# SGD Generalizes Better Than GD (And Regularization Doesn't Help)

**Idan Amir** IDANAMIR@MAIL.TAU.AC.IL
*Department of Electrical Engineering, Tel Aviv University*
**Tomer Koren** TKOREN@TAUEX.TAU.AC.IL
*Blavatnik School of Computer Science, Tel Aviv University and Google Research, Tel Aviv*
**Roi Livni** RLIVNI@TAUEX.TAU.AC.IL
*Department of Electrical Engineering, Tel Aviv University*

**Editors:** Mikhail Belkin and Samory Kpotufe

## Abstract

We give a new separation result between the generalization performance of stochastic gradient descent (SGD) and of full-batch gradient descent (GD) in the fundamental stochastic convex optimization model. While for SGD it is well-known that $O(1/\varepsilon^2)$ iterations suffice for obtaining a solution with $\varepsilon$ excess expected risk, we show that with the same number of steps GD may overfit and emit a solution with $\Omega(1)$ generalization error. Moreover, we show that in fact $\Omega(1/\varepsilon^4)$ iterations are necessary for GD to match the generalization performance of SGD, which is also tight due to recent work by Bassily et al. (2020). We further discuss how regularizing the empirical risk minimized by GD essentially does not change the above result, and revisit the concepts of stability, regularization, implicit bias and the role of the learning algorithm in generalization.

## 1. Introduction

The setting of Stochastic Convex Optimization (SCO) assumes a learner that observes a finite sample of convex functions drawn i.i.d. from some unknown distribution and in turn has to provide a parameter that minimizes the expected function with respect to the true and unknown distribution. This is a very simple and clean setting that has received considerable attention in the last two decades which culminated in remarkable bounds for both the statistical sample complexity as well as the optimization complexity.

The two most common and well-known optimization methods in SCO are *Gradient Descent* (GD) and *Stochastic Gradient Descent* (SGD). In the first method, one optimizes the empirical risk over a sample by computing iteratively the full-batch gradient; the second method is a "lighter" version that samples at each iteration a fresh new example that is used to form an unbiased estimate of the gradient. Perhaps surprisingly, even though SGD may seem like a noisy, inaccurate version of GD, it is well known (e.g., Hazan (2019)) that the former enjoys an optimal rate and converges after $O(1/\varepsilon^2)$ iterations to an $\varepsilon$-optimal solution with respect to the *true* underlying distribution, independently of the dimension of the problem. This in turn makes it highly suitable for large-scale optimization where the computational costs of the iterations should be taken into account (Bottou and Bousquet, 2011).

Even more surprisingly, while SGD is relatively well understood in terms of its sample complexity bounds, our understanding of full-batch GD is still lacking; in fact, it has remained an open question whether GD can obtain the same dimension-independent guarantees attained by SGD. While this question has been studied under various assumptions such as smoothness and strong convexity (Hardt et al., 2016; Bousquet and Elisseeff, 2002), in its general case it has remained largely unresolved.

**Our contributions.** In this work, we give a new separation result between the generalization performance of SGD and of full-batch GD in the context of SCO. We show that if one runs GD for $O(1/\varepsilon^2)$ iterations (with any fixed learning rate) the algorithm may overfit and exhibit a constant gap between empirical error and true error, and in fact, no less than $\Omega(1/\varepsilon^4)$ iterations are necessary for it to generalize to within $\varepsilon$. Interestingly, this last bound turns out to be tight and matches the upper bound implied by a recent stability analysis of GD due to Bassily et al. (2020). In contrast, as discussed above, SGD (with a suitable step size) generalizes after merely $O(1/\varepsilon^2)$ steps. Thus, SGD is not merely a "light" noisy version of GD—it is in fact a superior algorithm that enjoys improved generalization guarantees. To the best of our knowledge, this result is the first to provide such a quantitative separation in generalization performance between these two natural algorithms.

We then proceed to study the role of regularization in optimization. Regularization is known to be a key aspect in SCO: in particular, Shalev-Shwartz et al. (2009) demonstrated that, while an empirical risk minimizer (ERM) might overfit, regularized-ERMs do not. As such, it is natural to ask whether adding regularization to the optimization algorithm improves its performance. We show that applying GD to a regularized empirical risk (and choosing the natural learning rates for this setting) would in general require the learner to achieve empirical error $O(\varepsilon^2)$ in order to enjoy generalization error $\varepsilon$. Overall, then, order of $\Omega(1/\varepsilon^4)$ iterations are still necessary for full-batch GD even with added regularization.

Finally, as we further discuss in Section 5 below, our construction allows us to revisit and draw new insights on some of the existing notions and tools in theoretical machine learning such as regularization, stability, implicit bias, and their role towards generalization within the framework of SCO.

**Our techniques.** The technical heart of our work is a new generalization lower bound for GD that builds upon the two works of Shalev-Shwartz et al. (2009) and Bassily et al. (2020). The work of Shalev-Shwartz et al. (2009) was the first to demonstrate that in SCO an empirical risk minimizer may fail to learn. More formally, they showed that even though a regularized ERM can learn with dimension-independent sample complexity, $\Omega(\log d)$ examples are necessary so that any ERM will not overfit; This was later improved by Feldman (2016) to $\Omega(d)$. However, there is still a gap between showing that an abstract ERM can potentially fail and analyzing the performance of concrete algorithm such as GD. In detail, the result of Feldman (2016) shows a learning problem where there are some "bad" solutions but in contrast there are simpler and easier to find "good" solutions and it is not expected that a reasonable algorithm will overfit in that problem (in fact, in its most naive form the initialization at zero is an optimal point). The work of Shalev-Shwartz et al. (2009) demonstrates an example where there is *a unique* minimum. Hence, *any* empirical risk minimizer will fail to learn. However, even this result is limited and cannot be used to rule out the performance of GD. Indeed while the minimum is bad and unique, there still are many approximately good solutions and only at a very high level of accuracy the algorithm starts to fail. In fact, only at an exponentially small training accuracy we obtain a guarantee of overfitting. As such, a reasonable algorithm such as, say, *gradient descent* reaching to $O(1/\sqrt{n})$ optimization accuracy (which is the generalization error to begin with) will not fail. In fact, it will not fail as long as we run it for less than $2^{O(n)}$ iterations.

The second work we rely on is the work of Bassily et al. (2020) that demonstrated that GD may be an unstable algorithm (Bousquet and Elisseeff, 2002)—a necessary condition for an algorithm to overfit. Utilizing these two constructions we construct a new example where GD is unstable

and converges to one of the "bad" ERM solutions. We point out that mere instability and lack of uniform convergence are not sufficient for an algorithm to overfit. Indeed, Bassily et al. (2020) also demonstrated that SGD is unstable (on the same example on which GD is shown to be unstable), but at the same time, SGD comes with provable guarantees and does not overfit. Therefore, constructing such an example, even though utilizes previous constructions, does not follow some generic reduction.

## 2. Problem Setup and Background

We consider the standard setting of stochastic convex optimization. A learning problem consists of a fixed domain $\mathcal{W}_d \subseteq \mathbb{R}^d$ in $d$-dimensional Euclidean space, and a loss function $f : \mathcal{W}_d \times \mathcal{Z} \to \mathbb{R}$, parameterized by a parameter space $\mathcal{Z}$, where for each fixed $z \in \mathcal{Z}$ we assume the function $f(w; z)$ as a function of $w$ is $L$-Lipschitz and convex. We will treat throughout $L$ as a constant; in particular, in all our constructions $L$ will be fixed, and will not depend on other parameters of the problem (specifically, $d, \eta, T$ and $n$, as discussed next). We will normally choose $\mathcal{W}_d$ to be the unit-ball in $\mathbb{R}^d$. If the dimension is fixed, and there is no room for confusion we will also suppress dependence on $d$ and write

$$\mathcal{W} = \{w : \|w\| \leq 1 : w \in \mathbb{R}^d\}.$$

In this setting, a learner is provided with a sample $S = z_1, \ldots, z_n$ of i.i.d. examples drawn from an unknown distribution $D$ and needs to optimize the *true risk* (or *expected risk*, or *true loss*) which we define:

$$F(w) = \mathop{\mathbb{E}}_{z \sim D}[f(w; z)]. \tag{1}$$

More formally, given the sample $S$ the learner should return, in expectation, a parameter $w_S$ with $\varepsilon$-optimal true loss. Namely,

$$\mathop{\mathbb{E}}_{S \sim D^n}[F(w_S)] \leq \min_{w^\star \in \mathcal{W}} F(w^\star) + \varepsilon.$$

For high probability rates, note that $f$ is Lipschitz, hence bounded in the unit ball and we will mostly care about lower bounds. In turn, bounds in expectation can be readily turned into probability bounds using standard Markov inequality.

We also follow the standard algorithmic assumption in optimization which assumes the existence of a *first order oracle* for $f$. Namely, for any $w, z \in \mathcal{W} \times \mathcal{Z}$ the learner has access to a procedure that provides her with the value $f(w; z)$ and with the subgradient $\nabla f(w; z)$ with respect to $w$ (Nemirovsky and Yudin (1983); see also Hazan (2019); Bubeck (2015) for a more extensive background).

**Stochastic Gradient Descent.** Perhaps one of the most well studied optimization methods in SCO is *Stochastic Gradient Descent* (SGD). In this method, the algorithm iteratively chooses a parameter $w_t$ (we will always take $w_0 = 0$) and at step $t$ makes the update

$$w_{t+1}^{\mathrm{SGD}} = \Pi_{\mathcal{W}}\left(w_t^{\mathrm{SGD}} - \eta \nabla f(w_t^{\mathrm{SGD}}, z_{t+1})\right), \quad w_S^{\mathrm{SGD}} := \frac{1}{T}\sum_{t=1}^{T} w_t^{\mathrm{SGD}}.$$

where $\Pi_{\mathcal{W}}$ is the Euclidean projection over the set $\mathcal{W}$. It is well known (see for example Shalev-Shwartz and Ben-David (2014); Hazan (2019)) that if one runs SGD with a learning rate $\eta = \Theta(1/\sqrt{n})$ for $T = n$ iterations then the output $w_S^{\mathrm{SGD}}$ has:

$$\mathop{\mathbb{E}}_{S\sim D^n}[F(w_S^{\mathrm{SGD}})] \leq \min_{w^\star\in\mathcal{W}} F(w^\star) + O(1/\sqrt{n}). \tag{2}$$

In particular, for $\varepsilon > 0$ SGD succeeds to learn to within $\varepsilon$-accuracy with an order of $\Omega(1/\varepsilon^2)$ calls to a first order oracle, and $\Omega(1/\varepsilon^2)$ samples.

**Empirical risk.** An alternative method to stochastic gradient descent is to optimize the *empirical risk* over a sample $S$, defined next:

$$F_S(w) = \frac{1}{n}\sum_{i=1}^{n} f(w; z_i). \tag{3}$$

Using standard discretization and covering techniques one can show that if $n = \Omega(d/\varepsilon^2)$ then *any* algorithm that optimizes $F_S$ to accuracy $\varepsilon$ will also have roughly $O(\varepsilon)$ test error (e.g., Shalev-Shwartz and Ben-David (2014)). In fact, when $n = \Theta(d/\varepsilon^2)$ the empirical loss approximates the true loss uniformly, for all $w \in \mathcal{W}$. Shalev-Shwartz et al. (2009) showed that a dependence on $d$ in the uniform convergence rate is necessary, and Feldman (2016) proved that a linear dependence is in fact tight. We emphasize that the dimension dependence is necessary only for uniform convergence; indeed, SGD which does not rely on such arguments, does not exhibit dimension dependencies.

**Gradient Descent.** A concrete way to minimize the empirical risk in Eq. (3) is with (full-batch) *Gradient Descent*. We consider the following update rule

$$w_{t+1}^{\mathrm{GD}} = \Pi_{\mathcal{W}}\left(w_t^{\mathrm{GD}} - \eta\nabla F_S(w_t^{\mathrm{GD}})\right), \qquad w_S^{\mathrm{GD}} := \frac{1}{T}\sum_{t=1}^{T} w_t^{\mathrm{GD}}. \tag{4}$$

The output of GD is then given by the averaged sequence, $w_S^{\mathrm{GD}}$. The optimization error of GD is governed by the following equation for any choice of parameters $\eta$ and $T$ (see for example, Hazan (2019); Bubeck (2015)):

$$F_S(w_S^{\mathrm{GD}}) \leq \min_{w^\star\in\mathcal{W}} F_S(w^\star) + O\left(\eta + \frac{1}{\eta T}\right). \tag{5}$$

In particular, with a choice $\eta = O(\varepsilon)$ and $T = O(1/\varepsilon^2)$ we can optimize $F_S$ up to accuracy $\varepsilon > 0$. Using the naive dimension-dependent sample complexity bound, we have that if $n = O(d/\varepsilon^2)$, $T = O(1/\varepsilon^2)$, $\eta = O(\varepsilon)$ we can bound both the optimization error as well as generalization error and achieve true loss of order $\varepsilon$. Recently Bassily et al. (2020) provided the first, dimension-independent, generalization bound:

**Theorem** (Bassily et al., 2020, Thm 3.2)**.** *Let $D$ be an unknown distribution over $\mathcal{Z}$ and suppose $f(w; z)$ is $O(1)$-Lipschitz and convex w.r.t. $w \in \mathcal{W}$ where $\mathcal{W}$ is the unit ball in $\mathbb{R}^d$ then running GD over an i.i.d. sample $S$ with step size $\eta$ for $T$ rounds yields the following guarantee*

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

In particular, for $n = 1/\varepsilon^2$, choosing $\eta = \varepsilon^3$ and $T = \Omega(1/\varepsilon^4)$ provides:

$$\mathop{\mathbb{E}}_{S\sim D^n}[F(w_S^{\mathrm{GD}})] \leq \min_{w^\star\in\mathcal{W}} F(w^\star) + O(\varepsilon).$$

Notice the suboptimality in terms of $\varepsilon$. The above bound requires $T = \Omega(1/\varepsilon^4)$, which is suboptimal compared with the guarantee of Eq. (2) for SGD, as well as the dimension dependent generalization bound that requires $T = \Omega(1/\varepsilon^2)$. We will show that the above bound is in fact tight. Namely, if $n = O(\log d)$ then for any learning rate we need at least $T = \Omega(1/\varepsilon^4)$ iterations for GD to achieve $O(\varepsilon)$ true loss.

