 3 follows, Let $\mathbf{w}^*$ be the minimizer of $r(\mathbf{w})$ amongst all $\mathbf{w} \in \mathcal{W}$ with $F_S(\mathbf{w}) \leq F_S(\mathbf{w}_S)$ then by strong convexity

$$r(\mathbf{w}_S) \geq r(\mathbf{w}^*) + \frac{\lambda}{2}\|\mathbf{w}_S - \mathbf{w}^*\|^2.$$

Now if $\|\mathbf{w}_S - \mathbf{w}^*\| > \frac{1}{4}\cdot\|\mathbf{w}_r - \mathbf{w}_S\|$ we are done. If not, then

$$\|\mathbf{w}_S - \mathbf{w}^*\| \leq \frac{1}{4}\|\mathbf{w}_r - \mathbf{w}_S\| \leq \frac{1}{4}[\|\mathbf{w}_r - \mathbf{w}^*\| + \|\mathbf{w}_S - \mathbf{w}^*\|],$$

which leads to $\|\mathbf{w}_r - \mathbf{w}^*\| \geq \frac{3}{4}\cdot\|\mathbf{w}_r - \mathbf{w}_S\|$. Using this, we get by strong convexity:

$$r(\mathbf{w}_S) \geq r(\mathbf{w}_r) \geq r(\mathbf{w}^*) + \frac{\lambda}{2}\|\mathbf{w}_r - \mathbf{w}^*\|^2 \geq r(\mathbf{w}^*) + \frac{9\lambda}{32}\|\mathbf{w}_r - \mathbf{w}_S\|^2.$$

Proof of Theorem 13 Choose $d = 10\cdot T$ and let $\mathcal{W}$ be the unit ball of $\mathbb{R}^d$. Let D_0 be the distribution over convex functions in $\mathbb{R}^2$ whose existence follows from Corollary 12.1 with $c < 1/(4T^2)$.

We now define a distribution over convex functions in $\mathbb{R}^d$ as follows: at each iteration pick uniformly $\mathbf{z}$ from the set $\{\mathbf{z} = (z;i) : z \in \{-1,1\}, i = 1, ..., 5T\}$ and let:

$$\mathbf{f}(\mathbf{w};\mathbf{z}) = f((w_{2i-1}, w_{2i}); z).$$

To prove the result we proceed as follows: given a sample S drawn i.i.d from the distribution D, let us call a sample point $\mathbf{z}_t = (z_t, i_t)$ *good* if $t < T/2$ and if i_t appears only once in the sample (i.e. for any $t' \leq T$, $i_{t'} \neq i_t$). Denote by S_g the set of good samples.

Next for a sample S define a sample $S' = \{\mathbf{z}'_1, \ldots, \mathbf{z}'_T\}$ to be a sample that differ from S only at good sample points, and for every good sample point if $\mathbf{z}_t = (z_t, i_t)$ then $\mathbf{z}'_t = (z'_t, i_t) = (-z_t, i_t)$. It is not hard to see that S and S' are identically distributed (though dependent).

Now first, we want to show that $F_S(\mathbf{w}_S) = F_S(\mathbf{w}_{S'})$ w.p. 1 and that w.p. 0.2 we have that

$$\|\mathbf{w}_S - \mathbf{w}_{S'}\| > \frac{\sqrt{T}\eta}{22}.$$

If we can show that, then we are done. Indeed because, by symmetry, we have with probability 1/2 $r(\mathbf{w}_{S'}) \leq r(\mathbf{w}_S)$. We can then take $\mathbf{w}_r = \mathbf{w}_{S'}$. Taken together we have that with probability 0.1 all the requirements of the theorem hold.

