Shape Matters: Understanding the Implicit Bias of the Noise Covariance

Jeff Z. HaoChen   JHAOCHEN@STANFORD.EDU
Stanford University

Colin Wei   COLINWEI@STANFORD.EDU
Stanford University

Jason D. Lee   JASONLEE@PRINCETON.EDU
Princeton University

Tengyu Ma   TENGYUMA@STANFORD.EDU
Stanford University

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

The noise in stochastic gradient descent (SGD) provides a crucial implicit regularization effect for training overparameterized models. Prior theoretical work largely focuses on spherical Gaussian noise, whereas empirical studies demonstrate the phenomenon that parameter-dependent noise — induced by mini-batches or label perturbation — is far more effective than Gaussian noise. This paper theoretically characterizes this phenomenon on a quadratically-parameterized model introduced by Vaskevicius et al. and Woodworth et al. We show that in an over-parameterized setting, SGD with label noise recovers the sparse ground-truth with an arbitrary initialization, whereas SGD with Gaussian noise or gradient descent overfits to dense solutions with large norms. Our analysis reveals that parameter-dependent noise introduces a bias towards local minima with smaller noise variance, whereas spherical Gaussian noise does not.

Keywords: Implicit regularization, implicit bias, over-parameterization.

1. Introduction

One central mystery of deep artificial neural networks is their capability to generalize when having far more learnable parameters than training examples Zhang et al. (2016). To add to the mystery, deep nets can also obtain reasonable performance in the absence of any explicit regularization. This has motivated recent work to study the regularization effect due to the optimization (rather than objective function), also known as *implicit bias* or *implicit regularization* Gunasekar et al. (2017, 2018a,b); Soudry et al. (2018); Arora et al. (2019). The implicit bias is induced by and depends on many factors, such as learning rate and batch size Smith et al. (2017); Goyal et al. (2017); Keskar et al. (2016); Li et al. (2019b); Hoffer et al. (2017), initialization and momentum Sutskever et al. (2013), adaptive stepsize Kingma and Ba (2014); Neyshabur et al. (2015); Wilson et al. (2017), batch normalization Ioffe and Szegedy (2015) and dropout Srivastava et al. (2014).

Among these sources of implicit regularization, the SGD noise is believed to be a vital one (Le-Cun et al., 2012; Keskar et al., 2016). Previous theoretical works (e.g., Li et al. (2019b)) have studied the implicit regularization effect from the *scale of the noise*, which is directly influenced by learning rate and batch size. However, people have empirically observed that the *shape of the noise* also has a strong (if not stronger) implicit bias. For example, prior works show that mini-batch
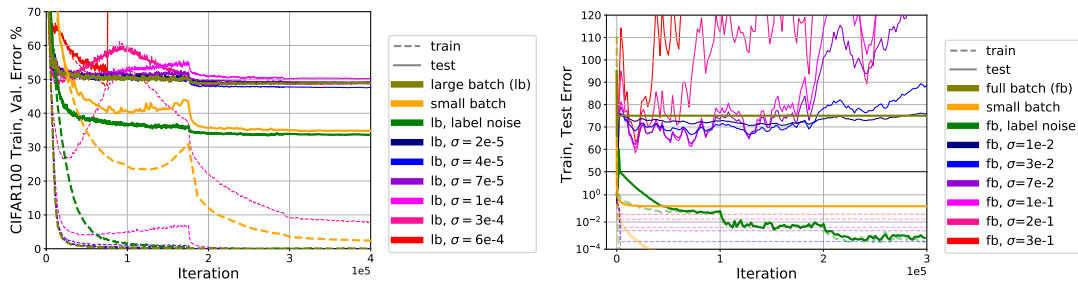
Figure 1: **The effect of noise covariance in neural network and quadratically-parameterized models.** We demonstrate that label noise induces a stronger regularization effect than Gaussian noise. In both real and synthetic data, adding label noise to large batch (or full batch) SGD updates can recover small-batch generalization performance, whereas adding Gaussian noise with optimally-tuned variance σ^2 cannot. **Left:** Training and validation errors on CIFAR100 for VGG19. Adding Gaussian noise to large batch updates gives little improvement (around 2%), whereas adding label noise recovers the small-batch baseline (around 15% improvement). **Right:** Training and validation error on a 100-dimensional quadratically-parameterized model defined in Section 2. Similar to deep models, label noise or mini-batch noise leads to better solutions than optimally-tuned spherical Gaussian noise. Moreover, Gaussian noise causes the parameter to diverge after sufficient mixing, as suggested by our negative result for Langevin dynamics (Theorem 2). More details are in Section A.

noise or label noise (label smoothing) – noise in the parameter updates from the perturbation of labels in training – is far more effective than adding spherical Gaussian noise (e.g., see (Shallue et al., 2018, Section 4.6) and Szegedy et al. (2016); Wen et al. (2019)). We also confirm this phenomenon in Figure 1 (left). Thus, understanding the implicit bias of the noise shape is crucial. Such an understanding may also apply to distributed training because synthetically adding noise may help generalization if parallelism reduces the amount of mini-batch noise (Shallue et al., 2018).

In this paper, we theoretically study the effect of the shape of the noise, demonstrating that it can provably determine generalization performance at convergence. Our analysis is based on a nonlinear quadratically-parameterized model introduced by (Woodworth et al., 2020; Vaskevicius et al., 2019), which is rich enough to exhibit similar empirical phenomena as deep networks. Indeed, Figure 1 (right) empirically shows that SGD with mini-batch noise or label noise can generalize with arbitrary initialization without explicit regularization, whereas GD or SGD with spherical Gaussian noise cannot. We aim to analyze the implicit bias of label noise and Gaussian noise in the quadratically-parametrized model and explain these empirical observations.

We choose to study label noise because it can replicate the regularization effects of minibatch noise in both real and synthetic data (Figure 1), and has been used to regularize large-batch parallel training (Shallue et al., 2018). Moreover, label noise is less sensitive to the initialization and the optimization history than mini-batch noise, which makes it more amenable to theoretical analysis. For example, in an extreme case, if we happen to reach or initialize at a solution that overfits the data exactly, then mini-batch SGD will stay there forever because both the gradient and the noise

vanish (Vaswani et al., 2019). In contrast, label noise will not accidentally vanish, so the analysis is more tractable. Understanding label noise may lead to understanding mini-batch noise or replacing it with other more robust choices.

In our setting, we prove that with a proper learning rate schedule, SGD with label noise recovers a sparse ground-truth classifier and generalizes well, whereas SGD with spherical Gaussian noise generalizes poorly. Concretely, SGD with label noise biases the parameter towards the low sparsity regime and exactly recovers the sparse ground-truth, even when the initialization is arbitrarily large (Theorem 1). In this same regime, noise-free gradient descent quickly overfits because it trains in the NTK regime (Jacot et al., 2018; Chizat and Bach, 2018). Adding Gaussian noise is insufficient to fix this, as this algorithm would end up sampling from a Gibbs distribution with infinite partition function and fail to converge to the ground-truth (Theorem 2). In summary, with not too small learning rate or noise level, label noise suffices to bias the parameter towards sparse solutions without relying on a small initialization, whereas Gaussian noise cannot.

Our analysis suggests that the fundamental difference between label or mini-batch noise and Gaussian noise is that the former is parameter-dependent, and therefore introduces stronger biases than the latter. The conceptual message highlighted by our analysis is that there are *two* possible implicit biases induced by the noise: 1. prior work (Keskar et al., 2016) shows that by escaping sharp local minima, noisy gradient descent biases the parameter towards more robust solutions (i.e, solutions with low curvature, or "flat" minima), and 2. when the noise covariance varies across the parameter space, there is another (potentially stronger) implicit bias effect toward parameters where the noise covariance is smaller. Label or mini-batch noise benefits from both biases, whereas Gaussian noise is independent of the parameter, so it benefits from the first bias but not the second. For the quadratically-parameterized model, this first bias is not sufficient for finding solutions with good generalization because there is a large set of overfitting global minima of the training loss with reasonable curvature. In contrast, the covariance of label noise is proportional to the scale of the parameter, inducing a much stronger bias towards low norm solutions which generalize well.

1.1. Additional Related Works

There has been a line of work empirically studying how noise influences generalization. Keskar et al. (2016) argued that large batch training will converge to "sharp" local minima which do not generalize well. Hoffer et al. (2017) argued that large batch size doesn't hurt generalization much if training goes on long enough and additional noise is added with a larger learning rate. Goyal et al. (2017) and Shallue et al. (2018) showed large batch training with proper learning rate and additional label noise can achieve similar generalization as small batch. Wei and Schwab (2019); Chaudhari and Soatto (2018); Yaida (2018) (heuristically) suggested that SGD may encourage solutions with smaller noise covariance. Martin and Mahoney (2018) used random matrix theory to analyze implicit regularization effects of noises. The noise induced by dropout has been shown to change the expected training objective, hence provides a regularization effect (Mianjy et al., 2018; Mianjy and Arora, 2019; Wei et al., 2020; Arora et al., 2020). Wei et al. (2020) showed that there also exists an implicit bias induced by dropout noise.

Several previous works have theoretically studied generalization bounds and training dynamics of SGD with state-dependent noises. Hardt et al. (2015) derived stability-based generalization bounds for mini-batch SGD based on training speed. Cheng et al. (2019) proved that SGD with state-dependent noises has iterate distribution close to the corresponding continuous stochastic dif-

ferential equation with the same noise covariance. Meng et al. (2020); Xie et al. (2020) showed that SGD with state-dependent noises escapes local minimum faster than SGD with spherical Gaussian noise .

Most closely related to our work, Blanc et al. (2019) and Zhu et al. (2019) also theoretically studied implicit regularization effects that arise due to shape, rather than scale, of the noise. However, they only considered the *local* effect of the noise near some local minimum of the loss. In contrast, our work analyzes the global effect of noise. For a more detailed comparison with (Blanc et al., 2019), see Section 2.2.

Woodworth et al. (2020); Vaskevicius et al. (2019) analyze the effect of initialization for the same model that we study, showing that large initialization trains in the NTK regime (shown to generalize poorly (Wei et al., 2019; Ghorbani et al., 2019)) whereas small initialization does not. We show that when the initialization is large, adding noise helps avoid the NTK regime (Li and Liang, 2018; Jacot et al., 2018; Du et al., 2018b; Woodworth et al., 2020) without explicit regularization.

Langevin dynamics or the closely-related stochastic gradient descent with spherical Gaussian noise has been studied in previous theoretical works (Welling and Teh, 2011; Teh et al., 2016; Raginsky et al., 2017; Zhang et al., 2017; Mou et al., 2017; Roberts et al., 1996; Ge et al., 2015; Negrea et al., 2019; Neelakantan et al., 2015; Mou et al., 2018). In particular, Raginsky et al. (2017) and Li et al. (2019a) provided generalization bounds for SGLD using algorithmic stability.

Several works have theoretically analyzed other types of implicit biases in simplified settings (Soudry et al., 2018; Gunasekar et al., 2018b; Ji and Telgarsky, 2018a). Gunasekar et al. (2017) and Li et al. (2017) showed that gradient descent finds low rank solutions in matrix completion. Gradient descent has also been shown to maximize the margin in linear and homogeneous models (Soudry et al., 2018; Ji and Telgarsky, 2018b; Nacson et al., 2018; Lyu and Li, 2019; Gunasekar et al., 2018a; Nacson et al., 2019; Poggio et al., 2017). Du et al. (2018a) showed that gradient descent implicitly balances the layers of deep homogeneous models. Other works showed that it may not always be possible to characterize implicit biases in terms of norm (Arora et al., 2019; Razin and Cohen, 2020). Gissin et al. (2019) showed that gradient descent dynamics exhibit different implicit biases based on depth. Li et al. (2019b) studied the implicit regularization effect of a large initial learning rate.

Recent works also suggest that explicit regularization may mitigate the lack of implicit regularization, especially in noisy or imbalanced settings. For example, Wei and Ma (2019) show that Lipschitz-ness regularization improves the performance in clean or noisy label setting when the learning rate is sub-optimal. Cao et al. (2019) show that additional regularization improves the generalization performance of rare classes. Nakkiran et al. (2020) show that explicit regularization can mitigate the double descent phenomenon in linear regression, which is caused by the fact that the implicit regularization of gradient descent with zero initialization is insufficient for the regime when the number of parameters is close to the number of datapoints.

2. Setup and Main Results

2.1. Setup and Backgrounds

Parameterization. We focus on the nonlinear model parametrization: $f_v(x) \triangleq \langle v^{\odot 2}, x \rangle$, where $v \in \mathbb{R}^d$ is the parameter of the model, $x \in \mathbb{R}^d$ is the data, and $v^{\odot 2}$ denotes the element-wise square of v. Prior works (Woodworth et al., 2020; Vaskevicius et al., 2019; Li et al., 2017) have studied this model because it is an interesting and informative simplification of nonlinear models. As SGD noise

exhibits many of the same empirical behaviors in this simplified model as in deep networks,[1] we use this model as a testbed to develop a mathematical understanding of various sources of implicit biases. As shown in Figure 1, both SGD with mini-batch noise and label noise generalize better than GD or SGD with spherical Gaussian noise.

Data distribution assumptions and overparametrization. We assume that there exists a ground-truth parameter $v^\star \in \mathbb{R}^d$ that generates the label $y = \langle v^{\star \odot 2}, x \rangle$ given a data point x, which is assumed to be generated from $\mathcal{N}(0, \mathcal{I}_{d \times d})$. A dataset $\mathcal{D} = \left\{ (x^{(i)}, y^{(i)}) \right\}_{i=1}^n$ of n i.i.d data points are generated from this distribution. The implicit bias is only needed in an over-parameterized regime, and therefore we assume that $n \ll d$. To make the ground-truth vector information-theoretically recoverable, we assume that the ground-truth vector $v^\star$ is r-sparse. Here r is much smaller than d, and casual readers can treat it as a constant. Because the element-wise square in the model parameterization is invariant to any sign flip, we assume $v^\star$ is non-negative without loss of generality. For simplicity, we also assume it only takes value in $\{0, 1\}$.[2] We use $S \subset [d]$ with $|S| = r$ to denote the support of $v^\star$ throughout the paper.

We remark that we can recover $v^\star$ by re-parameterizing $u = v^{\odot 2}$ and applying LASSO (Tibshirani, 1996) in the u-space when $n \geq \widetilde{O}(r)$, which is minimax optimal (Raskutti et al., 2012). However, the main goal of the paper, similar to several prior works (Woodworth et al., 2020; Vaskevicius et al., 2019; Li et al., 2017), is to prove that the implicit biases of non-convex optimization can recover the ground truth *without explicit regularization* in the over-parameterized regime when $n = \text{poly}(r) \ll d$.[3] We also assume throughout the paper that n, d are larger than some sufficiently large universal constant.

Loss function. We use the mean-squared loss denoted by $\ell^{(i)}(v) \triangleq \frac{1}{4} \left(f_v(x^{(i)}) - y^{(i)} \right)^2$ for the i-th example. The empirical loss is written as $\mathcal{L}(v) \triangleq \frac{1}{n} \sum_{i=1}^n \ell^{(i)}(v)$.

Initialization. We use a large initialization of the form $v^{[0]} = \tau \cdot \mathbb{1}$ where $\mathbb{1}$ denotes the all 1's vector, where we allow τ to be arbitrarily large (but polynomial in d).

Algorithm 1 Stochastic Gradient Descent with Label Noise

---

input: Number of iterations T, a sequence of step sizes $\eta^{[0:T]}$, noise level δ, initialization $v^{[0]}$
for $t = 0$ *to* $T - 1$ **do**

  Sample index $i_t \sim [n]$ uniformly and add noise $s_t \sim \{\pm\delta\}$ to $y^{(i_t)}$;
  Let $\tilde{\ell}^{(i_t)}(v) = \frac{1}{4}(f_v(x^{(i_t)}) - y^{(i_t)} - s_t)^2$;
  $v^{[t+1]} \leftarrow v^{[t]} - \eta^{[t]} \nabla \tilde{\ell}^{(i_t)}(v^{[t]})$;   // update with label noise

end

SGD with label noise. We study SGD with label noise as shown in Algorithm 1. We sample an example, add label noise sampled from $\{\pm\delta\}$ to the label, and apply the gradient update. Computing the gradient, we obtain the update rule written explicitly as:

$$v^{[t+1]} \leftarrow v^{[t]} - \eta^{[t]} \left((v^{[t] \odot 2} - v^{\star \odot 2})^\top x^{(i_t)} \right) x^{(i_t)} \odot v^{[t]} + \eta^{[t]} s_t x^{(i_t)} \odot v^{[t]}. \tag{1}$$

---

1. In contrast, the implicit bias of noise wouldn't show up in a simpler linear regression model.
2. Our analysis can be straightforwardly extended to $v^\star$ with other non-zero values.
3. We also remark that it's common to obtain only sub-optimal sample complexity guarantees in the sparsity parameters with non-convex optimization methods (Li et al., 2017; Ge et al., 2016; Vaskevicius et al., 2019; Chi et al., 2019) due to technical limitations.

Langevin dynamics/diffusion. We compare SGD with label noise to Langevin dynamics, which adds *spherical* Gaussian noise to gradient descent (Neal et al., 2011):

$$v^{[t+1]} \leftarrow v^{[t]} - \eta \nabla \mathcal{L}(v^{[t]}) + \sqrt{2\eta/\lambda} \cdot \xi, \tag{2}$$

where the noise $\xi \sim \mathcal{N}(0, \mathcal{I}_{d \times d})$ and $\lambda > 0$ controls the scale of noise. Langevin dynamics (LD) or its more computationally-efficient variant, stochastic gradient Langevin dynamics (SGLD), is known to converge to the Gibbs distribution $\mu(v) \propto e^{-\lambda \mathcal{L}(v)}$ under various settings with sufficiently small learning rate (Roberts et al., 1996; Dalalyan, 2017; Bubeck et al., 2018; Raginsky et al., 2017). In our negative result about Langevin dynamics/diffusion, we directly analyze the Gibbs distribution in order to disentangle the convergence and the generalization. In our negative result about Langevin dynamics/diffusion, we directly analyze the Gibbs distribution to disentangle the convergence and the generalization.

Notations. Unless otherwise specified, we use $O(\cdot), \Omega(\cdot), \Theta(\cdot)$ to hide absolute multiplicative factors and $\widetilde{O}(\cdot), \widetilde{\Theta}(\cdot), \widetilde{\Omega}(\cdot)$ to hide poly-logarithmic factors in problem parameters such as d and τ. For example, every occurrence of $\widetilde{O}(x)$ is a placeholder for a quantity $f(x)$ that satisfies that for some absolute constants $c_1, c_2 > 0$, $\forall x, |f(x)| \leq c_1 |x| \cdot \log^{c_2}(d\tau)$.

2.2. Main Results

Our main result can be summarized by the following theorem, which suggests that stochastic gradient descent with label noise can converge to the ground truth despite a potentially large initialization.

