Feature learning via mean-field neural networks and anisotropic features

Taiji Suzuki[1,2]
Joint work with Denny Wu[3], Atsushi Nitanda[2,4], Kazusato Oko[1,2]
[1]Graduate School of Information Science and Technology, the University of Tokyo,
[2]RIKEN Center for Advanced Intelligence Project,
[3]Center for Data Science, New York University,
[4]Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology.

Abstract

Neural network in the mean-field regime is known to be capable of *feature learning*, unlike the kernel (NTK) counterpart. Recent works have shown that mean-field neural networks can be globally optimized by a noisy gradient descent update termed the *mean-field Langevin dynamics* (MFLD). However, all existing guarantees for MFLD only considered the *optimization* efficiency, and it is unclear if this algorithm leads to improved *generalization* and sample complexity due to the presence of feature learning. To fill this gap, in this work we study the statistical and computational complexity of MFLD in learning a class of binary classification problems. Unlike existing margin bounds for neural networks, we avoid the typical norm control via the perspective that MFLD optimizes the *distribution* of parameters rather than the parameter itself; this leads to an improved analysis of the sample complexity and convergence rate. We apply our general framework to the learning of k-sparse parity functions, where we prove that unlike kernel methods, two-layer neural networks optimized by MFLD achieves a sample complexity where the degree k is "decoupled" from the exponent in the dimension dependence.

In addition to that, we consider a setting where the input feature has anisotropic property, that is, it has a covariance that is not the identity matrix. Then, we show that such an anisotropic property helps reduce generalization error and computational complexity for the MFLD. This implies that the anisotropic property of the real world dataset would help the training of neural network in terms of both statistical and computational performances.

1 Introduction

Mean-field Langevin dynamics. The optimization dynamics of two-layer neural networks in the mean-field regime can be described by a nonlinear partial differential equation of the distribution of parameters (Nitanda and Suzuki, 2017; Chizat and Bach, 2018; Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018; Sirignano and Spiliopoulos, 2020). Such a description has multiple advantages: (*i*) global convergence guarantees can be obtained by exploiting convexity of the loss function, and (*ii*) the parameters are allowed to evolve away from initialization and learn informative features, in contrast to the *neural tangent kernel* ("lazy") regime (Jacot et al., 2018).

Among the gradient-based optimization algorithms for mean-field neural networks, the *mean-field Langevin dynamics* (MFLD) (Mei et al., 2018; Hu et al., 2019) is particularly attractive due to the recently established *quantitative* optimization guarantees. MFLD arises from a noisy gradient descent update on the parameters, where Gaussian noise is injected to the gradient to encourage "exploration". It has been shown that MFLD globally optimizes an *entropy-regularized convex functional* in the space of measures, and for the infinite-width and continuous-time dynamics, the convergence rate is exponential under suitable isoperimetric conditions (Nitanda et al., 2022; Chizat, 2022). Furthermore, uniform-in-time estimates of the particle discretization error have also been established (Suzuki et al., 2023a; Chen et al., 2022), meaning that optimization guarantees for the infinite-dimensional problem can be effectively translated to a finite-width neural network.

However, existing analyses of MFLD only considered the *optimization* of neural networks; this alone does not demonstrate the benefit of mean-field regime nor the presence of feature learning. Therefore, an important problem is to characterize the *generalization* of the learned models, and prove efficient sample complexity guarantees. The goal of this work is to address the following question.

Can we show that neural network + simple noisy gradient descent (MFLD) efficiently learns an interesting class of functions, with a better rate of convergence compared to the "lazy" regime?

Learning sparse parity functions. One particularly relevant learning task is the k-sparse parity problem, where the response y is given by the sign of the product of k coordinates of the input (on hypercube); as a special case, setting $k = 2$ recovers the classical XOR problem. When $k \ll d$, this target function is low-dimensional, and hence we expect feature learning to be beneficial in that it can "zoom in" to relevant subspace. In contrast, for kernel methods (including neural networks in the lazy regime) which cannot adapt to such structure, it has been shown that a sample complexity of $n = \Omega(d^k)$ is unavoidable (Ghorbani et al., 2019; Hsu; Abbe et al., 2022).

For the XOR case ($k = 2$), recent works have shown that neural networks in the mean-field regime can achieve a sample complexity of $n = \mathcal{O}(d/\epsilon)$ (Wei et al., 2019; Chizat and Bach, 2020; Telgarsky, 2023), which indeed improves upon the NTK complexity (Ji and Telgarsky, 2019). However, all these results directly assumed convergence of the dynamics ($t \to \infty$) with no iteration complexity. Moreover, Wei et al. (2019); Chizat and Bach (2020) directly analyzed the infinite-width limit, and while Telgarsky (2023) provided a finite-width characterization, the dynamics is restricted to the low-rotation regime, and a very large number of particles $N = \mathcal{O}(d^d)$ is required. Lastly, these analyses are specialized to XOR, and do not directly generalize to the k-parity setting.

1.1 Our Contributions

In this work, we bridge the aforementioned gap by presenting a simple and general framework to establish sample complexity of MFLD in learning binary classification problems. We then apply this framework to the sparse k-parity problem, and obtain improved rate of convergence for the fully time- and space-discretized algorithm. More specifically, our contributions can be summarized as follows.

