NEURAL INFORMATION PROCESSING SYSTEMS

Particle Dual Averaging: Optimization of Mean Field Neural Network with Global Convergence Rate Analysis

Atsushi Nitanda[1,2,3] Denny Wu[4,5] 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

NeurIPS2021 (ONLINE)

Outline

Topic: Convergence analysis of mean field neural networks.

Mean field neural networks exhibit global convergence and adaptivity.

Outline

Topic: Convergence analysis of mean field neural networks.

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.

Outline

Topic: Convergence analysis of mean field neural networks.

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.

Contribution: We develop Particle Dual Averaging for KL-regularized problem.
We give quantitative convergence guarantees in discrete-time setting.

To obtain an ϵ-accurate solution,
Iteration complexity: $\tilde{O}(\epsilon^{-3})$, Particle complexity (# of neurons): $\tilde{O}(\epsilon^{-2})$.

Optimization for Two-layer NNs

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

$$\min_{g:\mathrm{2NN}} \mathbb{E}_{(X,Y)\sim\rho} l(g(X),Y),$$

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

Optimization for Two-layer NNs

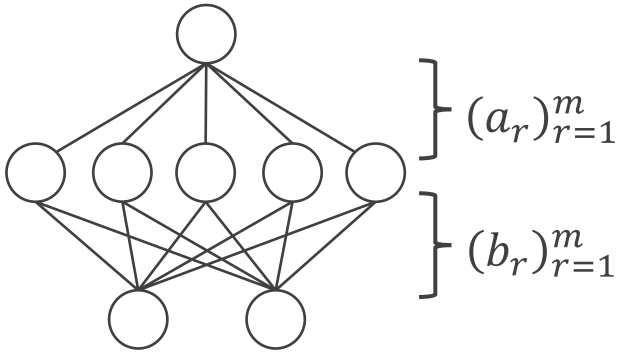
- **Risk minimization** $l(z, y)$: loss function,

$$\min_{g:2\text{NN}} \mathbb{E}_{(X,Y)\sim\rho} l(g(X), Y),$$

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

- **Two-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

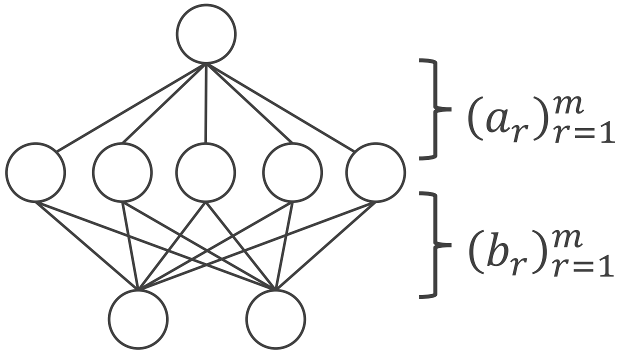
- **Risk minimization** $l(z,y)$: loss function,

$$\min_{g:2\mathrm{NN}} \mathbb{E}_{(X,Y)\sim\rho} l(g(X),Y),$$

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

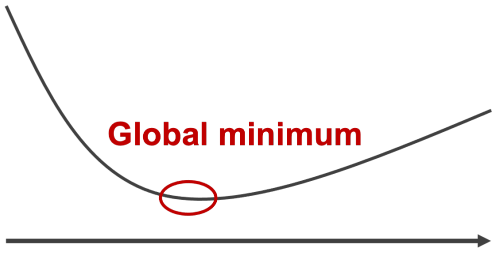
- **Two-layer neural networks** $\Theta = (a_r, b_r)_{r=1}^m$,

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

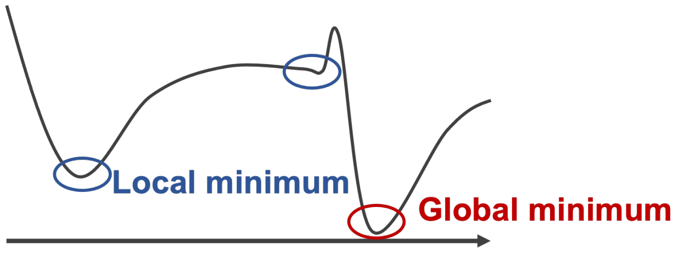


$(a_r)_{r=1}^m$ are fixed in the theory.

$\rightarrow$ **Nonconvex optimization problems**



Convex problem



Nonconvex problem

Gradient-based method converges to a stationary point : $\nabla_\Theta \mathcal{L}(\Theta) = 0$.

