Convex Analysis of Mean Field Langevin Dynamics

Atsushi Nitanda

This slide is based on the following paper:
A. Nitanda, D. Wu, and T. Suzuki. Convex Analysis of the Mean Field Langevin Dynamics. AISTATS, 2022.

Outline

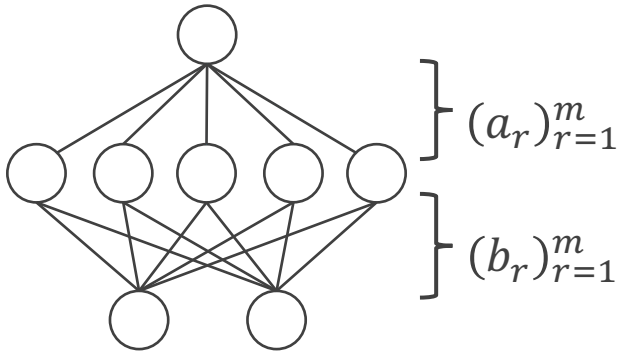
Topic: Convergence analysis of mean field Langevin dynamics.

Outline

Topic: Convergence analysis of mean field Langevin dynamics.

Application: optimization of neural networks in mean field regime.

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



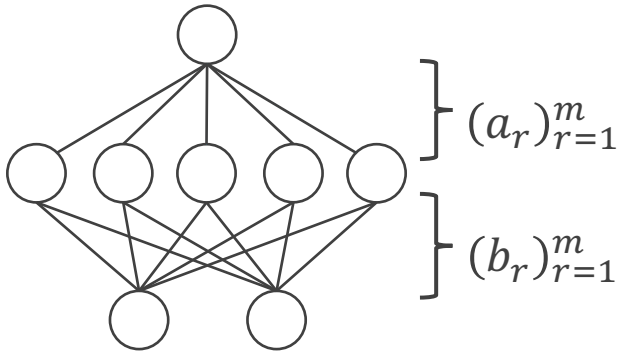
Mean field neural networks exhibit global convergence and adaptivity.

Outline

Topic: Convergence analysis of mean field Langevin dynamics.

Application: optimization of 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.

However, this model is difficult to optimize in general.
A structural assumption or regularization is needed for efficient optimization.

→ A noisy gradient descent optimizes KL-regularized loss.

Outline

A noisy gradient descent is analyzed via the mean-field Langevin dynamics.

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

Outline

A noisy gradient descent is analyzed via the mean-field Langevin dynamics.

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

The drift term involves the distribution unlike normal Langevin dynamics.

This dependence makes the convergence analysis difficult.

Outline

A noisy gradient descent is analyzed via the mean-field Langevin dynamics.

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

The drift term involves the distribution unlike normal Langevin dynamics.

This dependence makes the convergence analysis difficult.

Contribution:

- We resolve this difference by developing a simple proof with *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

A noisy gradient descent is analyzed via the mean-field Langevin dynamics.

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

The drift term involves the distribution unlike normal Langevin dynamics.

This dependence makes the convergence analysis difficult.

Contribution:

- We resolve this difference by developing a simple proof with *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

A noisy gradient descent is analyzed via the mean-field Langevin dynamics.

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

The drift term involves the distribution unlike normal Langevin dynamics.

This dependence makes the convergence analysis difficult.

Contribution:

- We resolve this difference by developing a simple proof with *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.

Mean-Field Neural Networks

Optimization for Two-layer NNs

- **Risk minimization** $\ell(z, y)$: loss function,

$$\min_{g:2\mathrm{NN}} \mathbb{E}_{(X,Y)\sim\rho}[\ell(g(X), Y)] + Reg,$$

squared loss: $\ell(z, y) = 0.5(z - y)^2$, logistic loss: $\ell(z, y) = \log(1 + \exp(-yz))$.

Optimization for Two-layer NNs

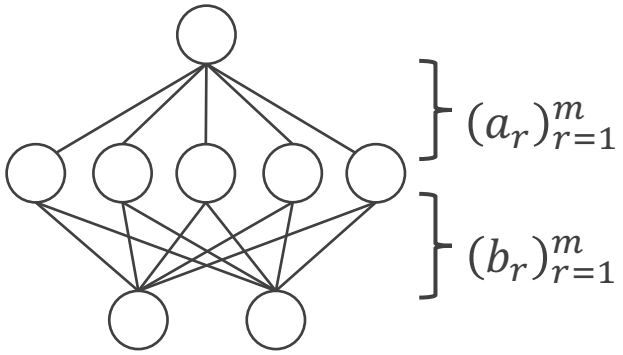
- **Risk minimization** $\ell(z, y)$: loss function,

$$\min_{g:2\mathrm{NN}} \mathbb{E}_{(X,Y)\sim\rho}[\ell(g(X), Y)] + Reg,$$

squared loss: $\ell(z, y) = 0.5(z - y)^2$, logistic loss: $\ell(z, y) = \log(1 + \exp(-yz))$.

- **A hypothesis function: 2-layer neural networks** $\Theta = (a_r, b_r)_{r=1}^m$,

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



$(a_r)_{r=1}^m$ are fixed in the theory.

Optimization for Two-layer NNs

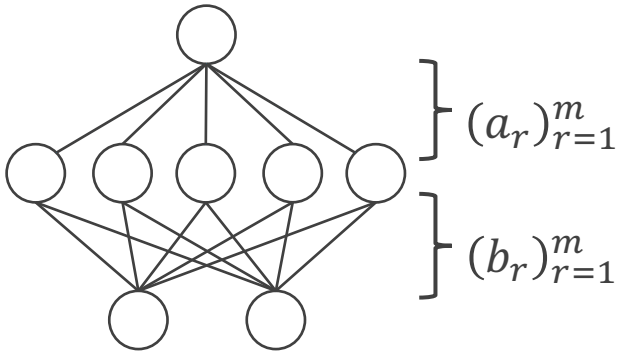
- **Risk minimization** $\ell(z, y)$: loss function,

$$\min_{g:2\mathrm{NN}} \mathbb{E}_{(X,Y)\sim\rho}[\ell(g(X), Y)] + Reg,$$

squared loss: $\ell(z, y) = 0.5(z - y)^2$, logistic loss: $\ell(z, y) = \log(1 + \exp(-yz))$.

