Mean field Langevin dynamics: convergence analysis and sparse parity feature learning

Pranik Chainani
Statistics & Data Science
Yale University
New Haven, CT
pranik.chainani@yale.edu

David Zhang
Statistics & Data Science
Yale University
New Haven, CT
david.zhang.ddz5@yale.edu

Abstract

Recent advances in the theoretical understanding of infinite-width neural network regime have shown that standard parameterizations often lead to kernel-like behavior, preventing feature learning. The Neural Tangent Kernel (NTK) regime captures this phenomenon, where weights tend to move little from initialization. In contrast, more recent work has presented mean-field Langevin dynamics (MFLD) as a promising alternative, incorporating stochasticity into weight updates to enable feature learning in the mean field regime [Suzuki et al., 2023]. In this paper, we present a primer into model training under mean field Langevin dynamics alongside convergence rate analysis as laid out in [Nitanda et al., 2022]. Furthermore, we also present a recent work exploring the feature learning capabilities of MFLD training for solving certain problems as shown in [Suzuki et al., 2023].

1 Introduction to mean field Langevin dynamics

In the infinite width limit of neural networks, two common regimes for describing model behavior are the neural tangent kernel [Jacot et al., 2018] where features appear to stay relatively fixed [Chizat et al., 2019] and the mean field regime where we view feature learning occurring through evolution in the space of probability distributions of your model parameters [Mei et al., 2019], [Yang and Hu, 2021]. More specifically, in the mean field regime as M, the model width, approaches infinity, we analogize to the "propagation of chaos" concept where we view our model as a system of independent particles interacting through a "mean field" ρ_t.

Hence, optimization and empirical risk analysis within the mean field regime typically follows using the probability measures of model neurons. Previous works have demonstrated that within the mean-field regime, two-layer neural networks achieve global convergence under certain structural assumptions [Mei et al., 2018] and maximum feature learning [Yang and Hu, 2021]. Recent works have proposed algorithmic modifications by incoporating Gaussian noise to gradient updates which has yielded global convergence with more relaxed conditions [Mei et al., 2019], [Guillin et al., 2019]. This is known as *mean field Langevin dynamics* which will the focus of this report. However, despite these existing works, we present a simpler perspective as laid out by [Nitanda et al., 2022] of the optimization analysis in the mean field Langevin dynamics regime that could be more suited for more general and practical machine learning settings. In particular, we highlight how mean field Langevin dynamics can be viewed from the lens of variational inference where by optimizing in the space of probability measures, [Nitanda et al., 2022] introduces a *proximal Gibbs distribution* p_q to prove linear convergence in continuous time.

Preprint. Under review.

1.1 Langevin Dynamics

Before examining mean field Langevin dynamics, let's recall the concepts of Langevin dynamics. First, recall that the Kullback-Leiber (KL) divergence objective can be used for optimizing in the space of measures where Langevin dynamics can be interpreted as gradient flow of the KL objective. From the definition of KL, we have

$$D_{\mathrm{KL}}(q\|p) = \int q \log \frac{q}{p} = \mathbb{E}_q[f] - H(q)$$

where $f = -\log p$, yielding the optimization problem

$$\min_q \{E_q[f] + E_q[\log q]\}$$

Hence when optimizing using gradient flow over the space of measures, we use the Wasserstein metric to obtain the Fokker-Planck PDE [Jordan et al., 1998]:

$$\frac{\partial q}{\partial t} = \nabla \cdot (q \nabla \log \frac{q}{p}) = \nabla \cdot (q \nabla f) + \Delta q$$

which is also translated to the following SDE of Langevin dynamics in sample space:

$$dX = -\nabla f(X) dt + \sqrt{2} dW$$

where $X = (X_t)_{t \geq 0}$ is a stochastic process and $W = (W_t)_{t \geq 0}$ refers to Brownian motion. If $X_t \sim p_t$, q_t evolves according to the Fokker-Planck equation above. In this sense, Langevin dynamics captures both a sample-level and distributional view of optimization. The latter enables the transition to mean-field settings.