Common Approach

Key: characterize the **function space** where optimization performs.

Convexity w.r.t the function

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

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

Common Approach

Key: characterize the **function space** where optimization performs.

Convexity w.r.t the function

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

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

- **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)]

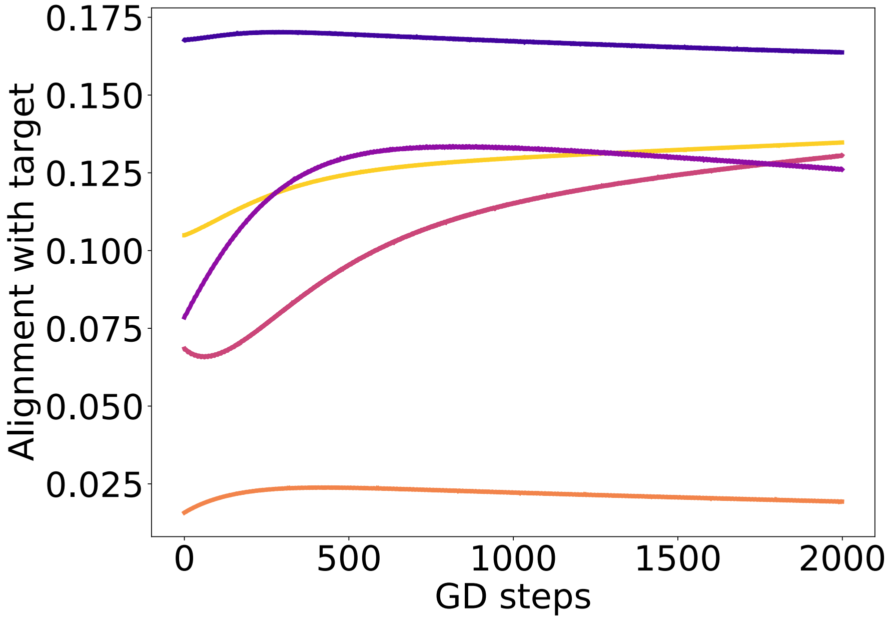
  Coefficient: $1/\sqrt{m}$, learning rate: $O(1)$.

  Function space: reproducing kernel Hilbert space (RKHS) associated with NTK.

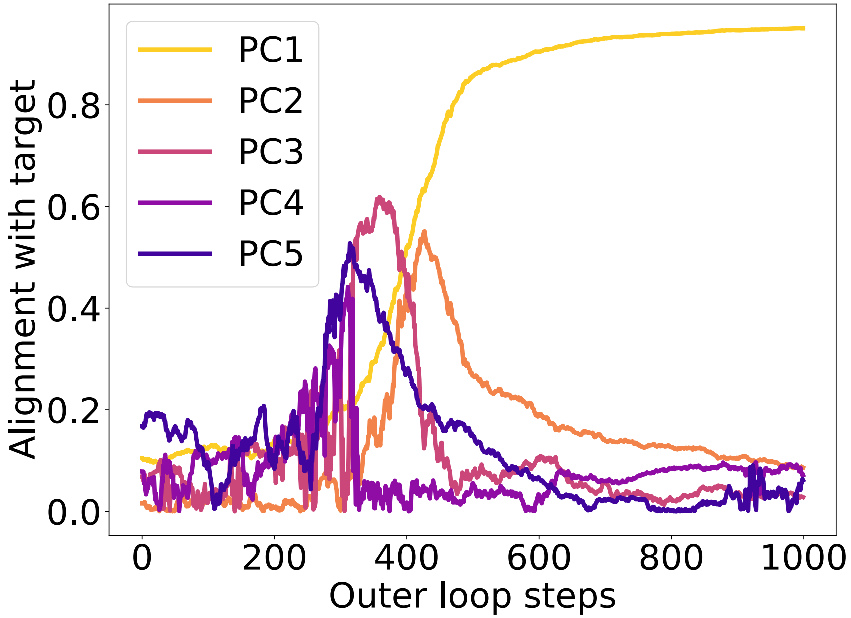
Adaptive Learning Aspect

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

The figure plots the cos similarity between w_* and top-5 singular vectors of the parameter.



NTK-regime



MF-regime

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

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 with ReLU and bounded smooth activations.

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 with ReLU and bounded smooth activations.
- **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)] KL-divergence regularization.
    Under strong regularization, Linear convergence of mean field Langevin.