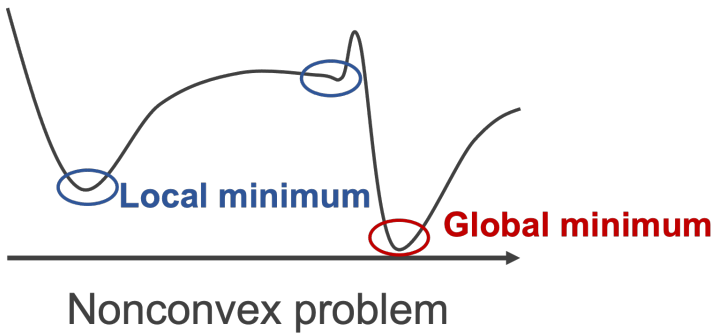
- **A hypothesis function: 2-layer neural networks** $\Theta = (a_r, b_r)_{r=1}^m$,

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



$(a_r)_{r=1}^m$ are fixed in the theory.

→ **Nonconvex optimization problems**



Gradient-based method converges to a stationary point, but over-parameterized neural networks often converges to a global minimum.

Common Approach

Key: mapping the optimization dynamics into the **function space and exploit the convexity** w.r.t the function:

$$\ell((g+\xi)(x), y) \geq \ell(g(x), y) + \partial_z \ell(z, y)|_{z=g(x)} \xi(x).$$

E.g.) squared loss: $\ell(z, y) = 0.5(z-y)^2$, logistic loss: $\ell(z, y) = \log(1+\exp(-yz))$.

Common Approach

Key: mapping the optimization dynamics into the **function space and exploit the convexity** w.r.t the function:

$$\ell((g+\xi)(x), y) \geq \ell(g(x), y) + \partial_z \ell(z, y)|_{z=g(x)} \xi(x).$$

E.g.) squared loss: $\ell(z, y) = 0.5(z-y)^2$, logistic loss: $\ell(z, y) = \log(1 + \exp(-yz))$.

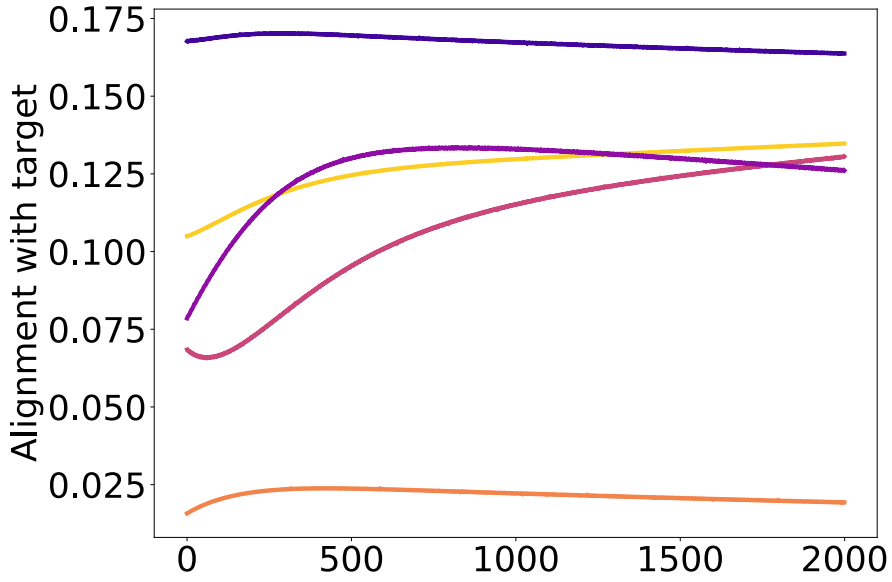
- **Mean field** [Nitanda & Suzuki (2017)], [Chizat & Bach (2018)], [Mei, Montanari, & Nguyen (2018)]
  Coefficient: $1/m$, learning rate: $O(m)$.
  Function space: probability measures.

- **Neural tangent kernel (NTK)** [Jacot, Gabriel, & Hongler (2018)]
  Coefficient: $1/\sqrt{m}$, learning rate: $O(1)$.
  Function space: reproducing kernel Hilbert space (RKHS) associated with NTK.

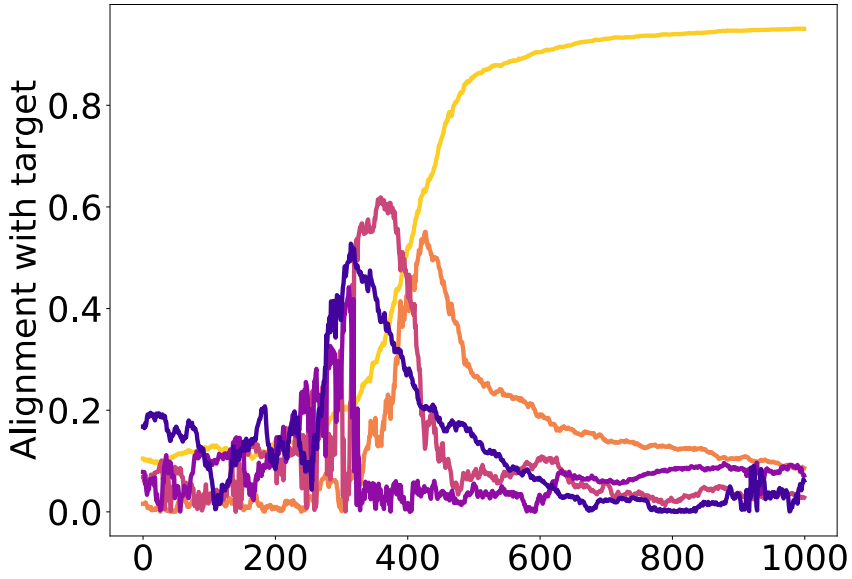
Adaptivity

The target function is a single neuron model with parameter w_*.

The figure plots the alignment between each vector of student model and w_*.