Theorem 1 *In the setting of Section 2.1, given a target error $\epsilon > 0$. Suppose we have $n \geq \widetilde{\Theta}(r^2)$ samples. For any label noise level $\delta \geq \widetilde{\Theta}(\tau^2 d^2)$, we run SGD with label noise (Algorithm 1) with the following learning rate schedule:*

1. *learning rate $\eta_0 = \widetilde{\Theta}(1/\delta)$ for $T_0 = \widetilde{\Theta}(1)$ iterations,*
2. *learning rate $\eta_1 = \widetilde{\Theta}(1/\delta^2)$ for $T_1 = \widetilde{\Theta}(1/\eta_1)$ iterations,*
3. *learning rate $\eta_2 = \widetilde{\Theta}(\epsilon^2/\delta^2)$ for $T_2 = \widetilde{\Theta}(1/\eta_2)$ iterations.*

Then, with probability at least 0.9, the final iterate $v^{[T]}$ at time $T = T_0 + T_1 + T_2$ satisfies

$$\|v^{[T]} - v^\star\|_\infty \leq \epsilon. \tag{3}$$

Here $\widetilde{\Theta}(\cdot)$ omits poly-logarithmic dependencies on $1/\epsilon$, d and τ.

In other words, with arbitrarily large initialization scale τ, we can choose large label noise level and the learning rate schedule so that SGD with label noise succeeds in recovering the ground truth. In contrast, when τ is large, gradient flow without noise trains in the "kernel" regime as shown by (Woodworth et al., 2020; Chizat and Bach, 2018). The solution in this kernel regime minimizes the RKHS distance to initialization, and in our setting equates to finding a zero-error solution with minimum $\|v^{\odot 2} - v^{[0] \odot 2}\|_2$. Such a solution could be arbitrarily far away when initialization scale τ is large and therefore have poor generalization. Figure 1 (right) confirms GD performs poorly with large initialization whereas SGD with minibatch or label noise works. We outline the proof of Theorem 1 in Section 3.

Blanc et al. (2019) also study the implicit bias of the label noise. For our setting, their result implies that when the iterate is near a global minimum for sufficient time, the iterates will *locally* move to the direction that reduces the ℓ_2-norm of v by a small distance (that is larger than random fluctuation). However, it does not imply the global convergence to a solution with good generalization with large (or any) initialization, which is what we prove in Theorem 1.[4] Moreover, our analysis captures the effect of the large noise or large learning rate – we require the ratio between the noise and the gradient, which is captured by the value $\eta\delta^2$, to be sufficiently large. This is consistent with the empirical observation that good generalization requires a sufficiently large learning rate or small batch (Goyal et al., 2017).

On the other hand, the following negative result for Langevin dynamics demonstrates that adding Gaussian noise fails to recover the ground truth even when $v^\star = 0$. This suggests that spherical Gaussian noise does not induce a strong enough implicit bias towards low-norm solutions.

Theorem 2 *Assume in addition to the setting in Section 2.1 that the ground truth $v^\star = 0$. When $n \leq d/3$, with probability at least 0.9 over the randomness of the data, for any $\lambda > 0$, the Gibbs distribution is not well-defined because the partition function explodes:*

$$\int_{\mathbb{R}^d} e^{-\lambda\mathcal{L}(v)}dv = \infty. \tag{4}$$

As a consequence, Langevin diffusion does not converge to a proper stationary distribution.

Theorem 2 helps explain the behavior in Figure 1, where adding Gaussian noise generalizes poorly for both synthetic and real data. In particular, in Figure 1 (right) adding Gaussian noise causes the parameter to diverge for synthetic data, and Theorem 2 explains this observation. A priori, the intuition regarding Langevin dynamics is as follows: as $\lambda \to +\infty$, the Gibbs distribution (if it exists) should concentrate on the manifold of global minima with zero loss. The measure on the manifold of global minima should be decided by the geometry of $\mathcal{L}(\cdot)$, and in particular, the curvature around the global minimum. As $\lambda \to +\infty$, the mass should likely concentrate at the flattest global minimum (according to some measure of flatness), which intuitively is $v^\star = 0$ in this case.

However, our main intuition is that when $n < d$, even though the global minimum at $v^\star$ is the flattest, there are also many bad global minima with only slightly sharper curvatures. The vast volume of bad global minima dominate the flatness of the global minimum at $v^\star = 0$ for any λ,[5] and hence the partition function blows up and the Gibbs distribution doesn't exist. An overview of the proof of Theorem 2 can be found in Section 4, while the full proof is deferred to Section F.

3. Analysis Overview of SGD with Label Noise (Theorem 1)

3.1. Warm-up: Updates with Only Parameter-dependent Noise

Towards building intuition and tools for analyzing the parameter-dependent noise, in this subsection we start by studying an extremely simplified random walk in one dimensional space. The random

---

4. It also appears difficult to generalize the local analysis directly to a global analysis because once the iterate leaves the local minimum, all the local tools do not apply anymore, and it's unclear whether the iterate will converge to a new local minimum or getting stuck at some region.

5. In fact, one can show that if this phenomenon happens for some $\lambda > 0$, then it happens for all other λ.

walk is purely driven by mean-zero noisy updates and does not involve any gradient updates:

$$v \leftarrow v + \eta\xi \cdot v, \text{ where } \xi \sim \{\pm 1\}. \tag{5}$$

Indeed, attentive readers can verify that when dimension $d = 1$, sample size $n = 1$, and $v^* = 0$, equation (1) degenerates to the above random walk if we omit the gradient update term (second to last term in equation (1)). We compare it with the standard Brownian motion (which is the analog of gradient descent with spherical Gaussian noise under this extreme simplification)

$$v \leftarrow v + \eta\xi, \text{ where } \xi \sim \mathcal{N}(0, 1). \tag{6}$$

We initialize at $v = 1$. We observe that both random walks have mean-zero updates, so the mean is preserved: $\mathbb{E}[v] = 1$. The variances of the two random walks are also both growing because any mean-zero update increases the variance. Moreover, the Brownian motion diverges because it has a Gaussian marginal with variance growing linearly in t, and there is no limiting stationary distribution.

However, the parameter-dependent random walk (5) has dramatically different behavior when $\eta < 1$: the random variable v will eventually converge to $v = 0$ with high probability (though the variance grows and the mean remains at 1.). This is because the variance of the noise depends on the scale of v. The smaller v is, the smaller the noise variance is, and so the random walk tends to get "trapped" around 0. This claim has the following informal but simple proof that does not strongly rely on the exact form of the noise and can be extended to more general high-dimensional cases.

Consider an increasing concave potential function $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\phi'' < 0$ (e.g., $\phi(v) = \sqrt{v}$ works). Note that when $\eta < 1$, the random variable v stays nonnegative. We can show that the expected potential function decreases after any update

$$\mathbb{E}[\phi(v + \eta\xi v)] \approx \mathbb{E}[\phi(v) + \phi'(v)\eta\xi v + \phi''(v)\eta^2\xi^2 v^2] \qquad \text{(by Taylor expansion)}$$
$$= \mathbb{E}[\phi(v)] + \mathbb{E}[\phi''(v)\eta^2 v^2] < \mathbb{E}[\phi(v)] \qquad \text{(by } \phi''(v) < 0 \text{ and } \mathbb{E}[\xi] = 0.)$$

With more detailed analysis, we can formalize the Taylor expansion and control the decrease of the potential function, and conclude that $\mathbb{E}[\phi(v)]$ converges to zero. Then, by Markov's inequality, with high probability, $\phi(v)$ is tiny and so is v.[6]

From the 1-D case to the high-dimensional case. In one dimension, the bias is introduced because of the varying scale of noise (i.e., the norm of the covariance). However, in the high dimensional case, the shape of the covariance also matters. For example, if we generalize the random walk (1) to high-dimensions by running d of the random walks in parallel, then we will observe the same phenomenon, but the noise variances in different dimensions are *not identical* — they depend on the current scales of the coordinates. (Precisely, the noise variance for dimension k is $\eta^2 v_k^2$.) However, suppose we instead add noise of the *same variance* to all dimensions. Even if this variance depends on the norm of v (say, $\eta^2\|v\|_2^2$), the implicit bias will be diminished, as the smaller coordinates will have relatively outsized noise and the larger coordinates will have relatively insufficient noise.

Outline of the rest of the subsections. We will give a proof sketch of Theorem 1 that consists of three stages. We first show in the initial stage of the training that label noise effectively decreases

---

6. The same proof strategy fails for the Brownian motion because v is not always nonnegative, and there is no concave potential function over the real that can be bounded from below.

the parameter on all dimensions, bringing the training from large initialization to a small initialization regime, where better generalization is possible (Section 3.2). Then, we show that when the parameter is decently small, with label noise and a *decayed learning rate*, the algorithm will increase the magnitude of those dimensions in support set of $v^\star$, while keep decreasing the norm of the rest of the dimensions(Section 3.3). Finally, with one more decay, the algorithm can recover the ground truth(Section 3.4).

3.2. Stage 0: Label Noise with Large Learning Rate Reduces the Parameter Norm

We first analyze the initial phase where we use a relatively large learning rate. When the initialization is of a decent size, GD quickly *overfits* to a bad global minimum nearest to the initialization. In contrast, we prove that SGD with label noise biases towards the small norm region, for a similar reason as the random walk example with parameter-dependent noise in Section 3.1.

Theorem 3 *In the setting of Theorem 1, recall that we initialize with $v^{[0]} = \tau \cdot \mathbb{1}$. Assume $n \geq \Theta(\log d)$. Suppose we run SGD with label noise with noise level $\delta \geq \widetilde{\Theta}(\tau^2 d^2)$ and learning rate $\eta_0 \in [\widetilde{\Theta}(\tau^2 d^2/\delta^2), \widetilde{\Theta}(1/\delta)]$ for $T_0 = \widetilde{\Theta}(1/(\eta^2\delta^2))$ iterations. Then, with probability at least 0.99 over the randomness of the algorithm,*

$$\|v^{[T_0]}\|_\infty \leq 1/d. \tag{7}$$

Moreover, the minimum entry of $v^{[T_0]}$ is bounded below by $\exp(-\widetilde{O}((\eta\delta)^{-1}))$.

We remark that our requirement of η being large is consistent with the empirical observation that a large initial learning rate helps generalization Goyal et al. (2017); Li et al. (2019b). We provide intuitions and a proof sketch of the theorem in the rest of the subsection and defer the full proof to Section B . Our proof is based on the construction of a *concave* potential function Φ similar to Section 3.1. We will show that, at every step, the noise has a second-order effect on the potential function and decrease the potential function by a quantity on the order of $\eta^2\delta^2$ (omitting the d dependency).[7] On the other hand, the gradient step may increase the potential by a quantity at most on the order of η (omitting d dependency again). Therefore, when $\eta^2\delta^2 \gtrsim \eta$, we expect the algorithm to decrease the potential and the parameter norm.

In particular, we define $\Phi(v) \triangleq \sum_{k=1}^d \phi(v_k) = \sum_{k=1}^d \sqrt{v_k}$. By the update rule 2.1, the update for a coordinate $k \in [d]$ can be written as

$$v_k^{[t+1]} \leftarrow v_k^{[t]} - \eta s_t x_k^{(i_t)} v_k^{[t]} - \eta^{[t]} \left((v^{[t]\odot 2} - v^{\star\odot 2})^\top x^{(i_t)} \right) x_k^{(i_t)} v_k^{[t]}, \tag{8}$$

where s_t is sampled from $\{-\delta, \delta\}$ and i_t is sampled from $[n]$. Let $g_k^{(i_t)} \triangleq ((v^{[t]\odot 2} - v^{\star\odot 2})^\top x^{(i_t)})x_k^{(i_t)}$ be the component coming from the stochastic gradient. Using the fact that $\phi(ab) = \phi(a)\phi(b)$ for any $a, b > 0$, we can evaluate the potential function at time $t + 1$,

$$\mathbb{E}\left[\phi(v_k^{[t+1]})\right] = \mathbb{E}\left[\phi(v_k^{[t]})\phi(1 - \eta s_t x_k^{(i_t)} - \eta g_k^{(i_t)})\right] = \phi(v_k^{[t]})\mathbb{E}\left[\phi(1 - \eta s_t x_k^{(i_t)} - \eta g_k^{(i_t)})\right]. \tag{9}$$

---

7. In general, any mean-zero noise has a second-order effect on any potential function. Therefore, when the noise level is fixed, as $\eta \to 0$, the effect of the noise diminishes. This is why a lower bound on the learning rate is necessary for the noise to play a role.

Here the expectation is over s_t and i_t. We perform Taylor-expansion on the term $\phi(1 - \eta s_t x_k^{(i_t)} - \eta g_k^{(i_t)})$ to deal with the non-linearity and use the fact that $\eta s_t x_k^{(i_t)}$ is mean-zero:

$$\mathbb{E}\left[\phi(1 - \eta s_t x_k^{(i_t)} - \eta g_k^{i_t})\right] \approx \phi(1) - \phi'(1)\eta\mathbb{E}\left[g_k^{i_t}\right] + \frac{1}{2}\phi''(1)\mathbb{E}\left[\left(\eta s_t x_k^{(i_t)} - \eta g_k^{(i_t)}\right)^2\right]$$

$$\leq \phi(1) - \phi'(1)\eta\mathbb{E}\left[g_k^{i_t}\right] + \frac{1}{2}\phi''(1)\mathbb{E}\left[\left(\eta s_t x_k^{(i_t)}\right)^2\right]$$

$$\leq \phi(1) - \phi'(1)\eta\mathbb{E}\left[g_k^{i_t}\right] - \Omega(\eta^2\delta^2). \tag{10}$$

In the second line we used $\phi''(1) < 0$ from the concavity and $\mathbb{E}[\eta s_t x_k^{(i_t)}] = 0$, and the third line uses the fact that $s_t \sim \{\pm\delta\}$ and $\mathbb{E}[x_k^{(i_t)2}] \approx 1$ (by the data assumption). The rest of the proof consists of bounding the second term in equation (10) from above to show the potential function is contracting.

We first note for every i_t, it holds that $|g_k^{i_t}| \leq \|v^{[t]\odot 2} - v^{\star\odot 2}\|_1 \|x^{(i_t)}\|_\infty^2 \leq (\|v^{[t]}\|_2^2 + r)\|x^{(i_t)}\|_\infty^2$. Furthermore, we can bound the ℓ_2 norm of $v^{[t]}$ with the following lemma:

Lemma 4 *In the setting of Theorem 3, for some failure probability $\rho > 0$, let $b_0 \triangleq 6\tau d/\rho$. Then, with probability at least $1 - \rho/3$, we have that $\|v^{[t]}\|_2 \leq b_0$ for any $t \leq T_0$.*

Note that $v^{[0]}$ has ℓ_2 norm $\tau\sqrt{d}$, and here we prove that the norm does not exceed τd with high probability. At the first glance, the lemma appears to be mostly auxiliary, but we note that it distinguishes label noise from Gaussian noise, which empirically causes the parameter to blow up as shown in Figure 1. The formal proof is deferred to Section B.

By Lemma 4 and the bound on $|g_k^{i_t}|$ in terms of $\|v^{[t]}\|_2$, we have $|g_k^{i_t}| \leq (b_0^2 + r)\|x^{(i_t)}\|_\infty^2 \leq \widetilde{O}(b_0^2 + r)$ with b_0 defined in Lemma 4 (up to logarithmic factors). Here we use again that each entry of the data is from $\mathcal{N}(0, 1)$. Plugging these into equation (10) we obtain

$$\mathbb{E}\left[\phi(1 - \eta s_t x_k^{(i_t)} - \eta g_k^{i_t})\right] \leq 1 + \eta\widetilde{O}(b_0^2 + r) - \Omega(\eta^2\delta^2) < 1 - \Omega(\eta^2\delta^2)$$

where in the last inequality we use the lower bound on η to conclude $\eta^2\delta^2 \gtrsim \eta\widetilde{O}(b_0^2 + r)$. Therefore, summing equation (9) over all the dimensions shows that the potential function decreases exponentially fast: $\mathbb{E}[\Phi(v^{[t+1]})] < (1 - \Omega(\eta^2\delta^2))\Phi(v^{[t]})$. After $T \approx \log(d)/(\eta^2\delta^2)$ iterations, $v^{[T]}$ will already converge to a position such that $\mathbb{E}[\Phi(v^{[T]})] \lesssim \sqrt{1/d}$, which implies $\|v^{[T]}\|_\infty \lesssim 1/d$ with probability at least $1 - \rho$ and finishes the proof.

3.3. Stage 1: Getting Closer to $v^\star$ with Annealed Learning Rate

Theorem 3 shows that the noise decreases the ∞-norm of v to $1/d$. This means that ℓ_1 or ℓ_2-norm of v is similar to or smaller than that of $v^\star$ if r is constant, and we are in a small-norm region where overfitting is less likely to happen. In the next stage, we anneal the learning rate to slightly reduce the bias of the label noise and increase the contribution of the signal. Recall that $v^\star$ is a sparse vector with support $S \subset [d]$. The following theorem shows that, after annealing the learning rate (from the order of $1/\delta^2$ to $1/\delta$), SGD with label noise increases entries in v_S and decreases entries in $v_{\bar{S}}$ simultaneously, provided that the initialization has ℓ_∞-norm bounded by $1/d$. (For simplicity and self-containedness of the statement, we reset the time step to 0.)

Theorem 5 *In the setting of Section 2.1, given a target error bound $\epsilon_1 > 0$, we assume that $n \geq \widetilde{\Theta}(r^2 \log^2(1/\epsilon_1))$. We run SGD with label noise (Algorithm 1) with an initialization $v^{[0]}$ whose entries are all in $[\epsilon_{\min}, 1/d]$, where $\epsilon_{\min} \geq \exp(-\widetilde{O}(1))$. Let noise level $\delta \geq \widetilde{\Theta}(\log(1/\epsilon_1))$ and learning rate $\eta = \widetilde{\Theta}(1/\delta^2)$, and number of iterations $T = \widetilde{\Theta}(\log(1/\epsilon_1)/\eta)$. Then, with probability at least 0.99, after T iterations, we have*

$$\|v_S^{[T]} - v_S^\star\|_\infty \leq 0.1 \text{ and } \|v_{\bar{S}}^{[T]} - v_{\bar{S}}^\star\|_1 \leq \epsilon_1. \tag{11}$$

We remark that even though the initialization is relatively small in this stage, the label noise still helps alleviate the reliance on small initialization. Li et al. (2017); Vaskevicius et al. (2019) showed that GD converges to the ground truth with sufficiently small initialization, which is required to be smaller than target error ϵ_1. In contrast, our result shows that with label noise, the initialization does not need to depend on the target error, but only need to have an ℓ_∞-norm bound on the order of $1/d$. In other words, v gets closer to $v^\star$ on both S and $\bar{S}$ in our case, whereas in Li et al. (2017); Vaskevicius et al. (2019) the $v_{\bar{S}}$ grows slowly.

The proof of this theorem balances the contribution of the gradient against that of the noise on S and $\bar{S}$. On S, the gradient provides a stronger signal than label noise, whereas on $\bar{S}$, the implicit bias of the noise, similar to the effect in Section 3.2, outweighs the gradient and reduces the entries to zero. The analysis is more involved than that of Theorem 3, and we defer the full proof to Section C.

3.4. Stage 2: Convergence to the ground-truth $v^\star$

The conclusion of Theorem 5 still allows constant error in the support, namely, $\|v_S - v_S^\star\|_\infty \leq 0.1$. The following theorem shows that further annealing the learning rate will let the algorithm fully converge to $v^\star$ with any target error ϵ.