The most relevant work.

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 with ReLU and bounded smooth activations.
- **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)] KL-divergence regularization.
    Under strong regularization, Linear convergence of mean field Langevin.

  The most relevant work.
- **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 with ReLU and bounded smooth activations.
- **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)] KL-divergence regularization.
    Under strong regularization, Linear convergence of mean field Langevin.

The most relevant work.

- **Convergence rate analysis in the discrete-time setting**
  - [Chizat (2019)], [Akiyama & Suzuki (2021)] Local linear convergence under structural assumption.

Convergence rate analysis is nontrivial and requires an additional assumption or regularization.

Remark: In parallel to our work, [Bou-Rabee and Eberle (2021)] shows a similar result on specific loss functions.

Mean field Models

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

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

Linear w.r.t. q.

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

Mean field Models

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

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

Linear w.r.t. q.

$$\begin{tikzcd}
h_\Theta(x) = \frac{1}{m}\sum_{r=1}^m h(\theta_r, x) \arrow[r, "m \to \infty"] \arrow[d, "\text{Loss}"] & h_q(x) = \int h(\theta, x) q(\theta)\mathrm{d}\theta \arrow[d, "\text{Loss}"] \\
\mathbb{E}\left[l(h_\Theta(X), Y)\right] \arrow[r, "m \to \infty"] & \mathbb{E}\left[l(h_q(X), Y)\right]
\end{tikzcd}$$

Nonconvex w.r.t. Θ. Convex w.r.t. q.

The diagram suggests the optimization in the space of probability measures.

Particle Based Approach [Nitanda & Suzuki (2017)]

Approach: Optimize a distribution via optimization of m-particles $(\theta_r)_{r=1}^m$ (random variables).
Optimization of the distribution is getting accurate as $m \to \infty$.

Particle Based Approach [Nitanda & Suzuki (2017)]

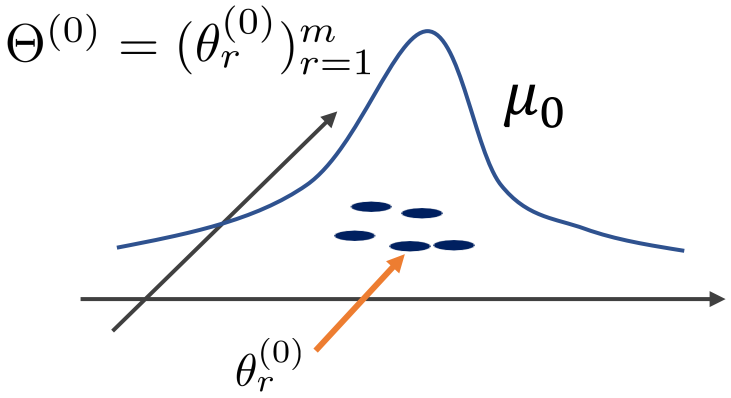
Approach: Optimize a distribution via optimization of m-particles $(\theta_r)_{r=1}^m$ (random variables).
Optimization of the distribution is getting accurate as $m \to \infty$.

Mean field model: $h_\Theta(x) = \frac{1}{m}\sum_{r=1}^m h(\theta_r, x)$, initialization: $\theta_r^{(0)} \sim \mu_0$.

Particle Based Approach [Nitanda & Suzuki (2017)]

Approach: Optimize a distribution via optimization of m-particles $(\theta_r)_{r=1}^m$ (random variables).
Optimization of the distribution is getting accurate as $m \to \infty$.

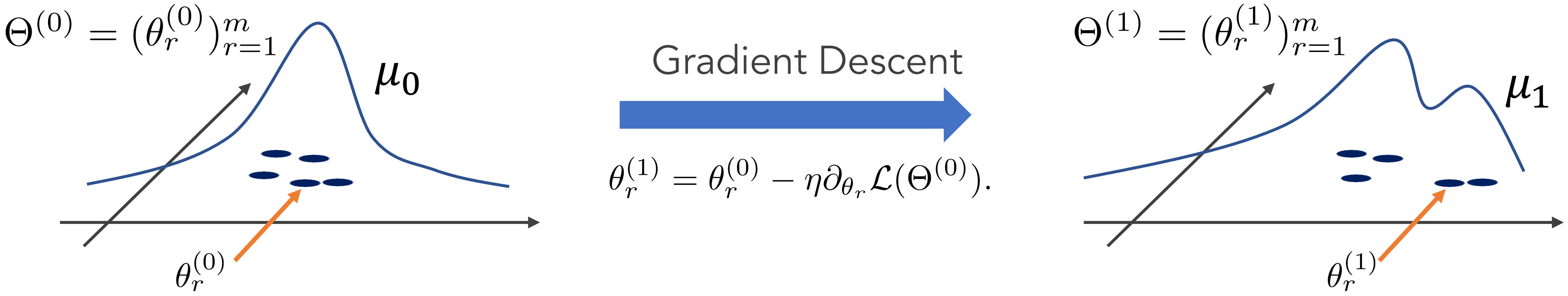
Mean field model: $h_\Theta(x) = \frac{1}{m}\sum_{r=1}^m h(\theta_r, x),$ initialization: $\theta_r^{(0)} \sim \mu_0.$