NTK-regime

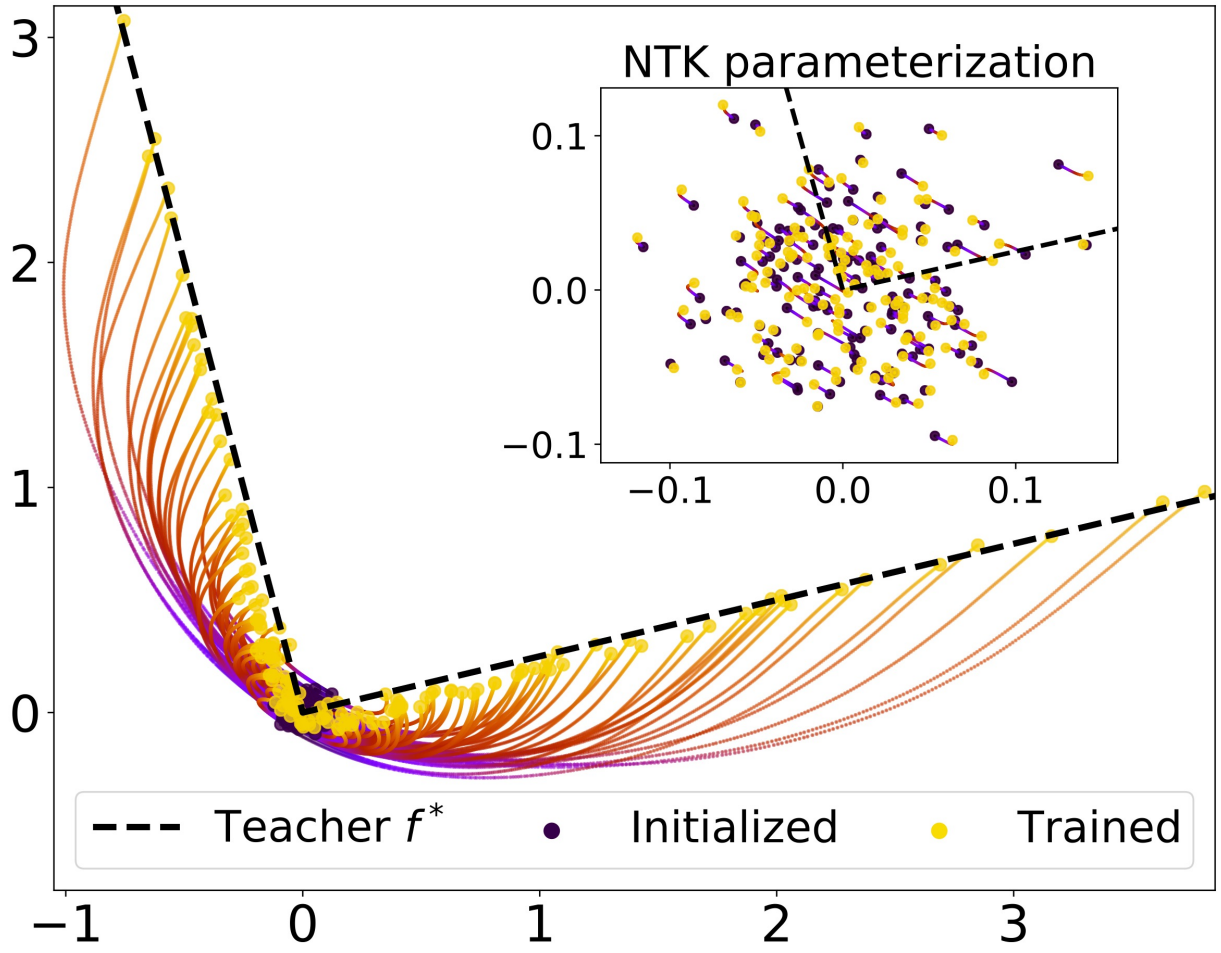


MF-regime

Mean field neural network shows the adaptivity to the low dimensional structure.

Adaptivity

Alignment of parameters can be observed in mean-field regime.



Basic Idea behind Mean field NNs

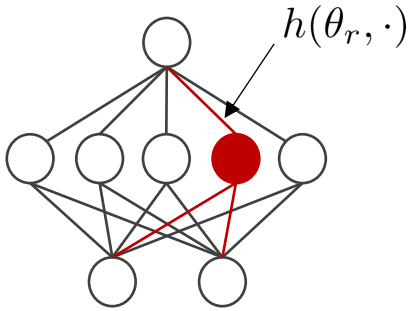
Element of mean field neural nets: $h(\theta, \cdot)$

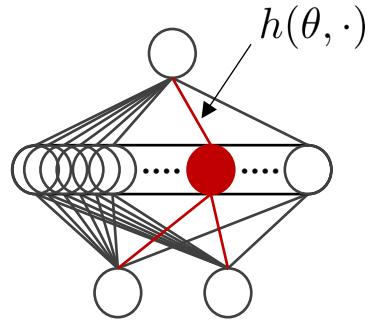
Parameter: $\Theta = (\theta_r)_{r=1}^m, \ (\theta_r \sim q(\theta)\mathrm{d}\theta)$

E.g.) $h(\theta, x) = a\sigma(b^\top x) \ \ (\theta = (a, b), \ or\ \theta = b).$

Linear w.r.t. q.

$$h_\Theta(x) = \frac{1}{m}\sum_{r=1}^{m} h(\theta_r, x) \xrightarrow{m \to \infty} h_q(x) = \int h(\theta, x) q(\theta)\mathrm{d}\theta$$





Infinitely many neurons are integrated with $q(\theta)$.

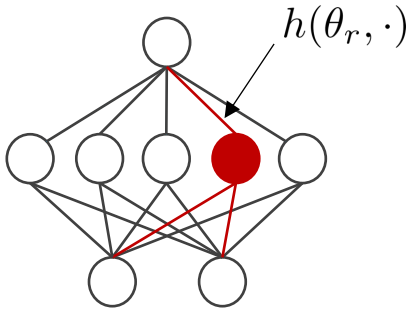
Basic Idea behind Mean field NNs

Element of mean field neural nets: $h(\theta, \cdot)$ E.g.) $h(\theta, x) = a\sigma(b^\top x) \quad (\theta = (a, b), \ or\ \theta = b).$

Parameter: $\Theta = (\theta_r)_{r=1}^m, \ (\theta_r \sim q(\theta)\mathrm{d}\theta)$

Linear w.r.t. q.

$$h_\Theta(x) = \frac{1}{m}\sum_{r=1}^{m} h(\theta_r, x) \xrightarrow{m \to \infty} h_q(x) = \int h(\theta, x) q(\theta)\mathrm{d}\theta$$



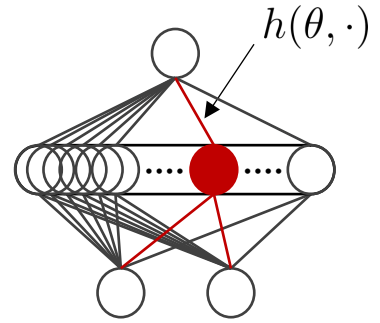
Loss

$$\mathbb{E}\left[\ell(h_\Theta(X), Y)\right] \xrightarrow{m \to \infty} \mathbb{E}\left[\ell(h_q(X), Y)\right]$$

Nonconvex w.r.t. Θ.

