Mean field Langevin dynamics: Convergence analysis and sparse parity learning

Pranik Chainani & David Zhang

Department of Statistics & Data Science, Yale University

April 16, 2025

Outline

- Introduction: Langevin dynamics and mean field regime
- Mean field Langevin dynamics
- Convergence analysis
- Feature learning of sparse parities

Why study the behavior of neural nets in the infinite width regime?

It is known that given sufficiently large number of hidden neurons, artifical neural networks can approximate any function[1]

However as $M \to \infty$,

- What solutions do we converge to in a highly non-convex loss landscape?
- What explains good generalizability in overparamterized regimes?

So in general, what is the behavior of

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

as

$$M \to \infty$$

[1] Hornik et al., 1989

Neural network behavior in infinite width regime

- Neural tangent kernel (NTK)
- Mean field regime

Neural tangent kernel

As $M \to \infty$, if we consider a scaling of $\frac{1}{\sqrt{M}}$, we enter what's known as lazy training where $\|\theta_t - \theta_0\|_2 \ll 1$ while achieving ERM_0. Hence we can view these models as approximate linear methods

$$f(x;\theta_t) \approx f(x;\theta_0) + \nabla_\theta f(x;\theta_0) \cdot (\theta_t - \theta_0)$$

Hence, we view the evolution of our model as kernel regression

$$\frac{d}{dt} f(x;\theta) = \nabla_\theta f(x;\theta_0)^T \frac{d\theta_t}{dt}$$

$$\frac{d}{dt} f(x;\theta) = -\sum_i^n (f(x_i;\theta_t) - y_i) \underbrace{\nabla_\theta f(x;\theta_0)^T \nabla_\theta f(x_i;\theta_0)}_{\text{NTK}}$$

Note that the NTK stays constant throughout training and prevents strong feature learning[2]

[2] Jacot et al., 2018

Mean Field Regime

Recall from lecture: our scaling of $\frac{\alpha_M}{M}$ requires

$$\alpha_M \lesssim \sqrt{\frac{n^2}{Md}}$$

to avoid lazy training. Hence, $\alpha_M \lesssim \frac{1}{\sqrt{M}}$.

We call this the mean-field scaling.

This scaling has been shown to maximize feature learning in two-layer neural networks[3].

In the mean field regime, analysis typically focuses on the evolution of the neuron distribution $\rho^{(M)}$, hence the name.

Model formulation under mean field scaling:

$$f(x_i;\theta) = \frac{1}{M}\sum_{j=1}^{M}\sigma(x_i;\theta_j), \quad \theta_j \in \mathbb{R}^d, \quad i \in [n]$$

[3] Yang et al., 2022

Mean Field with Measures

As $M \to \infty$, we can write $f(x_i; \theta_t) = f(x_i; \hat{\rho}_t^{(M)})$, where:

- $f(x_i; \rho) = \int \sigma(x_i; \theta)\, \rho(\theta)\, d\theta$
- $\hat{\rho}_t^{(M)} = \frac{1}{M} \sum_{j=1}^{M} \delta_{\theta_j^t}$
- $\hat{\rho}_t^{(M)} \to \rho_t$ for all t

We can view the evolution of f through the evolution of ρ, where our risk becomes:

$$R(\rho^{(M)}) = R_M(\theta)$$

So, gradient flow on M neurons becomes:

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

where the functional Ψ is:

$$\Psi(\theta_j; \hat{\rho}_t^{(M)}) = V(\theta_j) + \frac{1}{M} \sum_{i=1}^{M} U(\theta_j, \theta_i)$$

In the limit $M \to \infty$, the gradient flow becomes:

Mean Field with Measures

We can characterize the evolution of ρ_t through a **McKean–Vlasov** type PDE:

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

Hence, our question becomes:

- Find the optimal distribution ρ_*
- I.e., perform gradient/Wasserstein flow in the space of distributions

Langevin dynamics can be interpreted as gradient flow of KL divergence in the space of probability measures[4].

We consider the following optimization problem:

$$\min_{\rho} D_{\mathrm{KL}}(\rho \,\|\, q) = \int \rho \log \frac{\rho}{q} = \mathbb{E}_{\rho}[f] + \mathbb{E}_{\rho}[\log \rho]$$

where $f = -\log q$

This corresponds to the **Fokker-Planck PDE**:

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

This is the continuity equation for **Langevin dynamics**, corresponding to the SDE:

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

[4] Jordan et al., 1998

Mean field Langevin dynamics

Key question: Can Langevin dynamics help us find an optimal q_* in the **mean field regime**?
In other words, under noisy gradient descent, what does the convergence of our KL-regularized objective look like?

