The Sample Complexity of Gradient Descent in Stochastic Convex Optimization

Roi Livni
School of Electrical Engineering
Tel Aviv University
rlivni@tauex.tau.ac.il

Abstract

We analyze the sample complexity of full-batch Gradient Descent (GD) in the setup of non-smooth Stochastic Convex Optimization. We show that the generalization error of GD, with common choice of hyper-parameters, can be $\tilde{\Theta}(d/m + 1/\sqrt{m})$, where d is the dimension and m is the sample size. This matches the sample complexity of *worst-case* empirical risk minimizers. That means that, in contrast with other algorithms, GD has no advantage over naive ERMs. Our bound follows from a new generalization bound that depends on both the dimension as well as the learning rate and number of iterations. Our bound also shows that, for general hyper-parameters, when the dimension is strictly larger than number of samples, $T = \Omega(1/\varepsilon^4)$ iterations are necessary to avoid overfitting. This resolves an open problem by Amir, Koren, and Livni [3], Schliserman, Sherman, and Koren [20], and improves over previous lower bounds that demonstrated that the sample size must be at least square root of the dimension.

1 Introduction

Stochastic Convex Optimization (SCO) is a theoretical model that depicts a learner that minimizes a (Lipschitz) convex function, given finite noisy observations of the objective [22]. While often considered simplistic, in recent years SCO has become a focus of theoretical research, partly, because of its importance to the study of first-order optimization methods. But, also, it has become focus of study because it is one of few theoretical settings that exhibit *overparameterized learning* . In more detail, classical learning theory often focuses on the tension between number of samples, or training data, and the complexity of the model to be learnt. A common wisdom of classical theories [1, 7, 14, 24] is that, to avoid overfitting, the complexity of a model should be adjusted in proportion to the amount of training data. However, recent advances in Machine Learning have challenged this viewpoint. Evidently [18, 25], state-of-the-art algorithms generalize well but without, explicitly, controlling the capacity of the model to be learnt. In turn, today, it is one of the most emerging challenges, for learning theory, to understand learnability when the number of parameters in a learnt model exceeds the number of examples, and when, seemingly, nothing withholds the algorithm from overfitting.

Towards this, we look at SCO. In SCO, Shalev-Shwartz, Shamir, Srebro, and Sridharan [22] showed how algorithms can overfit with *dimension dependent* sample size. But, at the same time, it was known [8, 26] that there are algorithms that provably avoid overfitting with far fewer examples than dimensions. As such, SCO became a canonical model to study how a well-designed algorithm can avoid overfitting even when the number of examples is too small to guarantee generalization by an algorithmic-independent argument [2–5, 11, 12, 16, 20–22]. A step towards understanding *what* induces generalization is to identify *which* algorithms generalize. Then, we can ask what yields the separation. Surprisingly, for many well-studied algorithms this question is not always answered.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Perhaps the simplest algorithm, whose sample complexity is not yet understood, is Gradient Descent (GD). And we turn to the basic question of the sample complexity of gradient descent.

While this question remained open, there have been several advancements and intermediate answers: The first, dimension independent, generalization bound was given by Bassily, Feldman, Guzmán, and Talwar [6] that provided stability bounds [8]. The result of Bassily et al. demonstrated that, GD can have *dimension-independent* sample complexity rate. However, to achieve that, one has to use non-standard choice of hyperparameters which affects the efficiency of the algorithm. In particular, the number of rounds becomes quadratic in the size of the sample (as opposed to linear, with standard choice). On the other hand, a classical covering argument shows that linear dependence in the dimension is the worst possible, for any empirical risk minimizer, irrespective of properties such as stability.

In terms of lower bounds, Amir, Koren, and Livni [3] were the first to show that GD may have a dimension dependence in the sample complexity. They showed that, with natural hyperparameters, the algorithm must observe number of samples that is at least logarithmic in the dimension. This result was recently improved by Schliserman, Sherman, and Koren [20] that showed that at least square root of the dimension is required. Taken together, so far it was shown that either the algorithm's hyperparameters are tuned to achieve stability, at a cost in running time, or the algorithm must suffer *some* dimension dependence, linear at worst square root at best.

Here, we close the gap and show that linear dependence is necessary. Informally, we provide the following generalization error bound, in terms of dimension d, sample size, m, and hyperparameters of the algorithm, η and T (the learning rate and number of iterations). We show that when T is at most cubic in the dimension (see Theorem 1 for a formal statement):

$$\text{Generalization gap of GD} = \Omega\left(\min\left\{\frac{d}{m}, 1\right\} \cdot \min\left\{\eta\sqrt{T}, 1\right\}\right).$$

The first factor in the RHS describes the linear dependence of the generalization error in the dimension, and corresponds to the optimal sample complexity of empirical risk minimizers, as demonstrated by Carmon, Livni, and Yehudayoff [11]. The second term lower bounds the stability of the algorithm [6], and played a similar role in previous bounds [3, 20]. Each factor is optimal at a certain regime, and cannot be improved. Most importantly, for a standard choice of $\eta = O(1/\sqrt{T})$, the first term is dominant, and the aformentioned lower bound is complemented with the upper bound of Carmon et al. [11]. Our result implies, then, a sample complexity of $\tilde{\Theta}(d/m + 1/\sqrt{m})$. When $d \geq m$, the second factor is dominant. When running time is at most quadratic in number of examples, this term also governs the stability of the algorithm, hence the result of Bassily et al. [6] provides a complementary upper bound (see further discussion in Section 3.1).

2 Background

We consider the standard setup of Stochastic Convex Optimization (SCO) as in [22]. Set $\mathcal{W} = \{w : \|w\| \leq 1\}$, and let $\mathcal{Z}$ be an arbitrary, finite domain (our main result is a lower bound, hence finiteness of $\mathcal{Z}$ is without loss of generality). We assume that there exists a function $f(w, z)$ that is convex and L-Lipschitz in $w \in \mathcal{W}$ for every choice of $z \in \mathcal{Z}$. Recall that a function f is convex and L-Lipschitz if for any $w_1, w_2 \in \mathcal{W}$ and $0 \leq \lambda \leq 1$:

$$f(\lambda w_1 + (1-\lambda)w_2) \leq \lambda f(w_1) + (1-\lambda) f(w_2), \text{ and, } |f(w_1) - f(w_2)| \leq L\|w_1 - w_2\|. \quad (1)$$

First order optimization Algorithmically we require further assumptions concerning any interaction with the function to be optimized. Recall [19] that, for fixed z, the sub-gradient set of $f(w, z)$ at point w is the set:

$$\partial f(w, z) = \left\{g : f(w', z) \geq f(w, z) + g^\top (w' - w), \forall w' \in \mathcal{W}\right\}.$$

A first order oracle for f is a mapping $\mathcal{O}_z(w)$ such that $\mathcal{O}_z(w) \in \partial f(w, z)$. Our underlying assumption is that a learner has a first order oracle access. In other words, given a function $f(w, z)$, we will assume that there is a procedure $\mathcal{O}_z$ that calculates and returns a subgradient at every w for every z. Recall [10, 19] that when $|\partial f(w, z)| = 1$, the function is differentiable, at w, and in that case, the unique subdifferential is the gradient $\nabla f(w, z)$.

Learning

A learning algorithm A, in SCO, is any algorithm that receives as input a sample $S = \{z_1, \ldots, z_m\} \in \mathcal{Z}^m$ of m examples, and outputs a parameter w_S. An underlying assumption in learning is that there exists a distribution D, unknown to the learner A, and that the sample S is drawn i.i.d from D. The goal of the learner is to minimize the population loss:

$$F(w) = \mathop{\mathbb{E}}_{z\sim D}[f(w,z)],$$

More concretely, We will say that the learner has sample complexity $m(\varepsilon)$ if, assuming $|S| \geq m(\varepsilon)$, then w.p. $1/2$ (Again, because we mostly care about lower bounds, fixing the confidence will not affect the generality of our result):

$$F(w_S) - \min_{w\in\mathcal{W}} F(w) \leq \varepsilon. \tag{2}$$

Empirical Risk Minimization A natural approach to perform learning is by *Empirical Risk Minimization* (ERM). Given a sample S, the empirical risk is defined to be:

$$F_S(w) = \frac{1}{|S|}\sum_{z\in S} f(w,z).$$

An ε-ERM is any algorithm that, given sample S, returns a solution $w_S \in W$ that minimizes the empirical risk up to additive error $\varepsilon > 0$. Recently, Carmon et al. [11] showed that any ε-ERM algorithm has a sample complexity bound of

$$m(\varepsilon) = \tilde{O}\left(\frac{d}{\varepsilon} + \frac{1}{\varepsilon^2}\right), \tag{3}$$

The above rate is optimal up to logarithmic factor [12]. Namely, there exists a construction and an ERM that will fail, w.p. $1/2$, unless $m = \Omega(d/\varepsilon)$ examples are provided[1]. Importantly, though, there are algorithms that can learn with much smaller sample complexity. In particular SGD [26], stable-GD [6] and regularized ERMs [8].

Gradient Descent

We next depict Gradient Descent whose sample complexity is the focus of this work. GD depends on hyperparameters $T \in \mathbb{N}$ and $\eta \geq 0$ and operates as follows on the empirical risk. The algorithm receives as input a sample $S = \{z_1, \ldots, z_m\}$, defines $w_0 = 0$, and operates for T iterations according to the following recursion:

$$w_t = \Pi\left[w_{t-1} - \frac{\eta}{|S|}\sum_{z\in S}\mathcal{O}_z(w_t)\right] \Rightarrow\ w_S^{GD} := \frac{1}{T}\sum_{t=1}^{T} w_t, \tag{4}$$

where Π is the projection onto the unit ball, and $\mathcal{O}_z(w_t)$ is a subgradient of the loss function $f(w,z)$ at w_t. The final output, w_S^{GD}, of the algorithm is the averaged iterate (our result, though, can be generalized to other possible suffix-averages such as, say, outputting the last iterate, see Theorem 10). GD constitutes an ε-ERM. Concretely, it is known [10, 17] that GD minimizes the empirical risk and its optimization error is given by:

$$F_S(w_S^{GD}) - \min_{w\in W} F_S(w) = \Theta\left(\min\left\{\eta + \frac{1}{\eta T}, 1\right\}\right). \tag{5}$$

The above bound is tight irrespective of the dimension[2]. The population loss have also been studied, and Bassily et al. [6] demonstrated the following learning guarantee:

$$\mathop{\mathbb{E}}_{S\sim D^m}\left[F(w_S^{GD}) - \min_{w\in W} F(w)\right] = O\left(\eta\sqrt{T} + \frac{1}{\eta T} + \frac{\eta T}{m}\right). \tag{6}$$

The last two terms in the RHS follow from a stability argument, provided in [6], and the first term follows from the optimization error of GD as depicted in Eq. (5). Notice that there is always an $O(\eta\sqrt{T})$ gap between the generalization error and empirical error of gradient descent.