In regards to convergence-time guarantees, it is well known that if the target distribution satisfies an α-Log-Sobolev inequality (LSI),

$$D_{\mathrm{KL}}(q\|p) \leq \frac{1}{2\alpha} \int \left\| \nabla \log \frac{q}{p} \right\|^2 q(x) dx = \frac{1}{2\alpha} \mathcal{I}(q\|p)$$

where $\mathcal{I}(q\|p)$ is the relative Fisher information, then the Langevin dynamics enjoys exponential convergence to the target distribution in KL divergence [Bakry et al., 2014].

$$D_{\mathrm{KL}}(q_t\|p) \leq e^{-2\alpha t} D_{\mathrm{KL}}(q_0\|p)$$

Note that for discretized versions, convergence rates depend on various target assumptions [Nitanda et al., 2022].

1.2 Mean field regime[1]

Two of the main compelling regimes for analyzing neural networks in the infinite-width limit are 1) the neural tangent kernel and 2) the mean field regime. We can denote our formulation of a (two-layer) neural network as such:

$$f(x;\theta) = \frac{\alpha_M}{M} \sum_{j=1}^{M} a_j \sigma(\langle w_j, x \rangle)$$

where α_M denotes the scaling factor and M the model width. In the NTK regime, $\alpha_M = 1$ and to avoid lazy training, we require $\alpha_M \lesssim \sqrt{\frac{n^2}{Md}}$, hence for fixed $\frac{n}{d}$, we require $\alpha_M \lesssim \frac{1}{\sqrt{M}}$ which yields us global scaling of $\frac{1}{M}$ in the mean field regime.

Typically, analysis in the mean field regime views the model's distribution of neurons $\rho^{(M)}$ evolving through some PDE. We can formulate this starting with the empirical measure as

$$M \to \infty, f(x_i; \theta^t) = f(x_i; \hat{\rho}_t^{(M)})$$

where $f(x_i; \rho) = \int \sigma(x_i, \theta) \rho(\theta) d\theta$ and $\hat{\rho}^{(M)} = \frac{1}{M} \sum_{j=1}^{M} \delta_{\theta_i}$. Note that $\hat{\rho}_t^{(M)} \to \rho_t, \forall t > 0$. Additionally, using the fact that the empirical risk under the empirical measure yields the same result

[1] This subsection is primarily based on lecture notes on the mean field regime.

as computing from finite particles $R(\rho^{(M)} = R_M(\theta)$ which yields us this description of gradient flow

$$\frac{d}{dt}\theta_j^t = -\nabla_\theta \Psi(\theta_j; \hat{\rho}_t^{(M)})$$

where $\Psi(\theta_j; \hat{\rho}_t^{(M)}) = V(\theta_j) + \frac{1}{M}\sum_{i=1}^{M} U(\theta_j, \theta_i)$ and $V(\theta_j)$ is used to the denote the external potential of particle $\theta_j \in \mathbb{R}^Q$ and $U(\theta_j, \theta_i)$ the pairwise potential. From this, $\theta_j^t \sim \rho_t, \forall t > 0$ and we obtain the following **McKean-Vlasov** type PDE:

$$\partial_t \rho_t = \nabla_\theta \cdot [\rho_t \nabla_\theta \Psi(\theta; \rho_t)]$$

Prior works have demonstrated that, under the mean-field regime, convergence to the global optima is guaranteed under certain regularity assumptions [Chizat and Bach, 2018], [Mei et al., 2018]. However, as mentioned by [Nitanda et al., 2022], the convergence of mean field neural networks falls beyond the scope of traditional *linear* Fokker-Planck equations of Langevin dynamics. Instead, due to its *nonlinear* nature, convergence rates have been much harder to establish. Hence in the first half of this report, we present an analysis into MFLD performed by [Nitanda et al., 2022] through the properties of the proximal Gibbs distribution p_q

2 Preliminaries

In this section we introduce mean field Langevin dynamics alongside its optimization problem.

2.1 Problem setup

In the mean field limit, the neural network's trainable parameters $\theta \in \mathbb{R}^d$ are represented as a distribution $q(\theta) \in \mathcal{P}$, where $\mathcal{P}$ denotes the space of probability densities over $\mathbb{R}^d$ with finite entropy and second moment.

We consider the following optimization objective over the space of probability distributions:

$$\min_{q \in \mathcal{P}} \{\mathcal{L}(q) := F(q) + \lambda\, \mathbb{E}_q[\log q]\} \tag{1}$$

where $F(q)$ is a differentiable convex functional capturing the expected risk, and $\lambda > 0$ controls the strength of the entropy regularization. We now define the proximal Gibbs distribution around q in $\mathcal{P}$.