Claim. $F_S(\mathbf{w}_S) = F_S(\mathbf{w}_{S'})$

Proof. Indeed, fix a sample S. To avoid cumbersome notations, and because S, S' are fixed, we will denote here $\mathbf{w}_S = \bar{\mathbf{w}}$ and $\mathbf{w}_{S'} = \bar{\mathbf{w}}'$. Next, for a vector $\mathbf{w}$ and coordinate i_t let us also denote $\mathbf{w}(i_t) = (w_{2i_t-1}, w_{2i_t}) \in \mathbb{R}^2$. Note that for any $\mathbf{w}$, and $\mathbf{z}_t$, the value $\mathbf{f}(\mathbf{w};\mathbf{z}_t)$ depends only on $\mathbf{w}(i_t)$ (i.e. independent of the other coordinates). Also, for any i and t, we have that $\mathbf{w}^{(t)}(i)$ depends only on $\mathbf{z}_{t'}$'s such that $t' \le t$ and $i_{t'} = i$. In particular, for any $\mathbf{z}_t \notin S_g$ we have that $\bar{\mathbf{w}}(i_t) = \bar{\mathbf{w}}'(i_t)$, hence

$$f(\bar{\mathbf{w}};\mathbf{z}_t) = f(\bar{\mathbf{w}}(i_t); z_t) = f(\bar{\mathbf{w}}'(i_t); z_t) = f(\bar{\mathbf{w}}';\mathbf{z}_t).$$

Next, we want to show that for a good coordinate $\mathbf{z}_t$ we also have that $f(\bar{\mathbf{w}}; z_t) = f(\bar{\mathbf{w}}'; z_t)$. For this, as in Corollary 12.1 let us denote for any η and z by $\mathbf{v}_{z,\eta} = -\eta\nabla f(\mathbf{0}; z) \in \mathbb{R}^2$. Then, for any good coordinate we can show that

$$\bar{\mathbf{w}}(i_t) = -\frac{T-t}{T}\eta\nabla f(\mathbf{0}; z_t) = \mathbf{v}_{z_t,\eta'}, \tag{34}$$

$$\bar{\mathbf{w}}'(i_t) = -\frac{T-t}{T}\eta\nabla f(\mathbf{0}; -z_t) = \mathbf{v}_{-z_t,\eta'} \tag{35}$$

where $\eta' = \frac{T-t}{T}\eta > \frac{1}{2}\eta > c$. Indeed, recall that we chose $c = 1/(4T^2)$. Thus, from Corollary 12.1 we obtain that $f(\bar{\mathbf{w}}(i_t), z_t) = f(\bar{\mathbf{w}}'(i_t), z_t)$ and in particular

$$\mathbf{f}(\bar{\mathbf{w}};\mathbf{z}_t) = \mathbf{f}(\bar{\mathbf{w}}';\mathbf{z}_t).$$

□

Claim. *w.p. at least* 0.2 *we have that*

$$\|\mathbf{w}_S - \mathbf{w}_{S'}\| > \frac{\sqrt{T}\eta}{22},$$

Proof. Again we will use the notation $\bar{\mathbf{w}} = \mathbf{w}_S$ and $\bar{\mathbf{w}}' = \mathbf{w}_{S'}$. Note that by Corollary 12.1, as well as Eqs. (34) and (35) we have that

$$\|\bar{\mathbf{w}}(i_t) - \bar{\mathbf{w}}'(i_t)\| \ge \eta/8,$$

for any good sample point $\mathbf{z}_t$. Now:

$$\begin{aligned}\|\bar{\mathbf{w}}' - \bar{\mathbf{w}}\|^2 &= \sum_{i=1}^{5T} \|\bar{\mathbf{w}}(i) - \bar{\mathbf{w}}'(i)\|^2 \\ &\ge \sum_{i\in S_g} \|\bar{\mathbf{w}}(i) - \bar{\mathbf{w}}'(i)\|^2 \\ &\ge \frac{|S_g|\eta^2}{64}.\end{aligned} \tag{36}$$

Thus, we only need to show that $\mathbb{E}[|S_g|] > \frac{T}{5}$. Indeed, since $|S_g| < T/2$, we obtain by Markov's inequality that with probability 0.25, $|S_g| > T/7$

To show that $\mathbb{E}[|S_g|] > T/5$, for a sample S, let S_b contain all coordinates that collided (i.e. $\mathbf{z}_t$ such that for some $\mathbf{z}_{t'}$ we have that $i_t = i'_t$).
In order to calculate $|S_b|$ define $\chi_{t,t'} = I(i_t = i'_t)$ for every $t, t' \in [T]$. Note that $Pr(\chi_{t,t'} = 1) = \frac{1}{10T}$ and since there are at most $T(T-1)/2$ such pairs we get $\mathbb{E}[|S_b|] \le \sum_{t,t'} Pr(\chi_{t,t'} = 1) \le (T-1)/20$. Note that any coordinate i_t with $t < T/4$ that did not collide is a good coordinate, hence

$$\begin{aligned}\mathbb{E}[|S_g|] &\ge T/4 - \mathbb{E}[|S_b|] \\ &\ge T/5\end{aligned}$$