[1] the $\Omega(1/\varepsilon^2)$ sample complexity bound is more straightforward and follows from standard information-theoretic arguments

[2] For completeness, we demonstrate the lower bound for $d = 1$ at Appendix E

3 Main Result

Theorem 1. *For every $d \geq 4096, T \geq 10, m \geq 1$ and $\eta > 0$, there exists a distribution D, and a 4-Lipschitz convex function $f(w, z)$ in $\mathbb{R}^{d+1}$, such that for* any *first order oracle of $f(w, z)$, with probability $1/2$, if we run GD with η as a learning rate then:*

$$F(w_S^{GD}) - \min_{w \in \mathcal{W}} F(w) \geq \frac{1}{2 \cdot 272 \cdot 16^2} \cdot \min\left\{\frac{d}{1032m}, 1\right\} \cdot \min\left\{\eta\sqrt{\min\{\lfloor d^3/136 \rfloor, T\}}, 1\right\}.$$

We remark, that the above theorem is true for *any* suffix averaging (e.g. last iterate), and not restricted to the averaged iterate (see Theorem 10). We now specialize our bound for two interesting regimes. First, we improve previous dependence in the dimension in [3, 20] and obtain a generalization error bound for $d = \Omega(m + T^{1/3})$:

Corollary 2. *Fix η, and suppose $d = \Omega\left(m + T^{1/3}\right)$. There exists a distribution D, and an $O(1)$-Lipschitz convex function $f(w, z)$ in $\mathbb{R}^d$, such that for* any *first order oracle of $f(w, z)$, with probability $1/2$, if we run GD for T iterations, then:*

$$F(w_S^{GD}) - \min_{w \in \mathcal{W}} F(w) \geq \Omega\left(\min\left\{\eta\sqrt{T} + \frac{1}{\eta T}, 1\right\}\right). \tag{7}$$

The first term follows from Theorem 1, the second term follows from the optimization error in Eq. (5). Equation (7) does not hold for $d < m$, and the linear improvement over [2, 20] is tight. This can be seen from Eq. (5) that shows that GD achieves ε empirical excess error when $\eta = O(1/\sqrt{T})$ and $T = O(1/\varepsilon^2)$. Equation (7) becomes vacuous for such choice of parameters, but Carmon et al. [11] showed that the sample complexity of *any ERM* is bounded by $\tilde{O}((d + \sqrt{m})/m)$. However, as depicted next, this upper bound becomes tight and GD does not improve over a worst-case ERM:

Corollary 3. *Suppose $T = O(m^{1.5})$, and $\eta = \Theta(1/\sqrt{T})$. There exists a distribution D, and an $O(1)$-Lipschitz convex function $f(w, z)$ in $\mathbb{R}^d$, such that for any first order oracle of f, with probability $1/2$, if we run GD with η, for T iterations:*

$$F(w_S^{GD}) - \min_{w \in \mathcal{W}} F(w) \geq \Omega\left(\min\left\{\frac{d}{m} + \frac{1}{\sqrt{m}}, 1\right\}\right).$$

Corollary 3 complements Carmon et al. [11] upper bound, and improves over Feldman [12] lower bound that only showed existence of *some* ERM with the aforementioned sample complexity. To see that Corollary 3 follows from Theorem 1, notice that when $d \leq \sqrt{m}$, then $d/m < 1/\sqrt{m}$ and the bound is dominated by the second term, which is a well known-information theoretic lower bound for learning. When $d > \sqrt{m}$, and $T < m^{1.5}$ we have that $T \leq d^3$, plugging $\eta = O(1/\sqrt{T})$ yields the bound.

3.1 Discussion

Theorem 1 provides a new generalization error bound for GD. It shows that the worst case sample complexity for ERMs, derived by Feldman [12], is in fact applicable also to a very natural first order algorithm and not just to abstract ERMs. This Highlights the importance of choosing the right algorithm for learning in SCO. As discussed, the bound is tight in several regimes, nevertheless still there are unresolved open problems.

Stability in low dimension When GD is treated as a naive empirical risk minimizer, and $\eta = O(1/\sqrt{T})$, $T = O(m)$, there is no improvement, when using GD, over a worst-case ERM. In the other direction, for dimension that is linear in m, one cannot improve over the $\Omega(\eta\sqrt{T})$ term that governs stability. Our bound, though, provide a hope that stability in low dimension can yield an improved bound. In particular, consider the case where $\eta = 1/T^{1/4}$ and $d < m$. This is a case where we apply a stable algorithm in small dimensions. Our bound does not negate the possibility of an improved generalization bound. That would mean that, at least at some regime, GD can improve over the worst-case ERM behaviour. We leave it as an open problem for future study

Open Question 4. *Is there a generalization bound for GD such that*

$$\mathop{\mathbb{E}}_{S\sim D^m}\left[F(w_S^{GD}) - \min_{w\in\mathcal{W}} F(w)\right] = O\left(\frac{d\eta\sqrt{T}}{m} + \frac{1}{\sqrt{m}}\right).$$

Alternatively, can we prove an improved generalization error bound such that:

$$\mathop{\mathbb{E}}_{S\sim D^m}\left[F(w_S^{GD}) - \min_{w\in\mathcal{W}} F(w)\right] = \Omega\left(\min\left\{\frac{d}{m}, \eta\sqrt{T}, 1\right\}\right).$$

Late stopping Another regime where there is a gap between known upper bound and lower bound appears when $T = \Omega(m^2)$. Specifically, the stability upper bound for GD by Bassily et al. [6] gives

$$\mathop{\mathbb{E}}_{S\sim D^m}\left[F(w_S^{GD}) - \min_{w\in\mathcal{W}} F(w)\right] = O\left(\eta\sqrt{T} + \frac{\eta T}{m} + \frac{1}{\eta T}\right).$$

By Corollary 2, for large enough dimension:

$$\mathop{\mathbb{E}}_{S\sim D^m}\left[F(w_S^{GD}) - \min_{w\in\mathcal{W}} F(w)\right] = \Omega\left(\min\left\{\eta\sqrt{T} + \frac{1}{\eta T}, 1\right\}\right).$$

When $T = O(m^2)$, the two bounds coincide. Indeed, for the $\eta T/m$ term to dominate the $\eta\sqrt{T}$ term, we must have $T = \Omega(m^2)$. One has to take at least $T = O(m^2)$ iterations in order to generalize with GD (in fact, any full batch method [2]), however $T = O(m^2)$ iterations are sufficient. Nevertheless, the above gap does yield the possibility of an *unstable* GD method that does generalize. Particularly, if we just regulate the term $\eta\sqrt{T}$, but allow $\eta T/m = \Omega(1)$, then this may yield a regime where GD is unstable (and ERM bounds do not apply) and yet generalize.

Open Question 5. *Are there choices of η and T (that depend on m) such that $\eta T/m \in \Omega(1)$, but GD has dimension indpendent sample complexity?*

Notice that the $\eta T/m$ term also governs stability in the *smooth* convex optimization setup [13]. Recall that a function $f(w, z)$ is said to be β-smooth if for all z, $f(w, z)$ is differentiable, and the gradient is an β-Lipschitz mapping [10, 15]. For smooth optimization, even if $\eta\sqrt{T} = \Omega(1)$, GD is still stable. Hardt, Recht, and Singer [13] showed that the stability of GD in the smooth case is governed by $O\left(\frac{\eta T}{m}\right)$ for $\eta < 1/\beta$. Therefore, the question of generalization when $\eta T/m \in \Omega(1)$ remains open, even under smoothness assumptions:

Open Question 6. *Assume that $f(w, z)$ is $\Theta(1)$-smooth. What is the sample complexity of GD, when η and T are chosen so that $\eta + \frac{1}{\eta T} = o(1)$, but $\frac{\eta T}{m} = \Omega(1)$.*

4 Technical Overview

We next provide a high level overview of our proof technique. For simplicity of exposition we begin with the case $T = m = d$. We begin by a brief overview of previous construction by Amir et al. [3] that demonstrated Corollary 2 when $m = \Omega(\log d)$. The construction in [3] can be decomposed into three terms:

$$f(w, z) = g(w, z) + N_0(w) + h(w, z).$$

The function g has the property that an ERM may fail to learn, unless dimension dependent sample size is considered. Amir et al. [3] incorporated Shalev-Shwartz et al. [22] construction. Later, [20] used Feldman's function [12] to construct g. The shift from the construction depicted in Shalev-Shwartz et al. [22] to Feldman's function is the first step that allows to move from logarithmic to polynomial dependence in the dimension. In both constructions an underlying property of g is that there exists a distribution D such that, for small samples, there are overfitting minima. Concretely, there exists a $w_S \in \{0, 1/\sqrt{d}\}^d$ such that

$$\frac{1}{|S|}\sum_{z\in S} g(w_S, z) - \mathop{\mathbb{E}}_{z\sim D}[g(w_S, z)] = \Omega(1). \tag{8}$$

The challenge is then, to make gradient descent's trajectory move towards the point w_S. The idea can be decomposed into two parts:

Simplifying with an adversarial subgradient:

To simplify the problem, let us first ease the challenge and suppose we can choose our subgradient oracle in a way that depends on the observed sample. Let N_0 be the Nemirovski function [9]:

$$N_0(w) = \max\{-w(i), 0)\}.$$

Notice that N_0 is not differentiable and the choice of subgradient at certain points is not apriori determined. For example, notice that every standard basis vector $-e_i \in \partial N_0(0)$. More generally, given a sample S, let $I = i_1, \ldots, i_{d'}$ be exactly the set of indices such that w_S, from Eq. (8), $w_S(i) \neq 0$. Now assume by induction that $w_t(i) > 0$ exactly for $i = i_1, \ldots i_t$, then one can show that we can define the subgradient oracle of N_0:

$$\mathcal{O}(w_t) = -e_{i_{t+1}} \in \partial N_0(w_t).$$

In that case w_{t+1} will satisfy our assumption for i_{t+1} and we can continue to follow this dynamic for T steps.