**Regularization.** It is customary, when minimizing the empirical error, to add a regularization term in order to avoid overfitting. Concretely, given $S = \{z_1, \ldots, z_n\}$, we consider the regularized empirical loss

$$F_{\lambda,S}(w) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} f(w; z_i). \tag{7}$$

We will consider the following update rule of GD that is known to achieve fast optimal rates for strongly convex functions (in particular, regularized). At step $t$ we take the update rule:

$$w_{t+1}^{\lambda\mathrm{GD}} = \Pi_{\mathcal{W}}\left[w_t^{\lambda\mathrm{GD}} - \eta_{t+1}\nabla F_{\lambda,S}(w_t^{\lambda\mathrm{GD}})\right], \quad w_S^{\lambda\mathrm{GD}} := \sum_{t=1}^{T}\frac{2t}{T(T+1)}w_t^{\lambda\mathrm{GD}}, \tag{8}$$

where $\eta_t = \frac{2}{\lambda(t+1)}$.

The above learning rate was suggested by Lacoste-Julien et al. (2012) where they also demonstrated the optimization guarantee:

$$F_{\lambda,S}(w_S^{\lambda\mathrm{GD}}) \leq \min_{w^\star} F_{\lambda,S}(w^\star) + O\left(\frac{1}{\lambda T}\right). \tag{9}$$

As for the test error, utilizing Eq. (9) one can bound via the empirical error (see Shalev-Shwartz et al. (2009, Eq. (24))) (as well as relating the loss of the regularized objective and the non-regularized objective), and achieve the following bound:

$$\mathop{\mathbb{E}}_{S\sim D^n}[F(w_S^{\lambda\mathrm{GD}})] \leq \min_{w^\star\in\mathcal{W}} F(w^\star) + O\left(\frac{1}{\lambda\sqrt{T}} + \frac{1}{\lambda n} + \lambda\right). \tag{10}$$

Similar to before, if we wish to tighten the above bound, we need to choose $\lambda = O(1/\sqrt{n})$ and set $T = O(n^2) = O(1/\varepsilon^4)$. We will again show a matching lower bound.

## 3. Main Results

We proceed to present our main results which provide accompanying lower bounds to Eqs. (6) and (10) respectively.

### 3.1. Gradient Descent

The proof of the following result is provided in Appendix A.1.

**Theorem 3.1.** *Fix $\eta, T$ and $n$. For $d \geq T \cdot 2^{n+5} + 20 \cdot \eta^2 T^2$, there exists a Lipschitz convex function $f(w; z)$, and a distribution $D$ over $\mathfrak{Z}$, such that if $w_S^{GD}$ is defined as in Eq. (4), then:*

$$\mathop{\mathbb{E}}_{S\sim D^n}[F(w_S^{GD})] \geq \min_{w^\star \in \mathcal{W}} F(w^\star) + \Omega\left(\min\left\{\eta\sqrt{T} + \frac{1}{\eta T}, 1\right\}\right). \tag{11}$$

Tuning the parameters $\eta$ and $T$, we obtain that for *any* learning rate, to achieve $\varepsilon$ true risk we need at least $T = O(1/\varepsilon^4)$ iterations. Together with the upper bound in Eq. (6), we observe that $T = O(1/\varepsilon^4)$, and $n = O(1/\varepsilon^2)$ provide optimal rates. The main technical novelty of our work is in deriving the first term. Namely, we provide a novel $\Omega(\eta\sqrt{T})$ generalization lower bound. The second term follows from the standard optimization guarantees for GD, which we repeat in the proof.

### 3.2. Regularized Gradient Descent

We next turn to the question of regularized Gradient Descent. As discussed, it is known that while standard ERM may be liable to overfitting, regularization (and in particular strongly convex regularization) can induce stability which in turn allows learning. We then ask the question if an analogous result appears for algorithmic settings such as GD. Namely, if we optimize over the regularized objective do we guarantee improvement in the performance. Therefore, we now consider the performance of GD on the regularized objective as in Eq. (8). The proof is provided in Appendix A.2.

**Theorem 3.2.** *Fix $n$, $\lambda > 0$ and $T$, and assume $d \geq T \cdot 2^{n+5} \cdot n$. Suppose we run GD over the regularized objective as in Eq. (7) and we output $w^{\lambda GD}$ as defined in Eq. (8). Then there exist a distribution $D$ over convex Lipschitz functions, $f(w; z)$ such that:*

$$\mathop{\mathbb{E}}_{S\sim D^n}\left[F(w_S^{\lambda GD})\right] \geq \min_{w^\star \in \mathcal{W}} F(w^\star) + \Omega\left(\min\left\{\frac{1}{\lambda\sqrt{T}} + \lambda, 1\right\}\right). \tag{12}$$

*In particular, since $\Delta_{\lambda,S} := F_{\lambda,S}(w_S^{\lambda GD}) - \min_w F_{\lambda,S}(w) \leq \frac{1}{\lambda T}$ we obtain that*

$$\mathop{\mathbb{E}}_{S\sim D^n}\left[F(w_S^{\lambda GD})\right] \geq \min_{w^\star \in \mathcal{W}} F(w^\star) + \Omega\left(\min\left\{\sqrt{\frac{\Delta_{\lambda,S}}{\lambda}} + \lambda, 1\right\}\right). \tag{13}$$

Optimizing over the choice of $\lambda$ and $T$, we obtain, again, that at least $T = O(1/\varepsilon^4)$ iterations are needed to converge to an $\varepsilon$ test error, which is comparable to the guarantee provided for unregularized GD in Eq. (6).

Eq. (13) complements the upper bound of Shalev-Shwartz et al. (2009, Eq. 24). Taken together, we observe here that $O(\varepsilon)$-training error guarantees at best $O(\sqrt{\varepsilon})$-test error. Note that in contrast with Eq. (6) whose last term deteriorates from over-training, under regularization we obtain the reversed effect, and the generalization error stems from *under-training* (see Section 5, for further discussion). We also mention here the result of Sridharan et al. (2008) that showed that, in contrast to SCO, in (general) linear models, regularized objectives do enjoy a fast rate and the test error is linear in the train error.

## 4. Constructions and Proof Overview

In this section we give a brief overview over the proof techniques, deferring the complete proofs to Appendix A. As discussed above, the main technical contribution of our work is the first term in Eq. (11); namely, in showing that

$$F(w_S^{\mathrm{GD}}) - F(w^\star) = \Omega(\eta\sqrt{T}). \tag{14}$$

The other terms are standard terms that bound the optimization errors of GD. We therefore focus the exposition here on the derivation of Eq. (14).

The proof relies on two relevant constructions that were presented in Bassily et al. (2020) and Shalev-Shwartz et al. (2009): the former provides a lower bound for the stability of GD, while the latter demonstrates a case where uniform convergence fails. Naturally, since both phenomena are necessary to obtain a generalization error, our construction carefully tailors these two ingredients to obtain the final result.

Let us briefly overview the two constructions that we build upon. We begin with the work of Bassily et al. (2020).

**GD is unstable:** To demonstrate instability of GD, Bassily et al. (2020) constructed the following example that consists of the following two functions:

$$v(w) = \gamma v \cdot w, \quad \text{and} \quad u(w) = \max\left\{0, \max_{k\in[d]} w(k)\right\}, \tag{15}$$

where $v = (-1, -1, \ldots, -1)$ and $\gamma$ is an arbitrarily small scalar. Suppose that with some very small probability (order of $1/n$) we observe $v$, and note that the gradient of $v$ slightly perturbs GD from initialization (at zero) towards the positive orthant. The other function we observe is $u$ w.p. $1 - 1/n$.

Now to show instability, note that on a typical sample $v(w)$ will not appear with roughly probability $1/e$. If it is not observed, GD will not move from the origin. On the other hand, if we do observe the function $v(w)$ in the sample, then after the first iteration that perturbs us from zero, all coordinates become positive. At the second iteration, we will observe the gradient $\nabla_2 = \gamma v + e_1$. Taking $\gamma$ to be negligibly small, that means that $w_2 \approx w_1 - \eta e_1$, and in turn, since now $w_2(1) = -\eta \le 0$, we have that $\nabla_3 = \gamma v + e_2$, etc.[1] (Also note that for a sufficiently small $\gamma < 1/\sqrt{d}$, the Lipschitz property holds.) As such the algorithm will eventually converge to $w_T \approx -\sum_{t=1}^{T} \eta e_t$. Thus, changing one example leads to a solution that is $\eta\sqrt{T}$ far away and the algorithm is unstable if $\eta = \Omega(1/\sqrt{T})$. One can observe that averaging will not help.

Note though, that the different minima for which the algorithm converges to are all generalizing. In fact all minima are generalizing, hence the example alone is not enough to ensure overfitting.

**Uniform convergence fails:** The other construction we build upon is by Shalev-Shwartz et al. (2009) which demonstrates that ERM may overfit. Their idea is to consider a distribution over $z \in \{0,1\}^d$ where each coordinate $z(k)$ is 0 or 1 with equal probabilities and a loss function of the form:

$$g(w, z) = \sum_{k=1}^{d} z(k) w^2(k).$$

1. In our construction, we want to avoid subgradients hence we consider an alternative variant that ensures a well defined gradient at each point. But for the sake of exposition, let us assume that we are provided with the above subgradient oracle.

The main observation is that if $d$ is large enough then on a sample $\{z_1, \ldots, z_n\}$ of size $n$ (logarithmic in $d$) we expect to see at least one coordinate where $z_i(k) = 0$ for all $i$. We will refer to such a coordinate as a *bad* coordinate. Note that for any bad coordinate $k$, the solution $w = e_k$ will achieve zero training error, whereas it has expected loss of $1/2$. Here, however, note that gradient descent will be stable; in particular, the origin is already a minimum. We note that this can be remedied and Shalev-Shwartz et al. (2009) also show how this example can be altered to make sure the bad minimum is unique hence gradient descent will eventually converge to the bad minimum, but their construction will overfit only if we run gradient descent for an exponential number of steps. We, on the other hand, want to show that GD will fail even when it is tuned to achieve, say, $O(1/\sqrt{n})$ error.

**Putting both together:** For the sake of exposition, we will show a slightly easier, albeit suboptimal, lower bound of:

$$F(w_S^{\mathrm{GD}}) - F(w^\star) \geq \eta^2 T,$$

While the above lower bound doesn't match the upper bound of Bassily et al. (2020), note that it is still enough to show that gradient descent with step size $\eta = 1/\sqrt{T}$ might overfit. We next move on to show the above lower bound.

Since we need both overfitting minima as well as instability of GD, then naturally we would like to incorporate both constructions together. The most straightforward idea is consider

$$\tilde{f}(w, z) = g(w, z) + \gamma v \cdot w + u(w) = \sum_{k=1}^{d} z(k) w(k)^2 + \gamma v \cdot w + \max\left\{0, \max_{k \in [d]} w(k)\right\}.$$

As before, the first step drifts the vector $w$ to the positive orthant. Note, that whenever a bad coordinate is drifted by $u(w)$, we are inflicted a true loss by $g(w, z_i)$. However, if a good coordinate is drifted, the first and last term would counter-act: namely, at the second iteration $\nabla u$ drifts the first coordinate, then the gradient of $\nabla g$ forces (w.h.p.) the first coordinate back to zero unless its a bad coordinate (which will hapen with negligable probability). The construction, thus, fails.

In order to make the above construction work, we need to correlate the bad coordinates with the coordinates that are drifted from zero. This will ensure that the subgradient of $u$ pushes only coordinates on which $g$ is not active. Before we continue with this idea, we would like to point out here that even in this construction, if the first order oracle is allowed to see the whole sample in advance, and choose the subgradient of $u$ adversarially then GD could overfit on this example (In particular, the adversary can choose as subgradient any bad coordinate). In fact, against such a first-order oracle even SGD will fail. This is not allowed though, as the first-order oracle needs to return a subgradient given a single instance $f(w, z)$, without dependence on the sample.