□

C.2 Proof of Theorem 4

Again, let D_0 be the distribution from Corollary 12.1 with $c > \frac{1}{kT^2}$, for some constant k (to be determined later). We define a distribution D over $\mathbb{R}^d$, where we let $d = 100T \cdot k$. as follows: pick k r.v $\{z^{(1)}, ..., z^{(k)}\} \in \{-1, 1\}$ and k coordinates $\{i^{(1)}, \ldots, i^{(k)}\} \in [d/2]$, set

$$\mathbf{f}(\mathbf{w}; \mathbf{z}) = \frac{1}{k} \sum_{\ell=1}^{k} f((w_{2i^{(\ell)}-1}, w_{2i^{(\ell)}}), z^{(\ell)}).$$

As before, for a given sample S, let us define S_g to be the set of "good samples" as follows: a tuple $(\mathbf{z}_t, \ell)$ is said to be *good* if $t < T/2$ and i_t^ℓ did not collide. Namely, any other sampled coordinate, $i_{t'}^{(\ell')}$ with $\ell' \in [k]$ and $t' \in [T]$ we have that if $i_{t'}^{(\ell')} = i_t^{(\ell)}$, then $\ell' = \ell$ and $t' = t$.

Next, for every sample S define

$$\mathcal{S}(S) = \{S' = (\mathbf{z}'_1, \ldots, \mathbf{z}'_T) : {i'}_t^{(\ell)} = i_t^{(\ell)} \forall t, \ell \text{ and } \forall (\mathbf{z}_t, \ell) \notin S_g \ {z'}_t^{(\ell)} = z_t^{(\ell)} \ \}.$$

In words, $\mathcal{S}(S)$ includes all samples where at a good coordinate $(\mathbf{z}_t, \ell)$, z_t^ℓ may flip.

One can show that if we randomly pick S and then pick uniformly an elements from $S' \in \mathcal{S}(S)$ then S and S' are identically distributed. As a corollary, if we pick a sample S then w.p. 0.5 we have that

$$|\{\mathbf{w}_{S'} : S' \in \mathcal{S}(S), \ r(\mathbf{w}_{S'}) \leq r(\mathbf{w}_S)\}| \geq \frac{|\mathcal{S}(S)|}{2}.$$

We next wish to prove that

$$\{\mathbf{w}_{S'} : S' \in \mathcal{S}(S)\} \subseteq \{\mathbf{w} : F_S(\mathbf{w}) \leq F_S(\mathbf{w}_S)\}. \tag{37}$$

Once we show that, the result will follow from the following lemma (which we deter its proof to the end of the section).

Lemma 4. *For every S, let $A \subseteq \mathcal{S}(S)$ be a set such that $|A| > \frac{|\mathcal{S}(S)|}{2}$, then A is $(2T, 10^{-4}\eta\sqrt{T})$-statistically complex*

We proceed by showing that Eq. (37) holds. The proof is very similar to the analog case in Theorem 3, and we will use similar notations: fix $S' \in \mathcal{S}(S)$ and use the shorthand notation $\bar{\mathbf{w}}$ for $\mathbf{w}_S$ and $\bar{\mathbf{w}}'$ for $\mathbf{w}_{S'}$, we will also use $\mathbf{w}(i, \ell) = (w_{2i^{(\ell)}-1}, w_{2i^{(\ell)}}) \in \mathbb{R}^2$. Another notation we add, as in Corollary 12.1, is as follows: for η and z, $\mathbf{v}_{z,\eta} = -\eta \nabla f(0; z) \in \mathbb{R}^2$.

