Noisy Gradient Descent Converges to Flat Minima for Nonconvex Matrix Factorization

Tianyi Liu Georgia Tech

Yan Li Georgia Tech

Song Wei Georgia Tech

Enlu Zhou Georgia Tech

Tuo Zhao Georgia Tech

Abstract

Numerous empirical evidences have corroborated the importance of noise in nonconvex optimization problems. The theory behind such empirical observations, however, is still largely unknown. This paper studies this fundamental problem through investigating the nonconvex rectangular matrix factorization problem, which has infinitely many global minima due to rotation and scaling invariance. Hence, gradient descent (GD) can converge to any optimum, depending on the initialization. In contrast, we show that a perturbed form of GD with an arbitrary initialization converges to a global optimum that is uniquely determined by the injected noise. Our result implies that the noise imposes implicit bias towards certain optima. Numerical experiments are provided to support our theory.

1 Introduction

Nonconvex optimization has been widely adopted in various domains, including image recognition (Hinton et al., 2012; Krizhevsky et al., 2012), Bayesian graphical models (Jordan et al., 2004; Attias, 2000), recommendation systems (Salakhutdinov et al., 2007), etc. Despite the fact that solving a nonconvex problem is generally difficult, empirical evidences have shown that simple first order algorithms such as stochastic gradient descent (SGD), are able to solve a majority of the aforementioned nonconvex problems efficiently. The theory behind these empirical observations, however, is still largely unexplored.

In classical optimization literature, there have been fruitful results on characterizing the convergence of SGD to first-order stationary points for nonconvex problems. However, these types of results fall short of explaining the empirical evidences that SGD often converges to global minima for a wide class of nonconvex problems used in practice. More recently, understanding the role of noise in the algorithmic behavior of SGD has received significant attention. For instance, Jin et al. (2017) show that a perturbed form of gradient descent is able to escape from strict saddle points and converge to second-order stationary points (i.e., local minima). Zhou et al. (2019) further show that noise in the update can help SGD to escape from spurious local minima and converge to the global minima. We argue that, despite all these recent results on showing the convergence of SGD to global minima for various nonconvex problems, there is still an important question yet to be addressed. Specifically, the convergence of SGD in the presence of multiple global minima remains uncleared for many important nonconvex problems. For example, for over-parameterized neural networks, it has been shown that the nonconvex objective has multiple global minima (Kawaguchi, 2016), only few of which can yield good generalization. In addition, Allen-Zhu et al. (2018) show that for an over-parameterized network, SGD can converge to a global minimum in polynomial time. Combining with the empirical successes of training over-parameterized neural networks with SGD, these results strongly advocate that SGD not only can solve the nonconvex problem efficiently, but also implicitly biases towards solutions with good generalization ability. Motivated by this, this paper aims to provide more theoretical insights to the following question:

Does noise impose implicit bias towards certain minimizer in nonconvex optimization problems?

We answer this question through investigating a simple yet non-trivial problem – nonconvex matrix factorization, which serves as an important foundation for a wide spectrum of problems such as matrix sensing (Bhojanapalli et al., 2016; Zhao et al., 2015; Chen and Wainwright, 2015; Tu et al., 2015), matrix completion (Keshavan et al., 2010; Hardt, 2014; Zheng and Lafferty,

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

2016), and deep linear networks (Ji and Telgarsky, 2018; Gunasekar et al., 2018). Given a matrix $M \in \mathbb{R}^{d_1 \times d_2}$, the nonconvex matrix factorization aims to solve:

$$\min_{X\in\mathbb{R}^{d_1\times r}, Y\in\mathbb{R}^{d_2\times r}} \frac{1}{2}\|XY^\top - M\|_{\mathrm{F}}^2. \tag{1}$$

Despite its simplicity, (1) possesses several intriguing landscape properties: nonconvexity of the objective, all the saddle points satisfy the strict saddle property, and infinitely many global optima due to scaling and rotational invariance. Specifically, for any pair of global optimum (X^*, Y^*), $(\alpha X^*, \frac{1}{\alpha}Y^*)$ and (X^*R, Y^*R) are also global optima for any non-zero constant α and rotation matrix $R \in \mathbb{R}^{r\times r}$. This is different from symmetric matrix factorization, which only possesses rotational invariance. The scaling and rotational invariance also imply that the global minima of (1) are connected, a landscape property that is also shared by deep neural networks (Nguyen and Hein, 2017; Draxler et al., 2018; Nguyen and Hein, 2018; Venturi et al., 2018; Garipov et al., 2018; Liang et al., 2018; Nguyen, 2019; Kuditipudi et al., 2019).

Nonconvex matrix factorization (1) has been recently studied by Du et al. (2018), with focus on the algorithmic behavior of gradient descent (GD). Their results reveal an interesting algorithmic regularization imposed by gradient descent: (i) gradient flow (GD with an infinitesimal step size) has automatic balancing property, i.e., the difference of the squared norm $\|X\|_{\mathrm{F}}^2 - \|Y\|_{\mathrm{F}}^2$ stays constant during training. (ii) for properly chosen step size, GD converges asymptotically for rank-r case, linearly for rank-1 case, while maintaining approximate balancing property. However, Du et al. (2018) do not consider any noise in the update, hence their results can not provide further theoretical insights on understanding the role of noise when applying first order algorithms to nonconvex problems.

In this paper, we are interested in studying the algorithmic behavior of first order algorithms in the presence of noise. Specifically, we study a perturbed form of gradient descent (Perturbed GD) applied to the matrix factorization problem (1), which injects independent noise to iterates, and then evaluates gradient at the perturbed iterates. Note that our algorithm is different from SGD in terms of the noise. For our algorithm, we inject independent noise to the iterates $(X_t, Y_t)'s$ and use the gradient evaluated at the perturbed iterates. The noise of SGD, in contrast, comes from the training sample. As a consequence, the noise of SGD has very complex dependence on the iterate, which is difficult to analyze. See more detailed discussions in Sections 6.