Theorem 6 *[informal version of Theorem 18] Assume initialization $v^{[0]}$ satisfies $\|v_S^{[0]} - v_S^\star\|_\infty \leq 0.1$. Suppose we run SGD with label noise with any noise level $\delta \geq 0$ and small enough learning rate η for $T = \Theta(1/\eta)$ iterations. Then, with high probability over the randomness of the algorithm and data, there is $\|v_S^{[T]} - v_S^\star\|_\infty \leq \|v_S^{[0]} - v_S^\star\|_\infty/10$.*

The formal version of Theorem 6 and its proof can be found in Section D.

Proof of Theorem 1. In Section E of Appendix, we combine Theorem 3, Theorem 5, and Theorem 18 to prove our main Theorem 1.

4. Analysis Overview of Langevin Dynamics (Theorem 2)

To prove Theorem 2, recall that we would like to show that $\int_{\mathbb{R}^d} e^{-\lambda \mathcal{L}(v)}$ is infinite. Our approach will be to change variables to $u = v^{\odot 2}$ and compute this integral over u. First, we note that all such u must lie in the convex cone where each coordinate is positive. Second, we observe that the loss $\mathcal{L}(u)$ is invariant in the affine space $u + X^\perp$, where $X^\perp$ is the orthogonal subspace to the data.

Thus, for some fixed u', we have

$$\int_{v>0} e^{-\lambda\mathcal{L}(v)}dv > \int_{u>0,u\in u'+X^\perp} e^{-\lambda\mathcal{L}(u')}|\det\frac{\partial v}{\partial u}|du \tag{12}$$

$$= \frac{e^{-\lambda\mathcal{L}(u')}}{2^d}\int_{u>0,u\in u'+X^\perp}\prod_{i=1}^{d}\frac{1}{\sqrt{u_i}}du. \tag{13}$$

Thus, the aim is to show that with high probability over the data, for any choice of u', the integral $\int_{u>0,u\in u'+X^\perp}|\det\frac{\partial v}{\partial u}|du$ is infinite. To this end, we will perform another change of variables $u = u' + A\tilde{u}$ where $\tilde{u} \in \mathbb{R}^{d-n}$, and $A = [a^{(1)},\ldots,a^{(d-n)}] \in \mathbb{R}^{d\times d-n}$ is a specially constructed matrix whose columns form an orthogonal basis for $X^\perp$. We will select $a^{(1)} = \mu$ where $\mu \in X^\perp$ and μ is positive in every dimension. The existence of such μ is guaranteed with high probability, as shown in Section F. Now by construction of A, we will always have $u' + A\tilde{u} > 0$ if $\tilde{u}_1$ is sufficiently large. Thus, there exists a convex cone $\{|\tilde{u}_i| \leq c\tilde{u}_1, \forall 2 \leq i \leq d-n\}$, such that every $\tilde{u}$ in this cone satisfies $u = u' + A\tilde{u} > 0$. Integrating (13) over this cone is similar to integrating a polynomial with degree $-d/2$ for $d - n$ times, and this integral can be shown to be infinite when $d > 2n$. The full proof is in Section F.

5. Conclusion

In this work, we study the implicit bias effect induced by noise. For a quadratically-parameterized model, we theoretically show that the parameter-dependent noise has a strong implicit bias, which can help recover the sparse ground-truth from limited data. In comparison, our negative result shows that such a bias cannot be induced by spherical Gaussian noise. Our result explains the empirical observation that replacing mini-batch noise or label noise with Gaussian noise usually leads to degradation in the generalization performance of deep models.

Acknowledgements

References

Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

Raman Arora, Peter Bartlett, Poorya Mianjy, and Nathan Srebro. Dropout: Explicit forms and capacity control. *arXiv preprint arXiv:2003.03397*, 2020.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7411–7422, 2019.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. *arXiv preprint arXiv:1904.09080*, 2019.

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *Discrete & Computational Geometry*, 59(4):757–783, 2018.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1565–1576, 2019.

Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018.

Xiang Cheng, Dong Yin, Peter L Bartlett, and Michael I Jordan. Stochastic gradient and langevin processes. *arXiv preprint arXiv:1907.03215*, 2019.

Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

Lenaic Chizat and Francis Bach. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 8, 2018.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 651–676, 2017.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, pages 384–395, 2018a.

Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018b.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems*, pages 9108–9118, 2019.

Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. *arXiv preprint arXiv:1909.12051*, 2019.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018a.

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

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

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pages 1731–1741, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018a.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018b.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. *arXiv preprint arXiv:1902.00621*, 2019a.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. *arXiv preprint arXiv:1712.09203*, 2017.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, pages 11669–11680, 2019b.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *arXiv preprint arXiv:1810.01075*, 2018.

Qi Meng, Shiqi Gong, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. Dynamic of stochastic gradient descent with state-dependent noise. *arXiv preprint arXiv:2006.13719*, 2020.

Poorya Mianjy and Raman Arora. On dropout and nuclear norm regularization. *arXiv preprint arXiv:1905.11887*, 2019.

Poorya Mianjy, Raman Arora, and Rene Vidal. On the implicit bias of dropout. *arXiv preprint arXiv:1806.09777*, 2018.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. *arXiv preprint arXiv:1707.05947*, 2017.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638, 2018.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro HP Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. *arXiv preprint arXiv:1803.01905*, 2018.

Mor Shpigel Nacson, Suriya Gunasekar, Jason D Lee, Nathan Srebro, and Daniel Soudry. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. *arXiv preprint arXiv:1905.07325*, 2019.

Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. In *Advances in Neural Information Processing Systems*, pages 11013–11023, 2019.

Behnam Neyshabur, Russ R Salakhutdinov, and Nati Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2015.

Tomaso Poggio, Kenji Kawaguchi, Qianli Liao, Brando Miranda, Lorenzo Rosasco, Xavier Boix, Jack Hidary, and Hrushikesh Mhaskar. Theory of deep learning iii: explaining the non-overfitting puzzle. *arXiv preprint arXiv:1801.00173*, 2017.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13 (Feb):389–427, 2012.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *arXiv preprint arXiv:2005.06398*, 2020.

Gareth O Roberts, Richard L Tweedie, et al. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Christopher J Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E Dahl. Measuring the effects of data parallelism on neural network training. *arXiv preprint arXiv:1811.03600*, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2017.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for stochastic gradient langevin dynamics. *The Journal of Machine Learning Research*, 17(1):193–225, 2016.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. In *Advances in Neural Information Processing Systems*, pages 2968–2979, 2019.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204, 2019.

Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. In *Advances in Neural Information Processing Systems*, pages 9722–9733, 2019.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pages 9709–9721, 2019.

Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. *arXiv preprint arXiv:2002.12915*, 2020.

Mingwei Wei and David J Schwab. How noise affects the hessian spectrum in overparameterized neural networks. *arXiv preprint arXiv:1910.00195*, 2019.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

Yeming Wen, Kevin Luk, Maxime Gazeau, Guodong Zhang, Harris Chan, and Jimmy Ba. Interplay between optimization and generalization of stochastic gradient descent with covariance noise. *arXiv preprint arXiv:1902.08234*, 2019.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pages 4148–4158, 2017.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. *arXiv preprint arXiv:2002.09277*, 2020.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent escapes from sharp minima exponentially fast. *arXiv preprint arXiv:2002.03495*, 2020.

Sho Yaida. Fluctuation-dissipation relations for stochastic gradient descent. *arXiv preprint arXiv:1810.00004*, 2018.

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

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient langevin dynamics. *arXiv preprint arXiv:1702.05575*, 2017.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. 2019.

Appendix A. Experimental Details

A.1. Experimental Details for the Quadratically-Parameterized Model

In the experiment of our quadratically-parameterized model, we use a 100-dimensional model with $n = 40$ data randomly sampled from $\mathcal{N}(0, \mathcal{I}_{100 \times 100})$. We set the first 5 dimensions of the ground-truth $v^\star$ as 1, and the rest dimensions as 0. We always initialize with $v^{[0]} = \mathbb{1}$. We use a constant learning rate 0.01 for all the experiments except for label noise. For label noise, we start from 0.01 and then decay the learning rate by a factor of 10 after 1×10^5 and 2×10^5 iterations. For "full batch" experiment, we run full batch gradient descent without noise. For "small batch" experiment, in order to fully disentangle the effect of learning rate and mini-batch sgd noise (i.e., to avoid implicit biases from large lr rather than noise), we add small batch noise to full gradient via the following sampling method: for each iteration, we randomly sample two data i and j from $[n]$, and add $\delta(\nabla \ell^{(i)}(v) - \nabla \ell^{(j)}(v))$ to the full gradient (we set $\delta = 1.0$ in our experiment). For label noise, we randomly sample $i \in [n]$ and $s \in \{\delta, -\delta\}$ (we set $\delta = 1.0$ in our experiment), and add noise $\nabla \tilde{\ell}^{(i)}(v) - \nabla \ell^{(i)}(v)$ to full gradient, where $\tilde{\ell}^{(i)}(v) \triangleq \frac{1}{4}(f_v(x^{(i)}) - y^{(i)} - s)^2$. For Gaussian noise experiments, we add noise $\xi \sim \mathcal{N}(0, \sigma^2 \mathcal{I}_{d \times d})$ to full gradient every iteration, where the values of σ are shown in Figure 1. For experiments except for Gaussian noises, we train a total of 3×10^5 iterations. For a more generous comparison, we run all the Gaussian noise experiments for 4 times longer (i.e., 1.2×10^6 iterations) while plotting them in the same figure after scaling the x-axis by a factor of 4. The test error is measured by the square of ℓ_2 distance between $v^{\odot 2}$ and $v^{\star \odot 2}$, which is the same as the expectation of loss on a freshly randomly sampled data. The trianing and test error are plotted in Figure 1.

A.2. Experimental Details for Deep Neural Networks on CIFAR100

We train a VGG19 model (Simonyan and Zisserman, 2014) on CIFAR100, using a small and large batch baseline. We also experiment with adding Gaussian noise to the parameters after every gradient update as well as adding label noise in the following manner: with some probability that depends on the current iteration count, we replace the original label with a randomly chosen one.

To add additional mean-zero noise to the gradient which simulates the effect of label noise in the regression setting, we compute a noisy gradient of the cross-entropy loss ℓ_{ce} with respect to model output $f(x)$ as follows:

$$\tilde{\nabla}_{f(x)} \ell_{ce}(f(x), y) = \nabla_{f(x)} \ell_{ce}(f(x), y) + \sigma_{ln} z \tag{14}$$

where z is a 100-dimensional vector (corresponding to each class) distributed according to $\mathcal{N}(0, \mathcal{I}_{100 \times 100})$, and y is the (possibly flipped) label. We backpropagate using this noisy gradient when we compute the gradient of loss w.r.t. parameters for the updates. After tuning, we choose the initial label-flipping probability as 0.1, and reduce it by a factor of 0.5 every time the learning rate is annealed. We choose σ_{ln} such that $\sigma_{ln} \sqrt{\mathbb{E}[\|z\|_2^2]} = 0.1$, and also decrease σ_{ln} by a factor of 0.5 every time the learning rate is annealed.

To add spherical Gaussian noise to the parameter every update, we simply set $W \leftarrow W + \sigma z$ after every gradient update, where z is a mean-zero Gaussian whose coordinates are drawn independently from $\mathcal{N}(0, 1)$. We tune this σ over the values shown in Figure 1.

We turn off weight decay and BatchNorm to isolate the regularization effects of just the noise alone. Standard data augmentation is still present in our runs. Our small batch baseline uses a

batch size of 26, and our large batch baseline uses a batch size of 256. In runs where we add noise, the batch size is always 256. For all runs, we use an initial learning rate of 0.004. We train for 410550 iterations (i.e., minibatches), annealing the learning rate by a factor of 0.1 at the 175950-th and 293250-th iteration. Our models take around 20 hours to train on a single NVIDIA TitanXp GPU when the batch size is 256. The final performance gap between label noise or small minibatch training v.s. large batch or Gaussian noise is around 13% accuracy.

A.3. Additional Plots

Here we show empirical evidence that training with Gaussian noise fails to converge to a stationary distribution. We train VGG19 network on CIFAR100, and plot the norm of model weight along the training trajectory. As shown in Figure 2, the weigth norm of large batch (LB) and large batch with label noise (LB+LN) both converge to some finite value, while the weight norm of large batch with Gaussian noise (LB+GN) keeps increasing and fails to converge.
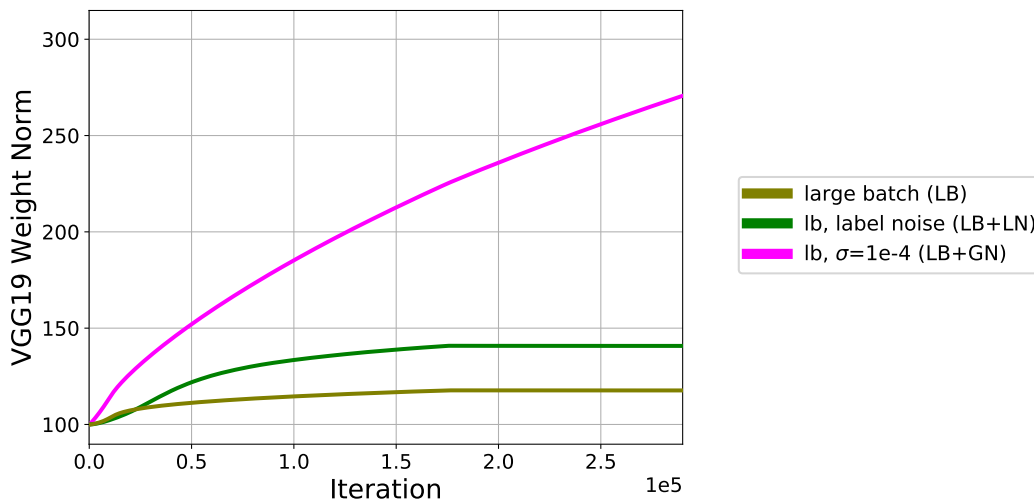

Figure 2: **The norm of model weight along training trajectory.** We demonstrate that the model weight fails to converge when training with Gaussian noise. In contrast, the weigth norm converges for training with label noise or without noise.

Appendix B. Proof of Theorem 3

In this section, we will first prove several lemmas on which the proof of Theorem 3 is built upon. Then we will provide a proof of Theorem 3.

Definition 7 *(b-bounded coupling) Let $v^{[0]}, v^{[1]}, \cdots, v^{[T]}$ be a trajectory of label noise gradient descent with initialization $v^{[0]}$. We call the following random sequence $\tilde{v}^{[t]}$ a b-bounded coupling of $v^{[t]}$: starting from $\tilde{v}^{[0]} = v^{[0]}$, for each time $t < T$, if $\left\|\tilde{v}^{[t]}\right\|_1 \leq b$, we let $\tilde{v}^{[t+1]} \triangleq v^{[t+1]}$; otherwise if $\left\|\tilde{v}^{[t]}\right\|_1 > b$ we don't update, i.e., $\tilde{v}^{[t+1]} \triangleq \tilde{v}^{[t]}$.*

Lemma 8 *In the setting of Theorem 3, assume $\left\|x^{(i)}\right\|_\infty \leq b_x$ for any $i \in [n]$. Let $\eta \leq \frac{\rho}{6Tb_x^2(b_0^2+r)}$, where $b_0 = \frac{6\tau d}{\rho}$. Let $\tilde{v}^{[t]}$ be the b_0-bounded coupling of $v^{[t]}$. If $\tilde{v}^{[t]}$ is always positive on each dimension, then with probability at least $1 - \frac{\rho}{3}$, there is*

$$\left\|\tilde{v}^{[T]}\right\|_1 \leq b_0. \tag{15}$$

Proof [Proof of Lemma 8] Recall the update at t-th iteration is:

$$v^{[t+1]} = v^{[t]} - \eta((v^{[t]\odot 2} - v^{\star\odot 2})^\top x^{(i_t)})x^{(i_t)} \odot v^{[t]} - \eta s_t x^{(i_t)} \odot v^{[t]}. \tag{16}$$

We first bound the increase of $\left\|\tilde{v}^{[t]}\right\|_1$ in expectation. When $\left\|\tilde{v}^{[t]}\right\|_1 \leq b_0$, there is:

$$\mathbb{E}\left[\tilde{v}_k^{[t+1]}\right] = \tilde{v}_k^{[t]} - \eta\mathbb{E}[((\tilde{v}^{[t]\odot 2} - v^{\star\odot 2})^\top x^{(i)})x_k^i \tilde{v}_k^{[t]}] \tag{17}$$

$$\leq \tilde{v}_k^{[t]} + \eta(\left\|\tilde{v}^{[t]\odot 2}\right\|_1 + \left\|v^{\star\odot 2}\right\|_1)b_x^2 \tilde{v}_k^{[t]} \tag{18}$$

$$\leq \tilde{v}_k^{[t]} + \eta(b_0^2 + r)b_x^2 \tilde{v}_k^{[t]} \tag{19}$$

where the first inequality is because we can separate the last term into $v^{[t]\odot 2}$ part and $v^{\star\odot 2}$ part and bound them with $\left\|v^{[t]\odot 2}\right\|_1$ and $\left\|v^{\star\odot 2}\right\|_1$ respectively, the second inequality is by $\left\|v^{[t]}\right\|_2^2 \leq \left\|v^{[t]}\right\|_1^2$ and sparsity of $v^\star$. So summing over all dimensions we have $\mathbb{E}\left[\left\|\tilde{v}^{[t+1]}\right\|_1\right] \leq \left\|\tilde{v}^{[t]}\right\|_1 + \eta b_0 b_x(b_0^2 + r)$. This bound is obviously also true when $\left\|\tilde{v}^{[t]}\right\|_1 > b_0$, in which case $\tilde{v}^{[t+1]} = \tilde{v}^{[t]}$.

