# Convex Analysis of Mean Field Langevin Dynamics

Atsushi Nitanda[1,2,3], Denny Wu[4,5], and Taiji Suzuki[2,6]

1 Kyutech Kyushu Institute of Technology
4 University of Toronto
6 The University of Tokyo
2 RIKEN AIP
3 PRESTO SAKIGAKE
5 Vector Institute

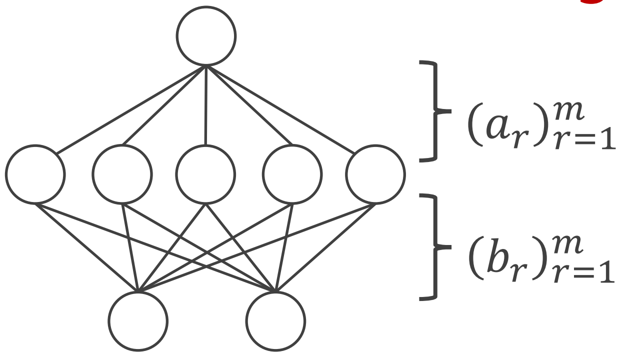
AISTATS2022 (Online)

# Outline

**Topic:** Convergence analysis of mean-field Langevin dynamics.

**Example:** noisy gradient descent for neural networks in mean-field regime:

$$h_\Theta(x) = \frac{1}{m}\sum_{r=1}^{m} a_r \sigma(b_r^\top x).$$

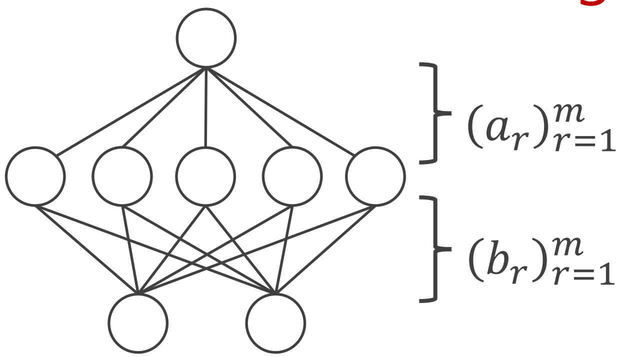


# Outline

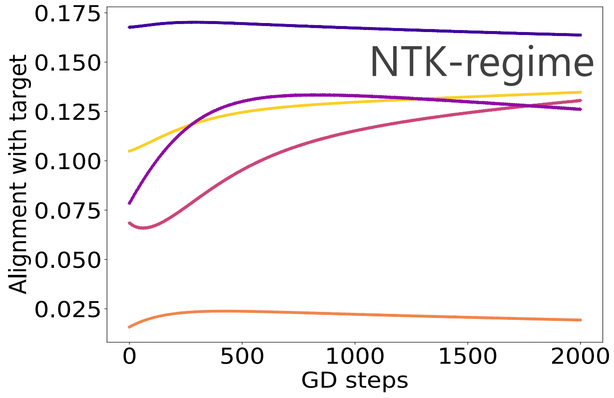
**Topic:** Convergence analysis of mean-field Langevin dynamics.

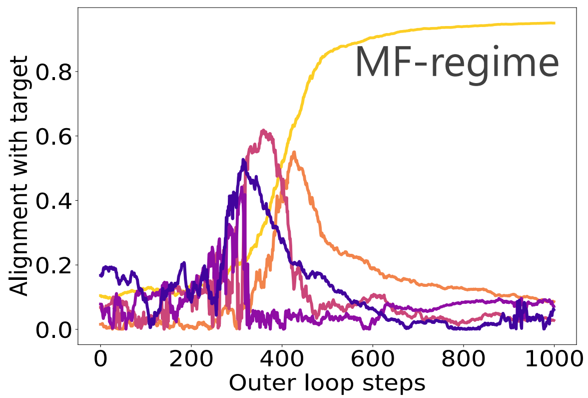
**Example:** noisy gradient descent for neural networks in mean-field regime:

$$h_\Theta(x) = \frac{1}{m}\sum_{r=1}^{m} a_r \sigma(b_r^\top x).$$




Mean field neural networks exhibit global convergence and adaptivity.