Notice that, in this case, GD will converge to w_S (if $\eta = 1/\sqrt{d}$ which we assume now for concreteness). One can also show that the output of GD (the averaged iterate) will overfit. The caveat is that our subgradient oracle depends on the sample S. In reality, the sample is drawn independent of the subgradient oracle. and all previous constructions, as well as ours need to handle this. This is discussed in the next section. But before that, let us review another challenge which is when $T \neq d$:

When $d \ll T$ Another challenge we face with the construction above is that it works when we assume that $T \approx d$. That is because, in Nemirovski's function, the number of iterates we can perform is bounded by the dimension. After d iterations we will end up with the vector $v = \sum_{t=1}^{d} \eta e_{i_t}$. If $T = \omega(d)$ then $\eta = o(1/\sqrt{d})$, and the dynamic will end up with a too small norm vector to induce a sizeable population loss. This strategy will provide, at best, with a factor of the form $\Omega\left(\eta\sqrt{\min\{d, T\}}\right)$. Such a factor may be unsatisfactory in a very natural setting where, say, $T = O(m)$, $\eta = O(1/\sqrt{m})$, and $d = \Omega(\sqrt{m})$. To obtain the d^3 dependence, we perform the following alternation over the Nemirovski function. Consider the function:

$$N(w) = \max\{0, \max_{i \leq d}\{-w(i)\}, \max_{i \leq j \leq d}\{w(j) - w(i)\}\}. \tag{9}$$

And suppose that at each iteration we return a subgradient as follows:

- If there is $i \leq d$, such that $w(i) = w(i+1) > \eta$, return subgradient $e_{i+1} - e_i$ and w is updated by $w_{t+1} = w_t - \eta e_{i+1} + \eta e_i$.
- If there is no such i, then take the minimal i (if exists) such that $w(i) = 0$, and return subgradient $-e_i$ and update $w_{t+1} = w_t + \eta e_i$.
- When non of the above is met, return subgradient 0.

The dynamic of the above scheme is depicted in Fig. 1, and solves the problem when $T \approx d^3$. One can show that GD will run for at least $d^3 \approx T$ iterations, and will increase $O(d)$ coordinates, each, on average, by an order of $O(\eta d)$. This is better than the increase of η in each coordinate that we get from Nemirovski's function. In this way we obtain the improved result of $\eta\sqrt{T}$, even when $T \approx d^3$.

When $T \ll d$, when the number of iterations is smaller than d we face a different challenge. The immediate solution is to embed in $\mathbb{R}^d$ a construction from $\mathbb{R}^T$, this will provide us with the $\Omega(\eta\sqrt{T})$ term but, on the other hand, such a construction will not yield a $\Omega(d/m)$ term. A different approach, that exploits the dimension to its fullest, is to consider blocks of coordinates and operate on those instead of single coordinates.

The conclusive outcome incorporates both ideas together, and we replace the Nemirovski function with a version of Eq. (9) that operates on $O(T^{1/3})$ blocks of coordinates. And this concludes our construction. We next move on to the challenge of replacing the data dependent oracle with a standard first order oracle.

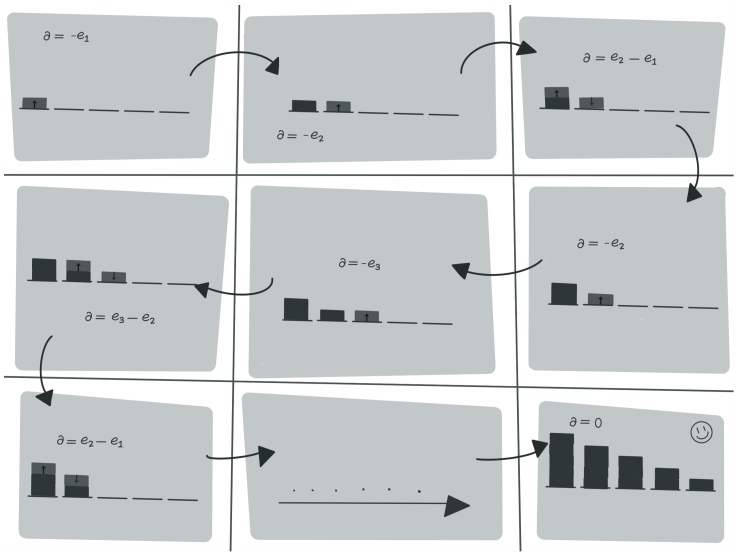

Figure 1: Depiction of the dynamics induced by Eq. (16) and our choice of sub-differentials

Reduction to sample dependent oracle:

As discussed, the construction above does not yield a lower bound as it relies on a subgradient oracle that is dependent on the whole sample. To avoid such dependence of the oracle on the sample, we observe that if we can infer the sample S from the trajectory, i.e. if the state w_t "encodes" the sample, then formally the subgradient is allowed to "decipher" the sample from the point w_t. In that way we achieve this behaviour of sample dependent subgradient oracle. This part becomes challenging and may depend on the way we choose g, and N. The simplest case, studied by Amir et al. [3], introduced the third function, h, which was a small perturbation function that elevates coordinates in I and inhibits coordinates not in I. The function h depends on z and not on S, hence it cannot know apriori I. But, an important observation is that, in Shalev-Shwartz et al. [22] construction, if $i \notin I$, there exists $z \in S$ that "certifies" that. In fact, each z can be thought of as a subset of indices, and if an index appears in z, then it cannot be in I. So we can build the perturbation in a way that every coordinate is elevated, unless z certifies $i \notin I$: In that case we define $h(w, z)$ so that its gradient will radically inhibit i.

The last observation is what becomes challenging in our case. As discussed, to achieve improved rate, we need to use Feldman's function. When using Feldman's function the coordinates cannot be ruled out, or identified, by a single z but one has to look at the whole sample to identify I. While Schliserman et al. [20] tackle a similar problem, we take a slightly different approach described next: For each z assign a random, positive, number $\alpha(z)$. We can think of this α as a hash function. Let us add another coordinate to the vector w, $w(d+1)$. Consider the function

$$h(w, z) = \gamma\alpha(z) \cdot w(d+1).$$

Then $\partial h(w, z) = \gamma\alpha(z)e_{d+1}$. Write $\alpha(S) = \frac{1}{|S|}\sum_{z\in S}\alpha(z)$ then in turn:

$$w_t(d+1) = w_{t-1}(d+1) - \partial\frac{\gamma}{|S|}\sum_{z\in S} h(w_{t-1}, z) = -t \cdot \gamma\alpha(S)e_{d+1}.$$

If $\gamma, \alpha(z)$ are chosen correctly, $\alpha(S)$ is a one to one mapping from samples to real numbers, and small γ ensures that the overall addition of h has negligible affect on the outcome. Then, we may define the subgradient oracle to be dependent on coordinate $d+1$ which encodes the whole sample. Our final construction will take a different h, which adds small strong convexity in this coordinate, for reasons next explained:

Working with any first order oracle Notice that our statement is slightly stronger than what we so far illustrated. Theorem 1 states that, for *any* subgradient oracle, GD will fail. For that, we need to be a little bit more careful, and we want to replace our function with a function that leads to the

same guaranteed trajectory, but at the same time it should be differentiable at visited points. This will ensure a unique derivative, making the construction independent of the choice of (sub)gradient oracle.

Towards this goal, we start with the construction depicted so far and consider the set of all values, gradients, and points $\{f_j, g_j, w_j\}_{j\in J}$ that our algorithm may visit, for any possible time step and any possible sample, with our construction. Notice that, while this set may be big and even exponential, it is nevertheless finite. What we want is to interpolate a new function through these triplets. In contrast with our original construction, we require a differentiable function at the designated points. Notice, that such an interpolation will have the exact same behaviour when implementing GD on it (with the added feature that the oracle is well defined and unique).

The problem of convex interpolation is well studied, for example Taylor et al. [23] shows sufficient and necessary conditions for interpolation of a smooth function. Our case is slightly easier as we do not care about the smoothness parameter. On the other hand we do require Lipschitzness of the interpolation. We therefore provide an elementary, self-contained, proof to the following easy to prove Lemma, (proof is provided in Appendix B)

Lemma 7. *Let $G = \{f_j, g_j, w_j\}_{j\in J} \subseteq \mathbb{R}\times\mathbb{R}^d\times\mathbb{R}^d$ be a triplet of values in $\mathbb{R}$, and gradients and points in $\mathbb{R}^d$, such that $\|g_j\| \le L$. Suppose that for every $i, j \in J$:*

$$f_i \ge f_j + g_j^\top (w_i - w_j), \tag{10}$$

and let

$$I_{diff} = \{i : f_i = f_j + g_j^\top (w_i - w_j) \Rightarrow g_i = g_j\}.$$

Then there exists a convex L-Lipschitz function $\hat{f}$ such that for all $j \in J$: $\hat{f}(w_j) = f_j$, and for all $i \in I_{diff}$, $\hat{f}$ is differentiable at w_i and:

$$\nabla f(w_i) = g_i.$$

With Lemma 7 at hand, consider the function

$$h(w, z) = \frac{1}{2}(w(d+1))^2 + \alpha(z)\cdot w(d+1).$$

The above function encodes in $w(d+1)$ the sample and time-step as before. Moreover, because it is slightly strongly convex (in coordinate $d+1$), $w_1(d+1) \ne w_2(d+1)$ ensures that

$$h(w_1, z) > h(w_2, z) + \nabla h(w_2, z)^\top (w_1 - w_2),$$

Then the term h in f ensures that the triples $\{f_j, g_j, w_j\}$ along the trajectory generate gradient vectors that satisfy strict inequality in Eq. (10) and in turn, our interpolation from Lemma 7 will be differentiable at these points. There's some technical subtlety because the interpolation needs to also take the averaged iterate into account, but this is handled in a similar fashion.

In the next two sections we provide more formal statements of the two main ingredients: First, we define a setup of optimization with a sample-dependent first order Oracle and state a lower bound for the generalization error in this setup. The second ingredient is a reduction from the standard setup of first order optimization.