We next move on to discuss how we correlate the bad coordinates in Shalev-Shwartz et al. (2009) with the drift in the construction of Bassily et al. (2020). At each example we draw $z \in \{0, 1\}^d$ as in Shalev-Shwartz et al. (2009) (w.p. $1/2$ each coordinate is 0 or 1) we then also draw a perturbing vector $v_z$ but now we make sure it perturbs to a positive value only coordinates $k$ on which $z_i(k) = 0$, on the other hand for all coordinates $z_i(k) = 1$ the vector $v_z$ will in fact have a strong and reverse effect.

Concretely, letting $\gamma$ be an arbitrarily small scalar we let $v_z(k) = -1$ for $z(k) = 0$ and $v_z(k) = n$ for all $k$ such that $z(k) = 1$. For this choice of $v_z$ it can be seen that for the average vector $v_S = \frac{1}{n}\sum_{z \in S} v_z$, we have that $v_S(k) < 0$ if and only if $z(k) = 0$ for every $z \in S$. Next, our distribution draws at each iteration the function

$$f(w; z) = g(w, z) + \gamma v_z \cdot w + u(w),$$

where again $\gamma$ should be thought of as negligibly small. This leads to the empirical loss:

$$F_S(w) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{d} z_i(k)w(k)^2 + \gamma v_S \cdot w + \max\left\{0, \max_{k\in[d]} w(k)\right\}.$$

Again, if $d$ is large enough then there are "many" bad coordinates on which $z_i(k) = 0$ for all $i$. We will need to choose $d$ to ensure that at least $T$ such coordinate exist. Our choice of $v_S$ ensures that at the first iteration, all coordinates where $\sum_{i=1}^{n} z_i(k) = 0$ are perturbed to a positive value. From that iteration on, the term $\max\{w_i, 0\}$ will induce over $w_t^{\mathrm{GD}}$ the dynamic depicted in the construction of Bassily et al. (2020): $w_{t+1}^{\mathrm{GD}} \approx w_t^{\mathrm{GD}} - \eta e_{i_t}$ where $i_t$ is the $t$'th bad coordinate. Eventually, taking $T$ iterations we obtain a true loss of $\frac{1}{2}\sum_{t=1}^{T}\eta^2 = \frac{1}{2}\eta^2 T$.

Since we actually want a loss of $\eta\sqrt{T}$ we need to alter the above construction and we choose a function that behaves more like $\sqrt{g(w, z_i)}$, the gradient of $\sqrt{g}$ is slightly less well behaved and may also cause instability in the "good" coordinates, so some consideration need to be taken care of. We refer the reader to Appendix A.1 for the full proof.

**Overview of Theorem 3.2:** The proof of Theorem 3.2 exploits roughly the same objective. Certain care need to be taken because of projections. In particular, because over the regularized objective the first iterations take steps of order $O(1/\lambda)$ we necessarily drive out of the unit ball and projections happen – in distinction from GD without regularization. Again for sake of exposition we will consider the last iterate and prove a weaker bound of $1/(\lambda^2 T)$.

Therefore, for simplicity of the analysis let us start by considering GD over the regularized objective without projections. This algorithm is in fact of interest of its own right and comes with comparable guarantees. Therefore let us consider the update rule

$$w_{t+1} = w_t - \eta_{t+1}\nabla F_{\lambda,S}(w_t) = (1 - \lambda\eta_{t+1})w_t + \eta_t \nabla F_S(w_t).$$

where $\eta_t = 2/(\lambda(t+1))$. A simple proof by induction yields the following update rule:

$$\begin{aligned} w_{t+1} &= (1 - \frac{2}{t+2})w_t + \frac{2}{\lambda(t+2)}\nabla F_S(w_t) \\ &= \frac{t}{t+2}\frac{2}{\lambda t(t+1)}\sum_{t'=0}^{t-1}(t'+1)\nabla F_S(w_{t'}) + \frac{2(t+1)}{\lambda(t+1)(t+2)}\nabla F_S(w_t) \qquad \text{(induction hyp.)} \\ &= \frac{2}{\lambda(t+1)(t+2)}\sum_{t'=0}^{t}(t'+1)\nabla F_S(w_{t'}). \end{aligned}$$

Therefore, again considering the last iterate we have that

$$F(w) \geq \mathbb{E}[g(w,z)] = \mathbb{E}\left[\sum_{i=1}^{d} z(i)w(i)^2\right] = \frac{1}{2}\|w_T\|^2 = \Theta\left(\frac{1}{\lambda^2 T^4}\sum_{t=0}^{T} t^2\right) = \Theta\left(\frac{1}{\lambda^2 T}\right).$$

As before, carefully replacing $g$ with a function that behaves closer to $\sqrt{g}$ leads to the tight bound.

Projections interfere with the above analysis as they contract the vectors and in turn reduce their norm. But if we scale the objective correctly, we can ensure that for enough (say half) of the iterations projections do not occur.

Finally, the analysis above is greatly simplified by the update step suggested in Lacoste-Julien et al. (2012). It might seem as if our proof greatly rely on this learning rate. We mention that a similar analysis can be done for learning rate $\eta_t = 1/\lambda t$ (and taking the average) as well as for fixed step size $\eta \approx 1/\lambda T$. Nevertheless we leave it as an open problem if there is some first-order optimization method that achieves rate of $T = O(1/\varepsilon^2)$ (see Section 5).

### 4.1. The Construction

We next provide in detail our main construction. We fix $n, d \geq 1$ and parameters $z = (\alpha, \varepsilon, \gamma) \in \{0,1\}^d \times \mathbb{R}^d \times \mathbb{R}^3$ are such that $0 < \varepsilon_1 < \ldots < \varepsilon_d$, $\alpha \in \{0,1\}^d$ and $\gamma_1, \gamma_2, \gamma_3 > 0$. Define a family of convex functions $f_{(16)} : \mathbb{R}^d \to \mathbb{R}$ as follows:

$$f_{(16)}(w; z) = \sqrt{\sum_{i \in [d]} \alpha(i) h_\gamma^2(w(i))} + \gamma_1 v_\alpha \cdot w + \gamma_3 r_\varepsilon(w), \quad \text{with} \quad v_\alpha(i) = \begin{cases} -\frac{1}{2n} & \text{if } \alpha(i) = 0; \\ +1 & \text{if } \alpha(i) = 1, \end{cases} \tag{16}$$

where $h_\gamma : \mathbb{R} \to \mathbb{R}$ and $r_\varepsilon : \mathbb{R}^d \to \mathbb{R}$ are defined as

$$h_\gamma(a) = \begin{cases} 0 & a \geq -\gamma_2; \\ a + \gamma_2 & a < -\gamma_2, \end{cases} \quad \text{and} \quad r_\varepsilon(w) = \max\left\{0, \max_{i \in [d]} \{w(i) - \varepsilon_i\}\right\}.$$

Observe that $f_{(16)}(w; z)$ are convex, as they are a vector composition of convex functions and since the $\ell_2$-norm is non-decreasing in each argument, (see e.g., (Boyd and Vandenberghe, 2014, p. 86)). Note also that $f_{(16)}(w; z)$ are 3-Lipschitz over the Euclidean unit ball for a sufficiently small $\gamma_1 \leq 1/\sqrt{d}$ and $\gamma_3 = 1$.

We will consider an $\alpha$ that is distributed uniformly over $\{0,1\}^d$; that is, we draw $\alpha \in \{0,1\}^d$ uniformly at random and pick the function $f_{(16)}(w; (\alpha, \varepsilon, \gamma))$. The corresponding expected population risk is then

$$F_{(16)}(w) = \mathop{\mathbb{E}}_{\alpha \sim D} [f_{(16)}(w; (\alpha, \varepsilon, \gamma))].$$

Now let $S$ be an i.i.d. sample of size $n$ drawn from this distribution; we think of $S$ as a multiset of items from $\{0,1\}^d$. Let $F_S$ be the associated empirical risk:

$$F_S(w) = \frac{1}{n} \sum_{\alpha \in S} f_{(16)}(w; (\alpha, \varepsilon, \gamma)). \tag{17}$$

We next provide the key Lemma we will use for the proof of Theorem 3.1 that describes the iteration of GD over Eq. (17) (an analogue results is used in the case of Theorem 3.2). For the Lemma, given a sample $S$, let us denote by $\mathfrak{I} = \{i : \forall \alpha \in S, \alpha(i) = 0\}$ and we will denote $\mathfrak{I} = \{i_1, \ldots, i_K\}$. We will also denote $\bar{v} = \frac{1}{n} \sum_{\alpha \in S} v_\alpha$.

**Lemma 4.1.** *Let $D$ be a distribution over $z = (\alpha, \varepsilon, \gamma)$ where $\alpha \in \{0,1\}^d$ is chosen uniformly and, suppose $\gamma_2 = 2\gamma_1 \eta T$, $0 < \varepsilon_1 < \ldots < \varepsilon_d < \frac{\gamma_1}{2n}\eta$, $\frac{\gamma_1}{2n} T < 1$, $\gamma_1 \leq \frac{1}{2\sqrt{d}\eta T}$ and $\gamma_3 = 1$ are all chosen deterministically and also that $K \leq \frac{3}{4\eta^2}$. Consider $F_S$ as in Eq. (17). Then for all $1 \leq t \leq \min\{T, K\}$:*

$$\nabla F_S(w_t^{GD}) = \gamma_1 \bar{v} + e_{i_t} \quad \textit{where,} \quad w_t^{GD} = -\eta t \cdot \gamma_1 \bar{v} - \eta \sum_{s=1}^{t-1} e_{i_s}.$$

*Assuming that $T > K$ then for all $K < t \leq T$:*

$$\nabla F_S(w_t^{GD}) = \gamma_1 \bar{v} \quad \textit{where,} \quad w_t^{GD} = -\eta t \cdot \gamma_1 \bar{v} - \eta \sum_{s=1}^{K} e_{i_s}.$$

Lemma 4.1 provides a description of the dynamics of GD over the above loss function with the distribution $D$. One can observe that if the set $\mathfrak{I}$ is "large" (which is the set of bad coordinates), then GD converges approximately to a vector $\|\eta \sum_{i \in \mathfrak{I}} e_i\| = \eta\sqrt{K}$. As such, if $K = O(T)$, then the algorithm is inflicted loss of $O(\eta\sqrt{T})$, as desired. The proof of Lemma 4.1 is provided in Appendix C.

## 5. Discussion

In this work we studied the role of the optimization algorithm in learning. We showed that while SGD successfully finds a "good" optima that also generalizes, GD minimizes the empirical risk but may suffer large test error. It is not by coincidence that we turned to stochastic convex optimization. Indeed, SCO is perhaps one of few learning models where such a phenomena can exist. Specifically, in setting such as PAC-learning, regression, and general linear models learning follows from uniform convergence. Namely, learnability requires sample complexity that ensures that every minimum of the empirical risk is also an approximate minimum of the true risk. In turn, learning is reduced to empirical risk optimization.[2]

In contrast, both in SCO as well as in practice, learning looks much different. In practice, it is a prevalent situation that the learner needs to observe *far less* examples than free parameters, and learning algorithms fully capable of overfitting still succeed to learn (Zhang et al., 2016; Neyshabur et al., 2014). Also, sometimes perfect-fitting and interpolation induce generalization (Belkin et al., 2019, 2018) and in other cases early stopping is the source of generalization (Prechelt, 1998; Cataltepe et al., 1999). Making the optimization algorithm a key component in the question of generalization.

While under "luckiness"-type distributional assumptions such phenomena can indeed be recreated even in the most simplistic settings of learning, SCO is a highly attractive theoretical model in this context, and one of few, that exhibits similar phenomena without distributional assumptions, and not less important using the same optimization algorithms as often invoked in practice. As such it is natural to try and study these phenomena in the setting of SCO and to understanding exactly the role of optimization algorithms/regularization/stability as well as perhaps implicit bias and such. We next discuss some of these conclusions as well as future work and open questions:

**The (dimension dependent) sample complexity of GD?** As discussed, it is well known that given $O(d/\varepsilon^2)$ examples, GD (or in fact any ERM algorithm) trained on the dataset will reach $\varepsilon$-test error. This work demonstrated that dependence on the dimension is necessary if we are provided with $O(\log d)$ examples.[3] Feldman (2016) showed that $\Omega(d)$ examples are necessary so that all ERM algorithms will succeed. This leaves an exponential gap and we leave it as an open question whether GD trained over $\Omega(\log d)$ examples may overfit. In particular, since GD is unstable (Bassily et al.,

2. Of course, even in these simplistic models under distributional assumptions one can emulate phenomena where the algorithm matters. Specifically if we allow to incorporate assumptions that the algorithm is luckily biased towards the right solution then indeed the algorithm matters, but here we try to focus on distribution independent generalization guarantees, and avoid such luckiness-type results.
3. Here we refer to GD as GD with iteration complexity $O(1/\varepsilon^2)$ and learning rate $O(\varepsilon)$.

2020), and uniform convergence does not apply (Feldman, 2016), such a result can potentially lead to a new proof technique for generalization. On the other hand, showing that GD overfits even with $O(d)$ examples will also be a significant improvement.

**Early stopping vs. perfect fitting.** As discussed, early stopping and perfect fitting are two (contradictory) important ingredients in the process of optimizing learning algorithms. In Theorem 3.2 we showed that GD, over a strongly convex objective, needs to be trained to $O(\varepsilon^2)$ train-accuracy in order to reach $O(\varepsilon)$-test accuracy. In contrast note that the upper bound in Eq. (6) contains a term, $O(\eta T/n)$, that deteriorates due to over training (i.e. $T \to \infty$). Remarkably, both terms rely on stability (i.e. for strongly convex optimization we need high training accuracy for stability, and in the general case, over-training deteriorates the stability).