We further analyze the convergence properties of our Perturbed GD algorithm for the rank-1 case. At the early stage, noise helps the algorithm to escape from regions with undesired landscape, including the strict saddle point. After entering the region with benign landscape, Perturbed GD behaves similarly to gradient descent, until the loss is sufficiently small. At the early stage, noise provides additional explorations that help the algorithm to escape from the strict saddle point. Then at the later stage, the noise dominates the update of Perturbed GD, and gradually rescales the iterates to a balanced solution that is uniquely determined by the injected noise. Specifically, the ratio of the norm $\|x_t\|_2/\|y_t\|_2$ is completely determined by the ratio of the variance of noise injected to (x_t, y_t). To the best of our knowledge, this is the first theoretical result towards understanding the implicit bias of noise in nonconvex optimization problems. Our analysis reveals an interesting characterization of the local landscape around global minima, which relates to the sharp/flat minima in deep neural networks (Keskar et al., 2016), and we will further discuss these connections in detail in Section 6. We believe that investigating the implicit bias of the noise in nonconvex matrix factorization can serve as a fundamental building block for studying stochastic optimization for more sophiscated nonconvex problems, including training over-parameterized neural networks.

Notations: Given a matrix A, $\mathrm{tr}(A)$ denotes the trace of A. For matrices $A, B \in \mathbb{R}^{n\times m}$, we use $\langle A, B\rangle$ to denote the Frobenius inner product, i.e., $\langle A, B\rangle = \mathrm{tr}(A^\top B)$. $\|A\|_{\mathrm{F}} = \sqrt{\langle A, A\rangle}$ denotes the Frobenius norm of A. $I_d \in \mathbb{R}^{d\times d}$ denotes the identity matrix.

2 Model and Algorithm

We first describe the nonconvex matrix factorization problem, and then present the perturbed gradient descent algorithm (Perturbed GD). For simplicity, we primarily focus on the rank-1 matrix factorization problem. Extensions to the rank-r case are provided in Section 4.

• **Rank-1 Matrix Factorization**. We consider the following nonconvex optimization problem:

$$\min_{x\in\mathbb{R}^{d_1}, y\in\mathbb{R}^{d_2}} \mathcal{F}(x, y) = \frac{1}{2}\|xy^\top - M\|_{\mathrm{F}}^2, \tag{2}$$

where $M \in \mathbb{R}^{d_1\times d_2}$ is a rank-1 matrix and can be factorized as follows: $M = u_* v_*^\top$, where $u_* \in \mathbb{R}^{d_1}$, $v_* \in \mathbb{R}^{d_2}$. Without loss of generality, we assume $\|u_*\|_2 = \|v_*\|_2 = 1$.

The optimization landscape of (2) has been well studied in the previous literature (Ge et al., 2016, 2017; Chi et al., 2019; Li et al., 2019a). Because of the bilinear form in $\mathcal{F}$, there exist infinitely many global minima to (2), including highly unbalanced ones, i.e., $xy^\top = M$ with $\|x\|_2 \gg \|y\|_2$ or $\|x\|_2 \ll \|y\|_2$ (see Definition 1). In Section 3, we will show that such unbalancedness essentially implies global minima with a large condition number. To address this issue, Tu et al. (2015); Ge et al. (2017) propose a regularizer of the form $(\|x\|_2^2 - \|y\|_2^2)^2$ to balance $\|x\|_2$ and $\|y\|_2$. Recently, Du et al. (2018) show that even without explicit regularization, gradient descent with small random initialization converges to balanced solutions with constant probability. Yet, all of the previous results assume noiseless updates. The algorithmic behavior of first order algorithms with noisy updates remains unclear for the nonconvex matrix factorization problem.

• **Perturbed Gradient Descent**. To study the effect of noise, we consider a perturbed gradient descent algorithm (Perturbed GD). At the t-th iteration, we first perturb the iterate (x_t, y_t) with independent Gaussian noise $\xi_{1,t} \sim N(0, \sigma_1^2 I_{d_1})$ and $\xi_{2,t} \sim N(0, \sigma_2^2 I_{d_2})$, respectively,

$$\widetilde{x}_t = x_t + \xi_{1,t} \quad \text{and} \quad \widetilde{y}_t = y_t + \xi_{2,t}.$$

We then update (x_t, y_t) with the gradient evaluated at the perturbed iterates,

$$x_{t+1} = x_t - \eta(\widetilde{x}_t \widetilde{y}_t^\top - M)\widetilde{y}_t,$$
$$y_{t+1} = y_t - \eta(\widetilde{y}_t \widetilde{x}_t^\top - M^\top)\widetilde{x}_t.$$

Please find the algorithm detail in Appendix 5.

Smoothing Effect. Using Perturbed GD to solve (2) can also be viewed as solving the following stochastic optimization problem:

$$\min_{x \in \mathbb{R}^{d_1}, y \in \mathbb{R}^{d_2}} \widetilde{\mathcal{F}}(x, y) = \mathbb{E}_{\xi_1, \xi_2} \mathcal{F}(x + \xi_1, y + \xi_2), \qquad (3)$$

where $\xi_1 \sim N(0, \sigma_1^2 I_{d_1})$ and $\xi_2 \sim N(0, \sigma_2^2 I_{d_2})$. Throughout this paper, we will refer to problem (3) as the smoothed problem. The expectation in (3) can be viewed as convoluting the objective function with a Gaussian kernel. In Section 3, we show that this convolution effectively smooths out unbalanced optima and yields a benign landscape.

Remark 1. The use of random noise to convolute with the objective function is also known as randomized smoothing, which is first proposed in Duchi et al. (2012). Zhou et al. (2019); Lu et al. (2019); Jin et al. (2017) further exploit this effect to explain the importance of noise in helping first order algorithms to escape from strict saddle points and spurious local optima.

3 Main Results

We study the algorithmic behavior of our proposed perturbed gradient descent (Perturbed GD) algorithm. We primarily focus on the rank-1 nonconvex matrix factorization. We first characterize the landscape of the original problem (2), and show that noise effectively smooths the original problem, yielding a smoothed problem (3) with benign landscape. We then provide a non-asymptotic convergence analysis of the Perturbed GD, and demonstrate the implicit bias induced by the noise. In particular, we consider the case where noise is balanced, see more details in Theorem 1. Due to space limit, we defer all the proofs to the appendix.

We analyze the landscape of the original problem (2) and the smoothed problem (3). Note that due to the bilinear form in $\mathcal{F}(x, y)$, we have for $\alpha \neq 0$, $\mathcal{F}\left(\alpha x, \alpha^{-1} y\right) = \mathcal{F}(x, y)$. The scaling invariance nature of (2) results in undesired landscape properties, and makes the analysis of first order algorithms particularly difficult. To facilitate further discussions, below we characterize the landscape of (2).