Preliminaries

Suppose $F : \mathcal{P} \to \mathbb{R}$ is a differentiable and convex functional. Our objective is to solve:

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

Definition 1 (Proximal Gibbs distribution). Let $p_q(\theta)$ denote the proximal Gibbs distribution with potential function $-\lambda^{-1} \, \delta F(q)/\delta q$. Then:

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

Assume that the functional derivative $\frac{\delta F}{\delta q}(q)(\theta)$ exists, is smooth, and that p_q satisfies a log-Sobolev inequality (LSI) with constant $\alpha > 0$:

$$D_{\mathrm{KL}}(q \,\|\, p_q) \leq \frac{1}{2\alpha} \, \mathbb{E}_q \left[\left\| \nabla \log \frac{q}{p_q} \right\|_2^2 \right]$$

Optimization dynamics

Recall, Langevin dynamics was formulated as:

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

In the mean field limit, we now have:

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

where $\theta_t \sim q_t(\theta)$ and $(W_t)_{t\geq 0}$ is Brownian motion.

Optimization Dynamics

The associated PDE for the distribution q_t is the nonlinear Fokker–Planck equation:

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

Using the **proximal Gibbs distribution** and taking gradients:

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

$$\nabla \frac{\delta F(q)}{\delta q}(\theta) = -\lambda \nabla \log p_q(\theta)$$

Substituting into the Fokker–Planck equation:

$$\begin{aligned} \frac{\partial q_t}{\partial t} &= \nabla \cdot (-\lambda q_t \nabla \log p_{q_t}) + \lambda \nabla \cdot \nabla q_t \\ &= \lambda \nabla \cdot \left(q_t \nabla \log \frac{q_t}{p_{q_t}} \right) \end{aligned}$$

Convergence Analysis

We begin by noting that the functional F is convex over the space of probability distributions:

$$F(q) = \mathbb{E}_{(X,Y)}[\ell(h_q(X), Y)] + \lambda' \, \mathbb{E}_{\theta\sim q}[r(\theta)]$$

Here:

- ℓ is a convex, smooth loss function
- $r(\theta)$ is a convex regularizer (e.g., $\|\theta\|^2$)
- $h_q(X) := \mathbb{E}_{\theta\sim q}[h_\theta(X)]$ is the neural network output in the mean field limit

The map $q \mapsto h_q(X)$ is linear, and the composition with a convex loss ensures $F(q)$ is convex in q.

Convergence Analysis

Furthermore, to guarantee exponential convergence of Langevin dynamics in KL divergence, we require the proximal Gibbs distribution p_q to satisfy a log-Sobolev inequality (LSI):

$$D_{\mathrm{KL}}(q \| p_q) \leq \frac{1}{2\alpha} \mathbb{E}_q \left[\left\| \nabla \log \frac{q}{p_q} \right\|^2 \right]$$

As shown in Nitanda et al., 2022, p_q satisfies LSI with constant

$$\alpha = \frac{2\lambda'}{\lambda} \cdot \exp(O(1/\lambda))$$

Optimality Condition and Proximal Gibbs Distribution

We aim to minimize the regularized functional:

$$\mathcal{L}(q) = F(q) + \lambda \, \mathbb{E}_q[\log q]$$

The first-order optimality condition is:

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

Solving this yields the optimal distribution q_*:

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

Convergence analysis

Under Assumption 1, the regularized objective

$$\mathcal{L}(q) = F(q) + \lambda \mathbb{E}_q[\log q]$$

satisfies the following properties:

1. **Functional Derivative via KL Divergence:**

$$\frac{\delta \mathcal{L}}{\delta q}(q) = \lambda \log \frac{q}{p_q}, \quad \text{where} \quad p_q(\theta) \propto \exp\left(-\frac{1}{\lambda}\frac{\delta F}{\delta q}(q)(\theta)\right)$$

2. **KL Surrogate Lower-Bounds Objective:** For all $q, q' \in \mathcal{P}$,

$$\mathcal{L}(q') \geq \mathcal{L}(q) + \int \frac{\delta \mathcal{L}}{\delta q}(q)(q' - q) + \lambda \mathrm{KL}(q' \| q)$$

Moreover, this lower bound is minimized at $q = p_q$.[5]

[5] Nitanda et al., 2022

Convergence analysis

3. Sandwich Inequality: If q_* is the minimizer of $\mathcal{L}(q)$, then for all $q \in \mathcal{P}$:

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

[6]

[6] Nitanda et al., 2022

Convergence analysis

Let $(q_t)_{t\geq 0}$ evolve via the MFLD:

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

We study convergence of the regularized loss:

$$\mathcal{L}(q) = F(q) + \lambda \, \mathbb{E}_q[\log q]$$

Step 1: Functional chain rule

$$\frac{d}{dt}\left(\mathcal{L}(q_t) - \mathcal{L}(q_*)\right) = \int \frac{\delta \mathcal{L}}{\delta q}(q_t)(\theta) \cdot \frac{\partial q_t}{\partial t}(\theta)\, d\theta$$