Definition 1. *(Proximal Gibbs Distribution)* Let $p_q(\theta)$ be the Gibbs distribution with potential function $-\lambda^{-1}\delta F(q)/\delta q$ with normalization constant $Z(q)$:

$$p_q(\theta) = \frac{\exp(-\frac{1}{\lambda}\frac{\delta F(q)}{\delta q}(\theta))}{Z(q)}$$

Also, assume that $\frac{\delta F}{\delta q}(q)(\theta)$ exists and is smooth in θ and $p_q(\theta)d\theta$ satisfies the log-Sobolev inequality with constant $\alpha > 0$,

$$D_{\mathrm{KL}}(q \| p_q) \leq \frac{1}{2\alpha} E_q[\|\nabla \log \frac{q}{p_q}\|_2^2]$$

2.2 Optimization dynamics

Now, let us consider mean field Langevin dynamics where we have the following SDE:

$$d\theta_t = -\nabla \frac{\delta F(q_t)}{\delta q}(\theta_t)\, dt + \sqrt{2\lambda}\, dW_t$$

The associated PDE describing the evolution of the parameter distribution q_t is a nonlinear Fokker-Planck equation:

$$\frac{\partial q_t}{\partial t} = \nabla \cdot \left(q_t \nabla \frac{\delta F(q_t)}{\delta q}\right) + \lambda \Delta q_t$$

Equivalently, this can be expressed in terms of a proximal Gibbs distribution $p_{q_t} \propto \exp\left(-\frac{1}{\lambda}\frac{\delta F(q_t)}{\delta q}\right)$ as:

$$\frac{\partial q_t}{\partial t} = \lambda \nabla \cdot \left(q_t \nabla \log \frac{q_t}{p_{q_t}}\right)$$

Finally, note that a discretized version of the above dynamics gives rise to a noisy gradient descent algorithm over the space of probability measures:

$$\theta^{(k+1)} = \theta^{(k)} - \eta \nabla \frac{\delta F(q^{(k)})}{\delta q}(\theta^{(k)}) + \sqrt{2\lambda\eta}\, \xi^{(k)}$$

where $\xi^{(k)} \sim \mathcal{N}(0, I)$, and $\theta^{(k)} \sim q^{(k)}$.

3 Convergence Analysis

In this section, we present the convergence analysis that was laid out in [Nitanda et al., 2022].

3.1 Basic properties and convexity

First note that our optimality condition in 1 works out to be

$$\frac{\delta \mathcal{L}}{\delta q}(q) = \frac{\delta F}{\delta q}(q) + \lambda \log q = 0$$

where our optimal distribution q_* is given by the proximal Gibbs distribution

$$q_*(\theta) = p_{q_*} \propto \exp\left(-\frac{1}{\lambda} \frac{\delta F_(q_*)}{\delta q}(\theta)\right)$$

This was shown in [Hu et al., 2019] and gives rise to the fact that the divergence between q and p_q can be viewed as an optimization problem. [Nitanda et al., 2022] now shows the following properties:

Proposition. Under the assumptions in 2.1, we have:

1. The functional derivative of the regularized objective $\mathcal{L}(q) = F(q) + \lambda \mathbb{E}_q[\log q]$ satisfies:

$$\frac{\delta \mathcal{L}}{\delta q}(q)(\theta) = \lambda \log \frac{q(\theta)}{p_q(\theta)}$$

2. For any probability distributions $q, q' \in \mathcal{P}$, the functional satisfies a three-point inequality:

$$\mathcal{L}(q) + \int \frac{\delta \mathcal{L}}{\delta q}(q)(\theta) \cdot (q'(\theta) - q(\theta))\, d\theta + \lambda \mathrm{KL}(q' \| q) \leq \mathcal{L}(q')$$

and the minimizer of the left-hand side in q' is exactly p_q, the proximal Gibbs distribution.

3. Let q_* be the minimizer of $\mathcal{L}(q)$. Then for any $q \in \mathcal{P}$, we have:

$$\lambda \mathrm{KL}(q \| p_q) \geq \mathcal{L}(q) - \mathcal{L}(q_*) \geq \lambda \mathrm{KL}(q \| q_*)$$

Note that the first property indicates the our gradient updates point in the direction of decreasing KL. The third property provides bounds on our optimality gap and can be recognized as Polyak–Łojasiewicz and quadratic growth inequalities and as the squared norm gradient at q and squared distance between q and q_* for our upper and lower bound, respectively.