The last term in Eq. (6) is the only term for which our main result Theorem 3.1 does not present a matching lower bound, and it is then an open question whether early-stopping is necessary in the setting of stochastic convex optimization. We observe though, that using a construction of Shalev-Shwartz et al. (2009) one can show that some sort of early stopping is indeed necessary. We provide a proof in Appendix A.3:

**Theorem 5.1** (informal, see Theorem A.8 for exact statement)**.** *For every $\eta, T$ and $n$, for $d \geq T \cdot 2^{n+5}$, there exists a distribution $D$ over convex functions such that*

$$\mathop{\mathbb{E}}_{S\sim D^n}[F(w_S^{GD})] - \min_{w\in\mathcal{W}} F(w) \geq \Omega\left(1 - \frac{2^{2n}}{\eta T}\right).$$

The "early stopping" terms in the above lower bound and in the upper bound presented in Eq. (6) leaves an exponential gap. It would be interesting to close this gap and to understand the exact effect of early stopping in convex optimization.

**The role of regularization.** We provided here a lower bound that shows that standard algorithms for minimizing regularized objectives don't have any advantage over GD in terms of generalization as long as both are tuned to induce stability. The lower bound we provide is for a specific choice of learning rate and averaging technique which are common and provide optimal guarantees for minimizing regularized objectives. It is an interesting question whether we can provide a similar lower bound for any choice of dynamic learning rate and averaging technique. More broadly, we would like to understand the limitations of first-order optimization methods over the empirical risk. The main take though of the theorem remains, that the lower bound of Eq. (13) is applicable not only to abstract regularized-ERM but in fact to a well–used optimization algorithm with explicit regularization.

**The implicit bias of Gradient Descent.** One of the most promising tools for understanding generalization in machine learning is the *implicit bias* or *implicit regularization* of optimization algorithms (Neyshabur et al., 2014; Gunasekar et al., 2018a,b,c). This term refers to the algorithms preference towards certain structured solutions which in turn seem to induce generalization.

We would like to revisit this paradigm in the context of SCO. In this work we showed that GD may overfit, but this is in contrast with Bassily et al. (2020)'s result that with a learning rate $\eta = O(\varepsilon^3)$ and $T = O(1/\varepsilon^4)$, it succeeds to learn. Moreover, having seen in Theorem 3.2 that adding regularization is not effective, it is natural, then, to conjecture that the conservative learning rate $\eta = O(\varepsilon^3)$ injects implicitly regularization, which in turn accounts for generalization.

However, the work of Dauber et al. (2020) demonstrated that for GD (with any learning rate that yields some optimization guarantee, in particular the above) there is no implicit-bias that accounts for the solution of the algorithm. In other words, no matter what is the learning rate and number of iterations, GD cannot be interpreted as minimizing some regularized version of the original loss function. This result, though, is true only if we don't take the distribution of the data into account and it is an interesting future study to understand if some distribution dependent implicit bias can explain the generalization of GD. We note, though, that in general, Dauber et al. (2020) did show that there are successful learning algorithms (in fact, SGD) that generalize but their performance does not stem from their bias (even if we take the distribution into account).

## Acknowledgments

The authors would like to thank Assaf Dauber, Vitaly Feldman and Kunal Talwar for helpful discussions. This work was partially supported by the Israeli Science Foundation (ISF) grants 2549/19 and 2188/20, by the Len Blavatnik and the Blavatnik Family foundation, by the Yandex Initiative in Machine Learning at Tel Aviv University, and partially funded by an unrestricted gift from Google. Any opinions, findings, and conclusions or recommendations expressed in this work are those of the author(s) and do not necessarily reflect the views of Google.

## References

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems 33*, 2020.

Mikhail Belkin, Daniel Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *arXiv preprint arXiv:1806.05161*, 2018.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

Léon Bottou and Olivier Bousquet. The tradeoffs of large-scale learning. *Optimization for machine learning*, page 351, 2011.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2014. ISBN 978-0-521-83378-3.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, 2015.

Zehra Cataltepe, Yaser S Abu-Mostafa, and Malik Magdon-Ismail. No free lunch for early stopping. *Neural computation*, 11(4):995–1009, 1999.

Assaf Dauber, Meir Feder, Tomer Koren, and Roi Livni. Can implicit bias explain generalization? stochastic convex optimization as a case study. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems*, 2020.

Vitaly Feldman. Generalization of ERM in stochastic convex optimization: The dimension strikes back. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3576–3584, 2016.

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1827–1836. PMLR, 2018a.

Suriya Gunasekar, Jason D Lee, Nathan Srebro, and Daniel Soudry. Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information Processing Systems*, 2018: 9461–9471, 2018b.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018c.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

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

Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an $o(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

Arkadiĭ Semenovich Nemirovsky and David Borisovich Yudin. *Problem complexity and method efficiency in optimization.* Wiley-Interscience, 1983.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, 2009.

Karthik Sridharan, Shai Shalev-Shwartz, and Nathan Srebro. Fast rates for regularized objectives. *Advances in neural information processing systems*, 21:1545–1552, 2008.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

## Appendix A. Proofs

### A.1. Proof of Theorem 3.1

The proof is an immediate corollary of the following two lower bounds. As one can pick the dominant term between the bounds and thus obtain the desired result. The first Theorem is the technical heart of our lower bound, and the rest of this section is devoted to prove it.

**Theorem A.1.** *For every $\eta > 0$, $T \geq 1$ and $n$, if $d \geq T \cdot 2^{n+5}$, then there exists a function $f(w;z) : \mathbb{R}^d \to \mathbb{R}$ convex and 3-Lipschitz in $w \in \mathbb{R}^d$ for every $z \in \mathcal{Z}$, and there exists a distribution $D$ over $\mathcal{Z}$ such that: if $S \sim D^n$ is an i.i.d sample drawn from the distribution $D^n$, then:*

$$\mathop{\mathbb{E}}_{S\sim D^n}[F(w_S^{GD})] - \min_{w\in\mathcal{W}} F(w) \geq \tfrac{1}{16}\min\{\eta\sqrt{T}, \tfrac{1}{3}\}.$$

The next lower bound is a well known consequence of the optimization error as well as a standard information-theoretic lower bound. We provide a detailed proof in Appendix B.1.

**Lemma A.2.** *There exists a function $f(w;z) : \mathbb{R}^d \to \mathbb{R}$ convex and 1-Lipschitz in $w \in \mathbb{R}^d$, and a distribution $D$ such that if $d > 18\eta^2T^2$ then:*

$$\mathop{\mathbb{E}}_{S\sim D^n}[F(w_S^{GD})] - \min_{w\in\mathcal{W}} F(w) \geq \tfrac{1}{36}\min\left\{\frac{1}{\eta T}, 9\right\}.$$

**Proof of Theorem A.1.** The proof is divided into two parts. The first, and central part, is for $\eta \leq \frac{1}{4\sqrt{3}}$ and the other is for $\eta > \frac{1}{4\sqrt{3}}$.

**Case 1 - Assume $\eta \leq \frac{1}{4\sqrt{3}}$:** Without loss of generality we assume that

$$2^n \cdot \max\{16, \min\{2T, \frac{1}{3\eta^2}\}\} \leq d \leq \frac{2^n}{2\eta^2}.$$

Indeed, we can assume this as we can always embed the example in any larger dimension.

Next, recall that $\mathfrak{I} = \{i : \forall \alpha \in S, \alpha(i) = 0\}$ and we denote $\mathfrak{I} = \{i_1, \ldots, i_K\}$ where $K$ is the cardinality of $\mathfrak{I}$. We start with a probabilistic claim on $K$.

**Claim A.3.** *Suppose that $\log(2\eta^2 d) \leq n \leq \min\{\log(\frac{d}{16}), \log(\frac{d}{\min\{2T, 1/(3\eta^2)\}})\}$. Then with probability at least $3/4$, it holds that $\min\{T, 1/6\eta^2\} \leq K \leq 3/4\eta^2$.*

**Proof.** The probability that a given index $i \in [d]$ is such that $\alpha_i = 0$ for all $\alpha \in S$ is $2^{-n}$. Thus, the expected number of such indices is $\mu = 2^{-n}d$ and the standard deviation is $\sigma = \sqrt{2^{-n}(1-2^{-n})d} \leq \sqrt{\mu}$. By an application of Chebyshev's inequality we obtain

$$\Pr(K \leq \tfrac{1}{2}\mu \bigcup K \geq \tfrac{3}{2}\mu)) \leq \Pr(K \leq \mu - 2\sigma \bigcup K \geq \mu + 2\sigma) \leq \tfrac{1}{4} \text{ for } \mu \geq 16.$$

This gives the claim since $\frac{1}{2}\mu \geq \min\{T, 1/(6\eta^2)\}$ and $\frac{3}{2}\mu \leq 3/(4\eta^2)$ whenever $\log(2\eta^2 d) \leq n \leq \log(\frac{d}{\min\{2T, 1/3\eta^2\}})$. Lastly, note that our application of Chebyshev's inequality holds for $\mu \geq 16$, thus we conclude that $n \leq \log(\frac{d}{16})$. ∎

Note that the condition of Claim A.3 is satisfied when $16 \leq 1/3\eta^2$, which holds for $\eta \leq \frac{1}{4\sqrt{3}}$. We can now lower bound the expected population risk of the GD iterates. It will be convenient to replace the sequence $w_t$ with the following approximating sequence: define a new sequence, $w'_1, w'_2, \ldots, w'_T$, by setting

$$w'_t = \begin{cases} -\eta \sum_{s=1}^{t-1} e_{i_s} & 1 \leq t \leq \min\{T, K\}; \\ -\eta \sum_{s=1}^{K} e_{i_s} & K < t \leq T. \end{cases}$$

Denote $w'_S = \frac{1}{T}\sum_{t=1}^{T} w'_t$. Using Lemma 4.1 it is clear that $w'_t - w_t = \eta t \cdot \gamma_1 \bar{v}$. Now observe that

$$\|w'_t - w_t\| \leq \gamma_1 \eta t \|\bar{v}\| \leq \gamma_1 \eta t \sqrt{d}, \tag{18}$$

where we have used the fact that $0 \leq |\bar{v}_i| \leq 1$ for any $i$. In addition, since $f_{(16)}(w)$ is 3-Lipschitz we have

$$\begin{aligned} F_{(16)}(w_S) &\geq F_{(16)}(w'_S) - 3\|w'_S - w_S\| && \text{(3-Lipschitz)} \\ &\geq F_{(16)}(w'_S) - \frac{3}{T}\sum_{t=1}^{T}\|w'_t - w_t\| && \text{(triangle inequality)} \\ &\geq F_{(16)}(w'_S) - 3\gamma_1 \eta T \sqrt{d}. && \text{(Eq. (18))} \end{aligned}$$

Note that for any $w \in \mathcal{W}$

$$\begin{aligned} F_{(16)}(w) &\geq \mathop{\mathbb{E}}_{\alpha\sim D}\left[\sqrt{\sum_{i\in[d]} \alpha(i) h_\gamma^2(w(i))}\right] + \gamma_1 \mathbb{E}[v_\alpha]\cdot w + r_\varepsilon(w) && (r_\varepsilon(w) \geq 0) \\ &\geq \mathop{\mathbb{E}}_{\alpha\sim D}\left[\sqrt{\sum_{i\in[d]} \alpha(i) h_\gamma^2(w(i))}\right] + \tfrac{1}{2}\gamma_1\left(1 - \tfrac{1}{2n}\right)\sum_{i\in[d]} w(i) && (\mathbb{E}[v_\alpha(i)] = \tfrac{1}{2}\left(1 - \tfrac{1}{2n}\right)) \\ &\geq \mathop{\mathbb{E}}_{\alpha\sim D}\left[\sqrt{\sum_{i\in[d]} \alpha(i) h_\gamma^2(w(i))}\right] - \tfrac{1}{2}\gamma_1\sqrt{d}, && (\textstyle\sum_{i=1}^{d} w(i) \geq -\sqrt{d}) \end{aligned}$$

where in the last inequality we used that $\sum_{i=1}^{d} w(i) \geq -\|w\|_1 \geq -\sqrt{d}\|w\|_2 \geq -\sqrt{d}$ for $w \in \mathcal{W}$. Putting both observations together this implies

$$F_{(16)}(w_S) \geq \mathop{\mathbb{E}}_{\alpha\sim D}\left[\sqrt{\sum_{i\in[d]} \alpha(i) h_\gamma^2(w'_S(i))}\right] - \tfrac{1}{2}\gamma_1\sqrt{d} - 3\gamma_1\eta T\sqrt{d}.$$

Applying the reverse triangle inequality we also have the inequality:

$$\sqrt{\sum_{i\in[d]} \alpha(i) h_\gamma^2(w'_S(i))} \geq \sqrt{\sum_{i\in[d]} \alpha(i)(w'_S(i))^2} - \sqrt{\sum_{i\in[d]} \alpha(i)(h_\gamma(w'_S(i)) - w'_S(i))^2}.$$