Lemma 1 (Landscape Analysis). The gradients of $\mathcal{F}$ with respect to x and y take the form:

$$\nabla_x \mathcal{F}(x, y) = (xy^\top - M)y, \ \nabla_y \mathcal{F}(x, y) = (xy^\top - M)^\top x.$$

Then $\mathcal{F}$ has two types of stationary points: (i) For any $\alpha \neq 0$, $\left(\alpha u_*, \alpha^{-1} v_*\right)$ is a global optimum; (ii) For any $x \in \mathbb{R}^{d_1}, y \in \mathbb{R}^{d_2}$ such that $x^\top u_* = y^\top v_* = 0$, $(x, 0)$ and $(0, y)$ are strict saddle.

The scaling-invariance leads to infinitely many global optima for (2), each taking the form $(\alpha u_*, \alpha^{-1} v_*)$. However, Lemma 2 shows that different values of α lead to significantly different local landscape around the global minima.

Lemma 2. The condition number of the Hessian matrix of $\mathcal{F}$ at the global optimum $(\alpha u_*, \alpha^{-1} v_*)$ is

$$\kappa\left(\nabla^2 \mathcal{F}\left(\alpha u_*, \alpha^{-1} v_*\right)\right) = \max\left\{\alpha^4, \alpha^{-4}\right\} + 1.$$

From Lemma 2 we know that when $|\alpha|$ is extremely large or small, the global minimum $(\alpha u_*, \alpha^{-1} v_*)$ is ill-conditioned (large condition number). The condition number also relates to the sharp/flat minima in neural networks, which we will discuss in detail in Section 6. Our previous discussion implies that any global optimum (x_*, y_*) to (2) will be ill-conditioned if the norm ratio $\|x_*\|_2 / \|y_*\|_2$ is close to zero or infinity. To facilitate further discussions, we define the balancedness property as follows.

Definition 1 (γ-balancedness)**.** We say that $(x, y) \in (\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$ is $\gamma-$balanced for some positive number γ if $\|x\|_2 = \gamma \|y\|_2$. Informally, we say (x, y) is unbalanced if γ is close to zero or infinity, and (x, y) is balanced if γ is close to 1.

Our next lemma shows that, in the presence of noise, the smoothed problem (3) has only balanced optima, and the balancedness is completely determined by the ratio of the variance of the noise injected to the iterates.

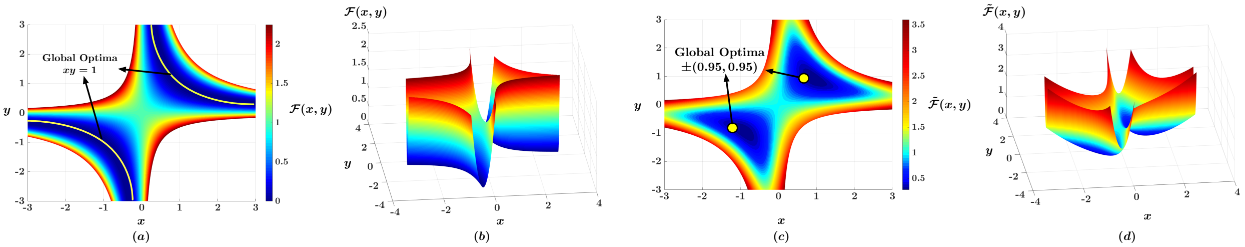


Figure 1: The visualization of objective functions $\mathcal{F}(x,y) = (1-xy)^2$ and $\widetilde{\mathcal{F}}(x,y)$ with $x, y \in \mathbb{R}$ and $\sigma_1^2 = \sigma_2^2 = 0.0975$. For $\mathcal{F}(x,y)$, any (x,y) that satisfies $xy = 1$ is a global minimum (as shown in (a)). $\widetilde{\mathcal{F}}(x,y)$ only has global minima close to $\pm(1,1)$ (as shown in (c)).

Lemma 3. We say the noise in the Perturbed GD is $\gamma-$balanced if $\mathbb{E}\left[\|\xi_1\|_2^2\right] = \gamma^2\mathbb{E}\left[\|\xi_2\|_2^2\right]$. For $\gamma = 1$, we say the noise is balanced. For any noise ratio $\gamma > 0$, if we take $\sigma^2 = \mathbb{E}\left[\|\xi_1\|_2^2\right] = \gamma^2\mathbb{E}\left[\|\xi_2\|_2^2\right] \leq \min\{\gamma, 1\}$, (3) has two global optima $(\widetilde{u}_*(\gamma), \widetilde{v}_*(\gamma)) = \pm\sqrt{\gamma - \sigma^2}\left(u_*, \gamma^{-1}v_*\right)$ and one saddle point $(0,0)$ with $\lambda_{\min}(\nabla^2\widetilde{F}(0,0)) < 0$.

We see that when the noise is γ-balanced and the noise level is sufficiently small, the global minima of the smoothed problem (3) can be arbitrarily close to the γ-balanced global minima of the original rank-1 matrix factorization problem (2). Compared with Lemma 1, Lemma 3 shows that noise effectively smooths the landscape of the original problem (2), the unbalanced optima (with ill-conditioned Hessian) to (2) are no longer optima after convolution. See Figure 1 for detailed illustration of the landscape for the original problem (2) and the smoothed problem (3).

Moreover, one can verify that the saddle point $(0,0)$ of problem (3) enjoys the strict saddle property, i.e., the smallest eigenvalue of Hessian at $(0,0)$ is negative. By the stable manifold theory proposed in Lee et al. (2019), Perturbed GD with an infinitesimal step size and vanishing noise avoids all the strict saddle points and converges to the global optima, asymptotically. However, this asymptotic result does not provide any finite time guarantee.

We next present the non-asymptotic convergence analysis of our proposed Perturbed GD algorithm. Our analysis shows that Perturbed GD has implicit bias determined by the injected noise. Specifically, we consider balanced noise, i.e., $\mathbb{E}\left[\|\xi_1\|_2^2\right] = \mathbb{E}\left[\|\xi_2\|_2^2\right]$, and show that Perturbed GD converges to the balanced $\pm(u_*, v_*)$ in polynomial time in Theorem 1.