[Nitanda, Denny, &Suzuki (NeurIPS2021)]

# Outline

The convergence analysis of this model is more challenging.

We consider noisy gradient descent for mean-field neural networks:

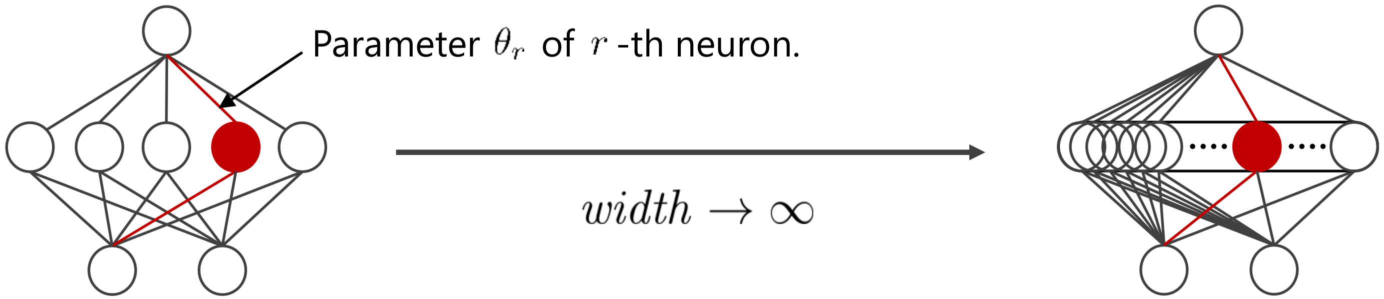
$$\theta_r^{(k+1)} \leftarrow \underbrace{(1-2\eta\lambda')\theta_r^{(k)}}_{L_2\text{-regularization}} - \eta\underbrace{\mathbb{E}[\partial_z \ell(h_{\Theta^{(k)}}(X), Y)\partial_{\theta_r} h(\theta_r^{(k)}, X)]}_{\text{Gradient of loss}} + \underbrace{\sqrt{2\eta\lambda}\zeta_r^{(k)}}_{\text{Gauss noise}}.$$

# Outline

The convergence analysis of this model is more challenging.

We consider noisy gradient descent for mean-field neural networks:

$$\theta_r^{(k+1)} \leftarrow \underbrace{(1-2\eta\lambda')\theta_r^{(k)}}_{L_2\text{-regularization}} - \underbrace{\eta\mathbb{E}[\partial_z \ell(h_{\Theta^{(k)}}(X), Y)\partial_{\theta_r} h(\theta_r^{(k)}, X)]}_{\text{Gradient of loss}} + \underbrace{\sqrt{2\eta\lambda}\zeta_r^{(k)}}_{\text{Gauss noise}}.$$



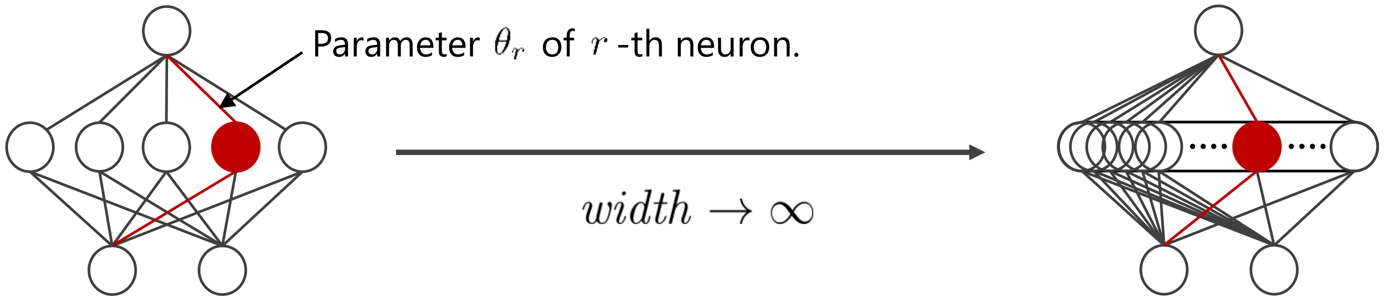

Infinitely wide networks parameterized by the probability distribution $q$.