4.1 Sample-dependent Oracle

As discussed, the first step in our proof is to consider a slightly weaker setup where the first-order oracle may depend on the whole sample. Let us formally define what we mean by that. Define

$$\mathcal{S}_m^T = \{\mathbf{S} = (S_1, \ldots, S_t), S_i \in \cup_{i=1}^m \mathcal{Z}^m, t \le T\},$$

the set of all subseqences of samples of size at most m. Given a function $f(w, z)$, a sample dependent oracle, $\mathcal{O}_S$, is a finite sequence of first order oracles

$$\mathcal{O}_S = \{\mathcal{O}^{(t)}(S; w, z)\}_{t=1}^T,$$

that each receive as input a finite sample S, as well as w and returns a subgradient:

$$\mathcal{O}^{(t)}(S, w, z) \in \partial f(w, z).$$

The sequence of samples can be thought of as the past samples that were observed by the algorithm. In the case of full-batch GD these will be the whole sample, and for SGD, each S provided to $\mathcal{O}^{(t)}$ will be all previously observed samples. Given $\mathbf{S} \in \mathcal{S}_m^T$ let us also denote

$$\mathcal{O}^{(t)}(\mathbf{S}, w) = \frac{1}{|S_t|} \sum_{z \in S_t} \mathcal{O}^{(t)}(S_{1:t-1}, w, z) \in \partial \left(\frac{1}{|S_t|} \sum_{z \in S_t} f(w, z) \right), \tag{11}$$

where we let $\mathcal{S}_{1:0} = \emptyset$, and $S_{1:t-1} = (S_1, \ldots, S_{t-1})$ is the concatenated subsample of all previously observed samples in the sequence. As discussed, working with a sample-dependent oracle is easier (for lower bounds). And indeed, our first result shows that, if the subgradient can be chosen in a way that depends on the sample, we can provide the desired lower bound. For fixed and known $\eta > 0, T$, a sample dependent first order oracle $\mathcal{O}_S$, and a sequence of samples $\mathbf{S} = (S_1, S_2, \ldots, S_T)$, define $w_0 = 0$ and inductively:

$$w_t^{\mathbf{S}} = \Pi \left[w_{t-1}^{\mathbf{S}} - \eta \mathcal{O}^{(t)}(\mathbf{S}, w_{t-1}^{\mathbf{S}}) \right],$$

and for every suffix $\mathbf{s} < T$:

$$w_{\mathbf{S},\mathbf{s}}^{GD} = \frac{1}{T - \mathbf{s}} \sum_{t=\mathbf{s}+1}^{T} w_t^{\mathbf{S}} \tag{12}$$

Lemma 8. *For every $m, d, T \geq 18$ and $\eta > 0$ there are a distribution D, a 3-Lipschitz convex function $f(w, z)$ in $\mathbb{R}^d$, as well as a sample dependent first order oracle $\mathcal{O}_S$ such that: if $\mathbf{S} = (S, S, \ldots S) \in \mathcal{S}_m^T$ for $S \sim D^m$ i.i.d, then w.p. $1/2$, for every suffix averaging $\mathbf{s}$:*

$$F(w_{\mathbf{S},\mathbf{s}}^{GD}) - F(0) \geq \frac{1}{\sqrt{2} \cdot 272 \cdot 16^2} \cdot \min\left\{ \frac{d}{1032m}, 1 \right\} \cdot \min\left\{ \eta \sqrt{\min\{\lfloor d^3/136 \rfloor, T\}}, 1 \right\}.$$

The proof of Lemma 8 is provided in Appendix A.1. We next move to describe the second ingredient of our proof.

4.2 Reduction to sample-dependent oracles

As discussed, the second ingredient of our proof is a reduction to the sample-dependent setup. Instead of using a perturbation function as in [3], we take a more black box approach and show that, given a sample dependent first order oracle, there exists a function that basically induces the same trajectory. Proof is provided in Appendix A.2.

Lemma 9. *Suppose $q \in \mathbb{R}^T$, $\|q\|_\infty \leq 1$. And suppose that $f(w, z)$ is a convex, L-Lipschitz, function over $w \in \mathbb{R}^d$, let $\eta > 0$, let $\mathcal{O}_S$ be a sample dependent first order oracle, and for every sequence of samples $\mathbf{S} = (S_1, S_2, \ldots, S_T)$ define the sequence $\{w_t^{\mathbf{S}}\}_{t=1}^T$ as in Eq. (12).*

Then, for every $\varepsilon > 0$ there exists an $L + 1$ Lipschitz convex function[3] $\bar{f}((w, x), z)$ over $\mathbb{R}^{d+1}$ (that depends on $q, f, T, \eta, m, \mathcal{O}_S, \varepsilon$).

such that for any first order oracle $\mathcal{O}_z$ for $\bar{f}$, define $u_0 = 0 \in \mathbb{R}^d$ and $x_0 = 0 \in \mathbb{R}$, and:

$$(u_t, x_t) = (u_{t-1}, x_{t-1}) - \frac{\eta}{|S_t|} \sum_{z \in S_t} \mathcal{O}_z((u_t, x_t))$$

then if we define:

$$u_q = \sum_{t=1}^{T} q(t) u_t, \text{ and, } \quad x_q = \sum_{t=1}^{T} q(t) x_t, \text{ and } w_q^{\mathbf{S}} = \sum_{t=1}^{T} q(t) w_t^{\mathbf{S}}.$$

then, we have that $u_q = w_q^{\mathbf{S}}$ and:

$$|\bar{f}((u_q, x_q), z) - f(w_q^{\mathbf{S}}, z)| \leq \varepsilon.$$

and,

$$|\bar{f}((0, 0), z) - f(0, z)| \leq \varepsilon.$$

[3] i.e. $w \in \mathbb{R}^d$ and $x \in \mathbb{R}$

Ackgnoweledgments The author would like to thank Tamar Livni for creating Figure 1. Tamar holds all copyrights to the artwork. The author would also like to thank Tomer Koren and Yair Carmon for several discussions. This research was funded in part by an ERC grant (FOG, 101116258), as well as an ISF Grant (2188 \ 20).

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

[2] I. Amir, Y. Carmon, T. Koren, and R. Livni. Never go full batch (in stochastic convex optimization). *Advances in Neural Information Processing Systems*, 34:25033–25043, 2021.

[3] I. Amir, T. Koren, and R. Livni. Sgd generalizes better than gd (and regularization doesn't help). In *Conference on Learning Theory*, pages 63–92. PMLR, 2021.

[4] I. Amir, R. Livni, and N. Srebro. Thinking outside the ball: Optimal learning with gradient descent for generalized linear stochastic convex optimization. *Advances in Neural Information Processing Systems*, 35:23539–23550, 2022.

[5] I. Attias, G. K. Dziugaite, M. Haghifam, R. Livni, and D. M. Roy. Information complexity of stochastic convex optimization: Applications to generalization and memorization. *arXiv preprint arXiv:2402.09327*, 2024.

[6] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

[7] 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.

[8] O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[9] S. Bubeck, Q. Jiang, Y.-T. Lee, Y. Li, and A. Sidford. Complexity of highly parallel non-smooth convex optimization. *Advances in neural information processing systems*, 32, 2019.

[10] S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[11] D. Carmon, R. Livni, and A. Yehudayoff. The sample complexity of erms in stochastic convex optimization. *arXiv preprint arXiv:2311.05398*, 2023.

[12] V. Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. *Advances in Neural Information Processing Systems*, 29, 2016.

[13] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

[14] D. Haussler and M. Warmuth. The probably approximately correct (pac) and other learning models. *The Mathematics of Generalization*, pages 17–36, 2018.

[15] E. Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[16] R. Livni. Information theoretic lower bounds for information theoretic upper bounds. *Advances in Neural Information Processing Systems*, 36, 2023.

[17] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[18] B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

[19] R. T. Rockafellar. *Convex Analysis:(PMS-28)*. Princeton university press, 2015.

[20] M. Schliserman, U. Sherman, and T. Koren. The dimension strikes back with gradients: Generalization of gradient methods in stochastic convex optimization. *arXiv preprint arXiv:2401.12058*, 2024.

[21] A. Sekhari, K. Sridharan, and S. Kale. Sgd: The role of implicit regularization, batch-size and multiple-epochs. *Advances In Neural Information Processing Systems*, 34:27422–27433, 2021.

[22] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *COLT*, volume 2, page 5, 2009.

[23] A. B. Taylor, J. M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161:307–345, 2017.

[24] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pages 11–30. Springer, 2015.

[25] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[26] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

A Proof of Theorem 1

The proof is an immediate corollary of Lemmas 8 and 9, which we prove in Appendices A.1 and A.2 respectively. To see that Theorem 1 indeed follows from these Lemmas, start with η, d, m, T that satisfy the conditions. Let $f(w,z)$ be the function and $\mathcal{O}_S$ the sample dependent first order oracle, whose existence follows from Lemma 8 with suffix $\mathbf{s} = 0$. And let $\bar{f}$ be the function whose existence follows from Lemma 9 to some arbitrarily small ε_0, with $q(t) = \frac{1}{T}$ for all t. It is easy to see that if we apply GD on $\bar{f}$ and define its output (u^{GD}, x^{GD}) then: $(u^{GD}, x^{GD}) = (u_q, x_q)$, and $w_q^{\mathbf{S}} = w_{\mathbf{S}}^{GD}$.

Then, we have that w.p. $1/2$:

$$\begin{aligned}
\bar{F}((u^{GD}, x^{GD})) - \bar{F}(0) &= \bar{F}(u_q, x_q) - \bar{F}(0) \\
&\geq F(w_{\mathbf{S}}^{GD}) - 2\varepsilon_0 - F(0) \\
&\geq \frac{1}{\sqrt{2} \cdot 272 \cdot 16^2} \cdot \min\left\{\frac{d}{1032m}, 1\right\} \cdot \min\left\{\eta\sqrt{\min\{\lfloor d^3/136 \rfloor, T\}}, 1\right\} - 2\varepsilon_0 \\
&\geq \frac{1}{2 \cdot 272 \cdot 16^2} \cdot \min\left\{\frac{d}{1032m}, 1\right\} \cdot \min\left\{\eta\sqrt{\min\{\lfloor d^3/136 \rfloor, T\}}, 1\right\}.
\end{aligned}$$

Where in the last equation, we assume ε_0 to be sufficiently small. Finally, note that $\bar{F}(0) \geq \min_{w \in \mathcal{W}} \bar{F}(w)$.