Theorem 1 (Convergence Analysis). Suppose $x_0 \in \mathbb{R}^{d_1}$, $y_0 \in \mathbb{R}^{d_2}$. For any $\epsilon > 0$ and for any $\delta \in (0,1)$, we take $\sigma^2 = \mathbb{E}\left[\|\xi_1\|_2^2\right] = \mathbb{E}\left[\|\xi_2\|_2^2\right] = \mathrm{poly}(\epsilon, (\log(1/\delta))^{-1})$,

$$\eta = \mathrm{poly}(\sigma, \epsilon, (d_1+d_2)^{-1}, (\log(d_1+d_2))^{-1}, (\log(1/\delta))^{-1}).$$

With probability at least $1-\delta$, we have $\|x_t - u_*\|_2 \leq \epsilon$ and $\|y_t - v_*\|_2 \leq \epsilon$. for all $t_1 \leq t \leq T_1 = O\left(1/\eta^2\right)$, where $t_1 = O\left(\eta^{-1}\sigma^{-2}\log(1/\eta)\log(1/\delta)\right)$.

Theorem 1 differs from the convergence analysis of GD in Du et al. (2018) in the following aspects: (i) Perturbed GD converges regardless of initializations, while GD requires a random initialization near $(0,0)$; (ii) Perturbed GD guarantees convergence with high probability, while GD converges with only constant probability over randomness of initializations; (iii) Perturbed GD converges to the balanced solutions $(\|x\|_2/\|y\|_2 = 1)$ while GD only maintains approximate balancedness, i.e., $c_0 \leq \|x\|_2/\|y\|_2 \leq C_0$ for some absolute constants $c_0, C_0 > 0$.

We provide a proof sketch that contains the essential ingredients of characterizing the convergence, since the proof of Theorem 1 is very technical and highly involved. See more details and the proof of technical lemmas in Appendix 3.

Proof Sketch. The convergence of Perturbed GD consists of three phases: **Phase I**: Regardless of initialization, within polynomial time the noise encourages Perturbed GD to escape from region with undesired landscape (e.g., strict saddle points) and enter the region with benign landscape. **Phase II**: Perturbed GD drives the loss to zero and approaches the set of global minima $\left\{(x,y)\middle|xy^\top = M\right\}$. **Phase III**: After the loss is sufficiently small, the injected noise dominates the update, which helps the Perturbed GD to balance x and y, and converge to a balanced optimum.

Before we proceed with our proof, we first define some notations. Let U and V denote the linear span of u_* and v_*, respectively, i.e., $\mathcal{U} = \{\alpha_1 u_* : \alpha_1 \in \mathbb{R}\}$, $\mathcal{V} = \{\alpha_1 v_* : \alpha_1 \in \mathbb{R}\}$. The corresponding orthogonal complement of $\mathcal{U}$ (or $\mathcal{V}$) in $\mathbb{R}^{d_1}$ (or $\mathbb{R}^{d_2}$) is denoted as $\mathcal{U}^\perp$ (or $\mathcal{V}^\perp$). Our analysis considers the convergence of Perturbed GD in $(\mathcal{U},\mathcal{V})$ and $(\mathcal{U}^\perp,\mathcal{V}^\perp)$, respectively. Specifically, we take the following orthogonal decomposition of x_t and y_t:

$$x_t = u_*^\top x_t u_* + (x_t - u_*^\top x_t u_*),$$
$$y_t = v_*^\top y_t v_* + (y_t - v_*^\top y_t v_*).$$

One can check that $(x_t - u_*^\top x_t u_*) \in \mathcal{U}^\perp$ and $(y_t - v_*^\top y_t v_*) \in \mathcal{V}^\perp$, respectively.

- **Phase I**: Regardless of initializations, the following lemma shows that $(x_t - u_*^\top x_t u_*)$ and $(y_t - v_*^\top y_t v_*)$ vanish after polynomial time.

Lemma 4. Suppose $\|x_t\|_2^2 + \|y_t\|_2^2 \le 2/\sigma^2$ holds for all $t > 0$. For any $\delta \in (0,1)$ we take

$$\eta \le \eta_2 = C_4\sigma^8 \left(\log((d_1+d_2)/\delta)\log(1/\delta)\right)^{-1},$$

where C_4 is some positive constant. Then with probability at least $1-\delta$, we have

$$\|x_t - u_*^\top x_t u_*\|_2^2 \le 2\eta C_2\sigma^{-2},$$
$$\|y_t - v_*^\top y_t v_*\|_2^2 \le 2\eta C_2\sigma^{-2}$$

for any $\tau_1 \le t \le T_1 = O(1/\eta^2)$, where $\tau_1 = O(\eta^{-1}\sigma^{-2}\log(1/\eta)\log(1/\delta))$ and $C_2 = (\sigma^2 + 1/\sigma^2)(2/\sigma^4 + 6/d_1 + 6/d_2 + 6\sigma^4)$.

Lemma 4 shows that with an arbitrary initialization, the projection of x_t (and y_t) onto $\mathcal{U}^\perp$ (and $\mathcal{V}^\perp$) vanishes. Thus, we only need to characterize the algorithmic behavior of Perturbed GD in subspace $(\mathcal{U},\mathcal{V})$. However, as the projection of x_t (and y_t) onto $\mathcal{U}^\perp$ (and $\mathcal{V}^\perp$) vanishes, Perturbed GD can possibly approach the region with undesired landscape, e.g., the small neighborhood around the saddle point. The next lemma shows that the Perturbed GD will escape such regions within polynomial time.

Lemma 5. Suppose $\|x_t - u_*^\top x_t u_*\|_2^2 \le 2\eta C_2\sigma^{-2}$ and $\|y_t - v_*^\top y_t v_*\|_2^2 \le 2\eta C_2\sigma^{-2}$ hold for all $t > 0$. For any $\delta \in (0,1)$, we take $\eta \le \eta_3 = C_3\sigma^{12}\left(\log((d_1+d_2)/\delta)\log(1/\delta)\right)^{-1}$, where C_3 is some positive constant. Then with probability at least $1-\delta$, we have

$$x_t^\top u_* v_*^\top y_t = x_t^\top M y_t \ge 1/4, \tag{4}$$

for all $\tau_2 \le t \le T_1 = O\left(1/\eta^2\right)$, where $\tau_2 = O\left(\eta^{-1}\sigma^{-2}\log(1/\eta)\log(1/\delta)\right)$.

Since the saddle point $(x,y) = (0,0)$ satisfies $x^\top M y = 0$, Lemmas 4 and 5 together imply that regardless of initializations, Perturbed GD is bounded away from the saddle point after polynomial time. The algorithm then enters Phase II and approaches the set of global optima $\left\{(x,y)\middle|xy^\top = M\right\}$.