We then bound the probability of $\left\|\tilde{v}^{[T]}\right\|_1$ being too large:

$$\Pr\left(\left\|\tilde{v}^{[T]}\right\|_1 > b_0\right) \leq \frac{\mathbb{E}\left[\left\|\tilde{v}^{[T]}\right\|_1\right]}{b_0} \tag{20}$$

$$\leq \frac{\tau d + T\eta b_0 b_x^2(b_0^2 + r)}{b_0} \tag{21}$$

$$\leq \frac{\rho}{3}, \tag{22}$$

where the first inequality is Markov Inequality, the second is by the previous equation, and the third is by assumption of η and the definition of b_0. ∎

Proof [Proof of Lemma 4] Notice that when $\left\|\tilde{v}^{[T]}\right\|_1 \leq b_0$, there is $v^{[T]} = \tilde{v}^{[T]}$, Lemma 4 naturally follows from Lemma 8. ∎

Definition 9 *(b-bounded potential function) For a vector v that is positive on each dimension, we define the b-bounded potential function $\Phi(v)$ as follows: if $\|v\|_1 \leq b$, we let $\Phi(v) \triangleq \sum_{k=1}^d \sqrt{v_k}$; otherwise $\Phi(v) \triangleq 0$.*

Lemma 10 *In the setting of Theorem 3, let $\epsilon_0 = 1/d$. Assume $\left\|x^{(i)}\right\|_\infty \leq b_x$ for $i \in [n]$ with some $b_x > 0$, $\mathbb{E}_i[(x_k^{(i)})^2] \geq \frac{2}{3}$ for all $k \in [d]$. Let $b_0 = \frac{6\tau d}{\rho}$. Assume $\eta \delta b_x + \eta(b_0^2 + r)b_x^2 \leq \frac{1}{16}$, $\eta \delta^2 \geq 32(b_0^2 + r)b_x^2$ and $T = \lceil \frac{32}{\eta^2 \delta^2} \log(\frac{\rho\sqrt{\epsilon_0}}{3d\sqrt{\tau}}) \rceil$. Let $\tilde{v}^{[t]}$ be the b_0-bounded coupling of $v^{[t]}$, and $\Phi(\cdot)$ is the b_0-bounded potential function. If $\tilde{v}^{[t]}$ is always positive on each dimension, then with probability at least $1 - \frac{\rho}{3}$, there is*

$$\Phi(\tilde{v}^{[T]}) \leq \sqrt{\epsilon_0}. \tag{23}$$

Proof [Proof of Lemma 10] We first show $\Phi(\tilde{v}^{[t]})$ decreases exponentially in expectation. If $\left\|\tilde{v}^{[t]}\right\|_1 \leq b_0$, we have:

$$\mathbb{E}\left[\Phi(\tilde{v}^{[t+1]})\right] \leq \sum_{k=1}^d \mathbb{E}\left[\sqrt{\tilde{v}_k^{[t+1]}}\right] \tag{24}$$

$$= \sum_{k=1}^d \mathbb{E}_{s_t, i_t}\left[\sqrt{\tilde{v}_k^{[t]} - \eta s_t x_k^{(i_t)} \tilde{v}_k^{[t]} - \eta((\tilde{v}^{[t] \odot 2} - v^{\star \odot 2})^\top x^{(i_t)}) x_k^{(i_t)} \tilde{v}_k^{[t]}}\right] \tag{25}$$

$$\leq \sum_{k=1}^d \sqrt{\tilde{v}_k^{[t]}} \mathbb{E}_{s_t, i_t}\left[\sqrt{1 + \eta s_t x_k^{(i_t)} + \eta(b_0^2 + r)b_x^2}\right] \tag{26}$$

$$\tag{27}$$

where the second inequality is because $\left\|\tilde{v}^{[t]}\right\|_2^2 = \left\|\tilde{v}^{[t]}\right\|_1^2 \leq b_0^2$. Toward bounding the expectation, we notice that by Taylor expansion theorem, there is for any general function $g(x) = \sqrt{1+x}$, there is

$$g(1+x) \leq g(1) + g'(1)x + \frac{1}{2}g''(1)x^2 + \frac{M}{6}|x|^3, \tag{28}$$

where M is upper bound on $|g'''(1+x')|$ for x' in 0 to x, which is less than 3 if $|x| \leq \frac{1}{2}$. So in our theorem if $\Delta \triangleq \eta s_t x_k^{(i_t)} + \eta(b_0^2 + r)b_x^2 \in [-\frac{1}{2}, \frac{1}{2}]$, we have

$$\sqrt{1 + \eta s_t x_k^{(i_t)} + \eta(b_0^2 + r)b_x^2} \leq 1 + \frac{1}{2}\Delta - \frac{1}{8}\Delta^2 + \frac{1}{2}|\Delta|^3. \tag{29}$$

Also since $\mathbb{E}_{s_t, i_t}[\Delta] = \eta(b_0^2 + r)b_x^2$, $\mathbb{E}_{s_t, i_t}[\Delta^2] \geq \eta^2 \delta^2 \mathbb{E}_{i_t}[(x_k^{(i_t)})^2] \geq \frac{2}{3}\eta^2 \delta^2$, we have when $|\Delta| \leq \frac{1}{16}$ and $\eta \delta^2 \geq 32(b_0^2 + r)b_x^2$, we have $\mathbb{E}_{s_t, i_t}[\sqrt{1+\Delta}] \leq 1 - \mathbb{E}_{s_t, i_t}[1 - \frac{1}{16}\Delta^2] \leq 1 - \frac{1}{32}\eta^2 \delta^2$. So

$$\mathbb{E}\left[\Phi(\tilde{v}^{[t+1]})\right] \leq (1 - \frac{1}{32}\eta^2 \delta^2)\Phi(\tilde{v}^{[t]}). \tag{30}$$

Also notice that when $\left\|\tilde{v}^{[t]}\right\|_1 > b_0$, there is $\Phi(\tilde{v}^{[t+1]}) = \Phi(\tilde{v}^{[t]}) = 0$, so obviously we have $\mathbb{E}[\Phi(\tilde{v}^{[t+1]})] \leq (1 - \frac{1}{32}\eta^2 \delta^2)\Phi(\tilde{v}^{[t]})$ always true.

Next we prove that $\Phi(\tilde{v}^{[T]}) \leq \sqrt{\epsilon_0}$ with probability more than $1 - \frac{\rho}{2}$. This is because:

$$\Pr\left(\Phi(\tilde{v}^{[T]}) > \sqrt{\epsilon_0}\right) \leq \frac{\mathbb{E}\left[\Phi(\tilde{v}^{[T]})\right]}{\sqrt{\epsilon_0}} \tag{31}$$

$$\leq \frac{(1 - \frac{1}{32}\eta^2\delta^2)^T d\sqrt{\tau}}{\sqrt{\epsilon_0}} \tag{32}$$

$$\leq \frac{\rho}{3}. \tag{33}$$

where the first inequality if by Markov Inequaltiy, the second inequality is by the previous inequality, and the last inequality is because $T = \lceil \frac{32}{\eta^2\delta^2}\log(\frac{3d\sqrt{\tau}}{\rho\sqrt{\epsilon_0}})\rceil$. $\blacksquare$

Proof [Proof of Theorem 3] Let $\rho = 0.01$, $\epsilon_0 = 1/d$. By Lemma 25, and Lemma 26, when $n \geq \Theta(\log d)$, with probability at least $1 - \frac{\rho}{3}$ there is $\left\|x^{(i)}\right\|_\infty \leq b_x$ for all $i \in [n]$ with some $b_x = \Theta(\sqrt{\log(nd)})$, and $\mathbb{E}_i[(x_k^{(i)})^2] \geq \frac{2}{3}$ for all $k \in [d]$.

Let $b_0 = \frac{6\tau d}{\rho}$. We try to define η and δ such that when $T = \lceil \frac{32}{\eta^2\delta^2}\log(\frac{3d\sqrt{\tau}}{\rho\sqrt{\epsilon_0}})\rceil$, the assumptions $\eta \leq \frac{\rho}{6Tb_x^2(b_0^2+r)}$ and $\tilde{v}^{[t]}$ always being positive in Lemma 8 and assumptions $\eta\delta b_x + \eta(b_0^2+r)b_x^2 \leq \frac{1}{16}$ and $\eta\delta^2 \geq 32(b_0^2 + r)b_x^2$ in Lemma 10 are satisfied.

Assume $\delta \geq 6\times 32^2 b_x^3 \frac{(b_0^2+r)}{\rho}\log(\frac{3d\sqrt{\tau}}{\rho\sqrt{\epsilon_0}})$, then we only need $\eta \in \left[\frac{6\times 32 b_x^2}{\rho\delta^2}(b_0^2 + r)\log(\frac{3d\sqrt{\tau}}{\rho\sqrt{\epsilon_0}}), \frac{1}{32\delta b_x}\right]$, and then all the above assumptions are satisfied.

Let $\tilde{v}^{[t]}$ be the b_0-bounded coupling of $v^{[t]}$. According to Lemma 10, we know with probability at least $1 - \frac{\rho}{3}$, $\Phi(\tilde{v}^{[T]}) \leq \sqrt{\epsilon_0}$, which means that either $\sum_{k=1}^d \sqrt{\tilde{v}_k^{[T]}} \leq \sqrt{\epsilon_0}$ or $\left\|\tilde{v}^{[T]}\right\|_1 > b_0$. According to Lemma 8, we know with probability at most $\frac{\rho}{3}$, $\left\|\tilde{v}^{[T]}\right\|_1 > b_0$. Combining these two statements, we know with probability at least $1 - \frac{2\rho}{3}$, $\left\|\tilde{v}^{[T]}\right\|_1 \leq b_0$ and $\sum_{k=1}^d \sqrt{\tilde{v}_k^{[T]}} \leq \sqrt{\epsilon_0}$. Notice that $\left\|\tilde{v}^{[T]}\right\|_1 \leq b_0$ implies $v^{[T]} = \tilde{v}^{[T]}$, while $\sum_{k=1}^d \sqrt{\tilde{v}_k^{[T]}} \leq \sqrt{\epsilon_0}$ implies $\tilde{v}_k^{[T]} \leq \epsilon_0$ for all dimension k, so we've finished the proof for the upper bound.

We then give a lower bound for each dimension of $\tilde{v}^{[T]}$. We can bound the decrease of any dimension k at time t:

$$\tilde{v}_k^{[t+1]} \geq (1 - \eta\delta - \eta(b_0^2 + r))\tilde{v}_k^{[t]} \tag{34}$$

$$\geq (1 - 2\eta\delta)\tilde{v}_k^{[t]}. \tag{35}$$

where the first inequality is by update rule and the second is because $\delta > (b_0^2 + r)$. Putting in the value of T, we have

$$\tilde{v}_k^{[T]} \geq (1 - 2\eta\delta)^T \tau \tag{36}$$

$$> \exp\left(-\frac{64}{\eta\delta}\log(\frac{3d\sqrt{\tau}}{\rho\sqrt{\epsilon_0}})\right). \tag{37}$$

$\blacksquare$

23

Appendix C. Proof of Theorem 5

In this section, we will first prove several lemmas on which the proof of Theorem 5 is built upon. Then we will provide a proof of Theorem 5.

Definition 11 $((b, \epsilon)$*-bounded coupling) Let* $v^{[0]}, v^{[1]}, \cdots, v^{[T]}$ *be a trajectory of label noise gradient descent with initialization* $v^{[0]}$. *Recall* $S \subset [d]$ *is the support set of* $v^\star$, *we notate* $\tilde{v}_S^{[t]}$ *a* r*-dimensional vector composed with those dimensions in* S *of* $\tilde{v}^{[t]}$, *and* $\tilde{v}_{\bar{S}}^{[t]}$ *the other* $d - r$ *dimensions. We call the following random sequence* $\tilde{v}^{[t]}$ *a* (b, ϵ)*-bounded coupling of* $v^{[t]}$: *starting from* $\tilde{v}^{[0]} = v^{[0]}$, *for each time* $t < T$, *if* $\left\| \tilde{v}_{\bar{S}}^{[t]} \right\|_1 \leq \epsilon$ *and* $\left\| \tilde{v}_S^{[t]} \right\|_\infty \leq b$, *we let* $\tilde{v}^{[t+1]} \triangleq v^{[t+1]}$; *otherwise* $\tilde{v}^{[t+1]} \triangleq \tilde{v}^{[t]}$.

Lemma 12 *In the setting of Theorem 5, let* $\rho \triangleq \frac{1}{100}$, $c_1 \triangleq \frac{1}{10}$, $\tilde{\epsilon}_1 \triangleq \frac{12}{\rho}$, $C_x \triangleq \max_{j \neq k} |\mathbb{E}_i[x_j^{(i)} x_k^{(i)}]|$. *Assume* $\left\| x^{(i)} \right\|_\infty \leq b_x$ *for* $i \in [n]$ *for some* $b_x > 0$, *and* $\mathbb{E}_i[(x_k^{(i)})^2] \geq \frac{2}{3}$ *for* $k \in [d]$. *Let* $\tilde{v}^{[t]}$ *be a* $(1 + c_1, \tilde{\epsilon}_1)$*-bounded coupling of* $v^{[t]}$. *Assume* $\frac{c_1^2}{8 \eta \delta^2 b_x^2} \geq \log \frac{6 r T^2}{\rho}$, $(\tilde{\epsilon}_1^2 + r) C_x b_x^2 \leq \frac{c_1}{20}$ *and* $\delta \geq b_x(\tilde{\epsilon}_1^2 + r)$. *Then, with probability at least* $1 - \frac{\rho}{6}$, *there is* $\left\| \tilde{v}_S^{[T]} \right\|_\infty \leq 1 + c_1$.

Proof [Proof of Lemma 12] For any fixed $1 \leq t_1 < t_2 \leq T$ and dimension $k \in S$, we consider the event that $\tilde{v}^{[t_1]} \in [1 + \frac{c_1}{3}, 1 + \frac{c_1}{2}]$, and at time t_2 it is the first time in the trajectory such that $\tilde{v}^{[t_2]} > 1 + c_1$. We first bound the probability of this event happens, i.e., the following quantity:

$$\Pr\left(\tilde{v}_k^{[t_2]} > 1 + c_1 \wedge \tilde{v}_k^{[t_1]} \leq 1 + \frac{c_1}{2} \wedge \tilde{v}_k^{[t_1:t_2]} \in [1 + \frac{c_1}{3}, 1 + c_1] \right), \tag{38}$$

where $\tilde{v}_k^{[t_1:t_2]} \in [1 + \frac{c_1}{3}, 1 + c_1]$ means that for all t such that $t_1 \leq t < t_2$, there is $1 + \frac{c_1}{3} \leq \tilde{v}_k^{[t]} \leq 1 + c_1$.

Notice that when $\left\| \tilde{v}_{\bar{S}}^{[t]} \right\|_1 \leq \tilde{\epsilon}_1$ and $\left\| \tilde{v}_S^{[t]} \right\|_\infty \leq 1 + c_1$ and $\tilde{v}_k^{[t_1:t+1]} \in [1 + \frac{c_1}{3}, 1 + c_1]$, there is

$$\mathbb{E}[\tilde{v}_k^{[t+1]} - 1] = \mathbb{E}_{s_t, i_t} \left[\left(1 + \eta s_t x_k^{i_t} - \eta((\tilde{v}^{[t] \odot 2} - v^{\star \odot 2})^\top x^{(i_t)}) x_k^{(i_t)} \right) \tilde{v}_k^{[t]} - 1 \right] \tag{39}$$

$$\leq (\tilde{v}_k^{[t]} - 1) - \frac{2}{3} \eta \tilde{v}_k^{[t]} (\tilde{v}_k^{[t]} + 1)(\tilde{v}_k^{[t]} - 1) + \eta(\tilde{\epsilon}_1^2 C_x + r C_x) b_x^2 \tilde{v}_k^{[t]} \tag{40}$$

$$\leq (1 - \eta)(\tilde{v}_k^{[t]} - 1). \tag{41}$$

where the first inequality is because $\left\| \tilde{v}_{\bar{S}}^{[t]} \right\|_2^2 \leq \tilde{\epsilon}_1^2$ and $\mathbb{E}_{i_t}[(x_k^{(i)})^2] \geq \frac{1}{2}$, the second inequality is because $(\tilde{\epsilon}_1^2 + r) b_x^2 C_x \leq \frac{c_1}{20}$. Also, we can bound the variance of this martingale as

$$\text{Var}\left[\tilde{v}_k^{[t+1]} - 1 \mid \tilde{v}^{[t]} - 1 \right] = \text{Var}\left[\eta s_t x_k^{(i_t)} \tilde{v}_k^{[t]} \right] + \text{Var}\left[\eta((\tilde{v}^{[t] \odot 2} - v^{\star \odot 2})^\top x^{(i_t)}) x_k^{(i_t)} \tilde{v}_k^{[t]} \right] \tag{42}$$

$$\leq (\eta \delta b_x (1 + c_1))^2 + \eta^2 (\tilde{\epsilon}_1^2 + r)^2 b_x^4 (1 + c_1)^2 \tag{43}$$

$$\leq 4 \eta^2 \delta^2 b_x^2, \tag{44}$$

where the first inequality is because $\eta s_t x_k^{(i_t)} \tilde{v}_k^{[t]}$ is mean-zero, the second inequality is by $\left\| x^{(i)} \right\|_\infty \leq b_x$, the third inequality is by $\delta \geq b_x(\tilde{\epsilon}_1^2 + r)$.

24

By Lemma 27, we have

$$\Pr(\tilde{v}_k^{[t_2]} - 1 > c_1) \tag{45}$$

$$\leq e^{\frac{-c_1^2}{8\eta^2\delta^2 b_x^2 \Sigma_{t=0}^{t_2-t_1-1}(1-\eta)^{2t}}} \tag{46}$$

$$\leq e^{\frac{-c_1^2}{8\eta\delta^2 b_x^2}}, \tag{47}$$

where the first inequality is by Lemma 27, the second inequality is by taking the sum of denominator.

Finally, we finish the proof with a union bound. Since if $\left\|\tilde{v}_S^{[T]}\right\|_\infty > 1 + c_1$, the event in Equation 38 has to happen for some $k \in S$ and $1 \leq t_1 < t_2 \leq T$, so we have

$$\Pr\left(\left\|\tilde{v}_S^{[T]}\right\|_\infty > 1 + c_1\right) \tag{48}$$

$$\leq \sum_{k \in S} \sum_{1 \leq t_1 < t_2 \leq T} \Pr\left(\tilde{v}_k^{[t_2]} > 1 + c_1 \wedge \tilde{v}_k^{[t_1]} \leq 1 + \frac{c_1}{2} \wedge \tilde{v}_k^{[t_1:t_2]} \in [1 + \frac{c_1}{3}, 1 + c_1]\right) \tag{49}$$

$$\leq rT^2 e^{\frac{-c_1^2}{8\eta\delta^2 b_x^2}} \tag{50}$$

$$\leq \frac{\rho}{6}, \tag{51}$$

where the last inequality is by assumption. ∎

Lemma 13 *In the setting of Lemma 12, assume $(\tilde{\epsilon}_1^2 + r)C_x \leq \frac{\rho}{12T\eta b_x^2}$. Then, with probability at least $1 - \frac{\rho}{6}$, there is $\left\|\tilde{v}_{\bar{S}}^{[T]}\right\|_1 \leq \tilde{\epsilon}_1$.*

Proof [Proof of Lemma 13] We first bound the increase of $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_1$ in expectation. When $\left\|\tilde{v}_S^{[t]}\right\|_\infty \leq 1 + c_1$ and $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_1 \leq \tilde{\epsilon}_1$, for any $k \notin S$, there is:

$$\mathbb{E}\left[\tilde{v}_k^{[t+1]}\right] = \tilde{v}_k^{[t]} - \eta\mathbb{E}_i\left[((\tilde{v}^{[t]\odot2} - v^{\star\odot2})^\top x^{(i)})x_k^{(i)}\tilde{v}_k^{[t]}\right] \tag{52}$$

$$\leq \tilde{v}_k^{[t]} + \eta(\tilde{\epsilon}_1^2 + r)C_x b_x^2 \tilde{v}_k^{[t]}. \tag{53}$$

because we can bound the dimensions in S and those not in S respectively. So summing over all dimensions not in S we have $\mathbb{E}\left[\left\|\tilde{v}_{\bar{S}}^{[t+1]}\right\|_1\right] \leq \left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_1 + \eta(\tilde{\epsilon}_1^2 + r)C_x b_x^2 \tilde{\epsilon}_1$. This bound is obviously also true when $\left\|\tilde{v}_S^{[t]}\right\|_\infty > 1 + c_1$ and $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_1 > \tilde{\epsilon}_1$.