Notice, that by the same argument, by taking any suffix $\mathbf{s} < T$, and setting $q(t) = 0$ for $t \leq \mathbf{s}$, and $q(t) = \frac{1}{T-\mathbf{s}}$ for $t \geq \mathbf{s}+1$, we can obtain the following stronger result for any suffix averaging:

Theorem 10. *For every $d \geq 4096, T \geq 10, m \geq 1$ and $\eta > 0$ and suffix $\mathbf{s} < T$, there exists a distribution D, and a 4-Lipschitz convex function $f(w,z)$ in $\mathbb{R}^d$, such that for* any *first order oracle of $f(w,z)$, with probability $1/2$, if we run GD with η as a learning rate then:*

$$F\left(\frac{1}{T-\mathbf{s}}\sum_{t=\mathbf{s}+1}^{T} w_t^{\mathbf{S}}\right) - F(0) \geq \frac{1}{2 \cdot 272 \cdot 16^2} \cdot \min\left\{\frac{d}{1032m}, 1\right\} \cdot \min\left\{\eta\sqrt{\min\{\lfloor d^3/136 \rfloor, T\}}, 1\right\}.$$

A.1 Proof of Lemma 8

For simplicity we will assume that $d = 2^n$ for some $n \in \mathbb{N}$, the final result will be obtained by embedding a construction in a subspace of size at least half the original dimension.

We start by recalling Feldman's construction [12]: There exists a set $\mathcal{V} \subseteq \{0,1\}^d$, such that:

$$\left(\forall v_1 \neq v_2 \in \mathcal{V},\ v_1 \cdot v_2 \leq \frac{5d}{16}\right), \text{and } \left(\|v\|^2 \geq \frac{7d}{16}\right), \text{and } \left(|\mathcal{V}| > e^{d/258}\right). \tag{13}$$

Indeed, suppose we pick randomly $w \in \{0,1\}^d$ according to probability P where each coordinate $P(w(i)) = 1$ with probability $1/2$ (independently). Then, by Hoeffding's inequality for two $w_1, w_2 \sim P$ independently:

$$P\left(w_1 \cdot w_2 > \frac{d}{4} + \frac{d}{16}\right) \leq e^{-\frac{d}{128}},$$
$$P\left(w_1 \cdot w_1 < \frac{d}{2} - \frac{d}{16}\right) \leq e^{-\frac{d}{128}}.$$

Thus, picking $\mathcal{V}$ elements i.i.d according to P, randomly, of size $|\mathcal{V}| \geq e^{d/258}$ we can show by union bound that with positive probability, all $|\mathcal{V}|^2$ pairs in $\mathcal{V}$ will satisfy the requirement. Next, we define a distribution D_ε supported on subsets of $\mathcal{V}$ such that for a random variable $V \subseteq \mathcal{V}$ each $v \in V$ w.p. ε independently. We start by assuming that $T \leq \frac{1}{17}d^3$ (the case $T > d^3$ is handled at the end), and let $k \in \mathbb{N}$ be such that:

$$d \leq k\left(\frac{T}{17}\right)^{1/3} < 2d.$$

One can show that without loss of generality we can assume k is also a power of 2 (in particular, d is divisble by k for large enough T). We next follow the idea depicted in Section 4, but we want to handle the case $k \gg 1$. For that, we redefine the function in Eq. (9), and take blocks of coordinates. To simplify notations, let us define for two set of indices I, J of elements in $[d]$: $I = \{i_1, \ldots, i_k\}$, $J = \{j_1, \ldots, i_k\}$, $I \prec J$ if $\max\{i \in I\} < \min\{j \in J\}$. and we will also write:

$$e_I = \frac{1}{\sqrt{|I|}} \sum_{i \in I} e_i, \quad \text{and, } w(I) = w \cdot e_I = \frac{1}{\sqrt{I}} \sum_{i \in I} w(i).$$

then we define our final function as:

$$N(w) = \max\left\{0, \max_{|I|=k}\{-w(I)\}, \max\left\{-(w(I) - w(J)) : I \prec J, |I| = |J| = k\right\}\right\} \tag{14}$$

Define $\alpha = \min\{\frac{1}{\eta\sqrt{2T}}, 1\}$, and let:

$$f(w, V) = g(w, V) + \alpha N(w).$$

where N is defined in Eq. (14), and g is defined to be Feldman's function with a suitable choice of threshold:

$$g(w, V) = \frac{1}{\sqrt{d}} \max_{v \in V}\left\{\frac{45\eta\alpha d^2}{2 \cdot 16^2 k^{1.5}}, w \cdot v\right\}. \tag{15}$$

Notice that N is 2-Lipschitz, and g is 1-Lipschitz.

To obtain the trajectory, we next define a sample dependent oracle. We only define it for samples $\mathbf{S}$ such that there exists $v^\star \notin V_i$ for all $V_i \in S$ (define it arbitrarily to any other type of sample). Let $\mathcal{I} = \{i_1, \ldots, i_{d'}\}$ be a set of $\frac{7d}{16}$ indices such that $v^\star(i_j) \neq 0$. Divide the elements of $\mathcal{I}$ into d'/k subsets. Namely, let

$$I_j = \{i_{(j-1)\cdot k+1}, i_{(j-1)\cdot k+2}, \ldots, i_{j\cdot k}\}, \quad j = 1, \ldots, d'/k.$$

We start by defining only an oracle for the function αN. We will later show that the trajectory induced by this oracle stays in the minima of g, and that will show that, for our purposes, we can choose the

same oracle for the whole function f. We denote by $\mathcal{O}_{\alpha N}^{(t)}$ the sample dependent oracle for αN, and we first define:

$$\mathcal{O}_{\alpha N}^{(1)}(\emptyset, 0) = 0,$$

Next, we define for $t > 1$. For any w such that $0 \notin \partial N(w)$, define it arbitrarily. If $0 \in \partial N(w)$ we define it as follows:

- If there is a multi-index I_j that $w(I_j) = w(I_{j+1}) > \eta\alpha$: Then:

$$\mathcal{O}_{\alpha N}^{(t)}(S, w) = \alpha(e_{I_{j+1}} - e_{I_j}).$$

- If there is no such multi-index, and if I_j is the minimal multiindex such that $w(I_j) = 0$, set:

$$\mathcal{O}_{\alpha N}^{(t)}(S, w) = -\alpha(e_{I_j}).$$

- If both conditions cannot be met, then:

$$\mathcal{O}_{\alpha N}^{(t)}(S, w) = 0.$$

The trajectory of this dynamic is depicted in Fig. 1, for the case $k = 1$. In the general case, we can think of each coordinate in Fig. 1 as a block of size k. Now we assume that $\{w_t^{\mathbf{S}}\}_{t=1}^{T}$ follows the trajectory depicted in Eq. (12) with that choice of Oracle. It can be seen that the update step is such that after $T' = 1 + \sum_{t'=1}^{d'/k} \sum_{t''=1}^{t'} t''$ rounds, we will have that for every $i \in I_t$:

$$w_{T'}^{\mathbf{S}}(i) = \alpha\eta\sqrt{k}(d'/k + 1 - t), \tag{16}$$

and that for $t > T'$: $w_t^{\mathbf{S}} = w_{T'}^{\mathbf{S}}$

Moreover, one can show that w_t is non zero only in coordinates $i \in \mathcal{I}$, and that for any subset $I_B \subseteq I$ such that $|I| = B$:

$$\sum_{i \in I_B} w_t^{\mathbf{S}}(i) \leq \max\left\{ \sum_{i \in I_B} w_T^{\mathbf{S}}(i) : I_B \subseteq I, |I| = B \right\} \tag{17}$$

A formal proof is provided in Appendix D. Notice that:

$$1 + \sum_{t'=1}^{d'/k} \sum_{t''=1}^{t'} t'' < \frac{d'^3}{k^3} \leq \frac{1}{5}T.$$

Next, for any $v \neq v^\star$, we have that:

$$
\begin{aligned}
w_t^{\mathbf{S}} \cdot v &\le \sum_{i=1}^{d'/k} w_t^{\mathbf{S}}(i_t)\mathbf{1}\{v(i_t)=1\} \\
&\le \max\left\{\sum_{i\in I_B} w_T^{\mathbf{S}}(i) : I_B \subseteq I, |I| \le \frac{5d}{16}\right\} && Eqs.\ (13) and\ (17) \\
&\le \sum_{t=1}^{\frac{5d}{16k}} \sum_{i\in I_t} w_T^{\mathbf{S}}(i) \\
&\le \sum_{t=1}^{\frac{5d}{16k}} \sqrt{k}\eta\alpha((d'/k)+1-t) && Eq.\ (16) \\
&\le \sum_{t=0}^{\frac{5d}{16k}} \sqrt{k}\eta\alpha((d'/k)-t) \\
&\le \sqrt{k}\eta\alpha\left(\frac{5d}{16k}\frac{d'}{k} - \frac{1}{2}\left(\frac{5d}{16k}\right)^2\right) \\
&\le \sqrt{k}\eta\alpha\frac{45d^2}{2(16k)^2} && d' = \frac{7d}{16}
\end{aligned}
$$

.