- We present a general framework to analyze MFLD in the learning of binary classification tasks. Our framework has two main ingredients: (i) an annealing procedure that applies to common classification losses that removes the exponential dependence on regularization parameters in the *logarithmic Sobolev inequality*, and (ii) a novel local Rademacher complexity analysis for the distribution of parameters optimized by MFLD. As a result, we can obtain learning guarantees for the mean-field neural network in discrete-time and finite-width settings.
- We apply our general framework to the k-sparse parity problem, and derived learning guarantees with improved rate of convergence and dimension dependence. Specially, in the $n \asymp d^2$ regime we obtain exponentially converging classification error, whereas in the $n \asymp d$ regime we achieve linear dimension dependence. Note that this improves upon the NTK analysis (which gives a sample complexity of $n = \Omega(d^k)$) in that it "decouples" the degree k from the exponent in the dimension dependence. Our theoretical results are supported by empirical findings.

2 Problem Setting

Throughout this paper, we consider a classification problem given by the following model:

$$Y = \mathbf{1}_A(Z) - \mathbf{1}_{A^c}(Z) \in \{\pm 1\}$$

where $Z = (Z_1, \ldots, Z_d)$ is the input random variable on $\mathbb{R}^d$ and $\mathbf{1}_A$ is the indicator function corresponding to a measurable set $A \in \mathbb{B}(\mathbb{R}^d)$, i.e., $\mathbf{1}_A(Z) = 1$ if $Z \in A$ and $\mathbf{1}_A(Z) = 0$ if $Z \notin A$. Let P_Z be the distribution of Z. We are given input-output pairs $D_n = (z_i, y_i)_{i=1}^n$ independently identically distributed from this model as training data. Then, we construct a binary classifier that predicts the label for the test input data as accurate as possible. To achieve this, we learn a two-layer neural network model in the mean-field regime via the *mean-field Langevin dynamics*.

One important problem setting for our analysis is the k-sparse parity problem defined as follows.

Example 1 (k-sparse parity problem). *P_Z is the uniform distribution on the grid $\{\pm 1/\sqrt{d}\}^d$ and $A = \{\zeta = (\zeta_1, \dots, \zeta_d) \in \{\pm 1/\sqrt{d}\}^d \mid \zeta_1 \cdots \zeta_k > 0\}$*[1].

As a special case, $k = 2$ (XOR) has been extensively studied (Wei et al., 2019; Telgarsky, 2023).

Mean-field two-layer network. Given input z, let $h_x(z)$ be one neuron in a two-layer neural network with parameter $x = (x_1, x_2, x_3) \in \mathbb{R}^{d+1+1}$ defined as

$$h_x(z) = \bar{R}[\tanh(z^\top x_1 + x_2) + 2\tanh(x_3)]/3,$$

where $\bar{R} \in \mathbb{R}$ is a hyper-parameter determining the scale of the network. We place an extra tanh activation for the bias term $x_3 \in \mathbb{R}$ because the boundedness of h_x is required in the convergence analysis. Let $\mathcal{P}$ be the set of probability measures on $(\mathbb{R}^{\bar{d}}, \mathcal{B}(\mathbb{R}^{\bar{d}}))$ where $\bar{d} = d + 2$ and $\mathcal{B}(\mathbb{R}^{\bar{d}})$ is the Borel σ-algebra on $\mathbb{R}^{\bar{d}}$ and $\mathcal{P}_p$ be the subset of $\mathcal{P}$ such that its p-th moment is bounded: $\mathbb{E}_\mu[\|X\|^p] < \infty$ ($\mu \in \mathcal{P}$). The mean-field neural network is defined as an integral over neurons h_x,

$$f_\mu(\cdot) = \int h_x(\cdot)\mu(\mathrm{d}x),$$

for $\mu \in \mathcal{P}$. To evaluate the performance of f_μ, we define the empirical risk and the population risk as

$$L(\mu) := \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(y_i f_\mu(z_i)), \quad \bar{L}(\mu) := \mathbb{E}[\ell(Y f_\mu(Z))],$$

respectively, where $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a convex loss function. In particular, we consider the logistic loss $\ell(f, y) = \log(1 + \exp(-yf))$ for $y \in \{\pm 1\}$ and $f \in \mathbb{R}$. To avoid overfitting, we consider a regularized empirical risk $F(\mu) := L(\mu) + \lambda \mathbb{E}_{X\sim\mu}[\lambda_1 \|X\|^2]$, where $\lambda, \lambda_1 \geq 0$ are regularization parameters. One advantage of this mean-field definition is that f_μ is a linear with respect to μ, and hence the functional $L(\mu)$ becomes a convex functional.

Mean-field Langevin dynamics. We optimize the training objective via MFLD, which is given by the following stochastic differential equation:

$$\mathrm{d}X_t = -\nabla \frac{\delta F(\mu_t)}{\delta \mu}(X_t)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}W_t, \quad \mu_t = \mathrm{Law}(X_t), \tag{1}$$

where $X_0 \sim \mu_0$, $\mathrm{Law}(X)$ denotes the distribution of the random variable X and $(W_t)_{t\geq 0}$ is the d-dimensional standard Brownian motion. Readers may refer to Theorem 3.3 of Huang et al. (2021) for the existence and uniqueness of the solution. Here, $\frac{\delta F(\mu_t)}{\delta \mu}$ is the first variation of F.