Loss

Convex w.r.t. q.



Infinitely many neurons are integrated with $q(\theta)$.

The diagram suggests the convex analysis in the space of probability measures.

Regularized Risk Minimization

KL-regularized risk minimization (convex optimization problems):

$$\min_{q\in\mathcal{P}} \left\{ \mathcal{L}(q) = \mathbb{E}_{(X,Y)}[\ell(h_q(X), Y)] + \lambda' \mathbb{E}_q[\|\theta\|_2^2] + \lambda \mathbb{E}_q[\log(q(\theta))] \right\}.$$

Kullback-Leibler divergence to zero-mean Gaussian

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

$\mathcal{P}$: the set of probability densities with well-defined second moment and entropy.

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

Regularized Risk Minimization

KL-regularized risk minimization (convex optimization problems):

$$\min_{q\in\mathcal{P}}\left\{\mathcal{L}(q)=\mathbb{E}_{(X,Y)}[\ell(h_q(X),Y)]+\lambda'\mathbb{E}_q[\|\theta\|_2^2]+\lambda\mathbb{E}_q[\log(q(\theta))]\right\}.$$

Kullback-Leibler divergence to zero-mean Gaussian

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

$\mathcal{P}$: the set of probability densities with well-defined second moment and entropy.

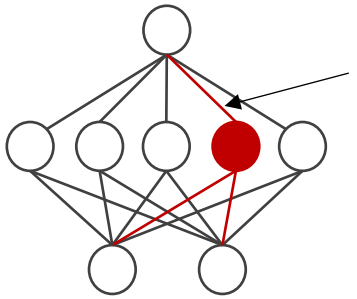
$\mathbb{E}_q$ denotes the expectation w.r.t $\theta \sim q(\theta)\mathrm{d}\theta$.

→ Global convergence rate analysis by exploiting the convexity of the loss function w.r.t. the probability density.

Optimization

- ($m < \infty$) Noisy particle gradient descent for $\mathbb{E}\left[\ell(h_\Theta(X), Y)\right]$ with L_2-reg.:

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

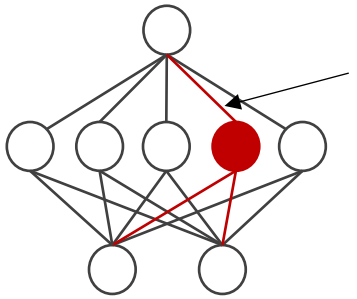


Optimization dynamics for parameter θ_r associated with $h(\theta_r, \cdot)$.

Optimization

- ($m < \infty$) Noisy particle gradient descent for $\mathbb{E}\left[\ell(h_\Theta(X), Y)\right]$ with L_2-reg.:

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



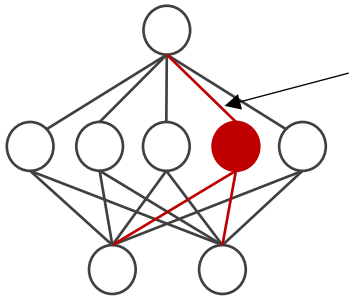
Optimization dynamics for parameter θ_r associated with $h(\theta_r, \cdot)$.

$$h_{\Theta^{(k)}}(x) = \frac{1}{m}\sum_{r=1}^{m} h(\theta_r^{(k)}, x) \to h_{q^{(k)}}(x) = \int h(\theta, x) q^{(k)}(\theta) \mathrm{d}\theta.$$

Optimization

- ($m < \infty$) Noisy particle gradient descent for $\mathbb{E}\left[\ell(h_\Theta(X), Y)\right]$ with L_2-reg.:

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

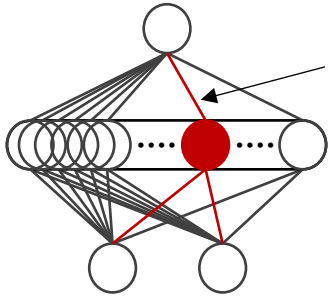


Optimization dynamics for parameter θ_r associated with $h(\theta_r, \cdot)$.

$$h_{\Theta^{(k)}}(x) = \frac{1}{m}\sum_{r=1}^{m} h(\theta_r^{(k)}, x) \to h_{q^{(k)}}(x) = \int h(\theta, x) q^{(k)}(\theta) \mathrm{d}\theta.$$

- ($m = \infty$) Mean-field Langevin dynamics (discrete-time): $q^{(k)}$ is a distribution of $\theta^{(k)}$,

$$\theta^{(k+1)} \leftarrow (1 - 2\eta\lambda')\theta^{(k)} - \eta\mathbb{E}[\partial_z \ell(h_{q^{(k)}}(X), Y)\partial_\theta h(\theta^{(k)}, X)] + \sqrt{2\eta\lambda}\zeta^{(k)}.$$



Optimization dynamics for parameter θ associated with $h(\theta, \cdot)$.

Discretization error w.r.t. m is well studied by related works (e.g., [Mei, Montanari, & Nguyen (2018)].

Optimization

Set $g_q(\theta) = \mathbb{E}[\partial_z \ell(h_q(X), Y) h(\theta, X)] + \lambda' \|\theta\|_2^2$.

- ($\eta > 0$) discrete-time mean-field Langevin dynamics:

$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta \nabla_\theta g_{q^{(k)}}(\theta^{(k)}) + \sqrt{2\eta\lambda}\zeta^{(k)}.$$

Optimization

Set $g_q(\theta) = \mathbb{E}[\partial_z \ell(h_q(X), Y) h(\theta, X)] + \lambda' \|\theta\|_2^2$.

- ($\eta > 0$) discrete-time mean-field Langevin dynamics:

$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta \nabla_\theta g_{q^{(k)}}(\theta^{(k)}) + \sqrt{2\eta\lambda}\zeta^{(k)}.$$

- ($\eta \to 0$) continuous-time mean-field Langevin dynamics:

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

Optimization

Set $g_q(\theta) = \mathbb{E}[\partial_z \ell(h_q(X), Y) h(\theta, X)] + \lambda' \|\theta\|_2^2$.