As such $0 \in \partial g(w_t^{\mathbf{S}}, V_i)$ for all V_i, and we can define $\mathcal{O}^{(t)}$ so that

$$\mathcal{O}^{(t)}(S, w_t^{\mathbf{S}}, V_i) = \mathcal{O}_{\alpha N}^{(t)}(S, w_t^{\mathbf{S}}).$$

And we have that, $\{w_t^{\mathbf{S}}\}$ is the trajectory obtained from Eq. (12) with respect to this oracle of g also. We also have:

$$w_{T'}^{\mathbf{S}} \cdot v^\star = \sqrt{k}\eta\alpha\sum_{t=1}^{d'/k}(d'/k+1-t) = \sqrt{k}\eta\alpha\sum_{t=1}^{d'/k} t \ge \frac{\sqrt{k}\eta\alpha}{2}\frac{d'^2}{k^2} \ge \sqrt{k}\eta\alpha\frac{49d^2}{2(16k)^2}.$$

And because $T' \le \frac{1}{17}T$, for every suffix $\mathbf{s} \in [T]$:

$$v^\star \cdot w_{\mathbf{S},\mathbf{s}}^{GD} \ge \frac{16}{17} w_{T'}^{\mathbf{S}} \cdot v^\star \ge \sqrt{k}\eta\alpha\frac{46d^2}{2(16k)^2}.$$

Recall that we assume $\frac{136d^3}{k^3} \ge T$:

$$F(W_{\mathbf{S},\mathbf{s}}^{GD}) - F_S(0) \ge \varepsilon\left(\frac{46\sqrt{k}\alpha\eta d^{1.5}}{2(16k)^2} - \frac{45\sqrt{k}\alpha\eta d^{1.5}}{2(16k)^2}\right) \ge \varepsilon\frac{\alpha\eta}{2\cdot 16^2}\left(\frac{d}{k}\right)^{1.5} \ge \frac{\varepsilon}{\sqrt{2}\cdot 272\cdot 16^2}\min\left\{\eta\sqrt{T}, 1\right\} \tag{18}$$

Eq. (18) lower bounds the generalization under the event that there exists $v \in \mathcal{V}$ such that $v \notin V_i$ for every $i = 1, \ldots, m$. Now assume $\varepsilon = \min\{\frac{d}{516m}, \frac{1}{4}\}$, then for every v, using the inequality $(1-\varepsilon) \le e^{-2\varepsilon}$ for $\varepsilon < 1/2$:

$$P(v \notin \cap_{i=1}^m V_i) = (1-\varepsilon)^m \ge e^{-2\varepsilon\cdot m} \ge e^{-2d/516},$$

and,

$$P(\exists v, \notin \cap_{i=1}^m V_i) \ge 1 - (1-e^{-2d/516})^{|V|} \ge 1 - (1-e^{-d/258})^{e^{d/258}} \ge 1 - e^{-1} \ge 1/2.$$

So far we assume that $d = 2^n$, notice that if $T < \frac{1}{8\cdot 17}d^3$, we can find a subspace of size $d_1 > \frac{d}{2}$ so that $T < \frac{1}{8\cdot 17}d_1^3$, and we obtain our final result by embedding our construction in this subspace.

When $T > \frac{1}{17\cdot 8}d^3$ Notice that if we take the construction with $T = \lfloor \frac{1}{17\cdot 8}d^3 \rfloor$ then:

$$\mathcal{O}^{(T)}(w_T^{\mathbf{S}}) = 0,$$

hence we can use the above construction for any $T' > T$, and the subdifferential oracle is defined for every iteration above T as returning 0, and we obtain a similar analysis to the case that $T = \lfloor d^3/136 \rfloor$, and our bounds yield as in Eq. (18):

$$F(w_{\mathbf{S},\mathbf{s}}^{GD}) - F_S(0) \geq \frac{\varepsilon}{\sqrt{2}\cdot 272\cdot 16^2}\min\left\{\eta\sqrt{\lfloor d^3/136\rfloor}, 1\right\}$$

A.2 Proof of Lemma 9

The main ingredient in our proof is the following claim whose proof is provided in Appendix C. To state the claim we will need a few notations. First, for any two tuples of sample-sequences in $\mathcal{S}_m^T$, and $t > 0$ $(\mathbf{S}, t)$ and $(\mathbf{S}', t)$, let us denote $(\mathbf{S}, t) \equiv (\mathbf{S}', t')$ if $t = t'$ and $S_{1:t} = S'_{1:t}$, namely the prefixes of the sample sequences agree up to point $t' = t$. Second, for a mapping $\alpha : \mathcal{Z} \to [0, 1]$, from sample points to real numbers, and a sample S, let us denote

$$\alpha(S) = \frac{1}{|S|}\sum_{z\in S}\alpha(z).$$

Claim 1. *For every η, γ and α define for every sequence of samples $\mathbf{S} = (S_1, \ldots, S_T)$, and t inductively: $x_0^{\mathbf{S}} = 0$ and*

$$x_t^{\mathbf{S}} = (1 - \gamma\eta)x_{t-1}^{\mathbf{S}} - \gamma\eta\alpha(S_t). \tag{19}$$

and define:

$$x_q^{\mathbf{S}} = \sum_{t=1} q(t)x_t^{\mathbf{S}} \tag{20}$$

Then, for any $\varepsilon > 0$, there is a choice of $\gamma < \varepsilon$ and α such that:

1. *For $(\mathbf{S}, t) \not\equiv (\mathbf{S}', t')$: $x_t^{\mathbf{S}} \neq x_{t'}^{\mathbf{S}'}$.*
2. *For every $\mathbf{S}, t$, such that $t < \max\{t' : q(t') \neq 0\}$ and $\mathbf{S}'$: $x_t^{\mathbf{S}} \neq \sum_{t=1}^T q(t)x_t^{\mathbf{S}'}$.*

The $x_t^{\mathbf{S}}$ represents the different states the trajectory can be in. $x_q^{\mathbf{S}}$ represent the output of the trajectory which can be an aggreagated sum. We require that each provide a signature for the state, and this will allow us to "code" the state of the trajectory along GD.

We now continue with the proof of Lemma 8. For every sample dependent first order oracle $\mathcal{O}^{(t)}$, we have that:

$$\mathcal{O}^{(t)}(\mathbf{S}; w) = \frac{1}{|S_t|}\sum_{z\in S_t}\mathcal{O}^{(t)}(S_{1:t-1}; w, z). \tag{21}$$

To simplify notations, let us denote:

$$\mathcal{O}^{(t)}(S_{1:t-1}, w_t^{\mathbf{S}}, z) = \mathcal{O}_{\mathbf{S},t,z},$$

as $\mathbf{S}, t, z$ completely determine the output. We next define

$$h_z(x) := \frac{1}{2}\gamma(x^2 - 2\alpha(z)x),$$

where $\gamma > 0$ is arbitrarily small[4].

and observe that, if $x_t^{\mathbf{S}}$ is defined as in Eq. (19):

$$x_t^{\mathbf{S}} = (1 - \gamma\eta)x_{t-1}^{\mathbf{S}} - \gamma\eta\alpha(S_t) = x_{t-1}^{\mathbf{S}} - \frac{\eta}{|S_t|}\sum_{z\in S_t}\nabla h_z(x_{t-1}^{\mathbf{S}}). \tag{22}$$

[4] $\gamma \leq \varepsilon/(\eta T)$, will suffice

To simplify notation, let us denote:

$$w^{\mathbf{S}}_{T+1} = w^{\mathbf{S}}_q, \text{ and } x^{\mathbf{S}}_{T+1} = x^{\mathbf{S}}_q,$$

and assume without loss of generality that $\max\{q(i) \neq 0\} = T$ (Otherwise, we look only at the sequence up to point $\max\{q(i) \neq 0\}$). Consider now the sets of triplets:

$$G(z) = \left\{(v, g, u) = \left(f(w^{\mathbf{S}}_t, z) + h_z(x^{\mathbf{S}}_t), (\mathcal{O}_{\mathbf{S},t,z}, \nabla h_z(x^{\mathbf{S}}_t)), (w^{\mathbf{S}}_t, x^{\mathbf{S}}_t)\right) : \mathbf{S}^m_T \in \mathcal{S}, t \leq T+1\right\},$$

where $\mathcal{O}_{\mathbf{S},T+1,z} \in \partial(f(w^{\mathbf{S}}_q) + h_z(x^{\mathbf{S}}_q))$ is chosen arbitrarily.

Convexity of $f + h_z$ ensure that the triplets in $G(z)$ satisfy Eq. (10) for all $t \leq T+1$, as in Lemma 7. To apply the Lemma, we also want to achieve differentiability at points such that $t < T$. Therefore, take any two triplets

$$(v_i, g_i, u_i) = \left(f(w^{\mathbf{S}}_{t_i}, z) + h_z(x^{\mathbf{S}}_{t_i}), \left(\mathcal{O}_{\mathbf{S},t,z}, \nabla h_z(x^{\mathbf{S}}_{t_i})\right), \left(w^{\mathbf{S}}_{t_i}, x^{\mathbf{S}}_{t_i}\right)\right), i = 1, 2,$$

where $t_1 < T$ and $t_2 \leq T+1$, and suppose $g_1 \neq g_2$. To simplify notations, let us write $w^{\mathbf{S}}_{t_i} = w_i$ and $x^{\mathbf{S}}_{t_i} = x_i$.

First, by convexity of f we have that:

$$\begin{aligned}
v_1 - v_2 + g_2^\top(u_2 - u_1) =& f(w_1, z) + h_z(x_1) - f(w_2, z) - h_z(x_2) - \left(\mathcal{O}_{\mathbf{S}_2,t_2,z}, \nabla h_z(x_2)\right)^\top ((w_1, x_1) - (w_2, x_2)) \\
=& f(w_1, z) - f(w_2, z) - \mathcal{O}^\top_{\mathbf{S}_2,t_2,z}(w_1 - w_2) + h_z(x_1) - h_z(x_2) - \nabla h_z(x_2)^\top (x_1 - x_2) \\
\geq& h_z(x_1) - h_z(x_2) - \nabla h_z(x_2)^\top (x_1 - x_2).
\end{aligned}$$

Next, because $g_1 \neq g_2$, either $\nabla h_z(x_1) \neq \nabla h_z(x_2)$, which implies $x_1 \neq x_2$, or $\mathcal{O}_{\mathbf{S}_1,t_1,z} \neq \mathcal{O}_{\mathbf{S}_2,t_2,z}$ which implies $(\mathbf{S}_1, t_1) \not\equiv (\mathbf{S}_2, t_2)$ which again implies $x_1 \neq x_2$ by Claim 1. In other words, if $g_1 \neq g_2$ then $x_1 \neq x_2$:

$$\begin{aligned}
h_z(x_1) - h_z(x_2) - \nabla h_z(x_2)^\top (x_1 - x_2) &= \gamma\left(x_1^2 - 2\alpha(z)x_1 - x_2^2 - 2\alpha(z)x_2 - (2x_1 - 2\alpha(z)) \cdot (x_1 - x_2)\right) \\
&= \gamma\left(x_1^2 + x_2^2 - 2x_1 \cdot x_2\right) \\
&= \gamma(x_1 - x_2)^2 \\
&> 0 \qquad\qquad x_1 \neq x_2
\end{aligned}$$