Definition 1. *For a functional $G : \mathcal{P} \to \mathbb{R}$, the first-variation $\frac{\delta G}{\delta \mu}(\mu)$ at $\mu \in \mathcal{P}$ is a continuous functional $\mathcal{P} \times \mathbb{R}^d \to \mathbb{R}$ satisfying $\lim_{\epsilon\to 0} \frac{G(\epsilon\nu + (1-\epsilon)\mu)}{\epsilon} = \int \frac{\delta G}{\delta \mu}(\mu)(x)\mathrm{d}(\nu - \mu)$ for any $\nu \in \mathcal{P}$.*

In our setting, we have $\frac{\delta F(\mu)}{\delta \mu}(x) = \frac{1}{n}\sum_{i=1}^n \ell'(y_i f_\mu(z_i)) y_i h_x(z_i) + \lambda(\lambda_1 \|x\|^2)$. It is known that the Fokker-Planck equation of the SDE (1) is given by[2]

$$\partial_t \mu_t = \lambda \Delta \mu_t + \nabla \cdot \left[\mu_t \nabla \tfrac{\delta F(\mu_t)}{\delta \nu}\right] = \nabla \cdot \left[\mu_t \nabla \left(\lambda \log(\mu_t) + \tfrac{\delta F(\mu_t)}{\delta \nu}\right)\right]. \tag{2}$$

Then, we can verify that this is equivalent to the Wasserstein gradient flow to optimize the following entropy regularized risk (Mei et al., 2018; Hu et al., 2019):

$$\mathcal{L}(\mu) = F(\mu) + \lambda \mathrm{Ent}(\mu) = L(\mu) + \lambda \mathrm{KL}(\nu, \mu) + (\text{const.}) \tag{3}$$

where $\mathrm{KL}(\nu, \mu) = \int \log(\mu/\nu)\mathrm{d}\mu$ is the KL divergence between ν and μ, and ν is the Gaussian distribution with mean 0 and variance $I/(2\lambda_1)$, i.e., $\nu = \mathcal{N}(0, I/(2\lambda_1))$.

[1] We present the axis-aligned setting for conciseness, but the same result holds under orthogonal transforms.

[2] This should be interpreted in a weak sense, that is, for any continuously differentiable function ϕ with a compact support, $\int \phi \mathrm{d}\mu_t - \int \phi \mathrm{d}\mu_s = -\int_s^t \int \nabla\phi \cdot (\nabla \log(\mu_t) - \nabla \frac{\delta F(\mu_t)}{\delta \nu})\mathrm{d}\mu_\tau \mathrm{d}\tau$.

For a practical algorithm, we need to consider a space- and time-discretized version of the MFLD, that is, we approximate the solution μ_t by an empirical measure $\mu_{\mathscr{X}} = \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$ corresponding to a set of finite particles $\mathscr{X} = (X^i)_{i=1}^N \subset \mathbb{R}^{\bar{d}}$. Let $\mathscr{X}_\tau = (X_\tau^i)_{i=1}^N \subset \mathbb{R}^{\bar{d}}$ be N particles at the τ-th update ($\tau \in \{0, 1, 2, \dots\}$), and define $\mu_\tau = \mu_{\mathscr{X}_\tau}$ as a finite particle approximation of the population counterpart. Then, the discretized MFLD is defined as follows: $X_0^i \sim \mu_0$, and $\mathscr{X}_\tau$ is updated as

$$X_{\tau+1}^i = X_\tau^i - \eta\nabla\frac{\delta F(\mu_\tau)}{\delta\mu}(X_\tau^i) + \sqrt{2\lambda\eta}\xi_\tau^i, \tag{4}$$

where $\eta > 0$ is the step size, and $\xi_\tau^i \sim_{i.i.d.} N(0, I)$. This is the Euler-Maruyama approximation of the MFLD with a discretized measure; we present the discretization error bounds in the next section.

3 Main Assumptions and Theoretical Tools

In this section, we introduce the basic assumptions and technical tools for our analysis.

Condition on the loss function To derive the convergence of the classification error, we assume that the loss function satisfies the following condition.

Assumption 1. *The convex loss function $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$ satisfies the following conditions:*

- *ℓ is first order differentiable, its derivative is Lipschitz continuous and its derivative is bounded by 1: $|\ell'(x) - \ell'(x')| \leq C|x - x'|$ and $\sup_x |\ell'(x)| \leq 1$.*
- *ℓ is monotonically decreasing, and is classification calibrated: $\ell'(0) < 0$ (Bartlett et al., 2006).*
- *$\psi(u)^{-1} := \ell(0) - (\ell(u) - u\ell'(u)) > 0$ for any $u > 0$.*

This standard assumption is satisfied by several loss functions such as the logistic loss. We remark that the first assumption is used to show the well-definedness of the mean-field Langevin dynamics and derive its discretization error, and also to obtain a uniform generalization error bound through the classical contraction argument (Boucheron et al., 2013; Ledoux and Talagrand, 1991). The second and third assumptions are used to show the convergence of classification error of our estimator.