- **Phase II**: In this phase, (x_t, y_t) is still away from the region $\left\{(x,y)\middle|xy^\top = M\right\}$. Thus, $\nabla\mathcal{F}(x_t,y_t)$ dominates the update of Perturbed GD. Perturbed GD behaves similarly to gradient descent while driving the loss to zero. The next lemma formally characterizes this behavior, showing that $xy^\top$ converges to M.

Lemma 6. Suppose $x_t^\top u_* v_*^\top y_t \ge \frac{1}{4}$ holds for all $t > 0$. For any $\epsilon > 0$ and for any $\delta \in (0,1)$, we choose $\sigma \le \sigma_1' = C_4\sqrt{\epsilon}$ and take $\eta \le \eta_4 = C_5\sigma^6\left(\log((d_1+d_2)/\delta)\log(1/\delta)\right)^{-1}$, where C_4, C_5 are some positive constants. Then with probability at least $1-\delta$, we have

$$\|x_t y_t^\top - M\|_{\mathrm{F}} \le \epsilon,$$

for all $\tau_3 \le t \le T_1 = O\left(1/\eta^2\right)$, where $\tau_3 = O\left(\eta^{-1}\log\frac{1}{\sigma}\log(1/\delta)\right)$.

- **Phase III**: After Phase II, Perturbed GD enters the region where $x_t y_t^\top \approx M$. Thus, the noise will dominate the update of the Perturb GD. Recall in Lemma 2, we show that the Hessian of unbalanced optima has a large condition number, hence a small perturbation will significantly change the gradient of the objective. This implies that the unbalanced optima would be unstable against the noise, and Perturbed GD will escape from such optima and continue iterating towards the balanced optima. The next lemma shows that x_t and y_t converge to u_*, v_*, respectively.

Lemma 7. For $\forall\epsilon > 0$, suppose $\|x_t y_t^\top - M\|_{\mathrm{F}} \le \epsilon$ holds for all $t > 0$. For any $\delta \in (0,1)$, we choose $\sigma \le \sigma_2' = C_6\left(\log(1/\delta)\right)^{-1/3}$ and take $\eta \le \eta_5 = C_7\sigma^{10}\epsilon$, where C_6, C_7 are some positive constants. Then with probability at least $1-\delta$, we have

$$\|x_t - u_*\|_2 \le \epsilon,\ \|y_t - v_*\|_2 \le \epsilon,$$

for all $\tau_3 \le t \le T_1 = O\left(1/\eta^2\right)$, where $\tau_4 = O\left(\eta^{-1}\sigma^{-2}\log\eta^{-1}\log(1/\delta)\right)$.

With all the lemmas in place, we take $\sigma \le \min\{\sigma_1', \sigma_2'\}$, $\eta \le \min\{\eta_1,\eta_2,\eta_3,\eta_4,\eta_5\}$ and $t_1 = \tau_1 + \tau_2 + \tau_3 + \tau_4$, then the claim of Theorem 1 follows immediately. □

4 Extension to Rank-r Matrix Factorization

We extend our results to the rank-r matrix factorization, which solves the following problem:

$$\min_{X\in\mathbb{R}^{d_1\times r}, Y\in\mathbb{R}^{d_2\times r}} \mathcal{F}(X,Y) = \frac{1}{2}\|XY^\top - M\|_{\mathrm{F}}^2, \tag{5}$$

where $M \in \mathbb{R}^{d_1\times d_2}$ is a rank-r matrix. Let $M = A\Sigma B^\top$ be the SVD of M. Let $U_* = A\Sigma^{\frac{1}{2}}$ and $V_* = B\Sigma^{\frac{1}{2}}$, then (U_*, V_*) is a global minimum for problem (5). Similar to the rank-1 case, using Perturbed GD to solve problem

(5) can be viewed as solving the following smoothed problem:

$$\min_{X,Y} \widetilde{\mathcal{F}}(X,Y) = \mathbb{E}_{\xi_1,\xi_2}\mathcal{F}(X+\xi_1, Y+\xi_2), \tag{6}$$

where $\xi_1 \in \mathbb{R}^{d_1\times r}$, $\xi_2 \in \mathbb{R}^{d_2\times r}$ have i.i.d. elements drawn from $N(0,\sigma_1^2)$ and $N(0,\sigma_2^2)$, respectively.

The next theorem shows that the noise in Perturbed GD effectively addresses the scaling invariance issue of (5). The smoothed problem (6) only has balanced global minima.

Theorem 2. Let $\sigma_{\min}(M)$ be the smallest singular value of M. Suppose

$$\mathbb{E}\left[\|\xi_1\|_{\mathrm{F}}^2\right] = \gamma^2\mathbb{E}\left[\|\xi_2\|_{\mathrm{F}}^2\right] = r\gamma^2\sigma^2,$$

and $\gamma\sigma^2 < \sigma_{\min}(M)$. Then for $\forall(U,V) \in (\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$ such that $\nabla\widetilde{\mathcal{F}}(U,V) = 0$, we have $U^\top U = \gamma^2 V^\top V$. Moreover, denote $(\widetilde{U},\widetilde{V}) = \left(\sqrt{\gamma}A(\Sigma-\gamma\sigma^2 I_r)^{\frac{1}{2}}, \gamma^{-1/2}B(\Sigma-\gamma\sigma^2 I_r)^{\frac{1}{2}}\right)$, then the set $\{(\widetilde{U},\widetilde{V})R \big| R\in\mathbb{R}^{r\times r}, RR^\top = R^\top R = I_r\}$ contains all the global optima. All other stationary points are strict saddles, i.e., $\lambda_{\min}\left(\nabla^2\widetilde{\mathcal{F}}(U,V)\right) < 0$.

Theorem 2 shows that when noise is balanced, (6) only has balanced global optima and strict saddle points. We can invoke Lee et al. (2019) again and show Perturbed GD with an infinitesimal step size and vanishing noise converges to the balanced global optima, asymptotically.

Note that compared to the rank-1 case, in addition to scaling invariance, the objective is also rotation invariant, i.e., $\widetilde{\mathcal{F}}(X,Y) = \widetilde{\mathcal{F}}(XR,YR)$, where $R\in\mathbb{R}^{r\times r}$ is an orthogonal matrix. Thus, we can only recover U_* and V_* up to a rotation factor. To establish the non-asymptotic convergence result, the optimization error should be measured by the following metric that is rotation invariant:

$$\mathrm{dist}_{\mathcal{R}}(D_1,D_2) = \min_{R\in\mathbb{R}^{r\times r}:RR^\top=I_r} \|D_1 - D_2R\|_{\mathrm{F}}.$$

However, it is more challenging and involved to handle the complex nature of the distance $\mathrm{dist}_{\mathcal{R}}(\cdot,\cdot)$. As our results for the rank-1 case already provide insights on understanding the implicit bias of noise, we leave the non-asymptotic analysis of rank-r matrix factorization for future investigation.