3.2 Convergence rates

We now analyze the convergence of the mean field Langevin dynamics in **continuous time**. Recall our nonlinear Fokker–Planck equation:

$$\frac{\partial q_t}{\partial t} = \lambda \nabla \cdot \left(q_t \nabla \log \frac{q_t}{p_{q_t}}\right)$$

Theorem 1 [Nitanda et al., 2022]. Let $\{q_t\}_{t\geq 0}$ follow the dynamics above. Under our assumptions of convexity of F and log-Sobolev inequality with constant α, for any $t \geq 0$, we have:

$$\mathcal{L}(q_t) - \mathcal{L}(q_*) \leq \exp(-2\alpha\lambda t)(\mathcal{L}(q_0) - \mathcal{L}(q_*))$$

Furthermore in the **discrete setting**, [Nitanda et al., 2022] show that exponential convergence holds up to a certain error. Recall, in the discrete-time version of the mean field Langevin dynamics (MFLD). This corresponds to the noisy gradient descent update:

$$\theta^{(k+1)} = \theta^{(k)} - \eta \nabla_\theta \frac{\delta F}{\delta q}(q^{(k)})(\theta^{(k)}) + \sqrt{2\lambda\eta}\,\xi^{(k)}, \quad \xi^{(k)} \sim \mathcal{N}(0, I_d)$$

where $q^{(k)}$ is the distribution of $\theta^{(k)}$ at iteration k, and $\eta > 0$ is the step size.

We denote the distribution of $\theta^{(k+1)}$ by $q^{(k+1)}$, and interpret this as the time-η solution to a stochastic differential equation initialized from $q^{(k)}$.

Theorem 2 [Nitanda et al., 2022]. Under our assumptions of convexity and differentiability of F and our log-Sobolev inequality with constant $\alpha > 0$ and suppose there exists a constant δ_η such that the discretization error $\delta_{q^{(k)},\eta} \leq \delta_\eta$ for all k. Then, we have:

$$\mathcal{L}(q^{(k)}) - \mathcal{L}(q_*) \leq \frac{\delta_\eta}{2\alpha\lambda} + \exp(-\alpha\lambda\eta k)\left(\mathcal{L}(q^{(0)}) - \mathcal{L}(q_*)\right)$$

Note that we define our discretization error as

$$\delta_{q^{(k)},t} := \mathbb{E}_{(\theta^{(k)},\theta_t^{(k+1)})}\left[\left\|\nabla_\theta \frac{\delta F}{\delta q}(q^{(k)})(\theta^{(k)}) - \nabla_\theta \frac{\delta F}{\delta q}(q_t^{(k+1)})(\theta_t^{(k+1)})\right\|_2^2\right]$$

3.3 Primal-Dual Perspective and Duality Gap

We now briefly describe the primal-dual structure underlying the mean field Langevin dynamics (MFLD). The original optimization problem (primal) seeks a distribution $q \in \mathcal{P}$ that minimizes a regularized empirical risk:

$$\mathcal{L}(q) = \frac{1}{n}\sum_{i=1}^{n} \ell(h_q(x_i), y_i) + \lambda' \mathbb{E}_{\theta\sim q}[\|\theta\|^2] + \lambda \mathbb{E}_q[\log q]$$

By introducing a dual variable $g \in \mathbb{R}^n$ and interpreting the regularization through convex duality, [Nitanda et al., 2022] define a dual objective $D(g)$ and show that the gap between the primal and dual objectives is given by the KL divergence:

$$\mathcal{L}(q) - D(g_q) = \lambda\, \mathrm{KL}(q \| p_q)$$

where $g_q = \{\partial_z \ell(h_q(x_i), y_i)\}_{i=1}^n$ and p_q is the proximal Gibbs distribution associated with q.

This identity reveals that the duality gap provides a natural certificate of optimality: when $q = p_q$, the gap is zero.

Convergence. Under suitable assumptions, the duality gap also converges exponentially over time. Specifically, for continuous-time MFLD:

$$\inf_{s\in[0,t]} \{\mathcal{L}(q_s) - D(g_{q_s})\} \leq \frac{e^{-2\alpha\lambda(t-1)}}{2\alpha\lambda}\left(\mathcal{L}(q_0) - \mathcal{L}(q_*)\right)$$

Hence, MFLD not only minimizes the primal objective, but also yields near-optimality via the decreasing KL divergence to the proximal distribution.