Logarithmic Sobolev inequality. Nitanda et al. (2022); Chizat (2022) showed that the convergence of MFLD crucially relies on properties of the *proximal Gibbs distribution* whose density is given by $p_\mu(X) \propto \exp\left(-\frac{1}{\lambda}\frac{\delta F(\mu)}{\delta\mu}(X)\right)$, where $\mu \in \mathcal{P}$. By the smoothness of the loss function (Assumption 1) and the tanh activation, we can show that the objective $\mathcal{L}$ has a unique solution μ^* which is also a proximal Gibbs measure of itself (Proposition 2.5 of Hu et al. (2019)). The next question is how fast the solution μ_t converges to the optimal solution μ^*. As we will see, the convergence of MFLD heavily depends on a *logarithmic Sobolev inequality* (LSI) on p_μ.

Definition 2 (Logarithmic Sobolev inequality)**.** *Let μ be a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. μ satisfies the LSI with constant $\alpha > 0$ if for any smooth function $\phi : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}_\mu[\phi^2] < \infty$, $\mathbb{E}_\mu[\phi^2 \log(\phi^2)] - \mathbb{E}_\mu[\phi^2]\log(\mathbb{E}_\mu[\phi^2]) \leq \frac{2}{\alpha}\mathbb{E}_\mu[\|\nabla\phi\|_2^2]$.*

This is equivalent to the condition that the KL divergence from μ is bounded by the Fisher divergence: $\int \log(\mathrm{d}\nu/\mathrm{d}\mu)\mathrm{d}\nu \leq \frac{2}{\alpha}\int \|\nabla\log(\mathrm{d}\nu/\mathrm{d}\mu)\|^2\mathrm{d}\mu$, for any $\nu \in \mathcal{P}$ which is absolutely continuous with respect to μ. The LSI of proximal Gibbs measure can be established via standard perturbation criteria. For $L(\mu)$ with bounded first-variation, we may apply the classical Bakry-Emery and Holley-Stroock arguments (Bakry and Émery, 1985; Holley and Stroock, 1987) (Corollary 5.7.2 and 5.1.7 of Bakry et al. (2014)): If $\|\frac{\delta L(\mu)}{\delta\mu}\|_\infty \leq B$ is satisfied for any $\mu \in \mathcal{P}_2$, then μ^* and $p_{\mathscr{X}}$ satisfy the LSI with

$$\alpha \geq \lambda_1 \exp\left(-4B/\lambda\right). \tag{5}$$

With the LSI condition on the proximal Gibbs distribution, it is known that the MFLD converges to the optimal solution in an exponential order by using a so-called *Entropy sandwich* technique.

Proposition 1 (Entropy sandwich (Nitanda et al., 2022; Chizat, 2022)). *Suppose that μ_0 satisfies $\mathcal{L}(\mu_0) < \infty$ and the proximal Gibbs measure p_{μ_t} corresponding to the solution μ_t has the LSI constant α for all $t \geq 0$, then the solution μ_t of MFLD satisfies*

$$\lambda \mathrm{KL}(\mu^*, \mu_t) \leq \mathcal{L}(\mu_t) - \mathcal{L}(\mu^*) \leq \exp(-2\alpha\lambda t)(\mathcal{L}(\mu_0) - \mathcal{L}(\mu^*)),$$

where $\mu^* = \mathrm{argmin}_{\mu \in \mathcal{P}} \mathcal{L}(\mu)$.

Hence we know that time horizon $T = O(\frac{1}{\lambda\alpha}\log(1/\tilde{\epsilon}))$ is sufficient to achieve $\tilde{\epsilon} > 0$ accuracy.

Convergence of the discretized algorithm. While Proposition 1 only established the convergence rate of the continuous dynamics, similar guarantee can be shown for the discretized setting. Let

$$\mathcal{L}^N(\mu^{(N)}) = N\mathbb{E}_{\mathscr{X} \sim \mu^{(N)}}[F(\mu_{\mathscr{X}})] + \lambda \mathrm{Ent}(\mu^{(N)}),$$

where $\mu^{(N)}$ is a distribution of N particles $\mathscr{X} = (X^i)_{i=1}^N \subset \mathbb{R}^{\bar{d}}$. Let $\mu_\tau^{(N)}$ be the distribution of the particles $\mathscr{X}_\tau = (X_\tau^i)_{i=1}^N$ at the τ-th iteration. Suzuki et al. (2023b) showed that, if $\lambda\alpha\eta \leq 1/4$ and $\eta \leq 1/4$, then for $\bar{B}^2 := \mathbb{E}[\|X_0^i\|^2] + \frac{1}{\lambda\lambda_1}\left[\left(\frac{1}{4} + \frac{1}{\lambda\lambda_1}\right)\bar{R}^2 + \lambda d\right] = O(d + \lambda^{-2})$ and $\delta_\eta := C_1\bar{L}^2(\eta^2 + \lambda\eta)$, where $\bar{L} = 2\bar{R} + \lambda\lambda_1 = O(1)$ and $C_1 = 8(\bar{R}^2 + \lambda\lambda_1\bar{B}^2 + d) = O(d + \lambda^{-1})$,

$$\frac{1}{N}\mathbb{E}[\mathcal{L}^N(\mu_\tau^{(N)})] - \mathcal{L}(\mu^*) \leq \exp\left(-\frac{\lambda\alpha\eta\tau}{2}\right)\left(\frac{\mathbb{E}[\mathcal{L}^N(\mu_0^{(N)})]}{N} - \mathcal{L}(\mu^*)\right) + \frac{4}{\lambda\alpha}\bar{L}^2C_1(\lambda\eta + \eta^2) + \frac{4C_\lambda}{\lambda\alpha N},$$