- ($\eta > 0$) discrete-time mean-field Langevin dynamics:

$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta \nabla_\theta g_{q^{(k)}}(\theta^{(k)}) + \sqrt{2\eta\lambda}\zeta^{(k)}.$$

- ($\eta \to 0$) continuous-time mean-field Langevin dynamics:

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

- Evolution of probability distributions.

$$\frac{\partial q_t}{\partial t} = \nabla \cdot (q_t \nabla_\theta g_{q_t}(\theta_t)) + \lambda \Delta q_t.$$

Optimization

Set $g_q(\theta) = \mathbb{E}[\partial_z \ell(h_q(X), Y) h(\theta, X)] + \lambda' \|\theta\|_2^2$.

- ($\eta > 0$) discrete-time mean-field Langevin dynamics:

$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta \nabla_\theta g_{q^{(k)}}(\theta^{(k)}) + \sqrt{2\eta\lambda}\zeta^{(k)}.$$

- ($\eta \to 0$) continuous-time mean-field Langevin dynamics:

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

- Evolution of probability distributions.

$$\frac{\partial q_t}{\partial t} = \nabla \cdot (q_t \nabla_\theta g_{q_t}(\theta_t)) + \lambda \Delta q_t.$$

These are very similar to Langevin dynamics and (linear) Fokker-Planck eq.
But the drift term depends on not only the parameter but also the distribution.

Optimization

Set $g_q(\theta) = \mathbb{E}[\partial_z \ell(h_q(X), Y) h(\theta, X)] + \lambda' \|\theta\|_2^2$.

- ($\eta > 0$) discrete-time mean-field Langevin dynamics:

$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta \nabla_\theta g_{q^{(k)}}(\theta^{(k)}) + \sqrt{2\eta\lambda}\zeta^{(k)}.$$

- ($\eta \to 0$) continuous-time mean-field Langevin dynamics:

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

- Evolution of probability distributions.

$$\frac{\partial q_t}{\partial t} = \nabla \cdot (q_t \nabla_\theta g_{q_t}(\theta_t)) + \lambda \Delta q_t.$$

These are very similar to Langevin dynamics and (linear) Fokker-Planck eq.
But the drift term depends on not only the parameter but also the distribution.

We propose a new proof techniques which resolves this distinction and extends the analsys of Langevin dynamics into the mean-field setting.

Related Method: Langevin Dynamics

Related Method: Langevin Dynamics

- Sampling method for Gibbs distributions $q_* \propto \exp(-f)$:

$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta \nabla_\theta f(\theta^{(k)}) + \sqrt{2\eta}\zeta^{(k)}.$$

(Gradient descent + Gaussian perturbation)

η : step size, $\zeta^{(k)} \sim \mathcal{N}(0, I)$: Gaussian noise.

Related Method: Langevin Dynamics

- Sampling method for Gibbs distributions $q_* \propto \exp(-f)$:

$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta \nabla_\theta f(\theta^{(k)}) + \sqrt{2\eta}\zeta^{(k)}.$$

(Gradient descent + Gaussian perturbation)

η : step size, $\zeta^{(k)} \sim \mathcal{N}(0, I)$: Gaussian noise.

- Langevin Monte Carlo is a discretization of the Langevin dynamics:

$$\mathrm{d}\theta_t = -\nabla f(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t.$$

Related Method: Langevin Dynamics

- Sampling method for Gibbs distributions $q_* \propto \exp(-f)$:

$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta \nabla_\theta f(\theta^{(k)}) + \sqrt{2\eta}\zeta^{(k)}.$$

(Gradient descent + Gaussian perturbation)

η : step size, $\zeta^{(k)} \sim \mathcal{N}(0, I)$: Gaussian noise.

- Langevin Monte Carlo is a discretization of the Langevin dynamics:

$$\mathrm{d}\theta_t = -\nabla f(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t.$$

- The evolution of distributions of $\{\theta_t\}_{t\geq 0}$:

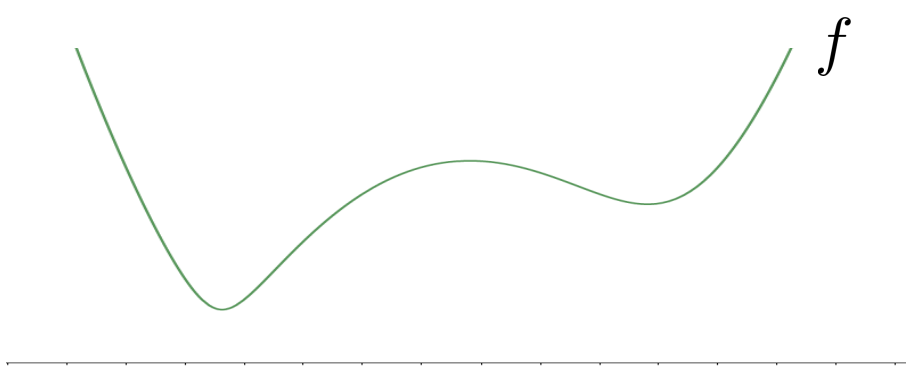
$$\frac{\partial q_t}{\partial t} = \nabla \cdot (q_t \nabla f) + \Delta q_t = \nabla \cdot \left(q_t \nabla \log \frac{q_t}{q_*} \right).$$

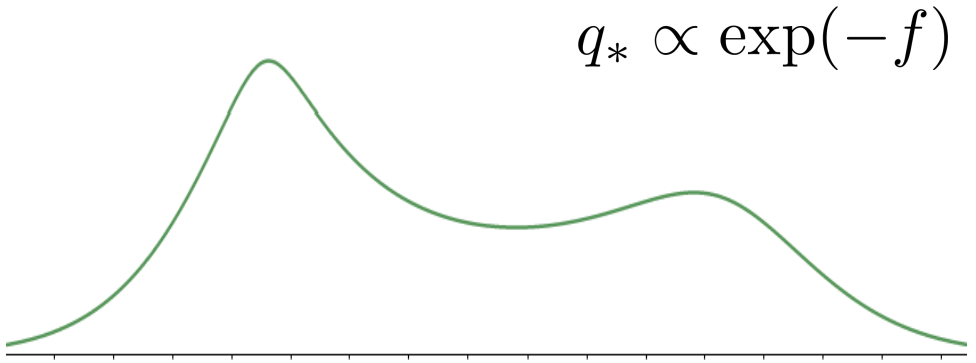
Related Method: Langevin Dynamics

- Langevin Monte Carlo: Gradient descent + Gaussian perturbation:

$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta \nabla_\theta f(\theta^{(k)}) + \sqrt{2\eta}\zeta^{(k)}.$$

- LMC converges not to a local minimum of f but to a distribution $q_* \propto \exp(-f)$ up to a discretization error in some sense.