We then bound the probability of $\left\|\tilde{v}_{\bar{S}}^{[T]}\right\|_1$ being too large:

$$\Pr\left(\left\|\tilde{v}_{\bar{S}}^{[T]}\right\|_1 > \tilde{\epsilon}_1\right) \leq \frac{\mathbb{E}\left[\left\|\tilde{v}_{\bar{S}}^{[T]}\right\|_1\right]}{\tilde{\epsilon}_1} \tag{54}$$

$$\leq \frac{1 + T\eta\tilde{\epsilon}_1(\tilde{\epsilon}_1^2 + r)C_x b_x^2}{\tilde{\epsilon}_1} \tag{55}$$

$$\leq \frac{\rho}{6}. \tag{56}$$

where the first inequality is Markov Inequality, the second is by $\left\|\tilde{v}_{\bar{S}}^{[0]}\right\|_1 \leq 1$ since every dimension is less than $1/d$, the third inequality is because $\tilde{\epsilon}_1 = \frac{12}{\rho}$ and $(\tilde{\epsilon}_1^2 + r)C_x \leq \frac{\rho}{12T\eta b_x^2}$. ∎

Lemma 14 *In the setting of Lemma 12, assume* $(\tilde{\epsilon}_1^2 + r)C_x b_x^2 < \frac{c_1}{12} - \frac{c_1^2}{4}$, $\eta\delta^2 \leq \frac{c_1}{8}$, $T\eta \geq \frac{16}{c_1}\log\frac{1}{\epsilon_{min}}$ *and* $\frac{T}{\delta^2} \geq \frac{2^9}{c_1^2}\log\frac{6r}{\rho}$. *Then, for any* $k \in S$, *with probability at least* $1 - \frac{\rho}{6r}$, *either* $\max_{t \leq T} \tilde{v}_k^{[t]} \geq 1 - \frac{c_1}{2}$, *or* $\left\|\tilde{v}_S^{[T]}\right\|_\infty > 1 + c_1$, *or* $\left\|\tilde{v}_{\bar{S}}^{[T]}\right\|_1 > \tilde{\epsilon}_1$.

Proof [Proof of Lemma 14] Fix $k \in S$. Let $\hat{v}^{[t]}$ be the following coupling of $\tilde{v}^{[t]}$: starting from $\hat{v}^{[0]} = \tilde{v}^{[0]}$, for each time $t < T$, if $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_1 \leq \tilde{\epsilon}_1$ and $\left\|\tilde{v}_S^{[t]}\right\|_\infty \leq 1 + c_1$ and $\tilde{v}_k^{[t]} \leq 1 - \frac{c_1}{2}$, we let $\hat{v}^{[t+1]} \triangleq \tilde{v}^{[t+1]}$; otherwise $\hat{v}^{[t+1]} \triangleq (1 + \frac{c_1}{2}\eta)\hat{v}^{[t]}$. Intuitively, whenever $\tilde{v}^{[t]}$ exceeds the proper range, we only times $\hat{v}^{[t]}$ by $1 + \frac{c_1}{2}\eta$ afterwards, otherwise we let it be the same as $\tilde{v}^{[t]}$.

We first show that $-t\log(1 + \frac{c_1}{2}\eta) + \log\hat{v}_k^{[t]}$ is a supermartingale, i.e., $\mathbb{E}[\log\hat{v}_k^{[t+1]} \mid \hat{v}^{[t]}] \geq \log(1 + \frac{c_1}{2}\eta) + \log\hat{v}_k^{[t]}$. This is obviously true if $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_1 > \tilde{\epsilon}_1$ or $\left\|\tilde{v}_S^{[t]}\right\|_\infty > 1 + c_1$ or $\tilde{v}_k^{[t]} > 1 - \frac{c_1}{2}$. Otherwise, there is

$$\mathbb{E}[\log\hat{v}_k^{[t+1]} \mid \hat{v}^{[t]}] = \mathbb{E}[\log\tilde{v}_k^{[t+1]} \mid \tilde{v}^{[t]}] \tag{57}$$

$$= \mathbb{E}_{s_t, i_t}\left[\log\left(1 + \eta s_t - \eta(\tilde{v}^{[t]\odot 2} - v^{\star\odot 2})^\top x^{(i_t)} x_k^{(i_t)}\right)\right] + \log\tilde{v}_k^{[t]} \tag{58}$$

$$\geq \mathbb{E}_{s_t}\left[\log\left(1 + \eta s_t + \frac{2}{3}\eta(1 - (\tilde{v}_k^{[t]})^2) - \eta(\tilde{\epsilon}_1^2 + r)C_x b_x^2\right)\right] + \log\tilde{v}_k^{[t]} \tag{59}$$

$$\geq \log(1 + \frac{c_1}{4}\eta) + \log\tilde{v}_k^{[t]}, \tag{60}$$

where the first inequality is by the update rule, the second inequality is because $(\tilde{\epsilon}_1^2 + r)C_x b_x^2 < \frac{c_1}{12} - \frac{c_1^2}{4}$ and $4\eta\delta^2 \leq \frac{c_1}{2}$ and $\delta \geq \tilde{\epsilon}_1^2 + r$. So by Azuma inequality, we have

$$\Pr\left(\hat{v}_k^{[T]} < 1 - \frac{c_1}{2}\right) \tag{61}$$

$$\leq e^{-\frac{2\left(T\log(1 + \frac{c_1}{4}\eta) + \log\epsilon_{min} - \log(1 - \frac{c_1}{2})\right)^2}{T(2\eta\delta)^2}} \tag{62}$$

$$\leq e^{-\frac{(\frac{1}{2}T\log(1 + \frac{c_1}{4}\eta))^2}{2T\eta^2\delta^2}} \tag{63}$$

$$\leq e^{-\frac{Tc_1^2}{2^9\delta^2}} \tag{64}$$

$$\leq \frac{\rho}{6r}. \tag{65}$$

where the first inequality is because Azuma inequality and $\mathrm{Var}[\log\hat{v}_k^{[t+1]} \mid \hat{v}^{[t]}] \leq (2\eta\delta)^2$, and the second inequality is because $T\log(1 + \frac{c_1}{4}\eta) \geq 2\log\frac{1}{\epsilon_{min}}$ which is true because $T\eta \geq \frac{16}{c_1}\log\frac{1}{\epsilon_{min}}$, the third inequality is because $\log(1 + \frac{c_1}{4}\eta) \geq \frac{c_1}{8}\eta$, the last inequality is because $\frac{T}{\delta^2} \geq \frac{2^9}{c_1^2}\log\frac{6r}{\rho}$. ∎

Lemma 15 *In the setting of Lemma 12, assume $(\tilde{\epsilon}_1^2 + r)C_x b_x^2 \leq \frac{c_1}{20}$ and $\frac{c_1^2}{8\eta\delta^2} \geq \log \frac{6rT^2}{\rho}$. Then, for any $k \in S$, with probability at least $1 - \frac{\rho}{6r}$, either $\max_{t<T} \tilde{v}_k^{[t]} < 1 - \frac{c_1}{2}$ or $\tilde{v}_k^{[T]} \geq 1 - c_1$.*

Proof [Proof of Lemma 15] For any fixed $1 \leq t_1 < t_2 \leq T$ and dimension $k \in S$, we consider the event that $\tilde{v}^{[t_1]} \in [1 - \frac{c_1}{2}, 1 - \frac{c_1}{3}]$, and at time t_2 it is the first time in the trajectory such that $\tilde{v}^{[t_2]} > 1 < c_1$. We first bound the probability of this event happens, i.e., the following quantity:

$$\Pr\left(\tilde{v}_k^{[t_2]} < 1 - c_1 \wedge \tilde{v}_k^{[t_1]} \geq 1 - \frac{c_1}{2} \wedge \tilde{v}_k^{[t_1:t_2]} \in [1 - c_1, 1 - \frac{c_1}{3}]\right), \tag{66}$$

where $\tilde{v}_k^{[t_1:t_2]} \in [1 - c_1, 1 - \frac{c_1}{3}]$ means that for all t such that $t_1 \leq t < t_2$, there is $1 - c_1 \leq \tilde{v}_k^{[t]} \leq 1 - \frac{c_1}{3}$.

Notice that when $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_1 \leq \tilde{\epsilon}_1$ and $\left\|\tilde{v}_S^{[t]}\right\|_\infty \leq 1 + c_1$ and $\tilde{v}_k^{[t_1:t+1]} \in [1 - c_1, 1 - \frac{c_1}{3}]$,

$$\mathbb{E}[1 - \tilde{v}_k^{[t+1]}] = \mathbb{E}_{s_t, i_t}\left[1 - (1 + \eta s_t x_k^{i_t} - \eta(\tilde{v}^{[t]\odot 2} - v^{\star\odot 2})^\top x^{(i_t)} x_k^{(i_t)})\tilde{v}_k^{[t]}\right] \tag{67}$$

$$\leq (1 - \tilde{v}_k^{[t]}) - \frac{2}{3}\eta\tilde{v}_k^{[t]}(\tilde{v}_k^{[t]} + 1)(1 - \tilde{v}_k^{[t]}) + \eta(\tilde{\epsilon}_1^2 + r)C_x b_x^2 \tilde{v}_k^{[t]} \tag{68}$$

$$\leq (1 - \eta)(1 - \tilde{v}_k^{[t]}). \tag{69}$$

where the first inequality is because $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_2^2 \leq \tilde{\epsilon}_1^2$, the second inequality is because $(\tilde{\epsilon}_1^2 + r)C_x b_x^2 \leq \frac{c_1}{20}$. Also, we can bound the variance of this martingale as

$$\text{Var}\left[1 - \tilde{v}_k^{[t]} \mid \tilde{v}_k^{[t]}\right] \leq (2\eta\delta)^2 . \tag{70}$$

By Lemma 27, we have

$$\Pr(1 - \tilde{v}_k^{[t_2]} > c_1) \tag{71}$$

$$\leq e^{\frac{-c_1^2}{8\eta^2\delta^2 \sum_{t=0}^{t_2-t_1-1}(1-\eta)^{2t}}} \tag{72}$$

$$\leq e^{\frac{-c_1^2}{8\eta\delta^2}}, \tag{73}$$

where the first inequality is by Lemma 27, the second inequality is by taking the sum of denominator.

Finally, we finish the proof with a union bound. Since if $\max_{t<T} \tilde{v}_k^{[t]} > 1 - \frac{c_1}{2}$ but $\tilde{v}_k^{[T]} < 1 - c_1$, the event in Equation 66 has to happen for some $1 \leq t_1 < t_2 \leq T$, so we have

$$\Pr\left(\max_{t<T} \tilde{v}_k^{[t]} > 1 - \frac{c_1}{2} \wedge \tilde{v}_k^{[T]} < 1 - c_1\right) \tag{74}$$

$$\leq \sum_{1 \leq t_1 < t_2 \leq T} \Pr\left(\tilde{v}_k^{[t_2]} < 1 - c_1 \wedge \tilde{v}_k^{[t_1]} \geq 1 - \frac{c_1}{2} \wedge \tilde{v}_k^{[t_1:t_2]} \in [1 - c_1, 1 - \frac{c_1}{3}]\right) \tag{75}$$

$$\leq T^2 e^{\frac{-c_1^2}{8\eta\delta^2}} \tag{76}$$

$$\leq \frac{\rho}{6r}. \tag{77}$$

$\blacksquare$

Definition 16 *$((b, \epsilon)$-bounded potential function) For a vector v that is positive on each dimension, we define the (b, ϵ)-bounded potential function $\Phi(v)$ as follows: if $\|v_{\bar{S}}\|_1 \leq \epsilon$ and $\|v_S\|_\infty \leq b$, we let $\Phi(v) \triangleq \sum_{k \notin S} \sqrt{v_k}$; otherwise $\Phi(v) \triangleq 0$.*

Lemma 17 *In the setting of Lemma 12, assume $\frac{2}{3}\eta\delta^2 > 32(\tilde{\epsilon}_1^2 + r)C_x b_x^2$ and $T\eta^2\delta^2 \geq 16\log\left(\frac{6\sqrt{d}}{\rho\sqrt{\epsilon_1}}\right)$. Then, with probability at least $1 - \frac{\rho}{6}$, there is $\Phi(\tilde{v}^{[T]}) \leq \sqrt{\epsilon_1}$.*

Proof [Proof of Lemma 17] We first show $\Phi(\tilde{v}^{[t]})$ decreases exponentially in expectation. For any $0 \leq t \leq T$, if $\left\|\tilde{v}_S^{[t]}\right\|_\infty \leq 1 + c_1$ and $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_1] \leq \tilde{\epsilon}_1$, we have:

$$\mathbb{E}\left[\Phi(\tilde{v}^{[t+1]})\right] \leq \sum_{k \notin S} \mathbb{E}\left[\sqrt{\tilde{v}_k^{[t+1]}}\right] \tag{78}$$

$$= \sum_{k \notin S}^d \mathbb{E}_{s_t, i_t}\left[\sqrt{\tilde{v}_k^{[t]} + \eta s_t x_k^{i_t} \tilde{v}_k^{[t]} - \eta(\tilde{v}^{[t]\odot 2} - v^{\star\odot 2})^\top x^{(i_t)} x_k^{(i_t)} \tilde{v}_k^{[t]}}\right] \tag{79}$$

$$\leq \sum_{k \notin S} \sqrt{\tilde{v}_k^{[t]}} \mathbb{E}_{s_t, i_t}\left[\sqrt{1 + \eta s_t x_k^{(i_t)} + \eta(\tilde{\epsilon}_1^2 + r)C_x b_x^2}\right] \tag{80}$$

$$\leq (1 - \frac{1}{16}\eta^2\delta^2)\Phi(\tilde{v}^{[t]}), \tag{81}$$

where the second inequality is because $\left\|\tilde{v}^{[t]}\right\|_2^2 \leq \tilde{\epsilon}_1^2$, the last inequality is by Taylor expansion and $\frac{2}{3}\eta\delta^2 > 32(\tilde{\epsilon}_1^2 + r)C_x b_x^2$. Also notice that when $\left\|\tilde{v}_S^{[t]}\right\|_\infty > 1 + c_1$ or $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_1 > \tilde{\epsilon}_1$, there is $\Phi(\tilde{v}^{[t+1]}) = p(\tilde{v}^{[t]}) = 0$, so obviously we have $\mathbb{E}[\Phi(\tilde{v}^{[t+1]})] \leq (1 - \frac{1}{16}\eta^2\delta^2)\Phi(\tilde{v}^{[t]})$ always true.

Next we bound the probability of $\Phi(\tilde{v}^{[T]}) \leq \sqrt{\epsilon_1}$:

$$\Pr(\Phi(\tilde{v}^{[T]}) > \sqrt{\epsilon_1}) \leq \frac{\mathbb{E}[\Phi(\tilde{v}^{[T]})]}{\sqrt{\epsilon_1}} \tag{82}$$

$$\leq \frac{(1 - \frac{1}{16}\eta^2\delta^2)^T \sqrt{d}}{\sqrt{\epsilon_1}} \tag{83}$$

$$\leq \frac{e^{-\frac{1}{16}T\eta^2\delta^2}\sqrt{d}}{\sqrt{\epsilon_1}} \tag{84}$$

$$\leq \frac{\rho}{6}. \tag{85}$$

where the first inequality if by Markov Inequaltiy, the second inequality is by the previous inequality and initially $\Phi(v^{[0]}) \leq \sqrt{d}$, the third is by $1 - x \leq e^{-x}$ for any $x \in \mathbb{R}$, and the last inequality is by $T\eta^2\delta^2 \geq 16\log\left(\frac{6\sqrt{d}}{\rho\sqrt{\epsilon_1}}\right)$. ∎

Proof [Proof of Theorem 5] Let $\rho = 0.01$, $c_1 = 0.1$, $\tilde{\epsilon}_1 \triangleq \frac{12}{\rho}$, $C_x \triangleq \max_{j \neq k} |\mathbb{E}_i[x_j^{(i)} x_k^{(i)}]|$. Let $b_x = \sqrt{2\log\frac{30d^2}{\rho}} = \widetilde{\Theta}(1)$. According to Lemma 25, when $n \leq d$, there is with probability at least $1 - \frac{\rho}{15}$ we have $\left\|x^{(i)}\right\|_\infty \leq b_x$ for $i \in [d]$.

Assume δ be positive number such that $\frac{16}{\delta^2} \log \frac{6\sqrt{d}}{\rho \epsilon_{min} \sqrt{\epsilon_1}} \leq 1$ and $\delta \geq b_x(\tilde{\epsilon}_1^2 + r)$. (since $\epsilon_{min} \geq \exp(-\widetilde{O}(1))$ this means $\delta \geq \widetilde{\Theta}(r + \log(1/\epsilon_1))$.) Let $P = \frac{c_1^2}{32\delta^2 b_x^2 \log \frac{5r}{\rho}}$, $Q = 2\log \frac{1}{P}$, $\eta = \min\{\frac{P}{Q}, \frac{16}{\delta^2}\} = \widetilde{\Theta}(\frac{1}{\delta^2})$, $T = \frac{16}{\eta^2 \delta^2} \log \frac{6\sqrt{d}}{\rho \epsilon_{min} \sqrt{\epsilon_1}} = \widetilde{\Theta}(\log(1/\epsilon_1)/\eta)$. Assume $C_x b_x^2(\tilde{\epsilon}_1^2 + r) \leq \min\left\{\frac{\eta\delta^2}{48}, \frac{\rho}{12T\eta}\right\} = \widetilde{\Theta}(\rho/\log(1/\epsilon_1))$. (this means $C_x \leq \widetilde{\Theta}(\frac{\rho}{r\log(1/\epsilon_1)})$.)

We show the assumptions in the previous lemmas are all satisfied. The assumption $\frac{c_1^2}{8\eta\delta^2 b_x^2} \geq \log \frac{6rT^2}{\rho}$ in Lemma 12 is satisfied by

$$\frac{c_1^2}{8\eta\delta^2 b_x^2} \geq \log \frac{6rT^2}{\rho} \tag{86}$$

$$\Leftarrow \frac{c_1^2}{8\eta\delta^2 b_x^2} \geq \log \frac{6r}{\rho} + 4\log \frac{1}{\eta} \tag{87}$$

$$\Leftarrow \eta \log \frac{1}{\eta} \leq \frac{c_1^2}{32\delta^2 b_x^2 \log \frac{6r}{\rho}} = P, \tag{88}$$

where the first is by $T \leq \frac{1}{\eta^2}$, the second is by $\log \frac{6r}{\rho} + 4\log \frac{1}{\eta} \leq 4\log \frac{6r}{\rho} \log \frac{1}{\eta}$, and the last line is true because

$$\eta \log \frac{1}{\eta} \leq \frac{P}{Q} \log \frac{Q}{P} \tag{89}$$

$$= P\left(\frac{\log Q}{Q} + \frac{\log 1/P}{Q}\right) \tag{90}$$

$$\leq P. \tag{91}$$

The assumption $\delta \geq b_x(\tilde{\epsilon}_1^2 + r)$ in Lemma 12 is satisfied by definition of δ. The assumption $(\tilde{\epsilon}_1^2 + r)C_x b_x^2 \leq \frac{c_1}{20}$ in Lemma 12 is satisfied by

$$C_x b_x^2(\tilde{\epsilon}_1^2 + r) \leq \frac{\eta\delta^2}{48} \leq \frac{c_1^2}{96} \leq \frac{c_1}{20}, \tag{92}$$

where we use

$$\eta\delta^2 \leq \delta^2 \frac{P}{Q} \leq \frac{c_1^2}{2}. \tag{93}$$

The assumption $(\tilde{\epsilon}_1^2 + r)C_x b_x^2 \leq \frac{\rho}{12T\eta}$ in Lemma 13 is satisfied by assumption of C_x.