where C_λ is a constant depending on λ. In particular, for a given $\tilde{\epsilon} > 0$, the right hand side can be bounded by $\tilde{\epsilon} + \frac{4C_\lambda}{\lambda\alpha N}$ after $T = O\left(\frac{\bar{L}^2C_1}{\alpha\tilde{\epsilon}} + \frac{\bar{L}\sqrt{C_1}}{\sqrt{\lambda\alpha\tilde{\epsilon}}}\right)\frac{1}{\lambda\alpha}\log(1/\tilde{\epsilon})$ iterations with the step size $\eta = O\big(\big(\frac{\bar{L}^2C_1}{\alpha\tilde{\epsilon}} + \frac{\bar{L}\sqrt{C_1}}{\sqrt{\lambda\alpha\tilde{\epsilon}}}\big)^{-1}\big)$. Furthermore, the convergence of the loss function can be connected to the convergence of the function value of the neural network as follows,

$$\begin{aligned}
&\mathbb{E}_{\mathscr{X}_\tau \sim \mu_k^{(N)}}\left[\sup_{z \in \mathrm{supp}(P_Z)}(f_{\mu_{\mathscr{X}_\tau}}(z) - f_{\mu^*}(z))^2\right] \\
&\leq \frac{4\bar{L}^2}{\lambda\alpha}\left(\frac{\mathcal{L}^N(\mu_\tau^{(N)})}{N} - \mathcal{L}(\mu^*)\right) + 2\mathbb{E}_{\mathscr{X}_* \sim (\mu^*)^{\otimes N}}\left[\sup_{z \in \mathrm{supp}(P_Z)}\left(\frac{1}{N}\sum_{i=1}^N h_{X_*^i}(z) - \int h_x(z)\mathrm{d}\mu^*(z)\right)^2\right].
\end{aligned}$$

Here the second term in the right hand side can be bounded by $\frac{32\bar{R}^2}{N}\left[1 + 2\left(\frac{2\bar{R}^2}{(\lambda\lambda_1)^2} + \frac{\bar{d}}{\lambda_1}\right)\right]$, if $\|z\| \leq 1$ for any $z \in \mathrm{supp}(P_Z)$ as in the k-sparse parity problem. Hence, by taking the number of particles as $N = \epsilon^{-2}[(\lambda\alpha)^{-2} + (\lambda\lambda_1)^{-2} + d/\lambda_1]$ and letting $\tilde{\epsilon} = \lambda\alpha\epsilon^2$ with the choice of T and η as described above, we have $\sup_{z \in \mathrm{supp}(P_Z)} |f_{\mu_{\mathscr{X}_T}}(z) - f_{\mu^*}(z)| = O_p(\epsilon)$.

Assumptions on the model specification. We restrict ourselves to the situation where a perfect classifier with margin c_0 is included in our model class, stated in the following assumption.

Assumption 2. *There exists $c_0 > 0$ and $R > 0$ such that the following conditions are satisfied:*

- *For some $\bar{R}$, there exists $\mu^* \in \mathcal{P}$ such that $\mathrm{KL}(\nu, \mu^*) \leq R$ and $L(\mu^*) \leq \ell(0) - c_0$.*
- *For any $\lambda < c_0/R$, the regularized expected risk minimizer $\mu_{[\lambda]} := \mathrm{argmin}\, L(\mu) + \lambda\mathrm{KL}(\nu, \mu)$ satisfies $Yf_{\mu_{[\lambda]}}(X) \geq c_0$ almost surely.*

Importantly, we can apply the same general analysis for different classification problems, as long as Assumption 2 is verified. The advantage of this generality is that we do not need to tailor our convergence proof for individual learning problems. Note that the convergence rate of MFLD is strongly affected by the values of $\bar{R}$ and R; therefore, it is crucial to establish this condition using the smallest possible values of $\bar{R}$ and R, in order to obtain a tight bound of the classification error. As an illustrative example, we now show that the k-sparse parity estimation satisfies the above assumption.

Example: k-sparse parity estimation. In the k-sparse parity setting (Example 1), Assumption 2 is satisfied with constants specified in the following propositions.

Proposition 2 (k-sparse parity)**.** *Under Assumption 1 and for $\bar{R} = k$, there exists $\mu^* \in \mathcal{P}$ such that*

$$\mathrm{KL}(\nu, \mu^*) \le c_1 k \log(k)^2 d \ (= R),$$

and $L(\mu^*) \le \ell(0) - c_2$, where $c_1, c_2 > 0$ are absolute constants.

Proposition 3. *Under Assumption 1 and the settings of R and $\bar{R}$ given in Proposition 2, if $\lambda < c_2/(2R)$, then $\mu_{[\lambda]}$ satisfies*

$$\max\{\bar{L}(\mu_{[\lambda]}), L(\mu_{[\lambda]})\} \le \ell(0) - c_2 + \lambda R < \ell(0) - \tfrac{c_2}{2},$$

and $f_{\mu_{[\lambda]}}$ is a perfect classifier with margin c_2, i.e., $Y f_{\mu_{[\lambda]}}(X) \ge \frac{c_2}{2}$.