Then for any sample $(\mathbf{z}_t, \ell) \in S_g$, in S_g, we can show that

$$\bar{\mathbf{w}}(i_t, \ell) = -\frac{T - t + 1}{kT} \eta \nabla f(0; z_t) = \mathbf{v}_{z_t^{(\ell)}, \eta'_t}, \tag{38}$$

$$\bar{\mathbf{w}}'(i_t, \ell) = -\frac{T - t + 1}{kT} \eta \nabla f(0; z_t) = \mathbf{v}_{{z'}_t^{(\ell)}, \eta'_t}, \tag{39}$$

Where $\eta' = \frac{T-t+1}{kT}\eta > c$. Next, for any coordinate $(\mathbf{z}_t, \ell) \notin S_g$ we can show that $\bar{\mathbf{w}}(i, \ell) = \mathbf{w}'(i, \ell)$, hence if $\mathbf{z}_t$ is such that $(i_t, \ell_t) = (i, \ell)$ we clearly have that $f(\bar{\mathbf{w}}, z_t^{(\ell)}) = f(\bar{\mathbf{w}}', z_t^{(\ell)})$. Now for $(\mathbf{z}_t, \ell) \in S_g$, from Corollary 12.1 we obtain that

$$\begin{aligned}
\mathbf{f}(\mathbf{w}, z_t^{(\ell)}) &= \frac{1}{k} \sum_{\ell=1}^{k} f(\bar{\mathbf{w}}(i_t, \ell); z_t^{(\ell)}) \\
&= \frac{1}{k} \sum_{\ell=1}^{k} f(\mathbf{w}_{\eta'_t, z_t^{(\ell)}}; z_t^{(\ell)}) && \textit{Eq. (38)} \\
&= \frac{1}{k} \sum_{\ell=1}^{k} f(\mathbf{w}_{\eta'_t, {z'}_t^{(\ell)}}; z_t^{(\ell)}) && \textit{Corollary 12.1} \\
&= \frac{1}{k} \sum_{\ell=1}^{k} f(\bar{\mathbf{w}}'(i_t, \ell); z_t^{(\ell)}) && \textit{Eq. (39)} \\
&= \mathbf{f}(\bar{\mathbf{w}}', z_t^{(\ell)})
\end{aligned}$$

Proof of Lemma 4 Fix S. Similar to the argument in Theorem 3, we have that with probability 0.2, that $|S_g| > T \cdot k/7$. We claim that if this event occurred, with correct choice of k, then every subset of size $|\mathcal{S}(S)|/2$ will be $(2T, 10^{-4}\eta\sqrt{T})$-statistically complex.

Indeed, let us index the coordinates of $\mathbb{R}^{|S_g|}$ by the elements of S_g. Then, for every element $\mathbf{w} \in \{\mathbf{w}_{S'} : S' \in \mathcal{S}\}$ we let $\mathbf{u}(\mathbf{w}) : \mathbb{R}^d \to \mathbb{R}^{|S_g|}$ be an affine projection such that: if $(\mathbf{z}_t, \ell) \in S_g$, then $\mathbf{u}(w)_{(\mathbf{z}_t,\ell)}$ satisfies the following:

$$\mathbf{u}(w)_{(\mathbf{z}_t,\ell)} = \begin{cases} \frac{1}{\sqrt{|S_g|}} & \mathbf{w}(i_t, \ell) = \mathbf{v}_{1,\eta'_t} \\ -\frac{1}{\sqrt{|S_g|}} & \mathbf{w}(i_t, \ell) = \mathbf{v}_{-1,\eta'_t} \end{cases}$$

It can be seen from Eq. (38) and Eq. (39) and Corollary 12.1 that $\|\mathbf{v}_{1,\eta'_t} - \mathbf{v}_{-1,\eta'_t}\| > \frac{T-t+1}{kT}\eta/4 > \frac{\eta}{12k}$, hence we can define $\mathbf{u}$ to be g-Lipschitz where

$$g = \frac{24k}{\eta\sqrt{T}}.$$

Combining this with Corollary 6.1, we get that there exists a distribution D over 1-Lipschitz convex functions such that, given $m = |S_g|/6 > Tk/62$ elements from D, with probability $1/4$ there is $\mathbf{w} \in A$ such that

$$\frac{1}{m}\sum_{i=1}^{m} f(\mathbf{w}, z_t) = 0.$$

but,

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim D} f(\mathbf{w}, z) > 3/(g \cdot 4) > 0.003\frac{\eta\sqrt{T}}{k}$$

Taking $k = 124$, we get that A is $(2T, 10^{-4}\eta\sqrt{T})$-Statistically complex.