Related Method: Langevin Dynamics

Assumption (log-Sobolev inequality): We say q_* satisfies LSI with a constant $\alpha > 0$ when it follows that for any smooth density q.

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

Related Method: Langevin Dynamics

Assumption (log-Sobolev inequality): We say q_* satisfies LSI with a constant $\alpha > 0$ when it follows that for any smooth density q.

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

Under LSI, we can show the convergence of Langevin dynamics.

Theorem [e.g., Vempala & Wibisono (2019)]

Suppose q_* satisfies LSI with $\alpha > 0$, then along the Langevin dynamics,

$$\mathrm{KL}(q_t\|q_*) \leq \exp(-2\alpha t)\mathrm{KL}(q_0\|q_*).$$

Remark: This result is generalized to discrete-time by evaluating the discretization error.

Related Method: Langevin Dynamics

Langevin MC approximately solve the entropy regularized linear functional:

$$\min_q \left\{ \mathbb{E}_q[f] + \mathbb{E}_q[\log(q)] = \mathbb{E}_q \left[\log \frac{q}{\exp(-f)} \right] \right\}.$$

This is a sort of strongly convex problems w.r.t. KL-divergence.

Related Method: Langevin Dynamics

Functional derivative (1st order variation) in q: $\frac{\delta}{\delta q}$.

Functional derivative of the negative entropy:

$$\frac{\delta}{\delta q}\mathbb{E}_q[\log q] = \log q$$

i.e., for τ such that $\int \tau(\theta)\mathrm{d}\theta = 0$,

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathbb{E}_{q+\epsilon\tau}[\log(q+\epsilon\tau)]|_{\epsilon=0} = \int \tau(\theta)\log q(\theta)\mathrm{d}\theta.$$

inner-product (coupling) between primal and dual spaces.

Related Method: Langevin Dynamics

Functional derivative (1st order variation) in q: $\frac{\delta}{\delta q}$.

Functional derivative of the negative entropy:

$$\frac{\delta}{\delta q}\mathbb{E}_q[\log q] = \log q$$

i.e., for τ such that $\int \tau(\theta)\mathrm{d}\theta = 0$,

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathbb{E}_{q+\epsilon\tau}[\log(q+\epsilon\tau)]|_{\epsilon=0} = \int \tau(\theta)\log q(\theta)\mathrm{d}\theta.$$

inner-product (coupling) between primal and dual spaces.

Strong convexity w.r.t. KL-divergence:

$$\mathbb{E}_{q'}[\log q'] = \mathbb{E}_q[\log q] + \int \frac{\delta}{\delta q}\mathbb{E}_q[\log q](\theta)(q'-q)(\theta)\mathrm{d}\theta + \mathrm{KL}(q'\|q).$$

Related Method: Langevin Dynamics

Therefore, the problem LMC solved is a strongly convex.

$$\min_q \{\mathcal{L}(q) = \mathbb{E}_q[f] + \mathbb{E}_q[\log(q)]\}.$$

Langevin dynamics can exploit this convexity and converge linearly under LSI.

Related Method: Langevin Dynamics

Therefore, the problem LMC solved is a strongly convex.

$$\min_q \{\mathcal{L}(q) = \mathbb{E}_q[f] + \mathbb{E}_q[\log(q)]\}.$$

Langevin dynamics can exploit this convexity and converge linearly under LSI.

However, the theory for Langevin MC cannot apply to learning mean-field NN because of the nonlinearity of the loss term:

$$\min_{q\in\mathcal{P}} \left\{\mathcal{L}(q) = \underbrace{\mathbb{E}_{(X,Y)}[\ell(h_q(X), Y)]}_{\text{nonlinear convex functional}} + \lambda'\mathbb{E}_q[\|\theta\|_2^2] + \lambda\mathbb{E}_q[\log(q(\theta))]\right\}.$$

Related Method: Langevin Dynamics

Basic properties of $\mathcal{L}(q) = \mathbb{E}_q[f] + \mathbb{E}_q[\log(q)]$.

Proposition

1. $\frac{\delta \mathcal{L}}{\delta q}(q) = \log \frac{q}{q_*}, \qquad q_* \propto \exp(-f)$
2. $\mathcal{L}(q) + \int \frac{\delta \mathcal{L}}{\delta q}(q)(\theta)(q' - q)(\theta) \mathrm{d}\theta + \mathrm{KL}(q' \| q) = \mathcal{L}(q'),$
3. $\mathcal{L}(q_*) + \mathrm{KL}(q' \| q_*) = \mathcal{L}(q').$

A similar property with a quadratic function in the finite-dimensional space.

Mean-Field Langevin Dynamics

Entropy Regularized Convex Functional

Consider the optimization of entropy regularized convex functional:

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

where $F : \mathcal{P} \to \mathbb{R}$ is a convex functional. That is,

$$F(q') \geq F(q) + \int \frac{\delta F}{\delta q}(q)(\theta)(q' - q)(\theta)\mathrm{d}\theta.$$

For τ such that $\int \tau(\theta)\mathrm{d}\theta = 0$,

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} F(q + \epsilon\tau)|_{\epsilon=0} = \int \tau(\theta) \frac{\delta F(q)}{\delta q}(\theta)\mathrm{d}\theta.$$

Entropy Regularized Convex Functional

Consider the optimization of entropy regularized convex functional:

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

where $F : \mathcal{P} \to \mathbb{R}$ is a convex functional. That is,

$$F(q') \geq F(q) + \int \frac{\delta F}{\delta q}(q)(\theta)(q' - q)(\theta)\mathrm{d}\theta.$$

For τ such that $\int \tau(\theta)\mathrm{d}\theta = 0$,

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} F(q + \epsilon\tau)|_{\epsilon=0} = \int \tau(\theta)\frac{\delta F(q)}{\delta q}(\theta)\mathrm{d}\theta.$$

Example (mean-field NNs): convex loss $\ell(z, y)$ in z.

$$F(q) = \mathbb{E}_{(X,Y)}[\ell(h_q(X), Y)] + \lambda' \mathbb{E}_q[\|\theta\|_2^2],$$

$$\frac{\delta F(q)}{\delta q}(\theta) = \mathbb{E}_{(X,Y)}[\partial_z \ell(h_q(X), Y)h(\theta, X)] + \lambda' \|\theta\|_2^2.$$

Optimization Dynamics

- Consider the following update to solve the problem:

$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta \nabla \frac{\delta F}{\delta q}(q^{(k)})(\theta^{(k)}) + \sqrt{2\lambda\eta}\zeta^{(k)}.$$

  where $q^{(k)}(\theta)\mathrm{d}\theta$ is a distribution of $\theta^{(k)}$.