In other words, Assumption 2 is achieved with $R = O(k \log(k)^2 d)$, $\bar{R} = k$ and $c_0 = c_2/2$. By substituting these values of R and $\bar{R}$ to our general results presented below, we can easily derive a bound for the classification error of the MFLD estimator.

4 Main Result

4.1 Generalization Error Analysis

We utilize the *local Rademacher complexity* (Mendelson, 2002; Bartlett et al., 2005; Koltchinskii, 2006; Giné and Koltchinskii, 2006) to obtain a faster generalization error rate. For the function class of mean-field neural networks, we introduce $\mathcal{F} := \{f_\mu \mid \mu \in \mathcal{P}\}$, and the KL-constrained model class $\mathcal{F}_M(\mu^\circ) := \{f_\mu \mid \mu \in \mathcal{P},\ \mathrm{KL}(\mu^\circ, \mu) \le M\}$ for $\mu^\circ \in \mathcal{P}$ and $M > 0$. The Rademacher complexity of a function class $\tilde{\mathcal{F}}$ is defined as

$$\mathrm{Rad}(\tilde{\mathcal{F}}) := \mathbb{E}_{\varepsilon_i, z_i}\left[\sup_{f \in \tilde{\mathcal{F}}} \tfrac{1}{n} \textstyle\sum_{i=1}^n \epsilon_i f(z_i)\right],$$

where $(z_i)_{i=1}^n$ are i.i.d. observations from P_Z and $(\varepsilon_i)_{i=1}^n$ is an i.i.d. Rademacher sequence ($P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$)). We have the following bound on the Rademacher complexity of the function class $\mathcal{F}_M(\mu^\circ)$.

Lemma 1 (Local Rademacher complexity of $\mathcal{F}_M(\mu^\circ)$, Chen et al. (2020) adapted)**.** *For any fixed $\mu^\circ \in \mathcal{P}$ and $M > 0$, it holds that* $\mathrm{Rad}(\mathcal{F}_M(\mu^\circ)) \le 2\bar{R}\sqrt{\frac{M}{n}}$.

Combining this local Rademacher complexity bound with the *peeling device* argument (van de Geer, 2000), we can roughly obtain the following estaimte (note that this is an informal derivation):

$$\underbrace{\bar{L}(\hat{\mu}) - \bar{L}(\mu^*) - (\hat{\mu} - \mu^*)\frac{\delta \bar{L}(\mu^*)}{\delta \mu}}_{\text{(II)}} + \underbrace{\lambda \mathrm{KL}(\mu^*, \hat{\mu})}_{\text{(I)}} \lesssim \sqrt{\frac{\mathrm{KL}(\mu^*, \hat{\mu})}{n}} \lesssim \frac{1}{n\lambda} + \lambda \mathrm{KL}(\mu^*, \hat{\mu}), \tag{6}$$

with high probability, where $\hat{\mu} = \mathrm{argmin}_{\mu \in \mathcal{P}} \mathcal{L}(\mu)$, $\mu^* = \mathrm{argmin}_{\mu \in \mathcal{P}} \bar{L}(\mu) + \lambda \mathrm{KL}(\nu, \mu)$, R and $\bar{R}$ are regarded as constants, and the last inequality is by the AM-GM relation. Observe that on the left hand side, we have two non-negative terms (I) and (II). Corresponding to each term, we obtain different types of classification error bounds (Type I and Type II in the following subsection, respectively). Note that there appears a $O(1/n)$ factor in the right hand side, which cannot be obtained by a vanilla Rademacher complexity evaluation because it only yields an $O(1/\sqrt{n})$ bound. In other words, localization is essential to obtain our fast convergence rate.

We remark that a local Rademacher complexity technique is also utilized by Telgarsky (2023) to derive a $O(1/n)$ rate. They adapted a technique developed for a smooth loss function by Srebro et al. (2010), which requires the training loss $L(\hat{\mu})$ to be sufficiently small, that is, of order $L(\hat{\mu}) = O(1/n)$. In our setting, to achieve such a small training loss, we need to take large $\bar{R}$ such as $\bar{R} = \Omega(\log(n))$. Unfortunately, such a large $\bar{R}$ induces exponentially small log-Sobolev constant like $\alpha \lesssim \exp(-cd \log(n)) = n^{-cd}$. In contrast, our analysis focuses on the local Rademacher complexity around μ^*, and hence we do not require the training loss to be close to 0; instead, it suffices to have a training loss that is close to or smaller than that of μ^*.

Type I: Perfect Classification with Exponentially Decaying Error In the regime of $n = \Omega(1/\lambda^{(K)2})$, we can prove that the MFLD estimator attains a perfect classification with an exponentially converging probability by evaluating the term (I). From Eq. (6) we can establish $\mathrm{KL}(\mu^*, \hat{\mu}) \leq O_p(1/(n\lambda^{(K)2}))$; this KL divergence bound can be used to control the L^∞-norm between $f_{\hat{\mu}}$ and f_{μ^*}. Indeed, we can show that $\|f_{\hat{\mu}} - f_{\mu^*}\|_\infty^2 \leq 2\bar{R}^2 \mathrm{KL}(\mu^*, \hat{\mu})$ (see the proof of Theorem 1). Then, under the margin assumption of f_{μ^*} (Assumption 2), we have that $f_{\hat{\mu}}$ also yields a Bayes optimal classifier. More precisely, we have the following theorem.