D Proof of Theorem 5

We begin the construction by the definition of the distribution D:

$$f(\mathbf{w}; z = 1) = \begin{cases} w_1 & \text{if } \mathbf{w} \in A \\ 0 & \text{else} \end{cases}, \qquad f(\mathbf{w}; z = 3) = w_2,$$

$$f(\mathbf{w}; z = 2) = \begin{cases} -w_1 & \text{if } \mathbf{w} \in A \\ 0 & \text{else} \end{cases}, \qquad f(\mathbf{w}; z = 4) = -w_2$$

Where $z \sim Uniform([1, 2, 3, 4])$ and

$$A = \{(w_1, w_2) : |w_1|, |w_2| \leq \frac{1}{4}\}.$$

Note that by symmetry $F = E_z[f(\mathbf{w}, z)] = 0$, and indeed in expectation this is a convex function. For the proof we will define two "good" events, set $c = \eta\sqrt{T} = \Theta(1)$, and let:

$$E_1 : |w_2^S| > \frac{1}{4},$$

$$E_2(\beta) : |w_1^S| > \frac{\eta\sqrt{T}}{2} \cdot \beta$$

where we write $\mathbf{w}_S = (w_1^S, w_2^S)$, and β is a parameter sufficiently small so that.

$$\mathrm{erf}(\beta) \leq \sqrt{\mathrm{erf}\left(\frac{\sqrt{50}}{4c}\right) - \mathrm{erf}\left(\frac{\sqrt{50}}{4c}\right)}.$$

Note that β depends only on $c = \Theta(1)$.

Let us denote by $E(\beta) = E_1 \cap E_2(\beta)$, then we will rely on the following claim that lower bounds the probability of the event E. We deter the proof of the claim to the end of the section and continue with the proof:

Claim 14. *Let $E(\beta) = E_1 \cap E_2(\beta)$ and suppose that $z \sim D$ then, for our choice of β, and sufficiently large T*

$$P(E(\beta)) > 1 - \sqrt{erf\left(\frac{50}{4c}\right)}.$$

Proof. Taking Claim 14 into account, Fix a random sample S. Let β and T be as in Claim 14 and assume that event $E := E(\beta)$ occurred. Throughout, let us denote $c = \eta\sqrt{T}$.

To show that the statement holds, we define $\mathbf{w}_0^* = (0, \bar{w}_2^S)$ and $\mathbf{w}_{-1}^* = (-w_1^S, \bar{w}_2^S)$. We will show that for one of these candidate vectors the statement holds.

First we want to show that if $\eta = \Theta(1/\sqrt{T})$, then $\|\mathbf{w}_S - \mathbf{w}_0^*\| = \Theta(1)$. Indeed, note that since $E_2(\beta)$ occurred

$$\|\mathbf{w}_S - \mathbf{w}_0^*\|_2 \geq |w_1^S| \geq \frac{\eta\sqrt{T}}{2} \cdot \beta = \Theta(1).$$

Similarly $\|\mathbf{w}_S - \mathbf{w}_{-1}^*\| = \Theta(1)$.

Next we want to show that $F_S(\mathbf{w}_0^*) \leq F(\bar{\mathbf{w}})$, or $F_S(\mathbf{w}_{-1}^*) \leq F(\bar{\mathbf{w}})$. Note that for every $\mathbf{w}$ such that $|w_2| \geq \frac{1}{4}$, for every $z = \{1, 2, 3, 4\}$, $f(\mathbf{w}; z)$ depends only on the second coordinate, namely w_2. In particular, if $|w_2^S| \geq \frac{1}{4}$ we obtain by the construction that $F_S(\mathbf{w}_S) = F_S(\mathbf{w}_0^*) = F_S(\mathbf{w}_{-1}^*)$. Thus, due to event E_1 we obtain the desired result.