4 Introduction to feature learning with MFLD

Following the convergence properties of MFLD presented by [Nitanda et al., 2022], we present [Suzuki et al., 2023] which highlights the feature learning capabilities within the MFLD regime.

[Suzuki et al., 2023] starts by setting each neuron's parameters as a vector in $\mathbb{R}^{d+2}$ (recall that we will later have an input dimension d, plus additional bias terms). Rather than tracking individual vectors, we view the set of neurons as being sampled from a probability measure ρ defined on $\mathbb{R}^{d+2}$. This shift in perspective, from the usual finite-dimensional optimization, to optimization over the full space of probability measures allows the network to learn features as the overall distribution ρ changes.

5 Network and Problem Formulation

5.1 Sparse Parity Classification

The target task studied in [Suzuki et al., 2023] is the k-sparse parity problem. In this setting, the data is defined as follows:

- The input $z \in \left\{\pm\frac{1}{\sqrt{d}}\right\}^d$, i.e., each coordinate is $\pm 1/\sqrt{d}$.
- The label is computed as

$$y = \operatorname{sign}\Bigl(\prod_{i=1}^{k} z_i\Bigr)$$

  For the special case of $k = 2$, we recover the XOR problem.

Kernel-based methods (such as NTK) are not capable of adapting to the low-dimensional structure inherent in such problems (the sample complexity scales exponentially with respect to k). However, the mean-field approach, through learning a flexible distribution ρ, is shown to provide better sample complexity guarantees. Specifically, the dependence on the sparsity parameter k can be decoupled from the exponent in the dimension d, resulting in improved theoretical performance.

6 The Annealing Procedure

A major challenge in the analysis of MFLD is that the logarithmic Sobolev inequality (LSI) constant, denoted by α, appears in convergence rates and may depend exponentially on the regularization parameter λ. In our setting, if the functional derivative satisfies

$$\left\|\frac{\delta F(\rho)}{\delta\rho}\right\|_\infty \le B$$

then the Holley–Stroock perturbation argument gives an LSI constant of the form

$$\alpha \ge \lambda_1 \exp\left(-\frac{4B}{\lambda}\right)$$

where λ_1 is a constant linked to the reference measure ν. The exponential dependence on $1/\lambda$ implies that for a small regularization parameter, the convergence rate of the MFLD dynamics may deteriorate significantly.

To mitigate this, [Suzuki et al., 2023] introduces an *annealing procedure*. Instead of fixing λ, the regularization parameter is gradually reduced over rounds indexed by κ. Specifically, one sets:

$$\lambda^{(\kappa)} = 2^{-\kappa}\lambda^{(0)}$$

At each round κ, the dynamics are run until the loss

$$L^{(\kappa)}(\rho) = L(\rho) + \lambda^{(\kappa)}\,\mathrm{KL}(\nu, \rho)$$

is close to its optimum (up to an optimization error ε^*). And since the LSI constant depends on $\lambda^{(\kappa)}$ as

$$\alpha^{(\kappa)} \ge \lambda_1 \exp\left(-\frac{4B}{\lambda^{(\kappa)}}\right)$$

the system's "temperature" is lowered gradually, which is analogous to reducing the step size in iterative solvers to ensure convergence along the dominant eigen-directions, while still allowing the measure ρ to adjust its features.

7 Generalization Error via Local Rademacher Complexity

To quantify the generalization error of the learned function

$$f_\rho(z) = \int h_x(z)\, \rho(dx)$$

[Suzuki et al., 2023] develops bounds based on local Rademacher complexities. Instead of using a global complexity measure, the authors consider the function class

$$\mathcal{F}_M(\rho^\circ) = \{f_\rho \mid \rho \in \mathcal{P},\ \mathrm{KL}(\rho^\circ, \rho) \leq M\}$$

where ρ° is an optimal (or benchmark) measure. They then show (via a peeling argument) that the excess risk, measured in terms of the population risk $\bar{L}$, satisfies an inequality of the form:

$$\bar{L}(\hat{\rho}) - \bar{L}(\rho^*) \lesssim \sqrt{\frac{\mathrm{KL}(\rho^*, \hat{\rho})}{n}}$$

where:

- $\bar{L}$ denotes the population risk,
- n is the number of samples,
- ρ^* is the optimal measure,
- $\hat{\rho}$ is the measure obtained via MFLD.