The assumption $\frac{T}{\delta^2} \geq \frac{2^9}{c_1^2} \log \frac{6r}{\rho}$ in Lemma 14 is satisfied by

$$\frac{T}{\delta^2} \geq \frac{16}{\eta^2 \delta^4} \log \frac{6r}{\rho} \geq \frac{2^6}{c_1^4} \log \frac{6r}{\rho} \geq \frac{2^9}{c_1^2} \log \frac{6r}{\rho}. \tag{94}$$

The other two assumptions $(\tilde{\epsilon}_1^2 + r)C_x b_x^2 < \frac{c_1}{12} - \frac{c_1^2}{4}$ and $\eta\delta^2 \leq \frac{c_1}{8}$ in Lemma 14 follows from $C_x b_x^2(\tilde{\epsilon}_1^2 + r) < \frac{\eta\delta^2}{48}$ and $\eta\delta^2 \leq \frac{c_1^2}{2}$. The assumption $T\eta \geq \frac{16}{c_1} \log \frac{1}{\epsilon_{min}}$ in Lemma 14 is satisfied by the definition of T.

The assumptions $(\tilde\epsilon_1^2 + r)C_x \le \frac{c_1}{20}$ and $\frac{c_1^2}{8\eta\delta^2} \ge \log\frac{6rT^2}{\rho}$ in Lemma 15 are satisfied by the same reason as that of Lemma 12. The assumption $T\eta^2\delta^2 \ge 16\log\left(\frac{6\sqrt{d}}{\rho\sqrt{\epsilon_1}}\right)$ in Lemma 17 is satisfied by the definition of T, the assumption $\frac{2}{3}\eta\delta^2 > 32(\tilde\epsilon_1^2 + r)C_x b_x^2$ in Lemma 17 is satisfied by the definition of C_x.

Since data are randomly from $\mathcal{N}(0, I)$, with $n \ge \widetilde\Theta((\frac{r\log(1/\epsilon_1)}{\rho})^2)$ data, there is with probability at least $1 - \frac{\rho}{18}$ there is $C_x b_x^2(\tilde\epsilon_1^2 + r) \le \min\left\{\frac{\eta\delta^2}{48}, \frac{\rho}{12T\eta}\right\} = \widetilde\Theta(\log(1/\epsilon_1)/\eta)$. Meanwhile, according to Lemma 26, with $n \ge \widetilde\Theta(1)$ data with probability at least $1 - \frac{\rho}{18}$ there is $\mathbb{E}_i[(x_k^{(i)})^2] \ge \frac{2}{3}$ for all $k \in [d]$. According to definition of b_x, we know when $n \le d$, with probability at least $1 - \frac{\rho}{18}$ there is also $\left\|x^{(i)}\right\|_\infty \le b_x$ for all $i \in [n]$. In summary, with $d \ge n \ge \widetilde\Theta((\frac{r\log(1/\epsilon_1)}{\rho})^2)$ data, with probability at least $1 - \frac{\rho}{6}$ there is $C_x b_x^2(\tilde\epsilon_1^2 + r) \le \min\left\{\frac{\eta\delta^2}{48}, \frac{\rho}{12T\eta}\right\}$ and $\mathbb{E}_i[(x_k^{(i)})^2] \ge \frac{2}{3}$ for all $k \in [d]$ and $\left\|x^{(i)}\right\|_\infty \le b_x$ for all $i \in [n]$.

Now we use these lemmas to finish the proof of the theorem. Let $\tilde v^{[t]}$ be a $(1 + c_1, \tilde\epsilon_1)$-bounded coupling of $v^{[t]}$, we only need to prove with probability at least $1 - \rho$, there is $\left\|\tilde v_S^{[T]} - 1\right\|_\infty \le c_1$ and $\left\|\tilde v_{\bar S}^{[T]}\right\|_1 \le \epsilon_1$, which follows from a union bound of the previous propositions. In particular, Lemma 12 and Lemma 13 tell us that probability of $\left\|\tilde v_S^{[T]}\right\|_\infty > 1 + c_1$ or $\left\|\tilde v_{\bar S}^{[T]}\right\|_1 > \tilde\epsilon_1$ is at most $\frac{\rho}{3}$. Lemma 14 and Lemma 15 tell us for any $k \in S$, probability of $\tilde v_k^{[T]} < 1 - c_1$ and $\left\|\tilde v_S^{[T]}\right\|_\infty \le 1 + c_1$ and $\left\|\tilde v_{\bar S}^{[T]}\right\|_1 \le \tilde\epsilon_1$ is at most $\frac{\rho}{3r}$. Lemma 17 tells us the probability of $\left\|\tilde v_{\bar S}^{[T]}\right\|_1 > \epsilon_1$ and $\left\|\tilde v_S^{[T]}\right\|_\infty \le 1 + c_1$ and $\left\|\tilde v_{\bar S}^{[T]}\right\|_1 \le \tilde\epsilon_1$ is at most $\frac{\rho}{6}$. Combining them together tells us that probability of $\left\|\tilde v_S^{[T]} - 1\right\|_\infty > c_1$ or $\left\|\tilde v_{\bar S}^{[T]}\right\|_1 > \epsilon_1$ is at most ρ. $\blacksquare$

Appendix D. Proof of Convergence to Ground Truth

The conclusion of Theorem 5 still allows constant error in the support, namely, $\|v_S - v_S^\star\|_\infty \leq 1/10$. The following theorem shows that further annealing the learning rate will let the algorithm fully converge to $v^\star$ with any target error. The end-to-end proof of the convergence to ground truth can be found in the next section (Section E).

Theorem 18 *Let $s \geq 2$ be the index of the current round of bootstrapping. Let constant $c_0 = 1/10$. In the setting of Theorem 1, assume $v^{[0]}$ is an initial parameter satisfying $\|v_S^{[0]} - v_S^\star\|_\infty \leq c_{s-1}$ and $\|v_{\bar{S}}^{[0]} - v_{\bar{S}}^\star\|_1 \leq \epsilon_{s-1}$, where $0 < \epsilon_{s-1} \leq c_{s-1} \leq c_0$. Given a failure rate $\rho > 0$. Assume $n \geq \widetilde{\Theta}(r^2)$. Suppose we run SGD with label noise with noise level $\delta \geq 0$ and learning rate $\eta \leq \widetilde{\Theta}(c_s^2/(\delta^2 + r^2))$ for $T = \log(4/c_0)/\eta$ iterations. Then, with probability at least $1 - \rho$ over the randomness of the algorithm and data, there is $\|v_S^{[T]} - v_S^\star\|_\infty \leq c_s \triangleq c_{s-1}c_0$ and $\|v_{\bar{S}}^{[T]} - v_{\bar{S}}^\star\|_1 \leq \epsilon_s \triangleq (4/c_0)^{2c_{s-1}}\epsilon_{s-1}$. Here $\widetilde{\Theta}(\cdot)$ omits poly logarithmic dependency on ρ.*

In the rest of this section, we will first prove several lemmas on which the proof of Theorem 18 is built upon. Then we will provide a proof of Theorem 18.

Definition 19 *((b, ϵ)-to-$v^\star$ coupling) Let $v^{[0]}, v^{[1]}, \cdots, v^{[T]}$ be a trajectory of label noise gradient descent with initialization $v^{[0]}$. Recall $S \subset [d]$ is the support set of $v^\star$. We call the following random sequence $\tilde{v}^{[t]}$ a (b, ϵ)-to-$v^\star$ coupling of $v^{[t]}$: starting from $\tilde{v}^{[0]} = v^{[0]}$, for each time $t < T$, if $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_1 \leq \epsilon$ and $\left\|\tilde{v}_S^{[t]} - 1\right\|_\infty \leq b$, we let $\tilde{v}^{[t+1]} \triangleq v^{[t+1]}$; otherwise $\tilde{v}^{[t+1]} \triangleq \tilde{v}^{[t]}$.*

Lemma 20 *In the setting of Theorem 18, let $C_x \triangleq \max_{j \neq k} |\mathbb{E}_i[x_j^{(i)} x_k^{(i)}]|$. Assume $\left\|x^{(i)}\right\|_\infty \leq b_x$ for $i \in [n]$ for some $b_x > 0$, and $\mathbb{E}_i[(x_k^{(i)})^2] \geq \frac{2}{3}$ for $k \in [d]$. Let $\tilde{v}^{[t]}$ be a $(2c_{s-1}, \epsilon_s)$-to-$v^\star$ coupling of $v^{[t]}$. Assume $\frac{c_{s-1}^2}{2\eta b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2)} \geq \log \frac{10rT^2}{\rho}$ and $(\epsilon_s^2 + 4rc_{s-1})C_x b_x^2 \leq \frac{c_{s-1}}{10}$. Then, with probability at least $1 - \frac{\rho}{5}$, there is $\left\|\tilde{v}_S^{[T]} - 1\right\|_\infty \leq 2c_{s-1}$.*

Proof [Proof of Lemma 20] For any fixed $1 \leq t_1 < t_2 \leq T$ and dimension $k \in S$, we consider the event that $\tilde{v}^{[t_1]} \in [1 + \frac{2}{3}c_{s-1}, 1 + c_{s-1}]$, and at time t_2 it is the first time in the trajectory such that $\tilde{v}^{[t_2]} > 1 + 2c_{s-1}$. We first bound the probability of this event happens, i.e., the following quantity:

$$\Pr\left(\tilde{v}_k^{[t_2]} - 1 > 2c_{s-1} \wedge \tilde{v}_k^{[t_1]} - 1 \leq c_{s-1} \wedge \tilde{v}_k^{[t_1:t_2]} \in [1 + \frac{2}{3}c_{s-1}, 1 + 2c_{s-1}]\right), \quad (95)$$

where $\tilde{v}_k^{[t_1:t_2]} \in [1 + \frac{2}{3}c_{s-1}, 1 + 2c_{s-1}]$ means that for all t such that $t_1 \leq t < t_2$, there is $1 + \frac{2}{3}c_{s-1} \leq \tilde{v}_k^{[t]} \leq 1 + 2c_{s-1}$.

Notice that when $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_1 \leq \epsilon_s$ and $\left\|\tilde{v}_S^{[t]} - 1\right\|_\infty \leq 2c_{s-1}$ and $\tilde{v}_k^{[t_1:t+1]} \in [1 + \frac{2}{3}c_{s-1}, 1 + 2c_{s-1}]$, there is

$$\mathbb{E}[\tilde{v}_k^{[t+1]} - 1] = \mathbb{E}_{s_t, i_t}\left[\left(1 + \eta s_t x_k^{i_t} - \eta(\tilde{v}^{[t]\odot 2} - v^{\star \odot 2})^\top x^{(i_t)} x_k^{(i_t)}\right)\tilde{v}_k^{[t]} - 1\right] \quad (96)$$

$$\leq (\tilde{v}_k^{[t]} - 1) - \frac{2}{3}\eta \tilde{v}_k^{[t]}(\tilde{v}_k^{[t]} + 1)(\tilde{v}_k^{[t]} - 1) + \eta(\epsilon_s^2 + 4rc_{s-1})C_x b_x^2 \tilde{v}_k^{[t]} \quad (97)$$

$$\leq (1 - \eta)(\tilde{v}_k^{[t]} - 1). \quad (98)$$

where the first inequality is because $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_2^2 \le \epsilon_s^2$ and properties of the data, the second inequality is because $(\epsilon_s^2 + 4rc_{s-1})b_x^2 C_x \le \frac{c_{s-1}}{10}$. Also, we can bound the variance of this martingale as

$$\mathrm{Var}\left[\tilde{v}_k^{[t+1]} - 1 \mid \tilde{v}^{[t]} - 1\right] = \mathrm{Var}\left[\eta s_t x_k^{(i_t)} \tilde{v}_k^{[t]}\right] + \mathrm{Var}\left[\eta((\tilde{v}^{[t]\odot 2} - v^{\star \odot 2})^\top x^{(i_t)}) x_k^{(i_t)} \tilde{v}_k^{[t]}\right] \quad (99)$$

$$\le (\eta \delta b_x (1 + 2c_{s-1}))^2 + \eta^2 (\epsilon_s^2 + r)^2 b_x^4 (1 + 2c_{s-1})^2 \quad (100)$$

$$\le 4\eta^2 b_x^2 (\delta^2 + b_x^2 (\epsilon_s^2 + r)^2), \quad (101)$$

By Lemma 27, we have

$$\Pr(\tilde{v}_k^{[t_2]} - 1 > 2c_{s-1} \wedge \tilde{v}_k^{[t_1]} - 1 \le c_{s-1} \wedge \tilde{v}_k^{[t_1:t_2]} \in [1 + \frac{2}{3}c_{s-1}, 1 + 2c_{s-1}]) \quad (102)$$

$$\le e^{\frac{-c_{s-1}^2}{2\eta^2 b_x^2 (\delta^2 + b_x^2 (\epsilon_s^2 + r)^2) \sum_{t=0}^{t_2 - t_1 - 1} (1-\eta)^{2t}}} \quad (103)$$

$$\le e^{\frac{-c_{s-1}^2}{2\eta b_x^2 (\delta^2 + b_x^2 (\epsilon_s^2 + r)^2)}}, \quad (104)$$

where the first inequality is by Lemma 27, the second inequality is by taking the sum of denominator.

Similarly, we bound

$$\Pr\left(1 - \tilde{v}_k^{[t_2]} > 2c_{s-1} \wedge 1 - \tilde{v}_k^{[t_1]} \le c_{s-1} \wedge \tilde{v}_k^{[t_1:t_2]} \in [1 - 2c_{s-1}, 1 - \frac{2}{3}c_{s-1}]\right). \quad (105)$$

Notice that when $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_1 \le \epsilon_s$ and $\left\|\tilde{v}_S^{[t]} - 1\right\|_\infty \le 2c_{s-1}$ and $\tilde{v}_k^{[t_1:t+1]} \in [1 - 2c_{s-1}, 1 - \frac{2}{3}c_{s-1}]$, there is

$$\mathbb{E}[1 - \tilde{v}_k^{[t+1]}] = \mathbb{E}_{s_t, i_t}\left[1 - (1 + \eta s_t x_k^{i_t} - \eta(\tilde{v}^{[t]\odot 2} - v^{\star \odot 2})^\top x^{(i_t)}) x_k^{(i_t)}) \tilde{v}_k^{[t]}\right] \quad (106)$$

$$\le (1 - \tilde{v}_k^{[t]}) - \frac{2}{3}\eta \tilde{v}_k^{[t]}(\tilde{v}_k^{[t]} + 1)(1 - \tilde{v}_k^{[t]}) + \eta(\epsilon_s^2 + 4rc_{s-1})C_x b_x^2 \tilde{v}_k^{[t]} \quad (107)$$

$$\le (1 - \eta)(1 - \tilde{v}_k^{[t]}). \quad (108)$$

where the first inequality is because $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_2^2 \le \epsilon_s^2$ and the properties of data, the second inequality is because $(\epsilon_s^2 + 4rc_{s-1})C_x b_x^2 \le \frac{c_{s-1}}{10}$. So

$$\Pr(1 - \tilde{v}_k^{[t_2]} > 2c_{s-1} \wedge 1 - \tilde{v}_k^{[t_1]} \le c_{s-1} \wedge \tilde{v}_k^{[t_1:t_2]} \in [1 - 2c_{s-1}, 1 - \frac{2}{3}c_{s-1}]) \quad (109)$$

$$\le e^{\frac{-c_{s-1}^2}{2\eta^2 b_x^2 (\delta^2 + b_x^2 (\epsilon_s^2 + r)^2) \sum_{t=0}^{t_2 - t_1 - 1} (1-\eta)^{2t}}} \quad (110)$$

$$\le e^{\frac{-c_{s-1}^2}{2\eta b_x^2 (\delta^2 + b_x^2 (\epsilon_s^2 + r)^2)}}, \quad (111)$$

where the first inequality is by Lemma 27, the second inequality is by taking the sum of denominator.

32

Finally, we finish the proof with a union bound. Since if $\left\| \tilde{v}_S^{[T]} - 1 \right\|_\infty > 2c_{s-1}$, either event in Equation 95 or in Equation 105 has to happen for some $k \in S$ and $1 \le t_1 < t_2 \le T$, so we have

$$\Pr\left(\left\| \tilde{v}_S^{[T]} - 1 \right\|_\infty > 2c_{s-1} \right) \tag{112}$$

$$\le \sum_{k \in S} \sum_{1 \le t_1 < t_2 \le T} \Pr\left(\tilde{v}_k^{[t_2]} - 1 > 2c_{s-1} \wedge \tilde{v}_k^{[t_1]} - 1 \le c_{s-1} \wedge \tilde{v}_k^{[t_1:t_2]} \in [1 + \frac{2}{3}c_{s-1}, 1 + 2c_{s-1}] \right) \tag{113}$$

$$+ \sum_{k \in S} \sum_{1 \le t_1 < t_2 \le T} \Pr\left(1 - \tilde{v}_k^{[t_2]} > 2c_{s-1} \wedge 1 - \tilde{v}_k^{[t_1]} \le c_{s-1} \wedge \tilde{v}_k^{[t_1:t_2]} \in [1 - 2c_{s-1}, 1 - \frac{2}{3}c_{s-1}] \right) \tag{114}$$

$$\le 2rT^2 e^{\frac{-c_{s-1}^2}{2\eta b_x^2 (\delta^2 + b_x^2 (\epsilon_s^2 + r)^2)}} \tag{115}$$

$$\le \frac{\rho}{5}, \tag{116}$$

where the first inequality is by union bound, the second inequality is by previous results, the third inequality is by assumption of this lemma. ∎

Lemma 21 *In the setting of Lemma 20, assume $(\epsilon_s^2 + 4c_{s-1}r)C_x b_x^2 \le c_{s-1}$, $\epsilon_s > (1 + \eta c_{s-1})^T \epsilon_{s-1}$ and $\frac{((1+\eta c_{s-1})^{-T}\epsilon_s - \epsilon_{s-1})^2}{2T\eta^2 b_x^2 (\delta^2 + b_x^2(\epsilon_s^2 + r)^2)\epsilon_s^2} \ge \log \frac{5}{\rho}$. Then, with probability at least $1 - \frac{\rho}{5}$, there is $\left\| \tilde{v}_{\bar{S}}^{[T]} \right\|_1 \le \epsilon_s$.*

Proof [Proof of Lemma 21] When $\left\| \tilde{v}_S^{[t]} - 1 \right\|_\infty \le 2c_{s-1}$ and $\left\| \tilde{v}_{\bar{S}}^{[t]} \right\|_1 \le \epsilon_s$, for any $k \notin S$, there is:

$$\mathbb{E}\left[\tilde{v}_k^{[t+1]} \right] = \tilde{v}_k^{[t]} - \eta \mathbb{E}_{i_t}\left[((\tilde{v}^{[t]\odot 2} - v^{\star \odot 2})^\top x^{(i_t)})x_k^{(i_t)} \tilde{v}_k^{[t]} \right] \tag{117}$$

$$\le \tilde{v}_k^{[t]} + \eta(\left\| \tilde{v}_{\bar{S}}^{[t]\odot 2} \right\|_1 + 4c_{s-1}r)C_x b_x^2 \tilde{v}_k^{[t]} \tag{118}$$

$$\le \tilde{v}_k^{[t]} + \eta(\epsilon_s^2 + 4c_{s-1}r)C_x b_x^2 \tilde{v}_k^{[t]} \tag{119}$$

$$\le (1 + \eta c_{s-1})\tilde{v}_k^{[t]}. \tag{120}$$

where the first inequality is because we can bound the dimensions in S and those not in S with $\left\| \tilde{v}_{\bar{S}}^{[t]\odot 2} \right\|_1 C_x b_x^2$ and $4c_{s-1}rC_x b_x^2$ respectively, the second inequality is by $\left\| \tilde{v}_{\bar{S}}^{[t]} \right\|_2^2 \le \left\| \tilde{v}_{\bar{S}}^{[t]} \right\|_1^2$, the third is because $(\epsilon_s^2 + 4c_{s-1}r)C_x b_x^2 \le c_{s-1}$. Summing over all $k \notin S$ we have $\mathbb{E}[\left\| \tilde{v}_{\bar{S}}^{[t+1]} \right\|_1] \le (1 + \eta c_{s-1})\left\| \tilde{v}_{\bar{S}}^{[t]} \right\|_1$. This bound is obviously also true when $\left\| \tilde{v}_S^{[t]} - 1 \right\|_\infty > 2c_{s-1}$ or $\left\| \tilde{v}_{\bar{S}}^{[t]} \right\|_1 > \epsilon_s$, in which case $\tilde{v}^{[t+1]} = \tilde{v}^{[t]}$.