Theorem 1. *Suppose Assumptions 1 and 2 hold. Let* $M_0 = (\epsilon^* + 2(\bar{R}+1))/\lambda^{(K)}$. *Moreover, suppose that* $\lambda^{(K)} < c_0/R$ *and*

$$Q := c_0^2 - \frac{4\bar{R}^2}{n\lambda^{(K)2}}\left[\lambda^{(K)}\left(4\bar{R} + \frac{\lambda^{(K)}}{32\bar{R}^2 n}\right) + 8\bar{R}^2(4 + \log\log_2(8n^2 M_0 \bar{R})) + n\lambda^{(K)}\epsilon^*\right] > 0,$$

then $f_{\hat{\mu}}$ *yields perfect classification, i.e.,* $P\left(Y f_{\hat{\mu}}(Z) > 0\right) = 1$, *with probability* $1 - \exp(-\frac{n\lambda^{(K)2}}{32\bar{R}^4}Q)$.

Type II: Polynomial Order Classification Error Next we evaluate the classification error bound from term (II) in Eq. (6). In this case, we do not require an L^∞-norm bound as in the Type I analysis above; this results in a milder dependency on $\lambda^{(K)}$ and hence a better sample complexity.

Theorem 2. *Suppose Assumptions 1 and 2 hold. Let* $\lambda^{(K)} < c_0/R$ *and* $M_0 = (\epsilon^* + 2(\bar{R}+1))/\lambda^{(K)}$. *Then, with probability* $1 - \exp(-t)$, *the classification error of* $f_{\mu^{(K)}}$ *is bounded as*

$$P(Y f_{\mu^{(K)}}(Z) \leq 0) \leq 2\psi(c_0)\left[\frac{8\bar{R}^2}{n\lambda^{(K)}}\left(4 + t + \log\log_2(8n^2 M_0 \bar{R})\right) + \frac{1}{n}\left(4\bar{R} + \frac{\lambda^{(K)}}{32\bar{R}^2 n}\right) + \epsilon^*\right].$$

We notice that the right hand side scales with $O(1/(n\lambda^{(K)}))$, which is better than $O(1/(n\lambda^{(K)2}))$ in Theorem 1; this implies that a sample size linear in the dimensionality is sufficient to achieve small classification error. The reason for such improvement in the λ^K-dependence is that the stronger L^∞-norm convergence is not used in the proof; instead, only the convergence of the loss is utilized. On the other hand, this analysis does not guarantee a perfect classification.

4.1.1 Computational Complexity of MFLD

From the general result in Section **??**, we can evaluate the computational complexity to achieve the statistical bounds derived above. In both cases (Theorems 1 and 2), we may set the optimization error $\epsilon^* = O(1/(n\lambda^{(K)}))$. Then, the total number of iteration can be

$$\sum_{\kappa=1}^{K} T_\kappa \leq O\left((d + \lambda^{(K)-1})n\exp(16c_{\mathrm{L}}\bar{R}(R+2))\log(n\lambda^{(K)})\right).$$

The width N (the number of particles) can be taken as $N = O((\epsilon^*\lambda^{(K)}\alpha)^{-2}) = O(n^2\alpha^{-2}) = O\left(n^2\exp(16c_{\mathrm{L}}\bar{R}(R+2))\right)$.

Corollary 1 (k-sparse parity setting)**.** *In the* k*-sparse parity setting, we may take* $R = O(k\log(k)^2 d)$, $\bar{R} = k$ *and* $\lambda^{(K)} = O(1/R) = O(1/(k\log(k)^2 d))$. *Therefore, the classification error is bounded by*

$$P(Y f_{\mu^{(K)}} < 0) \leq O\left(\frac{k^2\log(k)^2 d}{n}(\log(1/\delta) + \log\log(n))\right),$$

with probability $1 - \delta$. *Moreover, if* $n = \Omega(k^6\log(k)^4 d^2)$, *then* $P(Y f_{\mu^{(K)}} > 0) = 1$ *with probability*

$$1 - \exp(-\Omega(nk^6\log(k)^4/d^2)).$$

As for the computational complexity, we require $O(k\log(k)^2 dn\log(nd)\exp[O(k^2\log(k)^2 d)])$ *iterations, and the number of particles is* $O(\exp(O(k^2\log(k)^2 d)))$.

Comparison with prior results. In the 2-sparse parity setting, neural network in the kernel (NTK) regime achieves only $O(d^2/n)$ convergence in the classification error, as shown in Ji and Telgarsky (2019) and Telgarsky (2023, Theorem 2.3), whereas we demonstrate that mean-field neural network can improve the rate to $O(d/n)$ via feature learning. Telgarsky (2023) also analyzed the learning of 2-sparse parity problem beyond the kernel regime, and showed that 2-layer ReLU neural network can achieve the best known classification error $O(d/n)$. However, their analysis considered a low-rotation dynamics and assumed convergence at $t \to \infty$, whereas our framework also provides a concrete estimate of the computational complexity. Indeed, the number of iterations can be bounded as $O(dn \log(nd) \exp[O(d)])$. In addition, while we still require exponential width $N = O(n^2 \exp(O(d)))$, such a condition is an improvement over $N = O(d^d)$ in Telgarsky (2023).