Next, observe that $|h_\gamma(w'_S(i)) - w'_S(i)| \leq \gamma_2 = 2\gamma_1\eta T$ since $w'_S(i) \leq 0$ for any $i \in [d]$, thus:

$$F_{(16)}(w_S) \geq \mathop{\mathbb{E}}_{\alpha\sim D}\left[\sqrt{\sum_{i\in[d]} \alpha(i)(w'_S(i))^2}\right] - \tfrac{1}{2}\gamma_1\sqrt{d} - 5\gamma_1\eta T\sqrt{d}$$

$$\geq \frac{\|w'_S\|}{2} - \tfrac{1}{2}\gamma_1\sqrt{d} - 5\gamma_1\eta T\sqrt{d}, \qquad \text{(Jensen's inequality with } \mathbb{E}[\alpha(i)] = \tfrac{1}{2})$$

where a simple observation of $\alpha(i) = \alpha^2(i)$ ensures that the first term is convex. From the definition of $w'_t$ it is clear that for any $t_0 < \min\{K, T\}$ it holds that $w'_t(i_s) = -\eta$ for $s < t_0$ and $t > t_0$. Therefore, setting $t_0 = \frac{1}{2}\min\{K, T\}$ we have the following inequality

$$\forall s < \frac{1}{2}\min\{K, T\} : \qquad w'_t(i_s) \leq \begin{cases} -\eta & \frac{1}{2}\min\{T, K\} < t \leq T; \\ 0 & o.w. \end{cases}$$

Therefore, the average iterate holds $w'_S(i_s) \leq -\frac{1}{2}\eta$ for any $s < \frac{1}{2}\min\{K, T\}$. With this in hand, we can conclude:

$$\begin{aligned} F_{(16)}(w_S) &\geq \tfrac{1}{2\sqrt{2}}\eta\sqrt{\min\{K, T\}} - \tfrac{1}{2}\gamma_1\sqrt{d} - 5\gamma_1\eta T\sqrt{d} && (\|w'_S\| \geq \tfrac{1}{2}\eta\sqrt{\tfrac{1}{2}\min\{K, T\}}) \\ &\geq \tfrac{1}{2\sqrt{2}}\eta\sqrt{\min\{\tfrac{1}{6\eta^2}, T\}} - \tfrac{1}{2}\gamma_1\sqrt{d} - 5\gamma_1\eta T\sqrt{d} && \text{(using Claim A.3)} \\ &\geq \tfrac{1}{4}\min\{\eta\sqrt{T}, \tfrac{1}{3}\} - \tfrac{1}{2}\gamma_1\sqrt{d} - 5\gamma_1\eta T\sqrt{d}. && (2\sqrt{2} < 4 \text{ and } \sqrt{6} < 3) \end{aligned}$$

As the above inequality holds with probability at least 3/4 (Claim A.3), taking the expectation into account and the fact that in any case $F_{(16)}(w) \geq -\frac{1}{2}\gamma_1\sqrt{d}$ and that $\min_{w\in\mathcal{W}} F_{(16)}(w) \leq F_{(16)}(0) = 0$, we attain that

$$\begin{aligned} \mathop{\mathbb{E}}_{S\sim D^n}[F_{(16)}(w_S)] - \min_{w\in\mathcal{W}} F_{(16)}(w) &\geq \tfrac{3}{16}\min\{\eta\sqrt{T}, \tfrac{1}{3}\} - \gamma_1\sqrt{d} - 5\gamma_1\eta T\sqrt{d} \\ &\geq \tfrac{1}{8}\min\{\eta\sqrt{T}, \tfrac{1}{3}\}. \end{aligned}$$

For a sufficiently small $\gamma_1$ such that $\gamma_1(1 + 5\eta T)\sqrt{d} \leq \frac{1}{16}\min\{\eta\sqrt{T}, \frac{1}{3}\}$.

**Case 2 - Assume $\eta > \frac{1}{4\sqrt{3}}$:** To conclude the proof we are left to show a constant lower bound for the case of $\eta > \frac{1}{4\sqrt{3}}$. For that matter we define the deterministic convex 2-Lipschitz function $f_{(19)} : \mathbb{R} \to \mathbb{R}$:

$$f_{(19)}(w) = \begin{cases} |w - \frac{1}{4}\eta| & \eta \leq 1; \\ 2|w - \frac{2}{3}| & \eta > 1. \end{cases} \tag{19}$$

For the case $\eta \leq 1$ the gradients are given by $\nabla f_{(19)}(w) = \text{sign}(w - \frac{1}{6}\eta)$. The first GD iterate is then $w_1 = \eta$. Observe that $w_1 - \frac{1}{4}\eta > 0$. Therefore, the second GD iterate is $w_2 = w_1 - \eta = 0$ and we can deduce that $w_t = \eta$ for odd $t$ and $w_t = 0$ for even $t$. This implies that the average iterate holds $w_S \geq \frac{1}{2}\eta$ and we conclude that $f_{(19)}(w_S) \geq \frac{1}{2}\eta - \frac{1}{4}\eta \geq \frac{1}{4}\eta$. For the case $\eta > 1$, note that $\nabla f_{(19)}(0) = -2$ and therefore the first GD iterate after projection is then $w_1 = 1$. For the next iterate, observe that $\nabla f_{(19)}(1) = 2$, which implies that $w_2 = -1$. Examine the third iterate, since $\nabla f_{(19)}(-1) = -2$ we obtain that $w_3 = 1$. This entails that $w_t = 1$ for odd $t$ and $w_t = -1$ for even $t$. For even $T$ we get that the average iterate is $w_S = 0$ and therefore $f_{(19)}(w_S) = \frac{4}{3}$. On the other hand, when $T$ is odd we get that the average iterate is $w_S = \frac{1}{T}$ and therefore $f_{(19)}(w_S) = 2|\frac{1}{T} - \frac{2}{3}| \geq \frac{1}{3}$. Putting together both results (for $\eta > 1$ and $\eta \leq 1$) we obtain that for $\eta > \frac{1}{4\sqrt{3}}$

$$f_{(19)}(w_S) - \min_{w\in\mathcal{W}} f_{(19)}(w) \geq \tfrac{1}{4}\min\{\eta, \tfrac{4}{3}\} \geq \tfrac{1}{32}.$$

Note that this result is independent on the dimension as we can simply embed the function $f_{(19)}$ in the first coordinate of any large space. Namely, $f(w) = f_{(19)}(w(1))$ for $w \in \mathbb{R}^d$.

### A.2. Proof of Theorem 3.2

As before, the proof follows from the following two claim which divides the lower bound into two terms. The first term is dealt in Theorem A.4 and provides the main novelty of this section, and the rest of the section is devoted for its proof. Lemma A.5 accompanies the first result with the standard lower bounds that stems from optimization and sample complexity lower bound:

**Theorem A.4.** *Fix $n$, $\lambda < 3$, $T \geq 3$ and assume $d \geq T \cdot 2^{n+5}$. Suppose we run GD over the regularized objective as in Eq. (7) with learning rate $\eta_t = 2/(\lambda(t+1))$ and setting $w_S = \sum_{t=1}^{T} \frac{2t}{T(T+1)} w_t$. Then, there exists an $f(w,z)$ 3-Lipschitz and convex over $w \in \mathcal{W}$ and a distribution $D$ supported on $z$ such that:*

$$\mathop{\mathbb{E}}_{S\sim D^n}[F(w_S^{\lambda GD})] - F(w^\star) \geq \frac{3}{4}\min\left\{\frac{1}{8\lambda\sqrt{T+1}}, \frac{1}{16}\right\}.$$

The proof is now an immediate corollary of the following standard lower bound of $\Omega\left(\lambda + \min\left\{\frac{1}{\lambda n}, \frac{1}{\sqrt{n}}\right\}\right)$, we refer the reader to Appendix B.2 for complete proofs:

**Lemma A.5.** *For $\lambda > 0$ and fixed $n$, there exists a function $f(w,z) : \mathbb{R}^d \to \mathbb{R}$ and a distribution $D$, such that if we run GD over $F_{\lambda,S}$ as in Eq. (7), with $\eta_t = 2/(\lambda(t+1))$ and set $w_S = \sum_{t=1}^{T} \frac{t}{T(T+1)} w_t$ then*

$$\mathop{\mathbb{E}}_{S\sim D^n}[F(w_S^{\lambda GD})] - \min_{w^\star\in\mathcal{W}} F(w^\star) \geq \frac{1}{128}\min\left\{\min\left\{\frac{4}{\lambda n}, \frac{1}{\sqrt{n}}\right\} + 64\lambda, 128\right\}$$

**Proof of Theorem A.4.** We set $D$ to be a distribution over $z = (\alpha, \varepsilon, \gamma)$, similarly to Eq. (30), and again we define

$$f_{(20)}(w;z) = \sqrt{\sum_{i\in[d]} \alpha(i) h_\gamma^2(w(i))} + \gamma_1 v_\alpha \cdot w + \gamma_3 \cdot r_\varepsilon(w), \quad \text{with} \quad v_\alpha(i) = \begin{cases} -\frac{1}{2n} & \alpha(i) = 0; \\ +1 & \alpha(i) = 1, \end{cases} \tag{20}$$

where $\alpha \in \{0,1\}^d$ such that $\alpha(i) = 0$ w.p $1/2$, $\varepsilon_1 < \varepsilon_2 < \ldots < \varepsilon_d < \frac{\gamma_1}{6n(T+1)}$, and $\gamma_3 = \min\left\{\frac{\lambda}{2}\sqrt{T-2}, 1\right\}$. Finally we choose:

$$\gamma_2 \leq \frac{10^{-3}}{\sqrt{T}} \cdot \frac{\gamma_3}{4\sqrt{2}\lambda\sqrt{T+1}} \quad \gamma_1 \leq \min\left\{\frac{10^{-3}}{\sqrt{d}(3+\lambda)} \cdot \frac{\gamma_3}{4\sqrt{2}\lambda\sqrt{T+1}}, \frac{\gamma_2}{\sum_{t=1}^{T}\eta_t}, \left(\frac{\lambda}{3}\right)^{T+1} \frac{\gamma_3}{(T+1)}\right\}. \tag{21}$$

Observe that with these choice of parameters, $f_{(20)}$ is 3-Lipschitz. Because we only deal with regularized objectives, throughout this section we suppress dependence in the algorithm and write $w$ instead of $w^{\lambda\mathrm{GD}}$. Next, as in Eq. (7), we consider the regularized objective

$$F_{\lambda,S}(w) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} f_{(20)}(w;z_i),$$

The update rule is then given by:

$$w_{t+1} = \Pi_{\mathcal{W}}[w_t - \eta_{t+1}\nabla F_{\lambda,S}(w_t)] = \Pi_{\mathcal{W}}[(1-\lambda\eta_{t+1})w_t - \eta_{t+1}\nabla F_S(w_t)].$$

Next, let us set $\mathfrak{I} = \{j : \forall z_i \in S,\ \alpha_i(j) = 0\}$ and we will denote the element of $\mathfrak{I}$ as $i_1 \le i_2 \le \dots i_K$. By our assumption on $d$ we have that $n \le \min\left\{\frac{\log d}{16}, \frac{\log d}{2T}\right\}$ then, as in Claim A.3, we have that $K \ge T$ with probability $3/4$. We will show that if this event happens then:

$$F(w_S) - F(w^\star) \ge \min\left\{\frac{1}{8\lambda\sqrt{T+1}}, \frac{1}{16}\right\}.$$

The result in expectation then follows.