Step 2: Plug in MFLD and integration by parts

$$\frac{d}{dt}\left(\mathcal{L}(q_t) - \mathcal{L}(q_*)\right) = \lambda \int \frac{\delta \mathcal{L}}{\delta q}(q_t)(\theta) \nabla \cdot (q_t \nabla \log \frac{q_t}{p_{q_t}})$$

Convergence Analysis (Continuous Time)

$$\frac{d}{dt}\left(\mathcal{L}(q_t) - \mathcal{L}(q_*)\right) = -\lambda \int q_t(\theta) \nabla \frac{\delta \mathcal{L}}{\delta q}(q_t)(\theta)^T \nabla \log \frac{q_t}{p_{q_t}}(\theta)\, d\theta$$

Step 3: Apply Proposition 1

$$= -\lambda^2 \int q_t(\theta) \left\| \nabla \log \frac{q_t}{p_{q_t}}(\theta) \right\|_2^2 d\theta$$

Apply LSI (Assumption 2) and Proposition 3 (sandwich inequality):

$$\leq -2\alpha\lambda^2 \, \mathrm{KL}(q_t \,\|\, p_{q_t}) \leq -2\alpha\lambda \left(\mathcal{L}(q_t) - \mathcal{L}(q_*)\right)$$

Conclusion: Exponential Convergence

$$\frac{d}{dt}\left(\mathcal{L}(q_t) - \mathcal{L}(q_*)\right) \leq -2\alpha\lambda \left(\mathcal{L}(q_t) - \mathcal{L}(q_*)\right)$$

$$\Rightarrow \quad \mathcal{L}(q_t) - \mathcal{L}(q_*) \leq \left(\mathcal{L}(q_0) - \mathcal{L}(q_*)\right) e^{-2\alpha\lambda t}$$

Takeaways

- The term $\nabla \frac{\delta F(q)}{\delta q}(\theta)$ arises as the **mean field limit** of the gradient of wide neural networks.
- Adding noise to gradient descent yields a **distributional SDE**:

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

  whose marginal q_t evolves under **Mean Field Langevin Dynamics**.
- The associated objective $\mathcal{L}(q) = F(q) + \lambda \mathbb{E}_q[\log q]$ is convex, enabling:
  - Convergence guarantees via the Log-Sobolev Inequality
  - Exponential decay of $\mathcal{L}(q_t) - \mathcal{L}(q_*)$
- **MFLD simulates KL-regularized gradient flow** in the space of probability measures, and its discretization resembles noisy SGD.

Reference: Nitanda et al., 2022. "Convex Analysis of Mean Field Langevin Dynamics"

Outline

- Introduction: Motivation and Infinite-Width Regime
- Langevin Dynamics and Mean Field Langevin Dynamics (MFLD)
- Computing the Functional Derivative and Particle Update
- The Annealing Procedure
- Convergence Analysis via Local Rademacher Complexity (with Proof Ideas)
- Feature Learning of Sparse Parities
- Computational and Statistical Trade-offs
- Summary and Conclusions

Why Study Neural Networks in the Infinite-Width Regime?

- **High-Dimensional Vectors:** Each neuron's parameter is a vector $x \in \mathbb{R}^{d+2}$.
- **From Finite to Infinite:** Instead of tracking individual vectors, we study the evolution of a distribution μ over these vectors.
- **Simplified Analysis:** In the infinite-width limit, linearization (via Taylor expansion) and kernel methods become applicable.

Neural Network as an Integral:

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

with

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

Langevin Dynamics and MFLD

Standard Langevin Dynamics:

$$dX_t = -\nabla f(X_t)\, dt + \sqrt{2\lambda}\, dW_t,$$

where W_t is Brownian motion.

Mean Field Langevin Dynamics (MFLD):

$$dX_t = -\nabla \frac{\delta F(\mu_t)}{\delta \mu}(X_t)\, dt + \sqrt{2\lambda}\, dW_t, \quad \mu_t = \mathrm{Law}(X_t).$$

Practical Implementation: Approximate μ via particles and update them using Euler–Maruyama:

$$x^{\tau+1} = x^{\tau} - \eta\, \nabla_x \frac{\delta F(\mu_\tau)}{\delta \mu}(x^{\tau}) + \sqrt{2\lambda\eta}\, \xi^{\tau}.$$

Computing the Functional Derivative

Risk Functional:

$$F(\mu) = L(\mu) + \lambda\,\mathrm{KL}(\nu,\mu), \quad L(\mu) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(y_i\, f_\mu(z_i)\Big).$$

First Variation:

$$\frac{\delta L(\mu)}{\delta \mu}(x) = \frac{1}{n}\sum_{i=1}^{n} \ell'\Big(y_i\, f_\mu(z_i)\Big) y_i\, h_x(z_i).$$