- This is a discretization of the continuous dynamics:

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

Optimization Dynamics

- Consider the following update to solve the problem:

$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta \nabla \frac{\delta F}{\delta q}(q^{(k)})(\theta^{(k)}) + \sqrt{2\lambda\eta}\zeta^{(k)}.$$

  where $q^{(k)}(\theta)\mathrm{d}\theta$ is a distribution of $\theta^{(k)}$.

- This is a discretization of the continuous dynamics:

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

Example (mean-field Langevin dynamics):

$$\theta^{(k+1)} \leftarrow (1 - 2\eta\lambda')\theta^{(k)} - \eta\mathbb{E}[\partial_z \ell(h_{q^{(k)}}(X), Y)\partial_\theta h(\theta^{(k)}, X)] + \sqrt{2\eta\lambda}\zeta^{(k)}.$$

A continuous limit of the noisy gradient descent. $\nabla g_q(\theta) = \nabla \frac{\delta F}{\delta q}(q)(\theta).$

Proximal Gibbs Distribution

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

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

The proximal Gibbs p_q is the minimizer of the following functional in q':

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

This distribution plays the key role to resolve difference from Langevin dynamics.

Remark: When $F(q) = \mathbb{E}_q[f]$ with $\lambda = 1$, it follows that $p_q(\theta) \propto \exp(-f(\theta))$.

Basic Properties

Proposition:

1. $\frac{\delta \mathcal{L}}{\delta q}(q) = \lambda \log \frac{q}{p_q},$

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

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

Eq. 2 means the strong convexity of the problem.

Eq. 3 corresponds to the PL-condition and quadratic growth in finite-dim. convex analysis.

$\mathrm{KL}(q \| p_q)$ measures the optimization gap.

Comparison with linear case

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

Proposition (linear)

1. $\frac{\delta \mathcal{L}}{\delta q}(q) = \log \frac{q}{q_*}, \quad q_* \propto \exp(-f)$
2. $\mathcal{L}(q) + \int \frac{\delta \mathcal{L}}{\delta q}(q)(\theta)(q' - q)(\theta)\mathrm{d}\theta + \mathrm{KL}(q' \| q) = \mathcal{L}(q'),$
3. $\mathcal{L}(q_*) + \mathrm{KL}(q' \| q_*) = \mathcal{L}(q').$

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

Proposition (nonlinear)

1. $\frac{\delta \mathcal{L}}{\delta q}(q) = \lambda \log \frac{q}{p_q},$
2. $\mathcal{L}(q) + \int \frac{\delta \mathcal{L}}{\delta q}(q)(\theta)(q' - q)(\theta)\mathrm{d}\theta + \lambda \mathrm{KL}(q' \| q) \leq \mathcal{L}(q'),$
3. $\lambda \mathrm{KL}(q \| p_q) \geq \mathcal{L}(q) - \mathcal{L}(q_*) \geq \lambda \mathrm{KL}(q \| q_*).$

Properties for linear functional into the nonlinear (mean-field) settings via p_q.

Nonlinear Fokker-Planck Equation

- The evolution of distributions associated with $\mathrm{d}\theta_t = -\nabla \frac{\delta F}{\delta q}(q_t)(\theta_t)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}W_t$:

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

which can be reformulated with p_q :

$$\begin{aligned} \frac{\partial q_t}{\partial t} &= \lambda \nabla \cdot \left(q_t \nabla \log \exp \left(\frac{1}{\lambda} \frac{\delta F}{\delta q}(q_t) \right) + q_t \nabla \log q_t \right) \\ &= \lambda \nabla \cdot \left(q_t \nabla \log \frac{q_t}{p_{q_t}} \right). \end{aligned}$$

Recall: linear Fokker-Planck equation is $\frac{\partial q_t}{\partial t} = \nabla \cdot \left(q_t \nabla \log \frac{q_t}{q_*}. \right)$

This term is replaced with p_q.

Assumption

Assumption (modified LSI): For any q, the distribution p_q satisfies LSI with $\alpha > 0$.
That is, we suppose

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

Compared to the LSI used for the normal Langevin dynamics:

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

a density q_* is replaced with p_q.

When the objective (except regularization) is linear, these densities will match.
Hence, this LSI is a natural generalization of that for the normal Langevin dynamics.

Example: Mean-field NN with regularization

Lemma (Holley and Stroock (1987):

Let $q(\theta)\mathrm{d}\theta$ be a probability distribution which satisfies LSI with $\alpha > 0$. For any bounded function in proportion to $\exp(B(\theta))q(\theta)\mathrm{d}\theta$ satisfies LSI with $\alpha/\exp(4\|B\|_\infty)$.

Holley and Stroock (1987) argument guarantees the LSI of

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

when the potential $\frac{\delta F(q)}{\delta q}$ is the sum of strongly convex and bounded functions.

Example: Mean-field NN with regularization

Lemma (Holley and Stroock (1987):

Let $q(\theta)\mathrm{d}\theta$ be a probability distribution which satisfies LSI with $\alpha > 0$. For any bounded function in proportion to $\exp(B(\theta))q(\theta)\mathrm{d}\theta$ satisfies LSI with $\alpha/\exp(4\|B\|_\infty)$.

Holley and Stroock (1987) argument guarantees the LSI of

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

when the potential $\frac{\delta F(q)}{\delta q}$ is the sum of strongly convex and bounded functions.

Example (mean-field NN):

$$F(q) = \mathbb{E}_{(X,Y)}[\ell(h_q(X), Y)] + \lambda' \mathbb{E}_q[\|\theta\|_2^2],$$

$$\frac{\delta F(q)}{\delta q}(\theta) = \mathbb{E}_{(X,Y)}[\partial_z \ell(h_q(X), Y)h(\theta, X)] + \lambda'\|\theta\|_2^2.$$

Hence, if h is uniformly bounded, then p_q satisfies LSI with $\frac{2\lambda'}{\lambda \exp(C\lambda^{-1})}$.

Convergence Analysis (continuous-time)

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

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

[Chizat (2022)] also obtained the same result.