# Outline

The convergence analysis of this model is more challenging.

We consider noisy gradient descent for mean-field neural networks:

$$\theta_r^{(k+1)} \leftarrow \underbrace{(1-2\eta\lambda')\theta_r^{(k)}}_{L_2\text{-regularization}} - \underbrace{\eta\mathbb{E}[\partial_z \ell(h_{\Theta^{(k)}}(X),Y)\partial_{\theta_r} h(\theta_r^{(k)},X)]}_{\text{Gradient of loss}} + \underbrace{\sqrt{2\eta\lambda}\zeta_r^{(k)}}_{\text{Gauss noise}}.$$




Infinitely wide networks parameterized by the probability distribution $q$.

**Mean-field Langevin dynamics:**

$$\mathrm{d}\theta_t = -\nabla_\theta g_{q_t}(\theta_t)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}W_t.$$

$q_t(\theta_t)\mathrm{d}t$ is a probability distribution of $\theta_t$.

# Outline

We analyze the noisy gradient descent via mean-field Langevin dynamics:

$$\mathrm{d}\theta_t = -\underline{\nabla_\theta g_{q_t}(\theta_t)}\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}W_t$$

Drift term

The drift term involves the distribution unlike normal Langevin dynamics.
This difference makes the convergence analysis difficult.

# Outline

We analyze the noisy gradient descent via mean-field Langevin dynamics:

$$\mathrm{d}\theta_t = -\underbrace{\nabla_\theta g_{q_t}(\theta_t)}_{\text{Drift term}}\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}W_t$$

The drift term involves the distribution unlike normal Langevin dynamics.
This difference makes the convergence analysis difficult.

**Contribution:**

- We resolve this difference by proposing *proximal Gibbs distribution.*

# Outline

We analyze the noisy gradient descent via mean-field Langevin dynamics:

$$\mathrm{d}\theta_t = -\underline{\nabla_\theta g_{q_t}(\theta_t)}\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}W_t$$

Drift term

The drift term involves the distribution unlike normal Langevin dynamics.
This difference makes the convergence analysis difficult.

**Contribution:**

- We resolve this difference by proposing *proximal Gibbs distribution.*
- The proof is an extension of that for Langevin dynamics into mean-field settings, which mirrors the classical convex optimization theory.

# Outline

We analyze the noisy gradient descent via mean-field Langevin dynamics:

$$\mathrm{d}\theta_t = -\underbrace{\nabla_\theta g_{q_t}(\theta_t)}_{\text{Drift term}}\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}W_t$$

The drift term involves the distribution unlike normal Langevin dynamics.
This difference makes the convergence analysis difficult.

**Contribution:**

- We resolve this difference by proposing *proximal Gibbs distribution.*
- The proof is an extension of that for Langevin dynamics into mean-field settings, which mirrors the classical convex optimization theory.
- We show the global convergence with the rate for KL-regularized problems.

# Outline

We analyze the noisy gradient descent via mean-field Langevin dynamics:

$$\mathrm{d}\theta_t = -\underline{\nabla_\theta g_{q_t}(\theta_t)}\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}W_t$$

Drift term

The drift term involves the distribution unlike normal Langevin dynamics.
This difference makes the convergence analysis difficult.

**Contribution:**

- We resolve this difference by proposing *proximal Gibbs distribution.*
- The proof is an extension of that for Langevin dynamics into mean-field settings, which mirrors the classical convex optimization theory.
- We show the global convergence with the rate for KL-regularized problems.
- Prima-dual viewpoint of proximal Gibbs distribution.