We first utilize Lemma C.1 as before. Specifically, we want to show the following claim:

**Claim A.6.** *For every $t \ge 2$:*

$$\nabla F_S(w_t) = \gamma_1 \bar{v} + \gamma_3 e_{i_t}. \tag{22}$$

The proof of Claim A.6 is left to the end of this section and is provided in Appendix A.2 and we continue with the proof of Theorem A.4. It will be convenient to replace the sequence $w_t$ with the following approximating sequence: define a new sequence,$w'_1, w'_2, \dots, w'_T$, by setting $w'_1 = 0$ and for every $t \ge 2$

$$w'_{t+1} = \Pi_{\mathcal{W}}\left[(1 - \eta_{t+1}\lambda)w'_t - \eta_{t+1}\gamma_3 \cdot e_{i_t}\right], \tag{23}$$

and we set $w'_S = \sum_{t=1}^{T} \frac{2t}{T(T+1)} \cdot w'_t$. We next claim that

$$\|w_S - w'_S\| \le \frac{\gamma_1\sqrt{d}}{\lambda}. \tag{24}$$

We can prove the above by induction. We show that $\|w_t - w'_t\| \le \frac{\gamma_1\sqrt{d}}{\lambda}$, and then the result holds also for the averaged $w_S$. For $t = 1$ this is immediate from Lemma C.1 and the following calculation:

$$\|w_1 - w'_1\| = \|w_1\| = \|\eta_1 \nabla F_S(0)\| \le \frac{\gamma_1\sqrt{d}}{\lambda}.$$

Next we assume the statement holds for $t$ and prove for $t+1$:

$$\begin{aligned}
\|w'_{t+1} - w_{t+1}\| &= \left\|\Pi_{\mathcal{W}}\left[(1 - \eta_{t+1}\cdot\lambda)w'_t - \eta_{t+1}\gamma_3 \cdot e_{i_t}\right] - \Pi_{\mathcal{W}}\left[(1 - \eta_{t+1}\cdot\lambda)w_t - \eta_{t+1}\gamma_1\bar{v} - \eta_{t+1}\gamma_3 \cdot e_{i_t}\right]\right\| \\
&\le \|(1 - \eta_{t+1}\cdot\lambda)w'_t - \eta_{t+1}\gamma_3 \cdot e_{i_t} - (1 - \eta_{t+1}\cdot\lambda)w_t + \eta_{t+1}\gamma_1\bar{v} + \eta_{t+1}\gamma_3 \cdot e_{i_t}\| \\
&= \|(1 - \eta_{t+1}\cdot\lambda)(w'_t - w_t)\| + \|\eta_{t+1}\gamma_1\bar{v}\| \\
&\le \frac{t}{t+2}\|w'_t - w_t\| + \frac{2}{t+2}\frac{\gamma_1\sqrt{d}}{\lambda} \\
&\le \frac{t}{t+2} \cdot \frac{\gamma_1\sqrt{d}}{\lambda} + \frac{2}{t+2}\frac{\gamma_1\sqrt{d}}{\lambda} \\
&= \frac{\gamma_1\sqrt{d}}{\lambda}.
\end{aligned}$$

This establishes Eq. (24). Next, we formalize the final claim that we will need:

**Claim A.7.** *For $t_0 \ge T/2$, we have that*

$$w'_{t_0} = \frac{2}{\lambda t_0(t_0+1)}\left(\frac{\lambda T/2(T/2+1)}{2} w'_{T/2} - \sum_{t=T/2+1}^{t_0} \gamma_3 t e_{i_t}\right).$$

*and in particular for any* $T/2 < t_0 \leq 3T/4$

$$w'_S(i_{t_0}) = -\gamma_3 \sum_{t=t_0}^{T} \frac{2t}{T(T+1)} \frac{2t_0}{\lambda t(t+1)} \leq -\gamma_3 \sum_{t=t_0}^{T} \frac{2}{\lambda T(T+1)} \leq -\frac{\gamma_3}{2\lambda(T+1)} \tag{25}$$

The proof of Claim A.7 is again deferred to the end at Appendix A.2, and we proceed with the proof.

We will refer to $f_{(20)}(w;z)$ as $f(w;z)$ for the rest of the proof. First, we derive a generic lower bound for the generalization error. Since $\mathbb{E}_{z\sim D}[f(w^\star;z)] \leq \max_z f(0;z) = 0$, we have that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{z\sim D}[f(w_S;z)] - \mathop{\mathbb{E}}_{z\sim D}[f(w^\star;z)] \\
&= \mathop{\mathbb{E}}_{z\sim D}[f(w_S;z)] - \mathop{\mathbb{E}}_{z\sim D}[f(w'_S;z)] \\
&\quad + \mathop{\mathbb{E}}_{z\sim D}[f(w'_S;z)] - \mathop{\mathbb{E}}_{z\sim D}[f(w^\star;z)] \\
&\geq \mathop{\mathbb{E}}_{z\sim D}[f(w'_S;z)] - \mathop{\mathbb{E}}_{z\sim D}[f(w^\star;z)] - 3\frac{\gamma_1\sqrt{d}}{\lambda} && \text{(3-Lipschitzness of f) \& Eq. (24)} \\
&\geq \mathop{\mathbb{E}}_{z\sim D}[f(w'_S;z)] - 3\frac{\gamma_1\sqrt{d}}{\lambda} && \mathbb{E}[f(w^\star,z)] \leq 0 \\
&\geq \mathop{\mathbb{E}}_{\alpha\sim D}\left[\sqrt{\sum_{j=1}^{T} \alpha^2(i_j) h_\gamma^2(w'_S(i_j))}\right] - \left(\frac{3}{\lambda}+1\right)\sqrt{d}\gamma_1 && |v_\alpha \cdot w'_S| \leq \|v_\alpha\| \leq \sqrt{d} \\
&\geq \frac{1}{2}\sqrt{\sum_{j=1}^{T} h_\gamma^2(w'_S(i_j))} - \left(\frac{3}{\lambda}+1\right)\sqrt{d}\gamma_1 && \text{convexity of norm}
\end{aligned}$$

Next, note that for every $i$ and $t$ we have that $w'_t(i) \leq 0$. In particular we have that $|h_\gamma(w'_S(i)) - w'_S(i))| \leq \gamma_2$, hence:

$$\begin{aligned}
\frac{1}{2}\sqrt{\sum_{j=1}^{T} h_\gamma^2(w'_S(i_j))} &\geq \frac{1}{2}\sqrt{\sum_{j=1}^{T} w'^2_S(i_j)} - \frac{1}{2}\sqrt{\sum_{j=1}^{T} (w'_S(i_j) - h_\gamma(w_S(i_j))^2} && \|v\| \geq \|u\| - \|u-v\| \\
&\geq \frac{1}{2}\sqrt{\sum_{j=1}^{T} w'^2_S(i_j)} - \sqrt{T}\cdot\gamma_2 && |h_\gamma(w'_S(i)) - w'_S(i))| \leq \gamma_2 \\
&\geq \frac{1}{2}\sqrt{\sum_{j=T/2+1}^{3T/4} w'^2_S(i_j)} - \sqrt{T}\cdot\gamma_2 \\
&\geq \sqrt{\sum_{j=T/2}^{3T/4} \left(\frac{\gamma_3}{4\lambda(T+1)}\right)^2} - \sqrt{T}\cdot\gamma_2 && Eq.\ (25) \\
&\geq \frac{\gamma_3}{4\lambda\sqrt{2(T+1)}} - \sqrt{T}\cdot\gamma_2
\end{aligned}$$

Taken together, and with our choice of $\gamma_1, \gamma_2$ in Eq. (21) we obtain the desired result.

**Proof of Claim A.6.** We, again, prove the statement by induction and we show that for $w_t$ we have that

$$w_t = \sum_{i \notin \mathcal{I}} \rho_i^{(t)} e_i + \mu^{(t)} \sum_{i > i_{t-1}, i \in \mathcal{I}} e_i + \sum_{i \leq i_{t-1}, i \in \mathcal{I}} \xi_i^{(t)} e_i, \tag{26}$$

where

- $-\gamma_2 \leq -\gamma_1 \sum_{i=1}^{t-1} \eta_t \leq \rho_i^{(t)} \leq 0$
- $\varepsilon_d < \mu^{(t)} \leq \frac{\gamma_1}{2\lambda n}$
- $\xi_i^{(t)} \leq 0$

We now assume that the above holds for $w_t$ and prove the statement for $w_{t+1}$. First, by Lemma C.1 we have that for every $t' \leq t$:

$$\nabla F_S(w_{t'}) = \gamma_1 \bar{v} + \gamma_3 \cdot e_{i_{t'}}.$$

Let us denote

$$\begin{aligned} w_{t+1/2} = & \sum_{i \notin \mathcal{I}} \underbrace{\left((1-\lambda\eta_t)\rho_i^{(t)} - \eta_t \gamma_1 \bar{v}_i\right)}_{\rho_i'} e_i \\ & + \sum_{i > i_t, i \in \mathcal{I}} \underbrace{\left((1-\lambda\eta_t)\mu^{(t)} + \gamma_1 \frac{\eta_t}{2n}\right)}_{\mu'} e_i \\ & + \underbrace{\left((1-\lambda\eta_t)\mu^{(t)} + \gamma_1 \frac{\eta_t}{2n} - \gamma_3 \eta_t\right)}_{\xi_{i_t}'} e_{i_t} \\ & + \sum_{i < i_t, i \in \mathcal{I}} \underbrace{\left((1-\lambda\eta_t)\xi_i^{(t)} + \gamma_1 \frac{\eta_t}{2n}\right)}_{\xi_i'} e_i. \end{aligned}$$

With this notation note that

$$w_{t+1} = \Pi_{\mathcal{W}}[w_{t+1/2}] = \frac{w_{t+1/2}}{\min\{\|w_{t+1/2}\|, 1\}}.$$

We next show that all three conditions are met.
For $\rho_i^{(t+1)} = \frac{\rho_i'}{\min\{\|w_{t+1/2}\|,1\}}$, note that since $0 < \bar{v}_i < 1$ as well as $0 < \lambda\eta_t < 1$,

$$0 \geq (1-\lambda\eta_t)\rho_i^{(t)} - \eta_t\gamma_1\bar{v} \geq -\gamma_1 \sum_{i=1}^{t-1} \eta_i - \eta_t\gamma_1 \geq -\gamma_1 \sum_{i=1}^{t} \eta_t.$$

In particular $-\gamma_1 \sum_{i=1}^{t} \eta_i \leq \frac{\rho_i'}{\min\{\|w_{t+1/2}\|,1\}} \leq 0$.

Next, we examine $\mu^{(t)} = \frac{\mu'}{\min\{\|w_{t+1/2}\|,1\}}$. First, note that because $f$ is 3-Lipschitz:

$$\|w_{t+1/2}\| = \|(1-\lambda\eta_t)w_t + \lambda\eta_t \frac{1}{\lambda}\nabla F_S(w_t)\| \leq \frac{3}{\lambda}.$$

Hence,

$$\mu^{(t+1)} \geq \frac{\lambda}{3}\cdot(1-\lambda\eta_t)\mu^{(t)} + \gamma_1\frac{\eta_t}{2n}) \geq \frac{\lambda}{3}\frac{\gamma_1}{2n\lambda(T+1)} \geq \frac{\gamma_1}{6n(T+1)} \geq \varepsilon_d$$

That $\mu^{(t+1)} \leq \gamma_1\frac{\eta_t}{2\lambda n}$, again follows by induction and the fact that $\lambda\eta_t \leq 1$. Finally, we consider $\xi^{(t+1)}$. We again use the fact that $\|w_{t+1/2}\| \leq \frac{3}{\lambda}$, and we claim by induction that for $t \geq j$:

$$\xi_{i_j}^{(t)} \leq \sum_{i=0}^{t-j}\frac{\lambda^t}{3^t}\frac{\gamma_1}{6n} - \left(\frac{\lambda}{3}\right)^{t-j}\frac{2\gamma_3}{3(T+1)}.$$

For $j = t$ we have that $\eta^{(t)} \leq \frac{\gamma_1}{2\lambda n}$ hence:

$$\begin{aligned}
\xi_{i_t}^{(t)} &\leq \frac{\lambda}{3}\left((1-\lambda\eta_t)\eta^{(t)} + \gamma_1\frac{\eta_t}{2n} - \gamma_3\eta_t\right) \\
&\leq \frac{\lambda}{3}\cdot\left(\frac{\gamma_1}{2\lambda n} - \frac{2\gamma_3}{\lambda(T+1)}\right) && \lambda\eta_t \leq 1 \\
&\leq \frac{\gamma_1}{6n} - \frac{2\gamma_3}{3(T+1)}
\end{aligned}$$

Next, for $t > j$

$$\begin{aligned}
\xi_{i_j}^{(t+1)} &\leq \frac{\lambda}{3}\left[(1-\lambda\eta_t)\xi_{i_j}^{(t)} + \gamma_1\frac{\eta_t}{2n}\right] \\
&\leq \frac{\lambda}{3}\left[\xi_{i_j}^{(t)} + \frac{\gamma_1}{2\lambda n}\right] && \lambda\eta \leq 1 \\
&= \frac{\lambda}{3}\xi_{i_j}^{(t)} + \frac{\gamma_1}{6n} \\
&\leq \sum_{i=0}^{t+1-j}\frac{\lambda^t}{3^t}\frac{\gamma_1}{6n} - \left(\frac{\lambda}{3}\right)^{t+1-j}\frac{2\gamma_3}{3(T+1)}
\end{aligned}$$

Finally, $\lambda < 1$ and our choice of

$$\gamma_1 \leq 6n\left(\frac{\lambda}{3}\right)^{T+1}\frac{4}{9}\frac{\gamma_3}{(T+1)}$$

ensures $\xi_i^{(t)} \leq 0$.