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 (discrete-time)

Theorem (informal): Let $\{q^{(k)}\}_{k\in\mathbb{Z}_+}$ be the evolution of noisy gradient descent with $\eta = O(\epsilon\alpha\lambda)$ in the mean-field limit. Under modified LSI assumption with $\alpha > 0$ and appropriate assumptions,

$$\mathcal{L}(q^{(k)}) - \mathcal{L}(q_*) = O(\epsilon) + \exp(-O(\epsilon\alpha^2\lambda^2 k))(\mathcal{L}(q^{(0)}) - \mathcal{L}(q_*)).$$

- A convergence rate for noisy gradient descent (discrete-time update) is derived by evaluating time-discretization error.
- An approximation error remains.
- The rate deteriorates because of small step size.

Primal-Dual Viewpoint

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

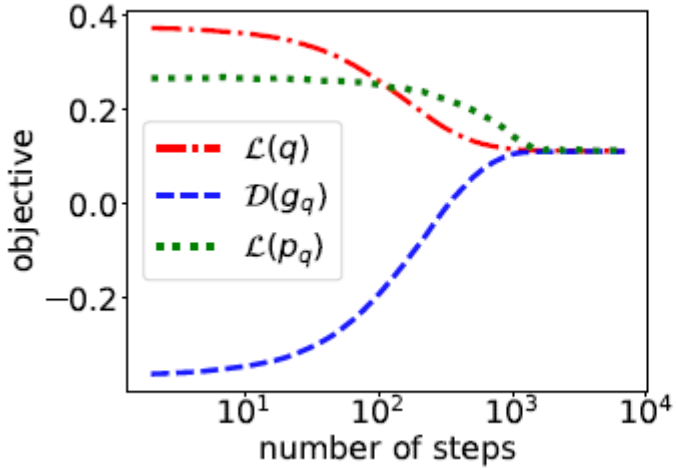
$$\min_q \left\{ \mathcal{L}(q) = \frac{1}{n}\sum_{i=1}^n \ell(h_q(x_i), y_i) + \lambda' \mathbb{E}_q[\|\theta\|_2^2] + \lambda \mathbb{E}_q[\log(q(\theta))] \right\}$$

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

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

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,

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



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

Related Work

- **Convergence analysis**
  - [Nitanda & Suzuki (2017)]: Relationship between the gradient descent and Wasserstein gradient flow.
  - [Chizat & Bach (2018)], [Mei, Montanari, & Nguyen (2018)]: Global convergence analysis for 2-NN.

Related Work

- **Convergence analysis**
  - [Nitanda & Suzuki (2017)]: Relationship between the gradient descent and Wasserstein gradient flow.
  - [Chizat & Bach (2018)], [Mei, Montanari, & Nguyen (2018)]: Global convergence analysis for 2-NN.
- **Convergence rate analysis in the continuous-time setting**
  - [Rotskoff, Jelassi, Bruna, & Vanden-Eijnden (2019)]: Sublinear convergence rate for the neuron birth-death dynamics.
  - [Javanmard, Mondelli, & Montanari (2019)]: Linear convergence rate for the strong concave target function.
  - [Hu, Ren, Siska, & Szpruch (2019)]: Linear convergence of mean field Langevin with strong KL-div. regularization.

Related Work

- **Convergence analysis**
  - [Nitanda & Suzuki (2017)]: Relationship between the gradient descent and Wasserstein gradient flow.
  - [Chizat & Bach (2018)], [Mei, Montanari, & Nguyen (2018)]: Global convergence analysis for 2-NN.
- **Convergence rate analysis in the continuous-time setting**
  - [Rotskoff, Jelassi, Bruna, & Vanden-Eijnden (2019)]: Sublinear convergence rate for the neuron birth-death dynamics.
  - [Javanmard, Mondelli, & Montanari (2019)]: Linear convergence rate for the strong concave target function.
  - [Hu, Ren, Siska, & Szpruch (2019)]: Linear convergence of mean field Langevin with strong KL-div. regularization.
- **Convergence rate analysis in the discrete-time setting**
  - [Chizat (2019)], [Akiyama & Suzuki (2021)]: Local linear convergence under structural assumption.

Related Work

- **Convergence analysis**
  - [Nitanda & Suzuki (2017)]: Relationship between the gradient descent and Wasserstein gradient flow.
  - [Chizat & Bach (2018)], [Mei, Montanari, & Nguyen (2018)]: Global convergence analysis for 2-NN.
- **Convergence rate analysis in the continuous-time setting**
  - [Rotskoff, Jelassi, Bruna, & Vanden-Eijnden (2019)]: Sublinear convergence rate for the neuron birth-death dynamics.
  - [Javanmard, Mondelli, & Montanari (2019)]: Linear convergence rate for the strong concave target function.
  - [Hu, Ren, Siska, & Szpruch (2019)]: Linear convergence of mean field Langevin with strong KL-div. regularization.
- **Convergence rate analysis in the discrete-time setting**
  - [Chizat (2019)], [Akiyama & Suzuki (2021)]: Local linear convergence under structural assumption.
- **Quantitative convergence rate analysis under KL-regularization with any strength**
  - [Nitanda, Denny, & Suzuki (2021)], [Oko, Suzuki, Nitanda, & Denny (2022)], [Bou-Rabee and Eberle (2021)].

Related Work

- **Convergence analysis**
  - [Nitanda & Suzuki (2017)]: Relationship between the gradient descent and Wasserstein gradient flow.
  - [Chizat & Bach (2018)], [Mei, Montanari, & Nguyen (2018)]: Global convergence analysis for 2-NN.
- **Convergence rate analysis in the continuous-time setting**
  - [Rotskoff, Jelassi, Bruna, & Vanden-Eijnden (2019)]: Sublinear convergence rate for the neuron birth-death dynamics.
  - [Javanmard, Mondelli, & Montanari (2019)]: Linear convergence rate for the strong concave target function.
  - [Hu, Ren, Siska, & Szpruch (2019)]: Linear convergence of mean field Langevin with strong KL-div. regularization.
- **Convergence rate analysis in the discrete-time setting**
  - [Chizat (2019)], [Akiyama & Suzuki (2021)]: Local linear convergence under structural assumption.
- **Quantitative convergence rate analysis under KL-regularization with any strength**
  - [Nitanda, Denny, & Suzuki (2021)], [Oko, Suzuki, Nitanda, & Denny (2022)], [Bou-Rabee and Eberle (2021)].

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)].