We showed then, that $v_1 - v_2 + g_2^\top(u_2 - u_1) = 0$, implies $g_1 = g_2$. We obtain, by Lemma 7, that there exists a function $\bar{f}((w, x), z)$ such that for all t and $\mathbf{S}$:

$$\bar{f}(w^{\mathbf{S}}_t, x^{\mathbf{S}}_t, z) = f(w^{\mathbf{S}}_t, z) + h_z(x^{\mathbf{S}}_t) \tag{23}$$

and for all $t \leq T$,

$$\nabla\bar{f}((w^{\mathbf{S}}_t, x^{\mathbf{S}}_t)) = (\mathcal{O}_{\mathbf{S},t,z}, \nabla h_z(x^{\mathbf{S}}_t)). \tag{24}$$

This proves Lemma 9. Indeed. By the Lipschitzness of h in the unit ball, we have that $|x_t| \leq \gamma\eta T$. For sufficiently small γ, from Eq. (23), since $\{(f(0, z) + h_z(0), (0, \nabla h_z(0)), (0, 0))\} \in G(z)$, we obtain that

$$|\bar{f}((0, 0), z) - f(0, z)| = \gamma|h_z(0)| \leq \gamma^2\eta T \leq \varepsilon.$$

$$|\bar{f}((w^{\mathbf{S}}_q, x^{\mathbf{S}}_q), z) - f((w^{\mathbf{S}}_q, z))| = \gamma|h_z(x^{\mathbf{S}}_q)| \leq \varepsilon.$$

Further, if we assume by induction that for every $t' \leq t-1$, $u_{t'} = w^{\mathbf{S}}_{t'}$, and $x_{t'} = x^{\mathbf{S}}_{t'}$, then from Eq. (24) we have that for any first order oracle:

$$\begin{aligned}
(u_t, x_t) &= (w^{\mathbf{S}}_{t-1}, x^{\mathbf{S}}_{t-1}) - \frac{\eta}{|S_t|}\sum_{z\in S_t} \mathcal{O}_z(w^{\mathbf{S}}_{t-1}, x^{\mathbf{S}}_{t-1}) \\
&= (w^{\mathbf{S}}_{t-1}, x^{\mathbf{S}}_{t-1}) - \frac{\eta}{|S_t|}\sum_{z\in S_t} \left(\mathcal{O}_{\mathbf{S},t,z}, \nabla h_z(x^{\mathbf{S}}_{t-1}, z)\right) \\
&= (w^{\mathbf{S}}_{t-1}, x^{\mathbf{S}}_{t-1}) - \frac{\eta}{|S_t|}\sum_{z\in S_t} \left(\mathcal{O}^{(t)}(S_{t-1}, w^{\mathbf{S}}_{t-1}, z), \nabla h_z(x^{\mathbf{S}}_{t-1}, z)\right) \\
&= (w^{\mathbf{S}}_{t-1}, x^{\mathbf{S}}_{t-1}) - \eta\left(\mathcal{O}(\mathbf{S}, w^{\mathbf{S}}_{t-1}), \frac{1}{|S_t|}\sum_{t=1}^{T} \nabla h_z(x^{\mathbf{S}}_{t-1}, z)\right) \\
&= \left(w^{\mathbf{S}}_{t-1} - \eta O(\mathbf{S}, w^{\mathbf{S}}_{t-1}), x^{\mathbf{S}}_{t-1} - \frac{\eta}{|S_t|}\sum \nabla h_z(x^{\mathbf{S}}_{t-1})\right) \\
&= (w^{\mathbf{S}}_t, x^{\mathbf{S}}_t), \qquad Eq.\ (22)
\end{aligned}$$

which proves, by linearity, that $u_q = w^{\mathbf{S}}_q$.

B Proof of Lemma 7

We choose

$$\hat{f}(w) = \max_{j\in J}\{f_j + g_j^\top(w - w_j\}.$$

$\hat{f}$ is indeed convex as it is the maximum of linear functions. Further, it is known [19] that at each point w:

$$\partial\hat{f}(w) = \text{conv}\{g_j : \hat{f}(w) = f_j + g_j^\top(w - w_j)\}. \tag{25}$$

It follows that, $\hat{f}$ is L-Lipschitz. Next, for any w_i, notice that our assumption implies:

$$f_i \geq \max_{j\in J}\{f_j + g_j^\top(w_i - w_j\} = \hat{f}(w_i).$$

On the other hand,

$$f_i = f_i + g_i^\top(w_i - w_i) \leq \hat{f}(w_i).$$

Hence $\hat{f}(w_i) = f_i$. Finally, to see the function is differentiable at designated points, take any $i \in I_{diff}$ and consider w_i. By Eq. (25), it is enough to show that if $g_j \in \partial f(w_i)$ then $g_i = g_j$, but this clearly follows from our assumption, and the fact that $\hat{f}(w_i) = f_i$.

C Proof of Claim 1

We first prove by induction that $x^{\mathbf{S}}_0 = 0$ and, for $t \geq 1$:

$$x^{\mathbf{S}}_t = \gamma\eta\sum_{z\in Z}\alpha(z)\left(\sum_{\{t'\leq t, z\in S_{t'}\}}\frac{(1-\gamma\eta)^{t-t'}}{|S_{t'}|}\right). \tag{26}$$

Indeed,

$$\begin{aligned}
x_t^{\mathbf{S}} &= x_{t-1}^{\mathbf{S}} - \gamma\eta\left(\eta x_{t-1}^{\mathbf{S}} - \alpha(\mathbf{S}_t)\right) \\
&= (1-\gamma\eta)x_{t-1}^{\mathbf{S}} + \gamma\eta\alpha(\mathbf{S}_t) \\
&= (1-\gamma\eta)\left(\sum_{z\in Z}\alpha(z)\sum_{\{t'<t:z\in S_{t'}\}}\frac{(1-\gamma\eta)^{t-1-t'}}{|S_{t'}|}\eta\gamma\right) + \frac{1}{|S_t|}\sum_{z\in S_t}\gamma\eta\alpha(z) \\
&= \sum_{z\in Z}\alpha(z)\sum_{\{t'<t:z\in S_{t'}\}}\frac{(1-\gamma\eta)^{t-t'}}{|S_{t'}|}\eta\gamma + \frac{\mathbf{1}[z\in\mathbf{S}_t]}{|S_t|}\gamma\eta \\
&= \eta\gamma\sum_{z\in Z}\alpha(z)\sum_{\{t'\le t:z\in S_{t'}\}}\frac{(1-\gamma\eta)^{t-t'}}{|S_{t'}|}
\end{aligned}$$

Now, for every $\mathbf{S}, t, z$, define a polynomial:

$$P_{\mathbf{S},t,z}(X) = \sum_{n=0}^{t-1}\frac{\mathbf{1}[z\in S_{t-n}]}{|S_{t-n}|}X^n,$$

and let r be a rational point, sufficinetly small, so that $1-r$ is *not* the root of any polynomial of the form $P_{\mathbf{S},t,z} - P_{\mathbf{S}',t',z'}$ that is distinct from 0. In other words, we choose r so that

$$P_{S,t,z}(1-r) = P_{S',t',z'}(1-r) \Leftrightarrow P_{S,t,z}(X) = P_{S',t',z'}(X).$$

and we also require that

$$\sum_{t=1}^{T} q(t)P_{S,t,z}(1-r) = P_{S',t',z}(1-r) \Leftrightarrow \sum_{t=1}^{T} q(t)P_{S,t,z}(X) = P_{S',t',z'}(X).$$

Notice that there are only finitely many polynomials of the above form, hence we can choose such r in any interval $(0,\varepsilon)$ for any $\varepsilon > 0$. Rewriting Eq. (26) we have:

$$x_t^{\mathbf{S}} = \gamma\eta\sum_{z\in Z}\alpha(z)P_{\mathbf{S},t,z}\,(1-\gamma\eta)$$

Now, suppose we choose $\{\alpha(z)\}$ to be reals, independent over the rationals, and suppose we choose $\gamma = r/\eta$.

Proof of Item 1 Assume that $x_t^{\mathbf{S}} = x_{t'}^{\mathbf{S}'}$. Because $\alpha(z)$ are independent over the rationals, and because $P(1-\gamma\eta)$ are always rationals, we have that $P_{\mathbf{S}',t',z}(1-r) = P_{\mathbf{S}',t',z}(1-r)$ for every z. But then,

$$\forall z \in Z : P_{\mathbf{S},t,z}(X) = P_{\mathbf{S}',t',z}(X),$$

by choice of r. But then, $t = t'$. Indeed, assume by contradiction and w.l.og assume $t < t'$: then for any $z \in S_1'$ $P_{\mathbf{S},t,z}$ is a $t'-1$-degree polynomial, on the other hand $P_{\mathbf{S},t,z}$ is at most of degree $t-1 < t'-1$. Next, if $S_i \neq S_i'$, for $i \le t$, then we can assume (w.l.o.g) that there is $z \in S_i$ such that $z \notin S_i'$ it follows that, by looking at the coefficient of the two polymials of the monomial X^{t-i} we have that:

$$P_{\mathbf{S}',t',z}(X) \neq P_{\mathbf{S},t,z}(X).$$

Overall then, we obtain that if $x_t^{\mathbf{S}} = x_{t'}^{\mathbf{S}'}$ then $(\mathbf{S},t) \equiv (\mathbf{S}',t')$..

Proof of Item 2: The proof of Item 2 is similar to how we proved $t = t'$. Notice that:

$$\sum_{t=1}^{T} q(t)x_t^{\mathbf{S}} = \gamma\eta\sum_{z\in Z}\alpha(z)\sum_{t=1}^{T} q(t)P_{\mathbf{S},t,z}(1-\eta\gamma).$$

Hence $x_t^{\mathbf{S}} = x^{\mathbf{S}'}$ implies for all $z \in Z$:

$$\sum_{t=1}^{T} q(t) P_{\mathbf{S}',t,z} = P_{\mathbf{S},t,z}.$$

For $z \in S_1'$ we have that $\sum_{t=1}^{T} q(t) P_{\mathbf{S},t,z}$ is a $\max\{t : q(t) \neq 0\}$-degree polynomial, but on the other hand we assume that $t < \max\{t : q(t) \neq 0\}$, hence $P_{\mathbf{S},t,z}$ is a lower degree polynomial.

D Proof of Eqs. (16) and (17)

D.1 Proof of Eq. (16)

To avoid cumbersome notations, we will supress dependence on $\mathbf{S}$ and write w_t instead of $w_t^{\mathbf{S}}$.