**Proof of Claim A.7.** To prove Claim A.7 we first show that for $t_0 > T/2$, we have that $w'_{t_0} = (1-\lambda\eta_t)w'_{t_0} - \eta_t e_{i_{t_0}}$. In other words, there are no projections after the $T/2$'th iteration. To show that we use the fact that $w'_t$ and $e_{i_t}$ are orthogonal for every $t$. This follows from the fact that $w'_t = \text{span}(e_{i_1},\ldots,e_{i_{t-1}})$. As such,

$$\|(1-\lambda\eta_{t+1})w'_t - \eta_{t+1}\cdot\gamma_3 e_{i_t}\|^2 = (1-\lambda\eta_{t+1})^2\|w'_t\|^2 + (\gamma_3\eta_{t+1})^2$$

$$
\begin{aligned}
&\leq \left(1 - \frac{2}{t+2}\right)^2 + \frac{4\gamma_3^2}{(\lambda(t+2))^2} \\
&= 1 - \frac{4}{t+2} + \frac{4}{(t+2)^2} + \frac{4\gamma_3^2}{(\lambda(t+2))^2} \\
&\leq 1 - \frac{4}{t+2} + \frac{8}{T(t+2)} + \frac{8\gamma_3^2}{\lambda^2 T(t+2)} && t+2 \geq T/2 \\
&\leq 1 + \left(8\frac{T/2}{T\cdot(t+1)} - \frac{4}{t+1}\right) && \gamma_3 \leq \frac{\lambda}{2}\sqrt{T-2} \\
&\leq 1.
\end{aligned}
$$

Using the fact that there are no projections taken, we prove by induction that at step $t$ for $t \geq T/2$,

$$
w'_t = \frac{2}{\lambda t(t+1)}\left(\frac{\lambda T/2(T/2+1)}{2} w'_{T/2} - \sum_{k=T/2+1}^{t} \gamma_3 k e_{i_k}\right).
$$

Indeed, set $t_0 > T/2$, and denote $c = \frac{\lambda T/2(T/2+1)}{2}$:

$$
\begin{aligned}
w'_{t_0} &= (1 - \lambda\eta_{t_0}) w'_{t_0-1} - \gamma_3 \eta_{t_0} e_{i_{t_0}} \\
&= \left(1 - \frac{2}{t_0+1}\right)\frac{2}{\lambda t_0(t_0-1)}\left(c w'_{T/2} - \gamma_3 \sum_{t=T/2+1}^{t_0-1} t e_{i_t}\right) - \gamma_3 \eta_{t_0} e_{i_{t_0}} \\
&= -\frac{t_0-1}{t_0+1}\cdot\frac{1}{\lambda(t_0-1)t_0}\left(c w'_{T/2} - \gamma_3 \sum_{t=T/2+1}^{t_0-1} t e_{i_t}\right) - \gamma_3 \frac{2}{\lambda t_0 + 1} e_{i_{t_0}} \\
&= \frac{2}{\lambda t_0(t_0+1)}\left(c w'_{T/2} - \gamma_3 \sum_{t=T/2+1}^{t_0} t e_{i_t}\right).
\end{aligned}
$$

### A.3. Proof of Theorem 5.1

We next state a lower bound for overtraining in the general (non strongly-convex) case. As discussed, the work of Bassily et al. (2020) showed that the stability of GD is governed by the $\eta\sqrt{T}$ term. The following result complements their work and gives a matching lower bound on the expected population risk via a construction a Shalev-Shwartz et al. (2009) with a unique empirical risk minimizer that overfits:

**Theorem A.8.** *For every $\eta, T$ and $n$, for $d \geq T \cdot 2^{n+5}$, there exists a function $f(w; z) : \mathbb{R}^d \to \mathbb{R}$ convex and 3-Lipschitz in $w \in \mathcal{W}$ for every $z$, and a distribution $D$ over $\mathcal{Z}$ such that if $S \sim D^n$ , then*

$$
\mathop{\mathbb{E}}_{S\sim D^n}[F(w_S^{GD})] - \min_{w\in\mathcal{W}} F(w) \geq \min\left\{\max\left\{\frac{1}{8} - \frac{2^{2n+2}}{4\eta T}, 0\right\}, \tfrac{1}{48}\right\}.
$$

The above suggests that for large $T$ training steps, GD might overfit. In particular, when $T = \Omega(2^{2n}/\eta)$, GD is susceptible to over-training.

**Proof of Theorem A.8** Define a family of convex functions $f_{(27)}(w;\alpha):\mathbb{R}^d\to\mathbb{R}$ parameterized by $\alpha\in\{0,1\}^d$

$$f_{(27)}(w;\alpha)=\sqrt{\sum_{i\in[d]}\alpha(i)w^2(i)}+\frac{1}{d^2}\sum_{i\in[d]}(1-w(i)), \tag{27}$$

Observe that the objective is convex and 2-Lipschitz. Fix $n\geq 1$ and consider the sequence $(\alpha_1,\ldots,\alpha_n)$. Then, we denote the empirical average over a sample $S$ of size $n$ as follows

$$F_S(w)=\frac{1}{n}\sum_{i=1}^{n}f_{(27)}(w;\alpha_i). \tag{28}$$

Similarly to our first construction we consider $D$ to be the uniform distribution over the functions $\{f(w;\alpha)\}_{\alpha\in\{0,1\}^d}$ and we denote

$$F_{(27)}(w_S)=\mathop{\mathbb{E}}_{\alpha\sim D}[f_{(27)}(w_S;\alpha)].$$

**Claim A.9.** *Running GD over the function $F_S(w)$ defined in Eq. (28) and denoting its output by $w_S$, then for $\eta\sqrt{T}\leq\frac{1}{2}$ and $d=2^{n+1}$ the following holds*

$$\mathop{\mathbb{E}}_{S\sim D^n}[F_{(27)}(w_S)]-\min_{w\in\mathcal{W}}F_{(27)}(w)\geq\max\Big\{\frac{1}{8}-\frac{2^{2n+2}}{4\eta T},0\Big\}.$$

Note that Claim A.9 holds for $\eta\sqrt{T}\leq\frac{1}{2}$. Using Theorem A.1 we know that for $\eta\sqrt{T}>\frac{1}{2}$, if $d\geq T\cdot 2^{n+5}$ there exist a function $f(w;z)$ and a distribution $D$ such that

$$\mathop{\mathbb{E}}_{S\sim D^n}\mathop{\mathbb{E}}_{z\sim D}[f(w_S;z)]-\min_{w\in\mathcal{W}}\mathop{\mathbb{E}}_{z\sim D}[f(w;z)]\geq\tfrac{1}{16}\min\{\eta\sqrt{T},\tfrac{1}{3}\}\geq\tfrac{1}{48}.$$

Combining both claims for $\eta\sqrt{T}>\frac{1}{2}$ and $\eta\sqrt{T}\leq\frac{1}{2}$, we conclude the desired result. We now proceed with proving Claim A.9.

**Proof of Claim A.9.** As $\alpha$ is distributed uniformly over $\{0,1\}^d$, we have that $\alpha(i)$ are i.i.d. uniform Bernoulli. Consider a sample $(\alpha_1,\ldots,\alpha_n)$, then the probability that a given index $i$ satisfies $\forall j\in[n]:\ \alpha_j(i)=0$ is $p=\frac{1}{2^{n+1}}$. Therefore, the probability of non-existence of such coordinate is then given by $(1-p)^d$. As a result, the probability that there exists such a coordinate is $1-(1-\frac{1}{2^{n+1}})^{2^{n+1}}\geq 1-e^{-1}\geq 1/2$. Recall that

$$F_S(w)=\frac{1}{n}\sum_{j=1}^{n}f_{(27)}(w;\alpha_j)=\frac{1}{n}\sum_{j\in[n]}\sqrt{\sum_{i\in[d]}\alpha_j(i)w^2(i)}+\frac{1}{d^2}\sum_{i\in[d]}(1-w(i)).$$

Suppose that the GD solution is in the interior of the domain, namely $\|w_S\|<1$. In addition, suppose that there exists an index $i^\star\in[d]$ such that $\forall j\in[n]:\ \alpha_j(i^\star)=0$. We can now propose a better alternative solution denoted by $\hat{w}_S$ and defined as,

$$\hat{w}_S(i)=\begin{cases}w_S(i)+1-\|w_S\| & i=i^\star;\\ w_S(i) & i\neq i^\star.\end{cases}$$

Observe that

$$F_S(w_S) - F_S(\hat{w}_S) \geq \frac{1}{d^2}(1 - \|w_S\|).$$

Using the well known optimization upper bound of GD on convex 2-Lipschitz functions (Bubeck, 2015, Theorem 3.2) we obtain for $\eta\sqrt{T} \leq \frac{1}{2}$:

$$F_S(w_S) - F_S(\hat{w}_S) \leq \frac{1}{2\eta T} + 2\eta \leq \frac{1}{\eta T},$$

where we used the fact that our domain is bounded in the Euclidean unit ball. This implies that

$$\|w_S\| \geq 1 - \frac{d^2}{\eta T}. \tag{29}$$

Consequently, we can conclude that with probability higher than $1/2$ we get

$$\begin{aligned}
F_{(27)}(w_S) - \min_{w\in\mathcal{W}} F_{(27)}(w) &\geq F_{(27)}(w_S) - \tfrac{1}{4} \\
&\geq \mathop{\mathbb{E}}_{\alpha\sim D}\Big[\sqrt{\sum_{i\in[d]} \alpha(i) w_S^2(i)}\Big] - \tfrac{1}{4} && (1 - w(i) \geq 0 \text{ for } w \in \mathcal{W}) \\
&\geq \tfrac{1}{2}\|w_S\| - \tfrac{1}{4} && (\text{Jensen's inequality with } \mathbb{E}[\alpha(i)] = \tfrac{1}{2}) \\
&\geq \frac{1}{4} - \frac{d^2}{2\eta T} && (\text{Eq. (29)})
\end{aligned}$$

where the first inequality stems from the observation that $f_{(27)}(0;\alpha) = 1/d = 2^{-(n+1)} \leq \frac{1}{4}$. For the third inequality, a simple observation of $\alpha(i) = \alpha^2(i)$ ensures that the first term is convex. Note that when $\|w_S\| = 1$ we get an even tighter lower bound of $1/4$. After taking expectation over $S \sim D^n$ we conclude that

$$\mathop{\mathbb{E}}_{S\sim D^n}[F_{(27)}(w_S)] - \min_{w\in\mathcal{W}} F_{(27)}(w) \geq \max\Big\{\frac{1}{8} - \frac{2^{2n+2}}{4\eta T}, 0\Big\}. \qquad \blacksquare$$

## Appendix B. Additional proofs for Theorem 3.1 and Theorem 3.2

### B.1. Proof of Lemma A.2

Without loss of generality assume $18\eta^2T^2 \leq d \leq 36\eta^2T^2$ (as we can always embed the below example in any larger space). Set parameters $0 < \varepsilon_1 < \cdots < \varepsilon_d < \frac{1}{2\sqrt{d}}$, and define the deterministic convex function $f_{(30)} : \mathbb{R}^d \to \mathbb{R}$ as follows:

$$f_{(30)}(w) = \Big\|w - \frac{1}{\sqrt{d}} + \varepsilon\Big\|_\infty, \tag{30}$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d)$. We begin the proof with the following claim that upper bounds the smallest coordinate:

**Claim B.1.** *There exists an $i \in [d]$ such that $w_S(i) \leq \frac{\eta T}{d}$.*

**Proof.** The update rule of GD states that

$$w_t = \Pi_{\mathcal{W}}[w_{t-1} - \eta \nabla f(w_{t-1})].$$

Note that

$$\nabla f(w) = -\operatorname{sign}(\frac{1}{\sqrt{d}} - w(i) - \varepsilon_i)e_i, \tag{31}$$

for an $i \in \arg\max_{j\in[d]}|\frac{1}{\sqrt{d}} - w(j) - \varepsilon_j|$. We will upper bound the $\ell_1$-norm of $w_{t+1}$. Observe that projection can only reduce the $\ell_1$-norm. Therefore,

$$\begin{aligned}
\|w_t\|_1 &\le \|w_{t-1} - \eta \nabla f(w_t)\|_1 \\
&\le \|w_{t-1}\|_1 + \eta\|\nabla f(w_t)\|_1 && \text{(triangle inequality)} \\
&\le \|w_{t-1}\|_1 + \eta && \text{(Eq. (31))} \\
&\le \eta t. && \text{(applying the claim recursively on } \|w_{t-1}\|_1)
\end{aligned}$$

This implies that the average iterate holds

$$\|w_S\|_1 \le \frac{1}{T}\sum_{t=1}^{T}\|w_t\|_1 \le \frac{1}{T}\sum_{t=1}^{T}\eta t \le \eta T. \tag{32}$$

If we assume by contradiction that for all $i \in [d]$ it holds that $w_S(i) > \eta T/d$, then we will get that $\|w_S\|_1 > \eta T$ which contradicts the claim in Eq. (32). ■

Using Claim B.1 the average iterate satisfies $w_S(i) \le \frac{\eta T}{d}$ for some $i$ and we can conclude

$$f(w_S) \ge \frac{1}{\sqrt{d}} - \frac{\eta T}{d} - \varepsilon_i \ge \frac{1}{2\sqrt{d}} - \frac{\eta T}{d},$$

since $\varepsilon_i \le \frac{1}{2\sqrt{d}}$. We obtain that for $18\eta^2T^2 \ge 1$

$$f(w_S) \ge \frac{1}{12\eta T} - \frac{1}{18\eta T} = \frac{1}{36\eta T}, \tag{33}$$

where we used the fact that $18\eta^2T^2 \le 36\eta^2T^2$. While for $18\eta^2T^2 \le 1$ we get that for $d = 1$

$$f(w_S) \ge \frac{1}{2} - \eta T \ge \frac{1}{4}.$$

Note also that $f(\sum_{i\in[d]}(\frac{1}{\sqrt{d}} - \varepsilon_i)e_i) = 0$, hence:

$$F(w_S) - \min_{w\in\mathcal{W}} F(w) \ge \min\left\{\frac{1}{36\eta T}, \frac{1}{4}\right\}.$$

Because we consider a deterministic function, the result also holds in expectation.

### B.2. Proof of Lemma A.5

Consider the function

$$h(w) = -\frac{\bar{\lambda}}{2}w(1),$$

where $\bar{\lambda} = \min\{1, \lambda\}$

We set the distribution $D$ to be deterministic, namely: $f(w, z) = h(w) = -\bar{\lambda}w(1)$ w.p. 1.