# Regularized Risk Minimization

KL-regularized empirical risk minimization over the probability space:

$$\min_{q \in \mathcal{P}} \left\{ \mathcal{L}(q) = \frac{1}{n} \sum_{i=1}^{n} \ell(h_q(x_i), y_i) + \underline{\lambda' \mathbb{E}_q[\|\theta\|_2^2] + \lambda \mathbb{E}_q[\log(q(\theta))]} \right\}.$$

$$\propto \lambda \mathrm{KL}\left(q \,\middle\|\, \mathcal{N}\left(0, \frac{\lambda}{2\lambda'} I\right)\right).$$

Kullback-Leibler divergence to zero-mean Gaussian.

$\mathcal{P}$ : the set of probability densities.

$\mathbb{E}_q$ : expectation w.r.t $\theta \sim q(\theta)\mathrm{d}\theta$.

$h_q(x) = \mathbb{E}_q[h(\theta, x)]$ : mean-field neural network.

$(h(\theta, x) = a\sigma(b^\top x), \theta = (a, b))$

# Proximal Gibbs Distribution

**Definition** (Proximal Gibbs distribution):
For a distribution $q$, we define $p_q$ as

$$p_q(\theta) \propto \exp\left(-\frac{1}{\lambda} g_q(\theta)\right).$$

$$g_q(\theta) = \frac{1}{n}\sum_{i=1}^{n} \partial_z \ell(h_q(X), Y) h(\theta, X)] + \lambda' \|\theta\|_2^2$$

$$= \frac{\delta}{\delta q}\left\{\frac{1}{n}\sum_{i=1}^{n} \ell(h_q(X), Y) + \lambda' \mathbb{E}_q[\|\theta\|_2^2]\right\}.$$

# Proximal Gibbs Distribution

**Definition** (Proximal Gibbs distribution):
For a distribution $q$, we define $p_q$ as

$$g_q(\theta) = \frac{1}{n}\sum_{i=1}^{n} \partial_z \ell(h_q(X), Y)h(\theta, X)] + \lambda' \|\theta\|_2^2$$

$$= \frac{\delta}{\delta q}\left\{\frac{1}{n}\sum_{i=1}^{n} \ell(h_q(X), Y) + \lambda' \mathbb{E}_q[\|\theta\|_2^2]\right\}.$$

$$p_q(\theta) \propto \exp\left(-\frac{1}{\lambda} g_q(\theta)\right).$$

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

- This formula is derived by the convex argument on the space of probability distributions.
- $p_q$ plays as a cushion to absorb difference from the analysis of normal Langevin dynamics.
- As a result, we can develop convergence analysis in the mean-field regime, which mirrors the classical convex optimization analysis.

# Assumption

**Assumption** (modified log Sobolev inequality (LSI)): Let $\alpha > 0$ be a constant.
For any $q$, the distribution $p_q$ satisfies the following with

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

# Assumption

**Assumption** (modified log Sobolev inequality (LSI)): Let $\alpha > 0$ be a constant.
For any $q$, the distribution $p_q$ satisfies the following with

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

Holley and Stroock (1987) argument guarantees the LSI of

$$p_q(\theta) \propto \exp\left( -\frac{1}{\lambda} g_q(\theta) \right)$$

when the potential $g_q$ is the sum of strongly convex and bounded functions.

**Example** (mean-field neural networks):
For uniformly bounded $h(\theta, x)$, the modified LSI is satisfied with the constant $\frac{2\lambda'}{\lambda \exp(C\lambda^{-1})}$.