5 Numerical Experiments

We present numerical results to support our theoretical findings. We compare our Perturbed GD algorithm with gradient descent (GD), and demonstrate that Perturbed GD with $\gamma-$balanced noise converges to $\gamma-$balanced optima, while the optima obtained by GD are highly sensitive to initialization and step size. We also show that the phase transition between Phase II and Phase III is not an artifact of the proof, and faithfully captures the true algorithmic behavior of Perturbed GD. All figures can be found in Appendix 6.

Rank-1 Matrix Factorization. We first consider the rank-1 matrix factorization problem. Without loss of generality, the matrix M to be factorized is given by $M = u_*v_*^\top$, where $u_* = (1,0,\ldots,0) \in \mathbb{R}^{d_1}$ and $v_* = (1,0,\ldots,0) \in \mathbb{R}^{d_2}$, with $d_1 = 20$ and $d_2 = 30$. We initialize iterates (x_0,y_0) with $x_0 \sim N(0,\sigma_x^2 I_{d_1})$, $y_0 \sim N(0,\sigma_y^2 I_{d_2})$. For all experiments, we use γ-balanced noise in Perturbed GD. Specifically, we choose $\xi_{1,t} \sim N(0,\sigma_1^2 I_{d_1}), \xi_{2,t} \sim N(0,\sigma_2^2 I_{d_2})$.

(1) Balanced Noise. We consider the case of balanced noise ($\gamma = 1$), where we take $\sigma_1 = \sqrt{1.5}\times 0.05$ and $\sigma_2 = 0.05$. One can verify that $\gamma^2 = d_1\sigma_1^2/(d_2\sigma_2^2) = 1$.

We first use balanced step size to compare with GD studied in Du et al. (2018). Specifically, we set $\eta_x = \eta_y = 10^{-2}$ for both Perturbed GD and GD. We further consider two initialization schemes: (i) small initializations (which is also adopted in Du et al. (2018)): $\sigma_x = \sigma_y = 10^{-2}$; (ii) large initializations: $\sigma_x = \sigma_y = 10^{-1}$. Fig. 2.(a, b, d, e) summarize the results of 100 repeated experiments in a box-plot. As can be seen, regardless of initializations, Perturbed GD always converges to the balanced optima (Fig. 2.(a, b)). In contrast, GD only converges to approximately balanced optima for small initializations (Fig. 2.(d)), and the large initialization yields a large variance in terms of the balancedness of the obtained solution (Fig. 2.(e)).

We also use unbalanced step size (different step sizes for updating x and y) and compare the convergence properties of Perturbed GD and GD. Specifically, we set $\eta_x = 0.5\eta_y = 5\times 10^{-3}$ for both Perturbed GD and GD. We adopt a small initialization scheme, with $\sigma_x = \sigma_y = 10^{-2}$. Fig. 2.(c, f) summarize the results of 100 repeated simulations in a box-plot. As can be seen, even with small initializations, GD with unbalanced step size converges to the approximately $\sqrt{0.5}$-balanced optima, instead of the 1-balanced optima (Fig. 2.(f)). In contrast, Perturbed GD is able to converge to the balanced optima with unbalanced step size (Fig. 2.(c)).

Our results suggest that the noise is the most important factor in determining the balancedness of the solutions obtained by Perturbed GD.

(2) Unbalanced Noise. We run Perturbed GD with unbalanced noise. We take $\sigma_1 = \sqrt{0.75}\times 0.05$ and $\sigma_2 = 0.05$ with $\gamma^2 = d_1\sigma_1^2/(d_2\sigma_2^2) = 0.5$. We use a small initialization: $\sigma_x = \sigma_y = 10^{-2}$, and balanced step size: $\eta_x = \eta_y = 10^{-2}$. Fig. 2.(g) summarizes the

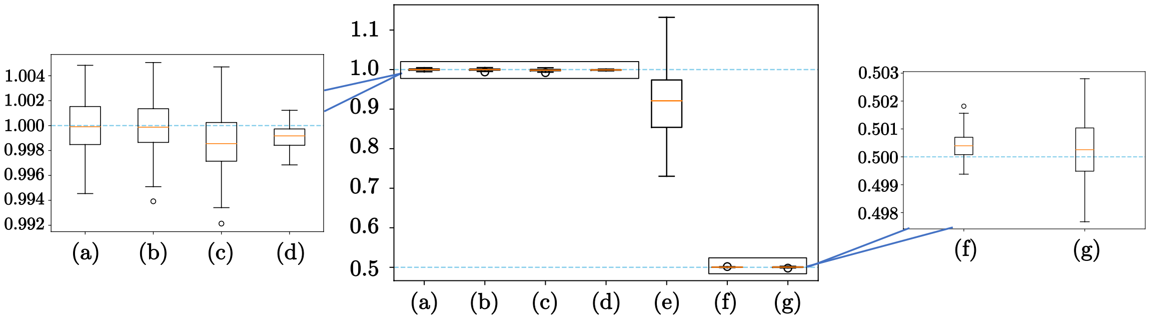


Figure 2: Perturbed GD with balanced noise (a, b and c), Perturbed GD with unbalanced noise (g) and GD (d,e and f) for the rank-1 matrix factorization problem. (a) and (d) use small initializations ($\sigma_x = \sigma_y = 10^{-2}$) and balanced step size ($\eta_x = \eta_y = 10^{-2}$). (b) and (e) use large initializations ($\sigma_x = \sigma_y = 10^{-1}$) and balanced step size. (c) and (f) use small initializations and unbalanced step size ($\eta_x = 0.5\eta_y = 5 \times 10^{-3}$).