Therefore we know $(1 + \eta c_{s-1})^{-t} \left\| \tilde{v}^{[t]} \right\|_1$ is a supermartingale. Also notice $| \left\| \tilde{v}_{\bar{S}}^{[t+1]} \right\|_1 - \mathbb{E}[\left\| \tilde{v}_{\bar{S}}^{[t+1]} \right\|_1]| \leq \eta \delta \epsilon_s$, By Azuma Inequality,

$$\Pr \left(\left\| \tilde{v}_{\bar{S}}^{[T]} \right\|_1 > \epsilon_s \right) \tag{121}$$

$$\leq e^{-\frac{((1+\eta c_{s-1})^{-T}\epsilon_s - \epsilon_{s-1})^2}{2T\eta^2 b_x^2 (\delta^2 + b_x^2 (\epsilon_s^2 + r)^2)\epsilon_s^2}}, \tag{122}$$

$$\leq \frac{\rho}{4}. \tag{123}$$

here we are using $\epsilon_s > (1 + \eta c_{s-1})^T \epsilon_{s-1}$ by assumption and the last step is by assumption. ∎

Lemma 22 *In the setting of Lemma 20, assume $(1 - \eta)^T 2c_{s-1} < \frac{c_s}{2}$, $\frac{c_{s-1}^2}{2\eta \delta^2} \geq \log \frac{5r}{\rho}$, and $(\epsilon_s^2 + 4c_{s-1}r)C_x b_x^2 \leq \frac{c_s}{10}$. Then, for any $k \in S$, with probability at least $1 - \frac{\rho}{5r}$, either $\min_{t \leq T} |\tilde{v}_k^{[t]} - 1| \geq \frac{c_s}{2}$, or $\left\| \tilde{v}_S^{[T]} - 1 \right\|_\infty > 2c_{s-1}$, or $\left\| \tilde{v}_{\bar{S}}^{[T]} \right\|_1 > \epsilon_s$.*

Proof [Proof of Lemma 22]

We first consider when $\tilde{v}_k^{[t]} \in [1 + \frac{c_s}{2}, 1 + 2c_{s-1}]$. For some $t < T_2$, if $\left\| \tilde{v}_S^{[t]} - 1 \right\|_\infty \leq c_{s-1}$ and $\left\| \tilde{v}_{\bar{S}}^{[t]} \right\|_1 \leq \epsilon_s$, there is

$$\mathbb{E}[\tilde{v}_k^{[t+1]} - 1] \tag{124}$$

$$= \tilde{v}_k^{[t]} - \eta \mathbb{E}_{i_t}[((\tilde{v}^{[t] \odot 2} - v^{\star \odot 2})^\top x^{(i_t)})x_k^{(i_t)}]\tilde{v}_k^{[t]} - 1 \tag{125}$$

$$\leq (\tilde{v}_k^{[t]} - 1) - \frac{2}{3}\eta \tilde{v}_k^{[t]}(\tilde{v}_k^{[t]} + 1)(\tilde{v}_k^{[t]} - 1) + \eta(\epsilon_s^2 + 4c_{s-1}r)C_x b_x^2 v_k^{[t]} \tag{126}$$

$$\leq (1 - \eta)(v_k^{[t]} - 1). \tag{127}$$

Here the first inequality is by assumption, the second inequality is because $(\epsilon_s^2 + 4c_{s-1}r)C_x b_x^2 \leq \frac{1}{10}c_s$ and $c_{s-1} \leq \frac{1}{10}$.

We define the event E_t as $\left\| \tilde{v}_S^{[t]} - 1 \right\|_\infty > 2c_{s-1}$ or $\left\| \tilde{v}_{\bar{S}}^{[t]} \right\|_1 > \epsilon_s$. Since $(1 - \eta)^{T_2} 2c_{s-1} < \frac{c_s}{2}$ by assumption, if $\tilde{v}_k^{[0]} \in [1 + \frac{c_s}{2}, 1 + 2c_{s-1}]$, by Lemma 28 we know:

$$\Pr \left(\min_{t \leq T} \tilde{v}_k^{[t]} > 1 + \frac{c_s}{2} \wedge \left\| \tilde{v}_S^{[T]} - 1 \right\|_\infty \leq 2c_{s-1} \wedge \left\| \tilde{v}_{\bar{S}}^{[T]} \right\|_1 \leq \epsilon_s \right) \tag{128}$$

$$\leq e^{-\frac{\left(\frac{1}{2}c_s(1-\eta)^{-T} - c_{s-1} \right)^2}{2\eta b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2)}} \tag{129}$$

$$\leq e^{-\frac{c_{s-1}^2}{2\eta b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2)}} \tag{130}$$

$$\leq \frac{\rho}{5r}, \tag{131}$$

where the second inequality is because of assumption.

34

Similarly, when $\tilde{v}_k^{[0]} \in [1 - 2c_{s-1}, 1 - \frac{c_s}{2}]$, there is

$$\Pr\left(\max_{t \leq T} \tilde{v}_k^{[t]} < 1 - \frac{c_s}{2} \wedge \left\|\tilde{v}_S^{[T_2]} - 1\right\|_\infty \leq 2c_{s-1} \wedge \left\|\tilde{v}_{\bar{S}}^{[T]}\right\|_1 \leq \epsilon_s\right) \tag{132}$$

$$\leq e^{-\frac{c_{s-1}^2}{2\eta b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2)}} \tag{133}$$

$$\leq \frac{\rho}{5r}. \tag{134}$$

Since $|\tilde{v}_k^{[0]} - 1| \leq c_{s-1}$ by assumption of Theorem 18, by bounding the probability for $\tilde{v}_k^{[0]} \in [1 + \frac{c_s}{2}, 1 + 2c_{s-1}]$ and $[1 - 2c_{s-1}, 1 - \frac{c_s}{2}]$ repectively we finished the proof. ∎

Lemma 23 *In the setting of Lemma 20, assume $\frac{c_s^2}{8\eta b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2)} \geq \log \frac{5rT^2}{\rho}$ and $(\epsilon_s^2 + 2rc_s)C_x b_x^2 \leq \frac{c_s}{20}$. Then, for any dimension $k \in S$, with probability at most $\frac{\rho}{5r}$, there is $\min_{t \leq T} |\tilde{v}_k^{[t]} - 1| \leq \frac{1}{2}c_s$ and $|\tilde{v}_k^{[T]} - 1| > c_s$ and $\left\|\tilde{v}_S^{[T]} - 1\right\|_\infty \leq 2c_{s-1}$ and $\left\|\tilde{v}_{\bar{S}}^{[T]}\right\|_1 \leq \epsilon_s$.*

Proof [Proof of Lemma 23] For any fixed $1 \leq t_1 < t_2 \leq T$, we consider the event that $\tilde{v}^{[t_1]} \in [1 + \frac{1}{3}c_s, 1 + \frac{1}{2}c_s]$, and at time t_2 it is the first time in the trajectory such that $\tilde{v}^{[t_2]} > 1 + c_s$. We first bound the probability of this event happens, i.e., the following quantity:

$$\Pr\left(\tilde{v}_k^{[t_2]} - 1 > c_s \wedge \tilde{v}_k^{[t_1]} - 1 \leq \frac{1}{2}c_s \wedge \tilde{v}_k^{[t_1:t_2]} \in [1 + \frac{1}{3}c_s, 1 + c_s]\right). \tag{135}$$

Notice that when $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_1 \leq \epsilon_s$ and $\left\|\tilde{v}_S^{[t]} - 1\right\|_\infty \leq 2c_{s-1}$ and $\tilde{v}_k^{[t_1:t+1]} \in [1 + \frac{1}{3}c_s, 1 + c_s]$, there is

$$\mathbb{E}[\tilde{v}_k^{[t+1]} - 1] = \mathbb{E}_{s_t, i_t}\left[\left(1 + \eta s_t x_k^{i_t} - \eta \mathbb{E}_{i_t}[(\tilde{v}^{[t]\odot 2} - v^{\star\odot 2})^\top x^{(i_t)})x_k^{(i_t)}]\right)\tilde{v}_k^{[t]} - 1\right] \tag{136}$$

$$\leq (\tilde{v}_k^{[t]} - 1) - \frac{2}{3}\eta\tilde{v}_k^{[t]}(\tilde{v}_k^{[t]} + 1)(\tilde{v}_k^{[t]} - 1) + \eta(\epsilon_s^2 + 2rc_s)C_x b_x^2 \tilde{v}_k^{[t]} \tag{137}$$

$$\leq (1 - \eta)(\tilde{v}_k^{[t]} - 1). \tag{138}$$

where the first inequality is because $\left\|\tilde{v}_{\bar{S}}^{[t]}\right\|_2^2 \leq \epsilon_s^2$, the second inequality is because $(\epsilon_s^2 + 2rc_s)C_x b_x^2 \leq \frac{c_s}{20}$. Also, we can bound the variance of this martingale as

$$\mathrm{Var}\left[\tilde{v}_k^{[t+1]} - 1 \mid \tilde{v}^{[t]} - 1\right] = \mathrm{Var}\left[\eta s_t x_k^{(i_t)}\tilde{v}_k^{[t]}\right] + \mathrm{Var}\left[\eta((\tilde{v}^{[t]\odot 2} - v^{\star\odot 2})^\top x^{(i_t)})x_k^{(i_t)}\tilde{v}_k^{[t]}\right] \tag{139}$$

$$\leq (\eta\delta b_x(1 + c_s))^2 + \eta^2(\epsilon_s^2 + r)^2 b_x^4(1 + c_s)^2 \tag{140}$$

$$\leq 4\eta^2 b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2), \tag{141}$$

By Lemma 27, we have

$$\Pr(\tilde{v}_k^{[t_2]} - 1 > c_s \wedge \tilde{v}_k^{[t_1]} - 1 \leq \frac{1}{2}c_s \wedge \tilde{v}_k^{[t_1:t_2]} \in [1 + \frac{1}{3}c_s, 1 + c_s]) \tag{142}$$

$$\leq e^{\frac{-c_s^2}{8\eta^2 b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2)\sum_{t=0}^{T-1}(1-\eta)^{2t}}} \tag{143}$$

$$\leq e^{\frac{-c_s^2}{8\eta b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2)}} \tag{144}$$

35

where the first inequality is by Lemma 27, the second inequality is by taking the sum of denominator. Similarly, we bound

$$\Pr(1 - \tilde{v}_k^{[t_2]} > c_s \wedge 1 - \tilde{v}_k^{[t_1]} \leq \frac{1}{2}c_s \wedge \tilde{v}_k^{[t_1:t_2]} \in [1 - c_s, 1 - \frac{1}{3}c_s]) \tag{145}$$

$$\leq e^{\frac{-c_s^2}{8\eta b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2)}}. \tag{146}$$

Finally, we finish the proof with a union bound:

$$\Pr\left(\min_{t \leq T} |\tilde{v}_k^{[t]} - 1| \leq \frac{1}{2}c_s \wedge |\tilde{v}_k^{[T]} - 1| > c_s \wedge \left\|\tilde{v}_S^{[T]} - 1\right\|_\infty \leq 2c_{s-1} \wedge \left\|\tilde{v}_{\bar{S}}^{[T]}\right\|_1 \leq \epsilon_s\right) \tag{147}$$

$$\leq \sum_{1 \leq t_1 < t_2 \leq T} \Pr(\tilde{v}_k^{[t_2]} - 1 > c_s \wedge \tilde{v}_k^{[t_1]} - 1 \leq \frac{1}{2}c_s \wedge \tilde{v}_k^{[t_1:t_2]} \in [1 + \frac{1}{3}c_s, 1 + c_s]) \tag{148}$$

$$+ \sum_{1 \leq t_1 < t_2 \leq T} \Pr(1 - \tilde{v}_k^{[t_2]} > c_s \wedge 1 - \tilde{v}_k^{[t_1]} \leq \frac{1}{2}c_s \wedge \tilde{v}_k^{[t_1:t_2]} \in [1 - c_s, 1 - \frac{1}{3}c_s]) \tag{149}$$

$$\leq T^2 e^{\frac{-c_s^2}{8\eta b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2)}} \tag{150}$$

$$\leq \frac{\rho}{5r}, \tag{151}$$

where the first inequality is by union bound, the second inequality is by previous results, the third inequality is by assumption of this lemma. ∎

Proof [Proof of Theorem 18] Let $C_x \triangleq \max_{j \neq k} |\mathbb{E}_i[x_j^{(i)} x_k^{(i)}]|$, $b_x = \sqrt{2 \log \frac{30d^2}{\rho}} = \widetilde{\Theta}(1)$. According to Lemma 25, when $n \leq d$, there is with probability at least $1 - \frac{\rho}{15}$ we have $\left\|x^{(i)}\right\|_\infty \leq b_x$ for $i \in [d]$.

Set η small enough such that $\frac{c_s^2}{8\eta b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2)} \geq \log \frac{10rT^2}{\rho}$. Obviously we only need $\eta \leq \widetilde{\Theta}(\frac{c_s^2}{\delta^2 + r^2})$, where $\widetilde{\Theta}(\cdot)$ omits poly logarithmic dependency on d and ρ. Assume $(\epsilon_s^2 + 4c_{s-1}r)C_x b_x^2 \leq \frac{c_s}{10}$, which can be represented as $C_x \leq \widetilde{\Theta}(\frac{1}{r})$. Recall $T = \frac{1}{\eta} \log \frac{4}{c_0}$, $\epsilon_s = e^{2c_{s-1} \log \frac{4}{c_0}} \epsilon_{s-1}$.

We first show that the additional assumptions in the previous lemmas are satisfied. There is

$$(1 + \eta c_{s-1})^T = (1 + \eta c_{s-1})^{\frac{1}{\eta} \log \frac{4}{c_0}} \tag{152}$$

$$\leq e^{c_{s-1} \log \frac{4}{c_0}} \triangleq P. \tag{153}$$

The assumption $\epsilon_s > (1 + \eta c_{s-1})^T \epsilon_{s-1}$ in Lemma 21 is therefore satisfied by definition of ϵ_s. The assumption $\frac{((1+\eta c_{s-1})^{-T}\epsilon_s - \epsilon_{s-1})^2}{2T\eta^2 b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2)\epsilon_s^2} \geq \log \frac{5}{\rho}$ in Lemma 21 is satisfied because:

$$\frac{((1 + \eta c_{s-1})^{-T}\epsilon_s - \epsilon_{s-1})^2}{2T\eta^2 b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2)\epsilon_s^2} \geq \frac{\epsilon_s^2(P^{-1} - P^{-2})^2}{2T\eta^2 b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2)\epsilon_s^2} \tag{154}$$

$$\geq \frac{(P - 1)^2}{2T\eta^2 b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2)} \tag{155}$$

$$\geq \frac{c_{s-1}^2}{2\eta b_x^2(\delta^2 + b_x^2(\epsilon_s^2 + r)^2)}, \tag{156}$$

36

which is larger than $\log \frac{5}{\rho}$ by the definition of η. The assumption $(1-\eta)^T 2c_{s-1} \leq \frac{c_s}{2}$ in Lemma 22 is satisfied because $(1-\eta)^T 2c_{s-1} \leq (\frac{1}{e})^{\log \frac{4}{c_0}} 2c_{s-1} = \frac{c_s}{2}$. All the other assumptions in Lemma 20, Lemma 21, Lemma 22 and Lemma 23 naturally follows from the definition of η and the requirement of C_x.

Since data are randomly from $\mathcal{N}(0, I)$, with $n \geq \widetilde{\Theta}(r^2)$ data, there is with probability at least $1 - \frac{\rho}{15}$ there is $(\epsilon_s^2 + 4c_{s-1}r)C_x b_x^2 \leq \frac{c_s}{10}$. Meanwhile, according to Lemma 26 with $n \geq \widetilde{\Theta}(1)$ data with probability at least $1 - \frac{\rho}{15}$ there is $\mathbb{E}_i[(x_k^{(i)})^2] \geq \frac{2}{3}$ for all $k \in [d]$. According to definition of b_x, we know when $n \leq d$, with probability at least $1 - \frac{\rho}{15}$ there is also $\left\| x^{(i)} \right\|_\infty \leq b_x$ for all $i \in [n]$. In summary, with $d \geq n \geq \widetilde{\Theta}(r^2)$ data, with probability at least $1 - \frac{\rho}{5}$ there is $(\epsilon_s^2 + 4c_{s-1}r)C_x b_x^2 \leq \frac{c_s}{10}$ and $\mathbb{E}_i[(x_k^{(i)})^2] \geq \frac{2}{3}$ for all $k \in [d]$ and $\left\| x^{(i)} \right\|_\infty \leq b_x$ for all $i \in [n]$.

Now we finish the proof with the above lemmas. Lemma 20 and Lemma 21 together tell us that with probability at least $1 - \frac{2\rho}{5}$, there is $\left\| \tilde{v}_S^{[T]} - 1 \right\|_\infty \leq 2c_{s-1}$ and $\left\| \tilde{v}_{\bar{S}}^{[T]} \right\|_1 \leq \epsilon_s$, in which case there is also $v^{[T]} = \tilde{v}^{[T]}$ by the definition of $\tilde{v}^{[T]}$. Lemma 22 and Lemma 23 together tell us the probability of $\left\| \tilde{v}_S^{[T]} - 1 \right\|_\infty \leq 2c_{s-1}$ and $\left\| \tilde{v}_{\bar{S}}^{[T]} \right\|_1 \leq \epsilon_s$ and $\left\| \tilde{v}_S^{[T]} - 1 \right\|_\infty > c_s$ is no more than $\frac{2\rho}{5}$. So together we know with probability at least $1 - \rho$, there is $\left\| v_S^{[T]} - 1 \right\|_\infty \leq c_s$ and $\left\| v_{\bar{S}}^{[T]} \right\|_1 \leq \epsilon_s$.