For the special case of the 2-sparse parity (XOR) problem, the analysis shows that if the sample size scales as

$$n = \Theta(d^2)$$

then (under the strong assumption leading to a uniform L_∞-bound) the excess classification error decays exponentially with n/d^2. Under a weaker assumption, the error bound becomes polynomial in $1/(n\lambda)$, i.e.,

$$\bar{L}(\hat{\rho}) - \bar{L}(\rho^*) = O\left(\frac{1}{n\lambda}\right)$$

This local analysis is crucial as it shows that if $\hat{\rho}$ is close to ρ^* in terms of KL divergence, then the excess risk, and hence the classification error, can be tightly controlled.

8 Computational and Statistical Trade-offs

[Suzuki et al., 2023] not only derives statistical generalization guarantees but also provides explicit bounds on the computational complexity required by the MFLD algorithm. In the analysis for the 2-sparse parity (XOR) problem, prior NTK-based approaches require a sample complexity of at least

$$n = \Omega(d^2),$$

and the number of iterations needed for convergence scales accordingly.

With MFLD, and employing the annealing strategy, the effective sample complexity can be improved. In particular:

- In the regime where $n \approx d^2$, the classification error decays exponentially fast (Type I result). In this case, the convergence guarantee can be written as

  $$P\left(Y\, f_{\hat{\rho}}(Z) \leq 0\right) \leq \exp\left(-\frac{n\lambda^2}{C}\right)$$

  for some constant C depending on the model parameters.
- Under weaker assumptions (Type II result), the error bound scales polynomially as

  $$P\left(Y\, f_{\hat{\rho}}(Z) \leq 0\right) = O\left(\frac{1}{n\lambda}\right)$$

The trade-off also involves the network width (i.e., the number of particles used to approximate the measure ρ). While a larger number of particles yields a better approximation, it increases computational cost. [Suzuki et al., 2023] shows that, compared to previous methods (which could require a width exponential in d), the proposed approach can achieve similar performance with a reduced number of particles, under appropriate conditions.

These results provide an explicit link between the statistical efficiency (generalization error) and the computational resources (number of iterations, particle width) of the MFLD algorithm, thus quantifying the benefit of feature learning in the mean-field regime.

9 Conclusion

We present a brief primer into the mean field Langevin dynamics training regime of infinite width neural networks, providing global convergence guarantees in both continuous and discrete-time [Nitanda et al., 2022] as well as applications to feature learning of sparse parity classificaton [Suzuki et al., 2023].

References

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2014. ISBN 978-3-319-00226-2. doi: 10.1007/978-3-319-00227-9. URL `https://doi.org/10.1007/978-3-319-00227-9`.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31, pages 3040–3050, 2018. URL `https://papers.nips.cc/paper_files/paper/2018/file/54fe9769c66ba74dc8631fa9f2c52b0c-Paper.pdf`.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL `https://arxiv.org/abs/1812.07956`.

Arnaud Guillin, Wei Liu, Liming Wu, and Chaoen Zhang. The kinetic fokker-planck equation with mean field interaction. *arXiv preprint arXiv:1912.02594*, 2019. URL `https://arxiv.org/abs/1912.02594`.

Kaifeng Hu, Zhong Ren, David Siska, and Lukasz Szpruch. Mean-field langevin dynamics and energy landscape of neural networks. *arXiv preprint arXiv:1905.07769*, 2019. URL `https://arxiv.org/abs/1905.07769`.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pages 8571–8580, 2018. URL `https://papers.nips.cc/paper_files/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html`.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. doi: 10.1137/S0036141096303359. URL `https://doi.org/10.1137/S0036141096303359`.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layers neural networks. *arXiv preprint arXiv:1804.06561*, 2018. URL `https://arxiv.org/abs/1804.06561`.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. *arXiv preprint arXiv:1902.06015*, 2019. URL `https://arxiv.org/abs/1902.06015`.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field langevin dynamics. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151, pages 2479–2497. PMLR, 2022. URL `https://proceedings.mlr.press/v151/nitanda22a.html`.

Taiji Suzuki, Denny Wu, Kazusato Oko, and Atsushi Nitanda. Feature learning via mean-field langevin dynamics: Classifying sparse parities and beyond. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL `https://openreview.net/forum?id=tj86aGVNb3`.

Greg Yang and Edward J. Hu. Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 11714–11725. PMLR, 2021. URL `https://arxiv.org/abs/2011.14522`.