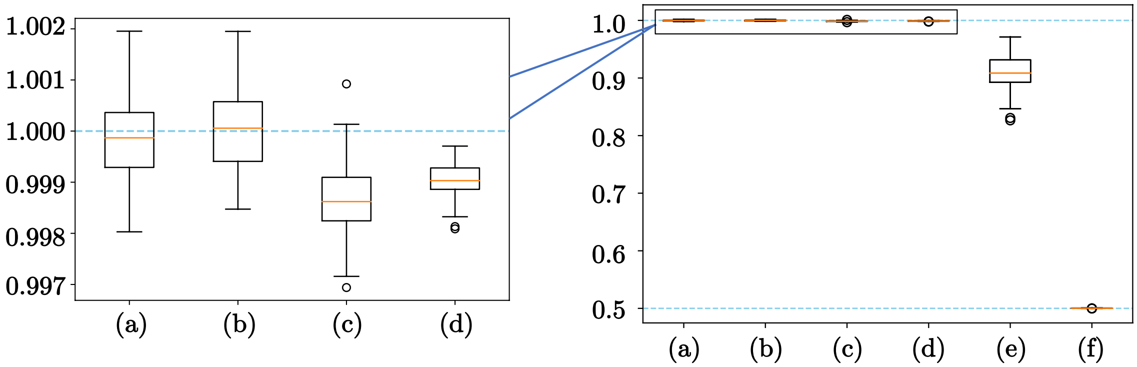


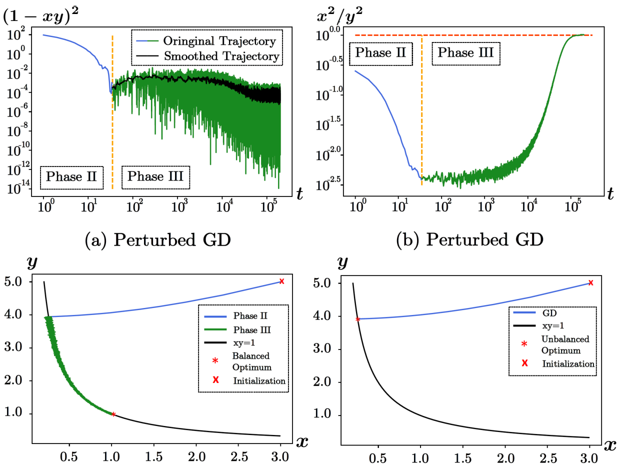
Figure 3: Perturbed GD with balanced noise (a, b and c) and GD (d,e and f) for rank-10 matrix factorization problem. (a) and (d) use small initializations ($\sigma_x = \sigma_y = 10^{-2}$) and balanced step size ($\eta_x = \eta_y = 10^{-2}$). (b) and (e) use large initializations ($\sigma_x = \sigma_y = 10^{-1}$) and balanced step size. (c) and (f) use small initializations and unbalanced step size ($\eta_x = 0.5\eta_y = 5 \times 10^{-3}$).

results of 100 repeated simulations in a box-plot. As can be seen, for $\gamma \neq 1$, the Perturbed GD converges to the γ-balanced optima.

Rank-10 Matrix Factorization. We then consider rank-10 nonconvex matrix factorization problem. The matrix M to be factorized is given by $M = U_* V_*^\top$, where $U_* = \left(I_{10}, 0\right)^\top_{d_1 \times 10}, V_* = \left(I_{10}, 0\right)^\top_{d_2 \times 10}$, with $d_1 = 20$ and $d_2 = 30$. We initialize iterates (X_0, Y_0) with all entries ${X_0}^{(i,j)}$'s and ${Y_0}^{(i,j)}$'s independently sampled from $N(0, \sigma_x^2)$ and $N(0, \sigma_y^2)$, respectively. For all experiments, we use γ-balanced noise in Perturbed GD. Specifically, we choose $\xi_{1,t}$ and $\xi_{2,t}$ with i.i.d. elements drawn from $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$ respectively. We repeat a similar set of experiments as in rank-1 case, and summarize the results in Fig. 3. For each of the experiments, we use the same set of $(\eta_x, \eta_y, \sigma_x, \sigma_y)$ as their counterpart in the rank-1 case. As can be seen, Perturbed GD always converges to the $\gamma-$balanced optima, regardless of initializations. Our experiments suggest that, for the rank-r matrix factorization problem, the noise still determines the balancedness of the optima obtained by Perturbed GD.

Phase Transition. We further demonstrate the transition between Phase II and Phase III in the Perturbed GD algorithm. Specifically, we consider 2-dimensional problem $f(x, y) = (1 - xy)^2$ with balanced optima $\pm(1, 1)$. We set $\sigma_1 = \sigma_2 = 0.05$, initialize $(x_0, y_0) = (3, 5)$, and use balanced step size $\eta_x = \eta_y = 0.01$. We repeat the experiments 50 times and summarize the result of one realization in Fig. 4, as the convergence properties of Perturbed GD are highly consistent across different realizations. We also use exponential moving average to smooth the loss trajectory to better illustrate the overall progress of the objective in Phase III.

As can be seen, in around the first 30 to 40 iterations, Perturbed GD and GD behave similarly. Both Perturbed GD and GD iterate towards the set of global optima $\{(x, y) \big| xy = 1\}$, while driving the loss to zero, and the squared norm ratio x_t^2/y_t^2 in Perturbed GD decreases from 0.36 to around 0.004. After that, GD converges to the unbalanced optimum. Since the loss is sufficiently small, the noise dominates the update of Perturbed GD. Then the squared norm ratio x_t^2/y_t^2 gradually increases from 0.004 to 1. Perturbed GD



(a) Perturbed GD

(b) Perturbed GD

(c) Perturbed GD

(d) GD

Figure 4: Algorithmic behaviors of Perturbed GD and GD. For Perturbed GD, phase transition happens around the first 30∼40 iterations, as shown in (a,b,c). GD does not show phase transitions.

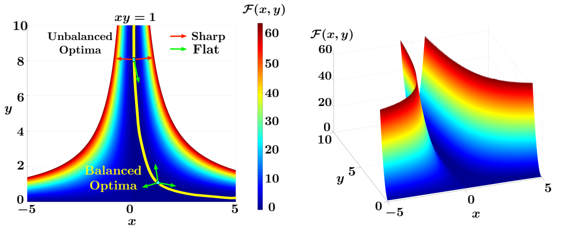


Figure 5: The visualization of objective $\mathcal{F}(x,y) = (1-xy)^2$. All the global optima are connected and form a path. The landscape around the path forms a valley. Around unbalanced optima, the landscape is sharp in some directions and flat in others. Around balanced optima, the landscape only contains flat directions.

iterates towards the balanced optimum while staying close to global optima. We remark that the phase transition phenomenon can also been observed for higher dimensional problems. Due to space limit, please refer to Figure 6 in Appendix 6 for the experimental result.