Finally, we want to show $\min\{r(\mathbf{w}_0^*), r(\mathbf{w}_{-1}^*)\} < r(\mathbf{w}_S)$, w.p probability at least $1/4$. First, assume that with probability $1/2$ we have that $r(\mathbf{w}_{-1}^*) \neq r(\mathbf{w}_S)$. By symmetry one can show that in this case we have that $r(\mathbf{w}_{-1}^*) < r(\mathbf{w}_S)$ with probability $1/2$. Next, assume that $r(\mathbf{w}_{-1}^*) = r(\mathbf{w}_S)$ with probability at least $1/2$. In this case, we obtain that:

$$\begin{aligned} r(\mathbf{w}_0^*) &= r(0.5 \cdot \mathbf{w}_S + 0.5 \cdot \mathbf{w}_{-1}^*) \\ &< \max(r(\mathbf{w}_S), r(\mathbf{w}_{-1}^*)) \\ &= r(\mathbf{w}_S) \end{aligned}$$

□

We are left with proving Claim 14

Proof of Claim 14 We will bound each event E_1, E_2 separately. We begin by bounding the event E_1:

Bounding E_1: For E_1 we claim the following:

$$Pr\Big(|w_2^S| \leq \frac{1}{4}\Big) \leq \text{erf}\left(\frac{\sqrt{50}}{4c}\right) + \sqrt{\frac{50^3}{T}} \tag{40}$$

where Φ is the CDF of a mean zero unit variate normally distributed random variable, and erf is the error function, namely $\text{erf}(x) = 1 - 2\Phi(-x)$.

Note that if $\eta = O(\frac{1}{\sqrt{T}})$, given the above bound, the probability that $|w_2^S| > \frac{1}{4}$ is a constant.

Proof. Recall that

$$w_2^S = \frac{1}{T}\sum \eta(T-t)\frac{\partial f(\mathbf{w}^{(t)}, z_t)}{\partial w_2},$$

and one can observe that $\frac{\partial f(\mathbf{w}^{(t)}, z_t)}{\partial w_2}$ equals 1 w.p. $1/4$, -1, w.p $1/4$ and 0 w.p $1/2$, independently of $z_{t'}$ for $t' \neq t$.

Hence, applying Lemma 2, with $c = \eta\sqrt{T}$, $k = 1$ and $a = \frac{\sqrt{50}}{4c}$ we obtain that

$$\begin{aligned} P(-\frac{1}{4} \leq w_2^S \leq \frac{1}{4}) =& P\Big(-\frac{\sqrt{50}}{4c}\frac{c}{\sqrt{50}} \leq w_2^S \leq \frac{\sqrt{50}}{4c} \cdot \frac{c}{\sqrt{50}}\Big) \\ &\leq \text{erf}\left(\frac{\sqrt{50}}{4c}\right) + \sqrt{\frac{50^3}{T}} \end{aligned} \tag{41}$$

□

We next move on to bound E_2

Bounding E_2: Let us consider a random sample $S' = \{z'_1, \ldots, z'_T\}$ that is generated by picking a random sample $S = z_1, \ldots, z_T$ i.i.d distributed according to D, and then for every z_t such that $z_t \in \{1,2\}$ with probability half we let $z'_t = 1$ and with probability half we let $z'_t = 2$. It can be seen that S' is an i.i.d sequence drawn according to the distribution D.

Next, let us denote $c = \eta\sqrt{T}$, and a parameter α (to be chosen later). Define the event

$$E_\tau : \{S : \min\{t : \mathbf{w}^{(t)} \notin A\} > \frac{T}{\alpha \cdot c}\}.$$

For our choice of $\beta > 0$ we claim that for every $\alpha > 0$

$$Pr\Big(|w_1^{S'}| < \frac{\beta}{\sqrt{50\alpha}}\Big|S, S' \in E_\tau\Big) \le 2\mathrm{erf}(\beta) + 2\sqrt{\frac{50^2 c^3 \alpha}{T}}$$

