SHAPE MATTERS: UNDERSTANDING THE IMPLICIT BIAS OF THE NOISE COVARIANCE

Anonymous authors
Paper under double-blind review

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.

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 (LeCun 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 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 be applicable 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 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.

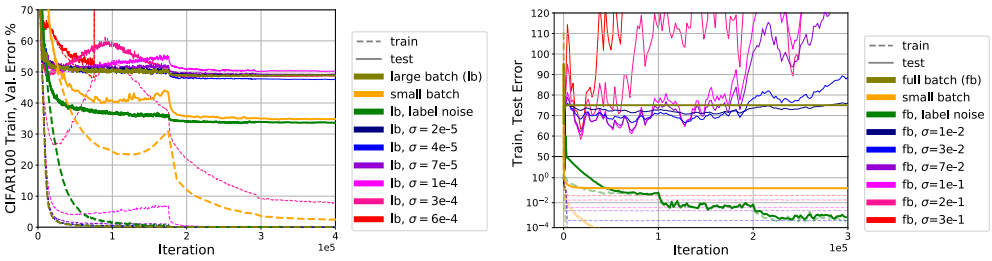

Figure 1: **The effect of noise covariance in neural network and quadratically-parameterized models.** We demonstrate that label noise induces a stronger regularization effect than Gaussian noise. In both real and synthetic data, adding label noise to large batch (or full batch) SGD updates can recover 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. Similarly 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 2.2). More details are in Section A.

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 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 2.1). 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 2.2). 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 solutions which are more robust (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 benefit 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. Agarwal et al. (2020) disentangled the effects of update direction and scale for a variety of optimizers. 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 exisits an implicit bias induced by dropout noise.

Blanc et al. (2019) and Zhu et al. (2019) also studied implicit regularization effects which arise due to shape, rather than scale, of the noise, but 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.

Langevin dynamics or the closely-related stochastic gradient descent with Gaussian noise, has been studied in previous 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. Cheng et al. (2019) proved that sgd with state-dependent noise has iterate distribution close to the corresponding continuous stochastic differential equation with same noise covariance, hence this distribution is mostly determined by the covariance of noise rather than any higher order statistics.

A number of works have theoretically analyzed implicit regularization 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 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 be always possible to characterize implicit biases in terms of some 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. Hardt et al. (2015) derived stability-based generalization bounds for SGD based on training speed. Li et al. (2019b) studied the impliciti regularization effect of large initial learning rate.

Woodworth et al. (2020); Vaskevicius et al. (2019) analyze the effect of initialization for the same model that we study, showing that a 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.

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

[1] In contrast, the implicit bias of noise wouldn't show up in a simpler linear regression model.

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

Require: Number of iterations T, a sequence of step sizes $\eta^{[0:T]}$, noise level δ, initialization $v^{[0]}$

1: **for** $t = 0$ to $T - 1$ **do**
2: Sample index $i_t \sim [n]$ uniformly and add noise $s_t \sim \{\pm\delta\}$ to $y^{(i_t)}$.
3: Let $\tilde{\ell}^{(i_t)}(v) = \frac{1}{4}(f_v(x^{(i_t)}) - y^{(i_t)} - s_t)^2$
4: $v^{[t+1]} \leftarrow v^{[t]} - \eta^{[t]}\nabla\tilde{\ell}^{(i_t)}(v^{[t]})$ ▷ update with label noise

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:

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

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 in order 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.

[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.

Theorem 2.1. *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 2.1 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 2.1.[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 empirical observation that good generalization requires 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 2.2. *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 2.2 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 2.2 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. The proof of Theorem 2.2 can be found in Section F.

[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.

3 ANALYSIS OVERVIEW OF SGD WITH LABEL NOISE (THEOREM 2.1)

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 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 $v^\star = 0$, 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. In fact, 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

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

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, it may appear that the varying scale of noise or norm of the covariance introduces the bias. 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 2.1 that consists of three stages. We first show in the initial stage of the training that label noise effectively decreases 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 in Section 3.3 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 dimensions. Finally, with one more decay, the algorithm can recover the ground truth.

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

[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.

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 3.1. *In the setting of Theorem 2.1, 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 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 $t+1$,

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

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:

$$\begin{aligned}
\mathbb{E}\left[\phi(1 - \eta s_t x_k^{(i_t)} - \eta g_k^{i_t})\right] &\approx \phi(1) - \phi'(1)\eta\mathbb{E}\left[g_k^{i_t}\right] + \frac{1}{2}\phi''(1)\mathbb{E}\Big[\left(\eta s_t x_k^{(i_t)} - \eta g_k^{(i_t)}\right)^2\Big] \\
&\leq \phi(1) - \phi'(1)\eta\mathbb{E}\left[g_k^{i_t}\right] + \frac{1}{2}\phi''(1)\mathbb{E}\Big[\left(\eta s_t x_k^{(i_t)}\right)^2\Big] \\
&\leq \phi(1) - \phi'(1)\eta\mathbb{E}\left[g_k^{i_t}\right] - \Omega(\eta^2\delta^2).
\end{aligned} \tag{10}$$

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 3.2. *In the setting of Theorem 3.1, 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 auxillary, 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.