6 Discussions

Connections to SGD. This paper studies the implicit bias of the noise in nonconvex optimization. Direct analysis on SGD is beyond current technical limit due to complex dependencies. Specifically, the noise in SGD comes from random sampling of the training data, and heavily depends on the iterate. This induces a complex dependency between iterate and the noise, and makes it difficult to characterize the distribution of the noise.

Our Perturbed GD can be viewed as a close variant of SGD. Moreover, the noise in Perturbed GD follows Gaussian distribution and is independent of the iterates. Hence, the analysis of Perturbed GD, though still highly non-trivial, is now technically manageable.

Biased Gradient Estimator. Different from SGD, Perturbed GD implements a biased gradient estimator, i.e., $\mathbb{E}_{\xi_1,\xi_2}\nabla\mathcal{F}(x+\xi_1, y+\xi_2) \neq \nabla\mathcal{F}(x,y)$. Such biased gradient also appears in training deep neural networks combined with computational heuristics. Specifically, Luo et al. (2018) show that with batch normalization, the gradient estimator in SGD is also biased with respect to the original loss. The similarity between this biased gradient and our perturbed gradient is worth future investigation.

Extension to Other Types of Noise. Our work considers Gaussian noise, but can be extended to analyzing other types of noise. For example, we can show that anisotropic noise will have different smoothing effects along different directions. The implicit bias will thus depend on the covariance of noise in addition to the noise level. For another example, heavy tailed distribution of noise will affect the probability of the convergence of Perturbed GD. It may be difficult to achieve high probability convergence as we have shown for light tailed distributions.

Sharp/Flat Minima and Phase Transition. Lemma 2 shows that the Hessian matrix of an unbalanced global optimum is ill-conditioned, and the landscape around such an optimum is sharp in some directions and flat in others. For nonconvex matrix factorization, all the global optima are connected and form a path. The landscape around the path forms a valley, which is narrow around unbalanced optima and wide around the balanced ones (See Figure 5). Our three-phase convergence analysis shows that the Perturbed GD first falls into the valley, and then traverses within the valley until it finds the balanced optima.

As we have mentioned earlier, people have shown that the local optima of deep neural networks are also connected. Thus, the phase transition also provides a new perspective to explain the plateau of training curves after learning rate decay in training neural networks. Our analysis suggests that after adjusting the learning rate, the algorithm enters a new phase, where the noise slowly re-adjusts the landscape. At the beginning of this phase, the loss decreases rapidly due to the reduced noise level. By the end of this phase, the algorithm falls into a region with benign landscape that is suitable for further decreasing the step size.

Related Literature. Implicit bias of noise has also been studied in HaoChen et al. (2020); Blanc et al. (2020). However, they consider perturbing labels while our work considers perturbing parameters. In a broader sense, our work is also related to Li et al. (2019b) which show that the noise scale of SGD may change the learning order of patterns. However, they do not study the implicit bias of noise towards certain optima.

References

Allen-Zhu, Z., Li, Y., and Song, Z. (2018). A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*.

Attias, H. (2000). A variational baysian framework for graphical models. In *Advances in neural information processing systems*, pages 209–215.

Bhojanapalli, S., Kyrillidis, A., and Sanghavi, S. (2016). Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582.

Blanc, G., Gupta, N., Valiant, G., and Valiant, P. (2020). Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on Learning Theory*, pages 483–513.

Chen, Y. and Wainwright, M. J. (2015). Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*.

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

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. A. (2018). Essentially no barriers in neural network energy landscape. *arXiv preprint arXiv:1803.00885*.

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

Duchi, J. C., Bartlett, P. L., and Wainwright, M. J. (2012). Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. (2018). Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems*, pages 8789–8798.

Ge, R., Jin, C., and Zheng, Y. (2017). No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1233–1242. JMLR. org.

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

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

HaoChen, J. Z., Wei, C., Lee, J. D., and Ma, T. (2020). Shape matters: Understanding the implicit bias of the noise covariance. *arXiv preprint arXiv:2006.08680*.

Hardt, M. (2014). Understanding alternating minimization for matrix completion. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 651–660. IEEE.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

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

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. (2017). How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org.

Jordan, M. I. et al. (2004). Graphical models. *Statistical science*, 19(1):140–155.

Kawaguchi, K. (2016). Deep learning without poor local minima. In *Advances in neural information processing systems*, pages 586–594.

Keshavan, R. H., Montanari, A., and Oh, S. (2010). Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998.

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

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kuditipudi, R., Wang, X., Lee, H., Zhang, Y., Li, Z., Hu, W., Ge, R., and Arora, S. (2019). Explaining landscape connectivity of low-cost solutions for multilayer nets. In *Advances in Neural Information Processing Systems*, pages 14574–14583.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. (2019). First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176(1-2):311–337.

Li, X., Lu, J., Arora, R., Haupt, J., Liu, H., Wang, Z., and Zhao, T. (2019a). Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization. *IEEE Transactions on Information Theory*, 65(6):3489–3514.

Li, Y., Wei, C., and Ma, T. (2019b). Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, pages 11674–11685.

Liang, S., Sun, R., Li, Y., and Srikant, R. (2018). Understanding the loss surface of neural networks for binary classification. *arXiv preprint arXiv:1803.00909*.

Lu, S., Hong, M., and Wang, Z. (2019). Pa-gd: On the convergence of perturbed alternating gradient descent to second-order stationary points for structured nonconvex optimization. In *International Conference on Machine Learning*, pages 4134–4143.

Luo, P., Wang, X., Shao, W., and Peng, Z. (2018). Towards understanding regularization in batch normalization. *arXiv preprint arXiv:1809.00846*.

Nguyen, Q. (2019). On connected sublevel sets in deep learning. *arXiv preprint arXiv:1901.07417*.

Nguyen, Q. and Hein, M. (2017). The loss surface of deep and wide neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2603–2612. JMLR. org.

Nguyen, Q. and Hein, M. (2018). The loss surface and expressivity of deep convolutional neural networks.

Salakhutdinov, R., Mnih, A., and Hinton, G. (2007). Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798. ACM.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. (2015). Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*.

Venturi, L., Bandeira, A. S., and Bruna, J. (2018). Spurious valleys in two-layer neural network optimization landscapes. *arXiv preprint arXiv:1802.06384*.

Zhao, T., Wang, Z., and Liu, H. (2015). A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567.

Zheng, Q. and Lafferty, J. (2016). Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*.

Zhou, M., Liu, T., Li, Y., Lin, D., Zhou, E., and Zhao, T. (2019). Towards understanding the importance of noise in training neural networks. *arXiv preprint arXiv:1909.03172*.