Taking the gradient with respect to x provides the update direction:

$$\nabla_x \frac{\delta F(\mu)}{\delta \mu}(x).$$

Particle Update:

$$x^{\tau+1} = x^\tau - \eta\,\nabla_x \frac{\delta F(\mu_\tau)}{\delta \mu}(x^\tau) + \sqrt{2\lambda\eta}\,\xi^\tau.$$

The Annealing Procedure

The convergence rate depends on the log-Sobolev inequality (LSI). If

$$\left\| \frac{\delta F(\mu)}{\delta \mu} \right\|_\infty \leq B,$$

then the LSI constant satisfies

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

Since a small λ would worsen the rate exponentially, an annealing schedule is adopted:

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

- At each round κ, the dynamics run until near convergence.
- Gradually reducing λ lowers the system's temperature, enabling finer tuning of μ.

Convergence Analysis via Local Rademacher Complexity

- **Goal:** Bound the excess population risk $\bar{L}(\hat{\mu}) - \bar{L}(\mu^*)$.
- Under appropriate conditions, one can prove:

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

  where n is the sample size.
- **Local Rademacher Complexity:** By localizing the function class around μ^*,

$$\mathfrak{R}_n\big(\mathcal{F}_M(\mu^*)\big) \le C\sqrt{\frac{M}{n}},$$

  with M being an upper bound on $\mathrm{KL}(\mu^*, \mu)$.

Local Rademacher Complexity: Definitions

- For a function class $\mathcal{F}$, the empirical Rademacher complexity is defined as
$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_{\sigma,z}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)\right],$$
where $\sigma_i \in \{\pm 1\}$ are independent Rademacher variables.
- To localize, we consider
$$\mathcal{F}_M(\mu^*) = \{f_\mu : \mathrm{KL}(\mu^*, \mu) \le M\}.$$

Bounding the Local Rademacher Complexity

- Under Lipschitz and boundedness assumptions,

$$\mathfrak{R}_n(\mathcal{F}_M(\mu^*)) \leq C\sqrt{\frac{M}{n}},$$

  where C is a constant.
- This bound reflects that if the KL divergence between μ and μ^* is small, the function values do not vary excessively.

Peeling Argument and Excess Risk Bound

- Divide $\mathcal{F}$ into layers:

$$\mathcal{F}_j = \{f \in \mathcal{F} : 2^{j-1}\varepsilon_0 \leq \mathrm{KL}(\mu^*, \mu) \leq 2^j \varepsilon_0\},$$

  for some small $\varepsilon_0 > 0$.
- Each layer satisfies

$$\mathfrak{R}_n(\mathcal{F}_j) \lesssim \sqrt{\frac{2^j \varepsilon_0}{n}}.$$

- Using concentration inequalities and a union bound,

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

Interpreting the Excess Risk Bound

- **Strong Assumptions:** In the 2-sparse parity problem, if

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

then the classification error decays exponentially.

- **Weaker Assumptions:** More generally,

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

- **Takeaway:** Close proximity (in KL) between the learned and optimal measures guarantees low excess risk.

Feature Learning of Sparse Parities

- **Sparse Parity Problem:** For inputs $z \in \{\pm 1/\sqrt{d}\}^d$, the target is

$$y = \mathrm{sign}\Big(\prod_{i=1}^{k} z_i\Big),$$

  with XOR corresponding to $k = 2$.
- **NTK vs. Mean Field:** NTK fixes features (requiring sample complexity $\Omega(dk)$) whereas the mean-field approach adapts μ.
- **Improved Complexity:** MFLD decouples k from the exponent in d, leading to better sample complexity.

Computational and Statistical Trade-offs

- **Statistical Efficiency:**
  - NTK: Sample complexity is $\Omega(d^2)$.
  - MFLD: Under favorable conditions, nearly linear dependence in d is achievable.
- **Computational Cost:** MFLD requires updating many particles and managing an annealing schedule.
- **Trade-off:** Feature learning via MFLD yields lower generalization error at the expense of higher computational demand compared to NTK.

Summary and Conclusions

- **Neural Network Representation:**

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

shifts the focus to optimizing the distribution μ.

- **MFLD Updates:** Update rule:

$$x^{\tau+1} = x^\tau - \eta\, \nabla_x \frac{\delta F(\mu_\tau)}{\delta \mu}(x^\tau) + \sqrt{2\lambda\eta}\, \xi^\tau.$$

- **Annealing:** Gradually reduce λ to control exponential dependencies.
- **Generalization:** Local Rademacher complexity analysis yields:

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

- **Feature Learning:** MFLD adapts features, improving sample complexity relative to NTK.

Takeaway: Controlled training dynamics in the mean-field regime enable true feature learning with enhanced generalization properties.

Questions?

Thank you for your attention!