Indeed, Given S, let $\tau = \min\{t : \mathbf{w}^{(t+1)} \notin A\}$ and set $S'_\tau = \{z'_1, \ldots, z'_\tau\}$ and denote

$$X_\tau = \frac{1}{\sqrt{T}} \sum_{t=1}^{\tau} c\frac{T-t}{T} x_t.$$

where x_t are i.i.d random variables such that w.p. $1/4$ equals 1, w.p. $1/4$ equals -1 and w.p. $1/2$ equals 0. Due to symmetry we have that:

$$Pr\Big(|w_1^{S'}| < \frac{\beta}{\sqrt{\alpha \cdot c}}\Big|S, S' \in E_\tau\Big) \le 2Pr\Big(|X_\tau| < \frac{\beta}{\sqrt{\alpha \cdot c}}\Big|S, S' \in E_\tau\Big)$$

One can observe that

$$X_\tau = \sum_{t \in I} \eta \frac{T-t}{T} x_t = \frac{1}{\sqrt{T}} \sum_{t \in I} c\frac{T-t}{T} x_t.$$

Thus applying again Lemma 2 with $c = \sqrt{T}\eta$, $I = \{1, \ldots, T/k\}$ with, $k = \alpha \cdot c^3/50$ and $a = \beta$, we obtain the desired result.

Next, we want to bound $P(E_\tau)$. Now assume that for some $t < T/(\alpha \cdot c)$, we have that $\mathbf{w}^{(t)} \notin A$.

Let $T_\alpha = T/(\alpha \cdot c)$ and let $Z_1, \ldots, Z_{T_\alpha}$, be i.i.d copies of a random variable such that $P(Z_t = 1) = P(Z_t = -1) = 1/2$. Then

$$\begin{aligned}
P(\neg E_\tau) &\le 4P\left(\min\{t : \eta\sum_{i=1}^{t} Z_i > \frac{1}{4}\} < \frac{T}{\alpha \cdot c}\right) \\
&\le 4P\left(\min\{t : \eta\sum_{i=1}^{t} Z_i > \frac{1}{4}\} < T_\alpha, \eta\sum_{i=1}^{T_\alpha} Z_i \ge \frac{1}{4}\right) + 4P\left(\min\{t : \eta\sum_{i=1}^{t} Z_i > \frac{1}{4}\} < T_\alpha, \eta\sum_{i=1}^{T_\alpha} Z_i \le \frac{1}{4}\right) \\
&= 8P\left(\eta\sum_{i=1}^{T_\alpha} Z_i \ge \frac{1}{4}\right) \qquad (42)
\end{aligned}$$

where the last inequality is by symmetry (reflection principle). Next, by applying Hoeffding's inequality we obtain that

$$P(\eta\sum_{t=1}^{T_\alpha} Z_t \ge \frac{1}{4}) = P(\frac{\alpha\eta}{\sqrt{T}}\sum_{t=1}^{T_\alpha} Z_t \ge \frac{\alpha}{4\sqrt{T}}) = P(\frac{\alpha c}{T}\sum_{t=1}^{T_\alpha} Z_t \ge \frac{\alpha}{4\sqrt{T}}) \le e^{-\frac{\alpha c}{32}}$$

Taken together we obtain that

$$P(\neg\, E_\tau) \le 8e^{-\frac{\alpha c}{32}},$$

and

$$\begin{aligned}
P(\neg E_2(\beta)) &\le P(\neg E_2|E_\tau)P(E_\tau) + P(\neg E_\tau) \\
&\le \mathop{\mathbb{E}}_{S}\left[P(|w_1^{S'}| < \frac{\beta}{\sqrt{\alpha c}}|S, S' \in E_\tau)\right] + P(\neg E_2) \\
&\le 2\mathrm{erf}(\beta) + 2\sqrt{\frac{50^2 c^3 \alpha}{T}} + 8e^{-\frac{\alpha c}{32}} \qquad (43)
\end{aligned}$$

which yields the desired result.

Bounding $E(\beta)$: Eqs. (41) and (43) yields then:

$$\begin{aligned} P(\neg E) &< P(\neg E_1) + P(\neg E_2(\beta)) \\ &\leq \text{erf}\left(\frac{\sqrt{50}}{4c}\right) + 2\text{erf}(\beta) + 3\sqrt{\frac{50^2(50 + c^3\alpha)}{T}} + 8e^{-\frac{\alpha \cdot c}{32}} \end{aligned}$$

Choosing β sufficiently small, one can see that for large enough α and T we obtain the desired result.