∎

Appendix E. Proof of Theorem 1

Proof [Proof of Theorem 1] Starting from initialization $\tau \cdot \mathbb{1}$, by Theorem 3, running SGD with label noise with noise level $\delta > \widetilde{O}(\frac{\tau^2 d^2}{\rho^3})$ and $\eta_0 = \widetilde{\Theta}(\frac{1}{\delta})$ for $T_0 = \widetilde{\Theta}(1)$ iterations gives us that with probability at least 0.99, $\epsilon_{min} \leq v_k^{[T_0]} \leq \frac{1}{d}$ where $\epsilon_{min} = exp(-\widetilde{\Theta}(1))$. Now $v^{[T_0]}$ satisfies the initial condition of Theorem 5.

Recall the final target precision is ϵ, set $\epsilon_1 = \frac{1}{40^3}\epsilon$. By Theorem 5, with $n \geq \widetilde{\Theta}(r^2 \log^2(1/\epsilon))$ data, after running SGD with label noise with learning rate $\eta_1 = \widetilde{\Theta}(\frac{1}{\delta^2})$ for $T_1 = \widetilde{\Theta}(\frac{\log(1/\epsilon)}{\eta_1})$ iterations, with probability at least 0.99, there is,

$$\left\| v_S^{[T_0+T_1]} - v_S^\star \right\|_\infty \leq c_1 \triangleq \frac{1}{10}, \tag{157}$$

and

$$\left\| v_{\bar{S}}^{[T_0+T_1]} - v_{\bar{S}}^\star \right\|_1 \leq \epsilon_1. \tag{158}$$

So we have $v^{[T_0+T_1]}$ satisfies the initial condition of Theorem 18.

Finally, set $\rho = 0.01/\lceil \log_{10}(1/\epsilon) \rceil$, and apply Theorem 18 for $n_s = \lceil \log_{10}(1/\epsilon) \rceil = \widetilde{\Theta}(1)$ rounds. Since c_s gets smaller by $1/10$ for each round, the final c_{n_s} satisfies $\frac{1}{10}\epsilon \leq c_{n_s} \leq \epsilon$. Since the requirement of η for round s is $\eta \leq \widetilde{\Theta}(\frac{c_s^2}{\delta^2+r^2})$, we can set $\eta_2 \leq \widetilde{\Theta}(\frac{\epsilon^2}{\delta^2})$ to satisfy all the rounds at the same time. Set T_2 be the total number of iterations in all of these rounds, obviously $T_2 = \widetilde{\Theta}(\frac{1}{\eta_2})$. Notice that $\epsilon_s \leq e^{\sum_{s=2}^\infty 2c_{s-1} \log \frac{4}{c_0}} \epsilon_1 \leq 40^3 \epsilon_1 = \epsilon$, we have with probability at least 0.99,

$$\left\| v_S^{[T_0+T_1+T_2]} - v_S^\star \right\|_\infty \leq c_1 \triangleq \epsilon, \tag{159}$$

and

$$\left\| v_{\bar{S}}^{[T_0+T_1+T_2]} - v_{\bar{S}}^\star \right\|_1 \leq \epsilon. \tag{160}$$

The total failure rate of above three stages is 0.03, so with probability at least 0.97, there is $\left\| v^{[T_0+T_1+T_2]} - v^\star \right\|_\infty \leq \epsilon$, which finishes the proof. ∎

Appendix F. Proof of Theorem 2

Lemma 24 *Assume $n \leq \frac{d}{2} - 9\sqrt{d}$. Let $C \subset \mathbb{R}^d$ be the convex cone where each coordinate is positive, K be a random subspace of dimension $d - n$. Then with probability at least 0.999, there is $K \cap C \neq \{0\}$*

Proof [Proof of Lemma 24] By Theorem 1 of Amelunxen et al. (2014), we only need to prove

$$\delta(C) + \delta(K) \geq d + 9\sqrt{d}, \tag{161}$$

where $\delta(\cdot)$ is the statistical dimension of a set. By equation (2.1) of Amelunxen et al. (2014), there is

$$\delta(K) = d - n. \tag{162}$$

To calculate $\delta(C)$, we use Proposition 2.4 from Amelunxen et al. (2014),

$$\delta(C) = \mathbb{E}[\|\Pi_C(g)\|^2], \tag{163}$$

where g is a standard random vector, Π_C is projection of g to C, the expectation is over g. Since C is the set of all points with element-wise positive coordinate, $\Pi_C(g)$ is simply setting all the negative dimension of g to 0 and keep the positive ones. Therefore,

$$\delta(C) = \mathbb{E}[\|\Pi_C(g)\|^2] = \frac{d}{2}. \tag{164}$$

Therefore we have

$$\delta(C) + \delta(K) = \frac{3}{2}d - n \geq d + 9\sqrt{d}. \tag{165}$$

$\blacksquare$

Proof [Proof of Theorem 2] Let $X^\perp$ be the subspace that is orthogonal to the subspace X spanned by data. Since data is random, with probability 1 the random subspace X is of n dimension. Therefore, according to the previous lemma, with probability at least 0.999, there is $X^\perp \cap C \neq \{0\}$, where C is the coordinate-wise positive cone. Let $\mu \in X^\perp$ be such a vector such that $\mu_i > 0$ for $\forall i \in [d]$, and we scale it such that $\|\mu\|_2 = 1$. We can construct the following orthonormal matrix

$$A = [a^1, \cdots, a^d] \in \mathbb{R}^{d \times d}, \tag{166}$$

such that $span\{a^1, \cdots a^n\} = X$ and $a^{n+1} = \mu$. Consider the following transformation

$$A\tilde{u} = u = v^{\odot 2}, \tag{167}$$

since only the projection of u to the span of data influences $L(v)$, we can write $L(v) = \tilde{L}(\tilde{u}_{1:n})$ as a function of the first n dimensions of $\tilde{u}$.

39

We can lower bound the partition function with

$$\int_{v \in \mathbb{R}^d} e^{-\lambda L(v)} dv \geq \int_{v > 0} e^{-\lambda L(v)} dv \tag{168}$$

$$= \int_{A\tilde{u} > 0} e^{-\lambda \tilde{L}(\tilde{u}_{1:n})} \det \frac{\partial v}{\partial u} \det \frac{\partial u}{\partial \tilde{u}} d\tilde{u} \tag{169}$$

$$= \frac{1}{2^d} \int e^{-\lambda \tilde{L}(\tilde{u}_{1:n})} \left(\int_{A\tilde{u} > 0} \prod_{i=1}^d \frac{1}{\sqrt{u_i}} d\tilde{u}_{n+1:d} \right) d\tilde{u}_{1:n}. \tag{170}$$

Here the inner loop is integrating over the last $d - n$ dimensions of $\tilde{u}$ in the set such that $A\tilde{u}$ is coordinate-wise positive. Now we prove that for each $\tilde{u}_{1:n}$ such that $S = \{\tilde{u}_{n+1:d} | A\tilde{u} > 0\}$ is not empty set, the inner loop integral is always $+\infty$.

Fix $\tilde{u}_{1:n}$, let $\tilde{u}_{n+1:d}^*$ be one possible solution such that $u^* = A\tilde{u} > 0$. Define constant

$$c = \min_{i \leq [1:d]} \max_{j \in [n+2:d]} \frac{a_i^{n+1}}{(d - n - 1)|a_i^j|}, \tag{171}$$

we can define the following set

$$S' = \{\tilde{u}_{n+1:d} | \tilde{u}_{n+1} \geq \tilde{u}_{n+1}^* \wedge |\tilde{u}_j - \tilde{u}_j^*| \leq c(\tilde{u}_{n+1} - \tilde{u}_{n+1}^*), \forall j \in [n+2, d]\} \tag{172}$$

In other words, this is a convex cone where constraint of $\tilde{u}_j$ is linear in $\tilde{u}_{n+1}$ for $j \in [n + 2 : d]$. By definition of c, it is easy to verify that S' is a subset of S. Also, for every $\tilde{u}_{n+1:d} \in S'$, u_i is upper bounded by

$$\left(A \begin{bmatrix} \tilde{u}_{1:n} \\ \tilde{u}_{n+1:d} \end{bmatrix} \right)_i = u_i^* + a_i^{n+1}(\tilde{u}_{n+1} - \tilde{u}_{n+1}^*) + \sum_{j=n+2}^d a_i^j(\tilde{u}_j - \tilde{u}_j^*) \tag{173}$$

$$\leq u_i^* + 2a_i^{n+1}(\tilde{u}_{n+1} - \tilde{u}_{n+1}^*). \tag{174}$$

Here the inequality is because of the definition of c.

Let $z = \tilde{u}_{n+1} - \tilde{u}_{n+1}^*$ we have

$$\int_{\tilde{u}_{n+1:d} \in S'} \prod_{i=1}^d \frac{1}{\sqrt{u_i}} d\tilde{u}_{n+1:d} \tag{175}$$

$$\geq \int_{z \geq 0} (2cz)^{d-n-1} \prod_{i=1}^d \frac{1}{\sqrt{u_i^* + 2a_i^{n+1}z}} dz \tag{176}$$

$$= + \infty. \tag{177}$$

Here the last step is because $n < d/2$, so the integrand is essentially a polynomial of z with degree larger than -1, so integrating it over all positive z has to be $+\infty$. So we finish the proof that $\int_{v \in \mathbb{R}^d} e^{-\lambda L(v)} dv = +\infty$. ∎

Appendix G. Extra Lemmas

Lemma 25 *Suppose $x^{(i)} \sim \mathcal{N}(0, \mathcal{I}_{d \times d})$ where $i \in [n]$ are random data. Then with probability at least $1 - \rho$, for every $i \in [n]$ there is*

$$\left\| x^{(i)} \right\|_\infty \leq \sqrt{2 \log \frac{2nd}{\rho}} \tag{178}$$

Proof By Gaussian tail bound, there is $\Pr\left(|x_k^{(i)}| > b_x \right) \leq 2 e^{-\frac{b_x^2}{2}}$. So by union bound we have $\Pr\left(\max_{i,k} |x_k^{(i)}| > b_x \right) \leq 2nd e^{-\frac{b_x^2}{2}}$. Let $b_x = \sqrt{2 \log \frac{2nd}{\rho}}$ we finished the proof. ■

Lemma 26 *Suppose $x^{(i)} \sim \mathcal{N}(0, \mathcal{I}_{d \times d})$ where $i \in [n]$ are random data. Then when $n > 24 \log \frac{d}{\rho}$, with probability at least $1 - \rho$, for every $k \in [d]$ there is*

$$\mathbb{E}_i[x_k^{(i)2}] \geq \frac{2}{3}. \tag{179}$$

Proof Since $\mathbb{E}_x[x_k^2] = 1$, $\mathbb{E}_x[x_k^4] = 3$, by Hoeffding inequality we have

$$\Pr\left(\frac{1}{n} \sum_{i=1}^n x_k^{(i)2} < \frac{2}{3} \right) \leq e^{\frac{-n}{24}}. \tag{180}$$

Therefore, when $n \geq 24 \log \frac{d}{\rho}$, by union bound we finish the proof. ■

Lemma 27 *Let $c > 0$, $1 > \gamma > 0$ be real constants. Let $A^{[0]}, A^{[1]}, \cdots, A^{[T]}$, be a series of random variables, such that given $A^{[0]}, \cdots, A^{[t]}$ for some $t < T$ with $A^{[t]} \in [\frac{c}{3}, c]$, there is either $A^{[t]} = A^{[t+1]} = \cdots = A^{[T]}$, or $\mathbb{E}[A^{[t+1]}] \leq (1 - \gamma) A^{[t]}$ with variance $\mathrm{Var}[A^{[t+1]} \mid A^{[0]}, \cdots, A^{[t]}] \leq a$. Then there is*

$$\Pr\left(A^{[T]} > c \wedge A^{[0]} \leq \frac{c}{2} \wedge A^{[0:T]} \in \left[\frac{c}{3}, c \right] \right) \leq e^{\frac{-c^2}{2a \sum_{t=0}^{T-1} (1-\gamma)^{2t}}}. \tag{181}$$

where $A^{[0:T]} \in \left[\frac{c}{3}, c \right]$ means for any $0 \leq t < T$, $A^{[t]} \in \left[\frac{c}{3}, c \right]$.

Proof [Proof of Lemma 27] We only need to consider when $A^{[0]} \leq \frac{c}{2}$. Let $\hat{A}^{[t]}$ be the following coupling of $A^{[t]}$: starting from $\hat{A}^{[0]} = A^{[0]}$, for each time $t < T$, if $\hat{A}^{[t]} = A^{[t+1]} = \cdots = A^{[T]}$ or there is $t' \leq t$ such that $A^{[t']} \notin [\frac{c}{3}, c]$, we let $\hat{A}^{[t+1]} \triangleq (1 - \gamma) \hat{A}^{[t+1]}$; otherwise $\hat{A}^{[t+1]} = A^{[t+1]}$. Intuitively, whenever $A^{[t]}$ stops updating or exceeds proper range, we only times $\hat{A}^{[t]}$ by $1 - \gamma$ afterwards, otherwise we let it be the same as $A^{[t]}$. Notice that if the event in Equation 181 happens, there has to be $\hat{A}^{[T]} = A^{[T]}$ (otherwise $A^{[t]}$ stops updating or exceeds range at some time, contradicting the event). So we only need to bound $\Pr\left(\hat{A}^{[T]} > c \right)$.

41

We notice that $(1-\gamma)^{-t}\hat{A}^{[t]}$ for $t = 0 \cdots T$ is a supermartingale, i.e., given history there is $\mathbb{E}[\hat{A}^{[t+1]}|\hat{A}^{[t]}] \le (1-\gamma)\hat{A}^{[t]}$. This is obviously true when $\hat{v}^{[t+1]} = (1-\gamma)\hat{A}^{[t]}$, and also true otherwise by assumption of the lemma. So we have

$$\Pr(\hat{A}^{[T]} > c) \tag{182}$$

$$= \Pr((1-\gamma)^{-T}\hat{A}^{[T]} > (1-\gamma)^{-T}c) \tag{183}$$

$$\le e^{-\frac{2\left(c(1-\gamma)^{-T}-\hat{A}^{[0]}\right)^2}{\sum_{t=0}^{T-1}(1-\gamma)^{-2t}a}} \tag{184}$$

$$\le e^{-\frac{c^2(1-\gamma)^{-2T}}{2\sum_{t=0}^{T-1}(1-\gamma)^{-2(t+1)}a}} \tag{185}$$

$$= e^{-\frac{c^2}{2a\sum_{t=0}^{T-1}(1-\gamma)^{2t}}}. \tag{186}$$

where the first inequality is because of Azuma Inequality and $\mathrm{Var}\left[(1-\gamma)^{-t+1}\hat{A}^{[t+1]} \mid \hat{A}^{[t]}\right] \le (1-\gamma)^{-2(t+1)}a.$, the second inequality is because $\hat{A}^{[0]} \le \frac{c}{2}$. Since the event in Equation 181 only happens when $\hat{A}^{[T]} > c$, we've finished the proof. ∎

Lemma 28 *Let $0 < c_1 < c_2$ be real constants. Let $A^{[0]}, A^{[1]}, \cdots, A^{[T]}$, be a series of random variables, such that given $A^{[0]}, \cdots, A^{[t]}$ for some $t < T$ with $A^{[t]} \in [c_1, c_2]$, there is either event E_t happens, or $\mathbb{E}[A^{[t+1]}] \le (1-\gamma)A^{[t]}$ with variance $\mathrm{Var}[A^{[t+1]} \mid A^{[0]}, \cdots, A^{[t]}] \le a$. Then when $A^{[0]} \in [c_1, c_2]$ and $(1-\gamma)^T c_2 < c_1$ there is*

$$\Pr\left(\min_{t \le T} A^{[t]} > c_1 \wedge \max_{t \le T} A^{[t]} \le c_2 \wedge \neg E_{[0:T]}\right) \le e^{-\frac{2\left(c_1(1-\gamma)^{-T}-A^{[0]}\right)^2}{\frac{1}{\gamma}a}}. \tag{187}$$

where $\neg E_{[0:T]}$ means for any $0 \le t < T$, E_t doesn't happen.

Proof [Proof of Lemma 28] Let $\hat{A}^{[t]}$ be the following coupling of $A^{[t]}$: starting from $\hat{A}^{[0]} = A^{[0]}$, for each time $t < T$, if exists $t' \le t$ such that $E_{t'}$ happens or $A^{[t']} \notin [c_1, c_2]$, we let $\hat{A}^{[t+1]} \triangleq (1-\gamma)\hat{A}^{[t+1]}$; otherwise $\hat{A}^{[t+1]} = A^{[t+1]}$. Intuitively, whenever $A^{[t]}$ exceeds proper range, we only times $\hat{A}^{[t]}$ by $1-\gamma$ afterwards, otherwise we let it be the same as $A^{[t]}$. Notice that if the event in Equation 187 happens, there has to be $\hat{A}^{[T]} = A^{[T]}$ (otherwise E_t happens sometimes or $A^{[t]} \notin [c_1, c_2]$, contradicting the event). So we only need to bound $\Pr\left(\hat{A}^{[T]} > c_1\right)$.

We notice that $(1-\gamma)^{-t}\hat{A}^{[t]}$ for $t = 0 \cdots T$ is a supermartingale, i.e., given history there is $\mathbb{E}[\hat{A}^{[t+1]}|\hat{A}^{[t]}] \le (1-\gamma)\hat{A}^{[t]}$. This is obviously true when $\hat{v}^{[t+1]} = (1-\gamma)\hat{A}^{[t]}$, and also true otherwise by assumption of the lemma. So we have

$$\Pr(\hat{A}^{[T]} > c_1) \tag{188}$$

$$= \Pr((1-\gamma)^{-T}\hat{A}^{[T]} > (1-\gamma)^{-T}c_1) \tag{189}$$

$$\le e^{-\frac{2\left(c_1(1-\gamma)^{-T}-A^{[0]}\right)^2}{\sum_{t=0}^{T-1}(1-\gamma)^{-2t}a}} \tag{190}$$

$$\le e^{-\frac{2\left(c_1(1-\gamma)^{-T}-A^{[0]}\right)^2}{\frac{1}{\gamma}a}} \tag{191}$$

where the first inequality is because of Azuma Inequality and $\text{Var}\left[(1-\gamma)^{-t+1}\hat{A}^{[t+1]} \mid \hat{A}^{[t]}\right] \leq$

$(1-\gamma)^{-2(t+1)}a$ and $\hat{A}^{[0]} \leq c_2 \leq e^{-\frac{2\left(c_1(1-\gamma)^{-T}-c_2\right)^2}{\sum_{t=0}^{T-1}(1-\gamma)^{-2t}a}}$. Since the event in Equation 187 only happens when $\hat{A}^{[T]} > c_1$, we've finished the proof. ∎