[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.

By Lemma 3.2 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 3.2 (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 3.1 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 3.3. *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, similarly 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 3.1, and we defer the full proof to Section C.

Stage 2: Convergence to the ground-truth $v^\star$: The conclusion of Theorem 3.3 still allows constant error in the support, namely, $\|v_S - v_S^\star\|_\infty \leq 1/10$. In Theorem D.1, we show that further annealing the learning rate will let the algorithm fully converge to $v^\star$ with any target error ϵ.

Proof of Theorem 2.1. In Section E of Appendix, we combine Theorem 3.1, Theorem 3.3, and Theorem D.1 to prove our main Theorem 2.1.

4 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 provides an explanation for 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.

REFERENCES

N. Agarwal, R. Anil, E. Hazan, T. Koren, and C. Zhang. Disentangling adaptive gradient methods from learning rates. *arXiv preprint arXiv:2002.11803*, 2020.

D. Amelunxen, M. Lotz, M. B. McCoy, and J. 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.

R. Arora, P. Bartlett, P. Mianjy, and N. Srebro. Dropout: Explicit forms and capacity control. *arXiv preprint arXiv:2003.03397*, 2020.

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

G. Blanc, N. Gupta, G. Valiant, and P. Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. *arXiv preprint arXiv:1904.09080*, 2019.

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

K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1565–1576, 2019.

P. Chaudhari and S. 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.

X. Cheng, D. Yin, P. L. Bartlett, and M. I. Jordan. Stochastic gradient and langevin processes. *arXiv preprint arXiv:1907.03215*, 2019.

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

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

A. 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.

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

S. S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018b.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

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

B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems*, pages 9108–9118, 2019.

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

P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

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

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

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

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

E. Hoffer, I. Hubara, and D. 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.

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

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

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

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

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

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

Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

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

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

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

Y. Li, C. Wei, and T. 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.

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

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

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

P. Mianjy, R. Arora, and R. Vidal. On the implicit bias of dropout. *arXiv preprint arXiv:1806.09777*, 2018.

W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. *arXiv preprint arXiv:1707.05947*, 2017.

W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638, 2018.

M. S. Nacson, J. Lee, S. Gunasekar, P. H. Savarese, N. Srebro, and D. Soudry. Convergence of gradient descent on separable data. *arXiv preprint arXiv:1803.01905*, 2018.

M. S. Nacson, S. Gunasekar, J. D. Lee, N. Srebro, and D. Soudry. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. *arXiv preprint arXiv:1905.07325*, 2019.

P. Nakkiran, P. Venkat, S. Kakade, and T. Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

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

A. Neelakantan, L. Vilnis, Q. V. Le, I. Sutskever, L. Kaiser, K. Kurach, and J. Martens. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015.

J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. In *Advances in Neural Information Processing Systems*, pages 11013–11023, 2019.

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

T. Poggio, K. Kawaguchi, Q. Liao, B. Miranda, L. Rosasco, X. Boix, J. Hidary, and H. Mhaskar. Theory of deep learning iii: explaining the non-overfitting puzzle. *arXiv preprint arXiv:1801.00173*, 2017.

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

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

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

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

C. J. Shallue, J. Lee, J. Antognini, J. Sohl-Dickstein, R. Frostig, and G. E. Dahl. Measuring the effects of data parallelism on neural network training. *arXiv preprint arXiv:1811.03600*, 2018.

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

S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2017.

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

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958, 2014.

I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Y. W. Teh, A. H. Thiery, and S. J. Vollmer. Consistency and fluctuations for stochastic gradient langevin dynamics. *The Journal of Machine Learning Research*, 17(1):193–225, 2016.

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

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

S. Vaswani, F. Bach, and M. 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.

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

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

C. Wei, S. Kakade, and T. Ma. The implicit and explicit regularization effects of dropout. *arXiv preprint arXiv:2002.12915*, 2020.

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

M. Welling and Y. 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.

Y. Wen, K. Luk, M. Gazeau, G. Zhang, H. Chan, and J. Ba. Interplay between optimization and generalization of stochastic gradient descent with covariance noise. *arXiv preprint arXiv:1902.08234*, 2019.

A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pages 4148–4158, 2017.

B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro. Kernel and rich regimes in overparametrized models. *arXiv preprint arXiv:2002.09277*, 2020.

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

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Y. Zhang, P. Liang, and M. Charikar. A hitting time analysis of stochastic gradient langevin dynamics. *arXiv preprint arXiv:1702.05575*, 2017.

Z. Zhu, J. Wu, B. Yu, L. Wu, and J. Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. 2019.