Barak et al. (2022) considered a learning method in which one-step gradient descent is performed for the purpose of feature learning, and then a network with randomly re-initialized bias units is used to fit the data. For the k-sparse parity problem, they derived a classification error bound of $O(d^{(k+1)/2}/\sqrt{n})$. In contrast, our analysis yields a much better statistical complexity of $O(k^2 \log(k)^2 d/n \wedge \exp(-\Omega(nk^6 \log(k)^4/d^2)))$, which "decouples" the degree k in the exponent of the dimension dependence.

5 Conclusion and Discussion

We provided a general framework to evaluate the classification error of a two-layer neural network trained by the mean-field Langevin dynamics. Thanks to the generality of our framework, an error bound for specific settings can be derived by directly specifying the parameters in Assumption 2 such as R, $\bar{R}$, and c_0. We also proposed an annealing procedure to alleviate the exponential dependencies in the LSI constant. As a special (but important) example, we investigated the k-sparse parity problem, for which we obtained more general and better sample complexity than existing works.

References

E. Abbe, E. B. Adsera, and T. Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

E. Abbe, E. Boix-Adsera, and T. Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. *arXiv preprint arXiv:2302.11055*, 2023.

J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, D. Wu, and G. Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=akddwRG6EGi.

J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, and D. Wu. Learning in the presence of low-dimensional structure: a spiked random matrix perspective. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

D. Bakry and M. Émery. Diffusions hypercontractives. In J. Azéma and M. Yor, editors, *Séminaire de Probabilités XIX 1983/84*, pages 177–206, Berlin, Heidelberg, 1985. Springer Berlin Heidelberg. ISBN 978-3-540-39397-9.

D. Bakry, I. Gentil, M. Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

B. Barak, B. L. Edelman, S. Goel, S. M. Kakade, eran malach, and C. Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=8XWP2ewX-im.

P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33:1487–1537, 2005.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

G. Ben Arous, R. Gheissari, and A. Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling. *Advances in Neural Information Processing Systems*, 35:25349–25362, 2022.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. ISBN 9780199535255. URL https://books.google.co.jp/books?id=koNqWRluhP0C.

F. Chen, Z. Ren, and S. Wang. Uniform-in-time propagation of chaos for mean field langevin dynamics. *arXiv preprint arXiv:2212.03050*, 2022.

S. Chen and R. Meka. Learning polynomials in few relevant dimensions. In J. Abernethy and S. Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1161–1227. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/chen20a.html.

Z. Chen, Y. Cao, Q. Gu, and T. Zhang. A generalized neural tangent kernel analysis for two-layer neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13363–13373. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/9afe487de556e59e6db6c862adfe25a4-Paper.pdf.

L. Chizat. Mean-field langevin dynamics : Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=BDqzLH1gEm.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems 31*, pages 3040–3050, 2018.

L. Chizat and F. Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

A. Damian, J. Lee, and M. Soltanolkotabi. Neural networks can learn representations with gradient descent. In P.-L. Loh and M. Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5413–5452. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/damian22a.html.

A. Daniely and E. Malach. Learning parities with neural networks. *arXiv preprint arXiv:2002.07400*, 2020.

B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019.

B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. When do neural networks outperform kernel methods? *arXiv preprint arXiv:2006.13409*, 2020.

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

R. Holley and D. Stroock. Logarithmic sobolev inequalities and stochastic ising models. *Journal of statistical physics*, 46(5-6):1159–1194, 1987.

D. Hsu. Dimension lower bounds for linear approaches to function approximation.

K. Hu, Z. Ren, D. Siska, and L. Szpruch. Mean-field langevin dynamics and energy landscape of neural networks. *arXiv preprint arXiv:1905.07769*, 2019.

X. Huang, P. Ren, and F.-Y. Wang. Distribution dependent stochastic differential equations. *Frontiers of Mathematics in China*, 16(2):257–301, 2021.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, pages 8580–8589, 2018.

Z. Ji and M. Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34:2593–2656, 2006.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces. Isoperimetry and Processes.* Springer, New York, 1991. MR1102015.

S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48:1977–1991, 2002.

A. Mousavi-Hosseini, D. Wu, T. Suzuki, and M. A. Erdogdu. Gradient-based feature learning under structured data. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

A. Nitanda and T. Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

A. Nitanda, D. Wu, and T. Suzuki. Convex analysis of the mean field langevin dynamics. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9741–9757. PMLR, 28–30 Mar 2022.

M. Refinetti, S. Goldt, F. Krzakala, and L. Zdeborova. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8936–8947. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/refinetti21b.html.

G. M. Rotskoff and E. Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *arXiv preprint arXiv:1805.00915*, 2018.

J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/76cf99d3614e23eabab16fb27e944bf9-Paper.pdf.

T. Suzuki, A. Nitanda, and D. Wu. Uniform-in-time propagation of chaos for the mean field gradient langevin dynamics. In *Submitted to The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=_JScUk9TBUn.

T. Suzuki, D. Wu, and A. Nitanda. Convergence of mean-field langevin dynamics: Time-space discretization, stochastic gradient, and variance reduction. *arXiv preprint*, 2023b.

M. Telgarsky. Feature selection and low test error in shallow low-rotation ReLU networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=swEskiem99.

S. van de Geer. *Empirical Processes in M-Estimation.* Cambridge University Press, 2000.

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 9712–9724, 2019.