As a first step, observe that at each iteration no projection is performed. Indeed, let us show by induction that:

$$\|w_{t+1}\|^2 = \|w_t - \eta \mathcal{O}^{(t)}(S, w, V)\|^2 \leq 2\eta^2\alpha^2(t+1).$$

The definition of α then implies that w_t are restricted to the unit ball without projections.

To see the above is true, let us consider the case where the first type of update is performed:

$$\begin{aligned}
\|w_{t+1}\|^2 &= \|w_t^{\mathbf{S}} - \eta \mathcal{O}^{(t)}(S, w, V)\|^2 \\
&= \|w_t^{\mathbf{S}} - \eta\alpha e_{I_{j+1}} + \eta\alpha e_{I_j}\|^2 \\
&= \sum_{s=1}^{j-1} (w_t(I_s))^2 + (w_t(I_j) + \eta\alpha)^2 + (w_t(I_{j+1}) - \eta\alpha)^2 + \sum_{s=j+2}^{\lfloor d'/k \rfloor} (w_t(I_s))^2 \\
&= \sum_{s=1}^{\lfloor d'/k \rfloor} (w_t(I_s))^2 + 2\eta\alpha(w_t(I_j) - w_t(I_{j+1})) + 2\eta^2\alpha^2 \\
&= \sum_{s=1}^{\lfloor d'/k \rfloor} (w_t(I_s))^2 + 2\eta^2\alpha^2 && w_t(I_{j+1}) = w_t(I_j) \\
&\leq 2\eta^2\alpha^2 \cdot t + \eta^2\alpha^2 \\
&= 2\eta^2\alpha^2 \cdot (t+1).
\end{aligned}$$

And if the second type of update is performed:

$$\begin{aligned}
\|w_{t+1}\|^2 &= \|w_t^{\mathbf{S}} - \eta \mathcal{O}^{(t)}(S, w, V)\|^2 \\
&= \|w_t^{\mathbf{S}} + \eta\alpha e_{I_j}\|^2 \\
&= \sum_{s \neq j} w_t(I_s) + \eta^2\alpha^2 && w_t(I_j) = 0 \\
&\leq 2\eta^2\alpha^2 \cdot t + \eta^2\alpha^2 \\
&\leq 2\eta^2\alpha^2 \cdot (t+1)
\end{aligned}$$

We now move on to prove that Eq. (16) holds by induction. Specifically, we will show that for $d_0 \leq \lfloor d'/k \rfloor$ that at time $T_{d_0} = 1 + \sum_{d=0}^{d_0} \sum_{k=0}^{d} k$, for any $t \leq d_0$:

$$w_{T_{d_0}}(I_t) = \alpha\eta \begin{cases} (d_0 + 1 - t) & t \leq d_0 \\ 0 & \text{o.w.} \end{cases}. \tag{27}$$

Eq. (16) then follows by plugging $d_0 = d'/k$, and noting that at every step we have that if $i \in I_t$ then $w_{T_{d_0}}(i) = \sqrt{k} w_{T_{d_0}}(I_t)$. Therefore, we are left with proving Eq. (27).

For $d_0 = 0$, $T_0 = 1$ and we have, indeed, $w_1 = 0$. Next assume we proved the statement for d_0, and we will prove it for $d_0 + 1$. Here too, we will use induction, and we prove that, for $d_1 \leq d_0 + 1$, at time

$$T_{d_0,d_1} = T_{d_0} + \sum_{k=0}^{d_1 - 1} (d_0 + 1 - k),$$

we have that:

$$w_{T_{d_0,d_1}}(I_t) = \alpha\eta \begin{cases} (d_0+2-t) & t \le d_1 \\ (d_0+1-t) & d_1 < t \le d_0 \\ 0 & \text{o.w.} \end{cases}$$

For the case d_1=0, $T_{d_0,d_1} = d_0$, and it follows from our outer-induction step. Assume the statement is true for d_1 and we will prove it for d_1+1, here, yet again, we use induction. And we will show that for $1 \le d_2 < d_0 + 2 - d_1$ we have at time

$$T_{d_0,d_1,d_2} = T_{d_0} + T_{d_1} + d_2,$$

we have that:

$$w_{T_{d_0,d_1,d_2}}(I_t) = \alpha\eta \begin{cases} (d_0+2-t) & t \le d_1 \\ (d_0+1-t) & d_1 < t < d_0+2-d_2 \\ (d_0+2-t) & t = d_0+2-d_2 \\ (d_0+1-t) & d_0+2-d_2 < t \le d_0 \\ 0 & \text{o.w.} \end{cases} \tag{28}$$

We start the induction with the case $d_2 = 1$, in that case notice that $T_{d_0,d_1,d_2} = T_{d_0,d_1} + 1$, and by induction hypothesis:

$$w_{T_{d_0,d_1}}(I_t) = \alpha\eta \begin{cases} (d_0+2-t) & t \le d_1 \\ (d_0+1-t) & d_1 < t \le d_0 \\ 0 & \text{o.w.} \end{cases}$$

In this case, note that there are no two consecutive coordinates that are equal, hence our choice of oracle is defined so that $\mathcal{O}^{(t)} = -e_{I_{d_0+1}}$. Hence, by our update rule (and the lack of projections which we proved at the beginning):

$$w_{T_{d_0,d_1,1}}(I_t) = \alpha\eta \begin{cases} (d_0+2-t) & t \le d_1 \\ (d_0+1-t) & d_0 < t \le d_0 \\ 1 & t = d_0+1 \\ 0 & \text{o.w.} \end{cases}$$

Which satisfies Eq. (28). Now assume that Eq. (28) holds for d_2, and take $d_2 + 1 < d_0 + 2 - d_1$ (otherwise, we are done). Notice that $T_{d_0,d_1,d_2+1} = T_{d_0,d_1,d_2} + 1$. Observe that $w_{T_0,d_1,d_2}(d_0 + 2 - d_2) = w_{T_0,d_1,d_2}(d_0 + 1 - d_2)$ (notice that $d_0 + 1 - d_2 > d_1$), and our update rule is such that $\mathcal{O}^{(t)} = e_{I_{d_0+2-d_2}} - e_{I_{d_0+1-d_2}}$ and we obtain then:

$$w_{T_{d_0,d_1,d_2+1}} = w_{T_{d_0,d_1,d_2}} - \eta\alpha e_{I_{d_0+2-d_2}} + \eta\alpha e_{I_{d_0+1-d_2}} = \alpha\eta \begin{cases} (d_0+2-t) & t \le d_1 \\ (d_0+1-t) & d_1 < t < d_0+2-(d_2+1) \\ (d_0+2-t) & t = d_0+2-(d_2+1) \\ (d_0+1-t) & d_0+2-(d_2+1) < t \le d_0 \\ 0 & \text{o.w.} \end{cases}$$

The most inner induction step is now complete. We now notice that $T_{d_0,d_1,d_0+1-d_1} = T_{d_0,d_1+1}$, which proves the middle-induction step. And we notice that $T_{d_0,d_0+1} = T_{d_0+1}$, which proves the whole induction argument.

D.2 Proof of Eq. (17)

We only need to show that the following quantity is increasing in t

$$X_t = \max\left\{ \sum_{i \in I_B} w_t^{\mathrm{S}}(i) : I_B \subseteq I, |I| = B \right\}.$$

But, as shown in Appendix D, the update rule is such that we don't perform projections. It then follows easily from our update step. Indeed if we increase a set of coordinates by $\alpha\eta$, then clearly the X_t only increases. Also, if we perform update of the form:

$$w_t^{\mathbf{S}}(I_j) = w_{t-1}^{\mathbf{S}}(I_j) + \alpha\eta, \quad w_t^{\mathbf{S}}(I_{j+1}) = w_t^{\mathbf{S}}(I_{j+1}) - \alpha\eta,$$

for two consecutive and equal coordinates, then for any I_B: If I_B includes same number of coordinates from I_j as in I_{j+1} then the magnitude doesn't change. If I_B contains more I_j then it increases, and if I_B contains more from i_{j+1} then consider $I_{B'}$ that swaps coordinates in I_j with I_{j+1} then we clearly have:

$$\sum_{i\in I_{B'}} w_t^{\mathbf{S}}(i) > \sum_{i\in I_{B'}} w_{t-1}^{\mathbf{S}}(i) > \sum_{i\in I_B} w_t^{\mathbf{S}}(i)$$

E Dimension independent lower bound for GD

In this section we prove that the optimization error of GD in Eq. (5) is optimal. The lower bound is an optimization error for first order methods and is well established (see [10]). The point here is to show that the bound is valid in any dimension, for GD.

Claim 2. *For every choice of η, T, there exists a convex and 1 Lipschitz function $f(x) : \mathbb{R} \to \mathbb{R}$ such that, if we run GD on f, then*

$$f(w^{GD}) - f(0) \geq \frac{\eta}{2} + \frac{1}{6\eta T}.$$

Proof. We divide the proof into two cases:

case1: $\eta \geq 1/\eta T$:

In this case we choose $f(x) = |x - \gamma|$, where $\gamma > 0$ is a arbitrarily small (may depend on η). It can be seen that for every even iteration, we have that:

$$\nabla f(x_{2t}) = -1,$$

at at every odd iteration

$$\nabla f(x_{2t+1}) = +1.$$

As such, for every even iteration we have that $x_{2t} = 0$, and at every odd iteration we have that $x_{2t+1} = \eta$. We thus have that $x^{GD} = \frac{\eta}{2}$ and

$$f(x^{GD}) - 0 = \frac{\eta}{2} - 0 \geq \frac{\eta}{4} + \frac{1}{2\eta T}.$$

case1: $\eta \leq 1/\eta T \leq 1$: In this case choose $f(x) = \alpha x$, where $\alpha = \frac{1}{(T+1)\eta} \leq 1$. Then, one can show that GD outputs

$$x^{GD} = -\frac{\eta}{T}\sum t \cdot \alpha = -\frac{(T+1)\eta}{2}\alpha,$$

and:

$$f(x^{GD}) - f(-1) = \alpha - \frac{(T+1)\eta}{2}\alpha^2 \geq \frac{1}{2(T+1)\eta} \geq \frac{1}{3T\eta} \geq \frac{1}{6T\eta} + \frac{\eta}{2}.$$

■

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.