# Convergence Analysis

**Theorem**: Let $\{q_t\}_{t\geq 0}$ be the evolution of mean-field Langevin dynamics. Under LSI assumption with $\alpha > 0$ and smoothness assumptions,

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

# Convergence Analysis

**Theorem**: Let $\{q_t\}_{t\geq 0}$ be the evolution of mean-field Langevin dynamics. Under LSI assumption with $\alpha > 0$ and smoothness assumptions,

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

Proof.
$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{L}(q_t) - \mathcal{L}(q_*)) &= \int \frac{\delta \mathcal{L}}{\delta q}(q_t)(\theta)\frac{\partial q_t}{\partial t}(\theta)\mathrm{d}\theta \\
&= \lambda \int \frac{\delta \mathcal{L}}{\delta q}(q_t)(\theta)\nabla \cdot \left(q_t(\theta)\nabla \log \frac{q_t}{p_{q_t}}(\theta)\right)\mathrm{d}\theta \\
&= -\lambda \int q_t(\theta)\nabla \frac{\delta \mathcal{L}}{\delta q}(q_t)(\theta)^\top \nabla \log \frac{q_t}{p_{q_t}}(\theta)\mathrm{d}\theta \\
&= -\lambda^2 \int q_t(\theta)\|\nabla \log \frac{q_t}{p_{q_t}}(\theta)\|_2^2\mathrm{d}\theta \\
&\leq -2\alpha\lambda^2\mathrm{KL}(q_t\|p_{q_t}) \\
&\leq -2\alpha\lambda(\mathcal{L}(q_t) - \mathcal{L}(q_*)).
\end{aligned}$$

The Grönwall's inequality finishes the proof.

# Convergence Analysis

**Theorem**: Let $\{q_t\}_{t \geq 0}$ be the evolution of mean-field Langevin dynamics. Under LSI assumption with $\alpha > 0$ and smoothness assumptions,

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

We also obtain a time-discretized version of this result. See our paper for details.

# Primal-Dual Viewpoint

Duality for empirical risk [Oko, Suzuki, Nitanda, and Denny (2022)] :

$$\min_{q} \left\{ \mathcal{L}(q) = \frac{1}{n} \sum_{i=1}^{n} \ell(h_q(x_i), y_i) + \lambda' \mathbb{E}_q[\|\theta\|_2^2] + \lambda \mathbb{E}_q[\log(q(\theta))] \right\}$$

$$= \max_{g \in \mathbb{R}^n} \left\{ \mathcal{D}(g) = -\frac{1}{n} \sum_{i=1}^{n} \ell_i^*(g_i) - \lambda \int q_g(\theta) \mathrm{d}\theta \right\}.$$

$\ell_i(z) = \ell(z, y_i),$

$q_g(\theta) = \exp\left(-\frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} h_\theta(x_i) g_i + \lambda' \|\theta\|_2^2\right)\right).$

# Primal-Dual Viewpoint

Duality for empirical risk [Oko, Suzuki, Nitanda, and Denny (2022)] :

$$\min_q \left\{ \mathcal{L}(q) = \frac{1}{n}\sum_{i=1}^n \ell(h_q(x_i), y_i) + \lambda' \mathbb{E}_q[\|\theta\|_2^2] + \lambda \mathbb{E}_q[\log(q(\theta))] \right\}$$

$$= \max_{g \in \mathbb{R}^n} \left\{ \mathcal{D}(g) = -\frac{1}{n}\sum_{i=1}^n \ell_i^*(g_i) - \lambda \int q_g(\theta) \mathrm{d}\theta \right\}.$$

$$\ell_i(z) = \ell(z, y_i),$$
$$q_g(\theta) = \exp\left(-\tfrac{1}{n}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n h_\theta(x_i) g_i + \lambda' \|\theta\|_2^2\right)\right).$$

**Theorem** (Duality Theorem): Set $g_q = \{\partial_z \ell(h_q(x_i), y_i)\}_{i=1}^n$.
Suppose $\ell(\cdot, y)$ is convex and differentiable. Then,

$$0 \le \mathcal{L}(q) - \mathcal{D}(g_q) = \lambda \mathrm{KL}(q \| p_q).$$

Through the round trip: $q \to g_q \to q_{g_q} \propto p_q$, $\mathrm{KL}(q\|p_q)$ exactly measures the duality gap.