Note that for our update step we have that

$$w_1 = -\frac{\bar{\lambda}}{2\lambda}\nabla h(w) = -\alpha \cdot e_1,$$

where $\alpha \leq \frac{1}{2}$ Since $\nabla F_{\lambda,S}(w_1) = 0$,we have that for every $t \geq 1$, $w_t = w_1$, in particular, $w_S = -e_1$. On the other hand:

$$h(\frac{\alpha \cdot \bar{\lambda}}{2}e_1) - h(e_1) \geq -\frac{\alpha\bar{\lambda}}{2} + \frac{\bar{\lambda}}{2} = \frac{\bar{\lambda}}{4} \geq \min\left\{\frac{\lambda}{4}, \frac{1}{4}\right\}.$$

## Appendix C. Proof of Lemma 4.1

Because we only analyze non-regularized objectives in this proof we will suppress the dependence on the algorithm and we will use $w_t$ for $w_t^{\mathrm{GD}}$. Before we proceed with the proof, we present a generic Lemma that this proof relies upon. The proof of Lemma C.1 can be found at Appendix C.1.

**Lemma C.1.** *Fix $\varepsilon \in \mathbb{R}^d$ and $\gamma \in \mathbb{R}^2$. Let $(z_1, \ldots z_n)$ be a sequence such that $z_j = (\alpha_j, \varepsilon, \gamma)$, and consider*

$$F(w) = \frac{1}{n}\sum_{j\in[n]} f_{(16)}(w; z_j).$$

*Denote $\nabla_i F(w)$ to be the $i$-th element of the gradient $\nabla F(w)$. If $\mathfrak{I} = \{i : \forall j \in [n],\ \alpha_j(i) = 0\}$, then for a choice $0 < \varepsilon_1 < \ldots < \varepsilon_d$ we have the following:*

1. $\nabla F(0) = \gamma_1 \bar{v}$.
2. *For every $i \in \mathfrak{I}$ then $\nabla_i F(0) = -\frac{\gamma_1}{2n}$.*
3. *If $i \notin \mathfrak{I}$ then $0 < \nabla_i F(0) \leq \gamma_1$.*
4. *Suppose for some $k \in \mathfrak{I}$:*

$$w = \sum_{i\notin\mathfrak{I}} \rho_i e_i + \sum_{i\geq k, i\in\mathfrak{I}} \mu e_i + \sum_{i<k, i\in\mathfrak{I}} \xi_i e_i,$$

*where $-\gamma_2 < \rho_i < 0$, $\xi_i \leq 0$ and, $\mu > \varepsilon_d$. Then:*

$$\nabla F(w) = \gamma_1 \bar{v} + \gamma_3 e_k,$$

*where $\bar{v} = \frac{1}{n}\sum_{j\in[n]} v_{\alpha_j}$ and $e_k$ is the $k$-th standard basis vector in $\mathbb{R}^d$.*

5. *Suppose:*

$$w = \sum_{i \notin \mathfrak{I}} \rho_i e_i + \sum_{i \in \mathfrak{I}} \xi_i e_i,$$

*where* $-\gamma_2 < \rho_i < 0$ *and* $\xi_i \leq 0$. *Then:*

$$\nabla F(w) = \gamma_1 \bar{v}.$$

First, we address the case of $t \leq \min\{T, K\}$. We will prove by induction on $t$ that

$$w_t = -\eta t \cdot \gamma_1 \bar{v} - \eta \sum_{s=1}^{t-1} e_{i_s}. \tag{34}$$

For $t = 1$ we know that the first GD step takes to $w_1 = -\eta \nabla F_S(0)$. Using Lemma C.1 we get that $w_1 = -\eta\gamma_1 \bar{v}$ which concludes the base of the induction. For the induction step we assume that $w_t$ is given by Eq. (34). Recall that $0 < \bar{v}_i \leq 1$ for $i \notin \mathfrak{I}$ and $\bar{v}_i = -\frac{1}{2n}$ for $i \in \mathfrak{I}$. Now observe that $w_t$ takes the following form

$$\begin{aligned} w_t &= \sum_{i \notin \mathfrak{I}} \rho_i^{(t)} e_i + \sum_{i \in \mathfrak{I}} \frac{\gamma_1}{2n} \eta t e_i + \sum_{i \leq i_{t-1}, i \in \mathfrak{I}} (-\eta) e_i \\ &= \sum_{i \notin \mathfrak{I}} \rho_i^{(t)} e_i + \sum_{i \geq i_t, i \in \mathfrak{I}} \frac{\gamma_1}{2n} \eta t e_i + \sum_{i < i_t, i \in \mathfrak{I}} (-\eta + \frac{\gamma_1}{2n} \eta t) e_i, \end{aligned}$$

where $-\gamma_1 \eta t \leq \rho_i^{(t)} < 0$. One can ensure that all the necessary conditions of the fourth claim in Lemma C.1 hold, under the assumptions of Lemma 4.1. Namely, that $-\gamma_2 < \rho_i^{(t)} < 0$, $\left(-\eta + \frac{\gamma_1}{2n} \eta t\right) \leq 0$ and $\frac{\gamma_1}{2n} \eta t > \varepsilon_d$. Therefore, we can apply Lemma C.1 and obtain that $\nabla F_S(w_t) = \gamma_1 \bar{v} + e_{i_t}$. The next iterate is then

$$\begin{aligned} w_{t+1} &= w_t - \eta\gamma_1 \bar{v} - \eta e_{i_t} \\ &= -\eta(t+1) \cdot \gamma_1 \bar{v} - \eta \sum_{s=1}^{t} e_{i_s}, \end{aligned}$$

which concludes the first part of the proof. We now address the case of $K < t \leq T$ when $T > K$. Similarly to the first part, we will prove by induction on $t$ that

$$w_t = -\eta t \cdot \gamma_1 \bar{v} - \eta \sum_{s=1}^{K} e_{i_s}. \tag{35}$$

Starting at $t = K + 1$ we know that $w_{K+1} = w_K - \eta \nabla F_S(w_K)$. From the first part of the claim we can deduce that $w_K$ holds Eq. (34) and $\nabla F_S(w_K) = \gamma_1 \bar{v} + e_{i_K}$. Therefore we conclude that

$$w_{K+1} = w_K - \eta\gamma_1 \bar{v} - \eta e_{i_K} = -\eta(K+1) \cdot \gamma_1 \bar{v} - \eta \sum_{s=1}^{K} e_{i_s}.$$

For the induction step we assume that $w_t$ holds Eq. (35). Taking advantage of the properties of $\bar{v}$ we have that $w_t$ takes the following form

$$\begin{aligned} w_t &= \sum_{i \notin \mathfrak{I}} \rho_i^{(t)} e_i + \sum_{i \in \mathfrak{I}} \frac{\gamma_1}{2n} \eta t e_i + \sum_{i \in \mathfrak{I}} (-\eta) e_i \\ &= \sum_{i \notin \mathfrak{I}} \rho_i^{(t)} e_i + \sum_{i \in \mathfrak{I}} (-\eta + \frac{\gamma_1}{2n} \eta t) e_i, \end{aligned}$$

where $-\gamma_1 \eta t \leq \rho_i^{(t)} < 0$. Again, one can ensure that all the necessary conditions of the fifth claim in Lemma C.1 hold, under the assumptions of Lemma 4.1. Namely, that $-\gamma_2 < \rho_i^{(t)} < 0$ and $\left(-\eta + \frac{\gamma_1}{2n} \eta t\right) \leq 0$. Applying Lemma C.1 we obtain that $\nabla F_S(w_t) = \gamma_1 \bar{v}$. We then conclude that

$$\begin{aligned} w_{t+1} &= w_t - \eta \gamma_1 \bar{v} \\ &= -\eta(t+1) \cdot \gamma_1 \bar{v} - \eta \sum_{s=1}^{K} e_{i_s}. \end{aligned}$$

Because we ignored projections throughout the proof we need to ensure that each $w_t$ for $t = 0, \ldots, T$ lies in the Euclidean unit ball. Observe that this is indeed the case, as we get

$$\|w_t\|^2 \leq \eta^2 \min\{t, K\} + d\gamma_1^2 \eta^2 T^2 \leq \eta^2 (K + d\gamma_1^2 T^2) \leq 1,$$

since $\gamma_1 \leq \frac{1}{2\sqrt{d}\eta T}$ and $K \leq \frac{3}{4\eta^2}$.

### C.1. Proof of Lemma C.1

Consider the gradient of $f_{(16)}(w; z_j)$ at $w = 0$,

$$\nabla f_{(16)}(0; z_j) = \nabla \left( \sqrt{\sum_{i \in [d]} \alpha_j(i) h_\gamma^2(w(i))} \right) \Bigg|_{w=0} + \gamma_1 v_{\alpha_j} + \gamma_3 \nabla r_\varepsilon(0).$$

Observe that $\nabla r_\varepsilon(w) = 0$ for any $w$ that satisfies $\forall i \in [d] : \; w(i) < \varepsilon_1$. In particular, this holds when $w = 0$. Now, note that for any $w(i) > -\gamma_2$ we have $h_\gamma(w(i)) = 0$. Since $\gamma_2 > 0$, this implies that

$$\forall j \in [n] : \quad \nabla \left( \sqrt{\sum_{i \in [d]} \alpha_j(i) h_\gamma^2(w(i))} \right) \Bigg|_{w=0} = 0,$$

and we obtain $\nabla f_{(16)}(0; z_j) = \gamma_1 v_{\alpha_j}$. This concludes the first claim proof, as we get

$$\nabla F(0) = \frac{\gamma_1}{n} \sum_{j \in [n]} v_{\alpha_j} = \gamma_1 \bar{v}.$$

For $i \in \mathfrak{I}$, note that $v_{\alpha_j}(i) = -\frac{1}{2n}$ and therefore $\bar{v}_i = -\frac{1}{2n}$. In addition, for $i \notin \mathfrak{I}$ there is at least one sample $j \in [n]$ such that $\alpha_j(i) = 1$. This entails that for any $i \notin \mathfrak{I}$ it holds $0 < \bar{v}_i = \frac{1}{n} \sum_{j \in [n]} v_{\alpha_j}(i) \leq 1$. Taking both cases conclude the second and third claim proofs. For the fourth claim we assume that for some $k \in \mathfrak{I}$

$$w = \sum_{i \notin \mathfrak{I}} \rho_i e_i + \sum_{i \geq k, i \in \mathfrak{I}} \mu e_i + \sum_{i < k, i \in \mathfrak{I}} \xi_i e_i.$$

Consider then the gradient at $w$

$$\nabla F(w) = \frac{1}{n}\sum_{j\in[n]} \nabla\left(\sqrt{\sum_{i\in[d]} \alpha_j(i) h_\gamma^2(w(i))}\right) + \frac{\gamma_1}{n}\sum_{j\in[n]} v_{\alpha_j}(i) + \gamma_3 \nabla r_\varepsilon(w).$$

Let us examine each term separately. We start with the first term,

$$\frac{1}{n}\sum_{j\in[n]} \nabla\left(\sqrt{\sum_{i\in[d]} \alpha_j(i) h_\gamma^2(w(i))}\right) = \frac{1}{n}\sum_{j\in[n]} \nabla\left(\sqrt{\sum_{i\notin\mathfrak{I}} \alpha_j(i) h_\gamma^2(w(i))}\right),$$

where we used the fact that for any $i \in \mathfrak{I}$ we have $\alpha_j(i) = 0$. First, note that this term is independent in $w(i)$ for $i \in \mathfrak{I}$. In addition, for any $i \notin \mathfrak{I}$ we have $w(i) = \rho_i > -\gamma_2$. This implies that

$$\frac{1}{n}\sum_{j\in[n]} \nabla\left(\sqrt{\sum_{i\in[d]} \alpha_j(i) h_\gamma^2(w(i))}\right) = 0. \tag{36}$$

The second term is trivially given by the definition of $\bar{v}$,

$$\frac{\gamma_1}{n}\sum_{j\in[n]} v_{\alpha_j} = \gamma_1 \bar{v}.$$

Recall that $r_\varepsilon(w) = \max\{0, \max_{i\in[d]}\{w(i) - \varepsilon_i\}\}$ and observe the following

$$w(i) = \begin{cases} \rho_i & i \notin \mathfrak{I}; \\ \mu & i \in \mathfrak{I}, i \geq k; \\ \xi_i & i \in \mathfrak{I}, i < k. \end{cases}$$

Since $\rho_i < 0$ and $\xi_i \leq 0$ the maximum of $w_i - \varepsilon_i$ can only be achieved for $i \in \mathfrak{I}, i \geq k$. Specifically, as $\varepsilon_i$ are strictly increasing and $\mu > \varepsilon_i$ for all $i \in [d]$, we get that the maximum is given in $i = k$. This concludes the fourth claim proof as $\nabla r_\varepsilon(w) = e_k$. For the last claim we assume

$$w = \sum_{i\notin\mathfrak{I}} \rho_i e_i + \sum_{i\in\mathfrak{I}} \xi_i e_i.$$

Following the same arguments as in the previous claim we get that Eq. (36) holds here as well. Therefore,

$$\nabla F(w) = \gamma_1 \bar{v} + \gamma_3 \nabla r_\varepsilon(w).$$

Observe that $w(i) \leq 0$ for any $i \in [d]$, since $\rho_i < 0$ and $\xi_i \leq 0$. This implies that $\nabla r_\varepsilon(w) = 0$, which concludes the proof.