# Primal-Dual Viewpoint

Duality for empirical risk [Oko, Suzuki, Nitanda, and Denny (2022)] :

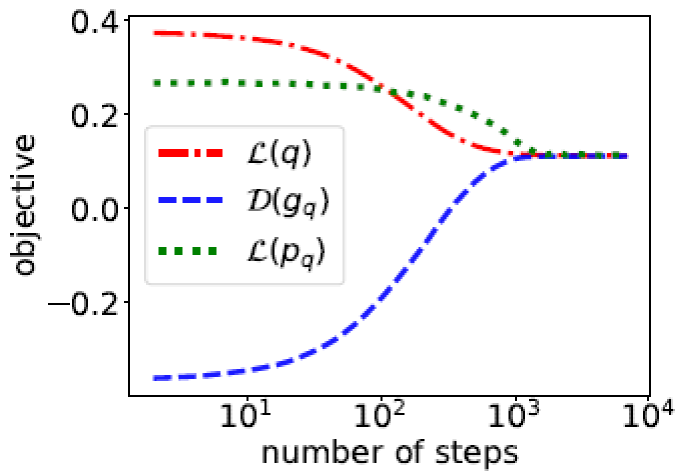
$$\min_q \left\{ \mathcal{L}(q) = \frac{1}{n}\sum_{i=1}^n \ell(h_q(x_i), y_i) + \lambda' \mathbb{E}_q[\|\theta\|_2^2] + \lambda \mathbb{E}_q[\log(q(\theta))] \right\}$$

$$= \max_{g \in \mathbb{R}^n} \left\{ \mathcal{D}(g) = -\frac{1}{n}\sum_{i=1}^n \ell_i^*(g_i) - \lambda \int q_g(\theta) \mathrm{d}\theta \right\}.$$

$$\ell_i(z) = \ell(z, y_i),$$
$$q_g(\theta) = \exp\left(-\tfrac{1}{n}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n h_\theta(x_i) g_i + \lambda' \|\theta\|_2^2\right)\right).$$

**Theorem** (Duality Theorem): Set $g_q = \{\partial_z \ell(h_q(x_i), y_i)\}_{i=1}^n$.
Suppose $\ell(\cdot, y)$ is convex and differentiable. Then,

$$0 \le \mathcal{L}(q) - \mathcal{D}(g_q) = \lambda \mathrm{KL}(q \| p_q).$$




Through the round trip: $q \to g_q \to q_{g_q} \propto p_q$, $\mathrm{KL}(q\|p_q)$ exactly measures the duality gap.

# Most Related Work

- **Convergence rate analysis in the continuous-time setting**

  [Hu, Ren, Siska, & Szpruch (2019)] shows the linear convergence of mean field Langevin with strong KL-regularization.

- **Quantitative convergence rate analysis under KL-regularization with any strength**

  [Nitanda, Denny, & Suzuki (2021)] is the first work that gives the quantitative convergence guarantees by proposing a method which exploits the convexity of the problem.
  [Oko, Suzuki, Nitanda, & Denny (2021)] gives an improved guarantee based on the similar idea.

**Contribution:** global convergence rate analysis for mean field Langevin dynamics with any strength KL-regularization.

# Most Related Work

- **Convergence rate analysis in the continuous-time setting**

  [Hu, Ren, Siska, & Szpruch (2019)] shows the linear convergence of mean field Langevin with strong KL-regularization.

- **Quantitative convergence rate analysis under KL-regularization with any strength**

  [Nitanda, Denny, & Suzuki (2021)] is the first work that gives the quantitative convergence guarantees by proposing a method which exploits the convexity of the problem.
  [Oko, Suzuki, Nitanda, & Denny (2021)] gives an improved guarantee based on the similar idea.

**Contribution:** global convergence rate analysis for mean field Langevin dynamics with any strength KL-regularization.

Concurrent work: [Chizat (2022)] also arrived at the same result in continuous time analysis.

Unique contributions in each paper: time-discretization and dual viewpoint in ours and annealed version in [Chizat (2022)].