Particle Based Approach [Nitanda & Suzuki (2017)]

Approach: Optimize a distribution via optimization of m-particles $(\theta_r)_{r=1}^m$ (random variables).
Optimization of the distribution is getting accurate as $m \to \infty$.

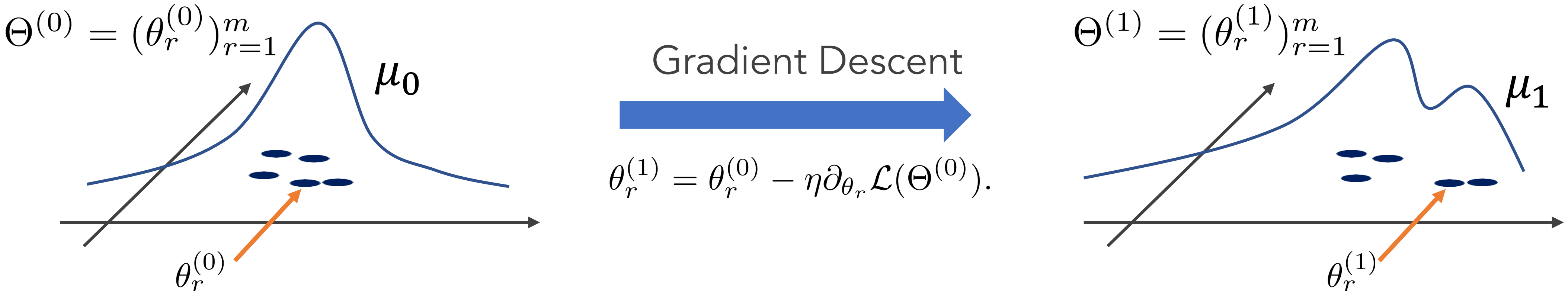
Mean field model: $h_\Theta(x) = \frac{1}{m}\sum_{r=1}^m h(\theta_r, x)$, initialization: $\theta_r^{(0)} \sim \mu_0$.



Particle Based Approach [Nitanda & Suzuki (2017)]

Approach: Optimize a distribution via optimization of m-particles $(\theta_r)_{r=1}^m$ (random variables).
Optimization of the distribution is getting accurate as $m \to \infty$.

Mean field model: $h_\Theta(x) = \frac{1}{m}\sum_{r=1}^m h(\theta_r, x)$,   initialization: $\theta_r^{(0)} \sim \mu_0$.



The update of parameter $\Theta^{(0)} \mapsto \Theta^{(1)}$ implicitly updates its distribution: $\mu^{(0)} \mapsto \mu^{(1)}$
$\to$ **GD on mean field model implicitly optimizes the parameter distribution:** $\min_\mu \mathcal{L}(\mu)$.

Regularized Empirical Risk Minimization

KL-regularized empirical risk minimization over the probability space:

$$\min_{q \in \mathcal{P}_2} \left\{ \frac{1}{n} \sum_{i=1}^{n} l(\mathbb{E}_q[h(\cdot, x_i)], y_i) + \underbrace{\lambda_1 \mathbb{E}_q[\|\theta\|_2^2] + \lambda_2 \mathbb{E}_q[\log(q(\theta))]}_{\text{Kullback-Leibler divergence to zero-mean Gaussian}} \right\}.$$

$\mathcal{P}_2$: the set of smooth positive 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 Empirical Risk Minimization

KL-regularized empirical risk minimization over the probability space:

$$\min_{q \in \mathcal{P}_2} \left\{ \frac{1}{n} \sum_{i=1}^{n} l(\mathbb{E}_q[h(\cdot, x_i)], y_i) + \underline{\lambda_1 \mathbb{E}_q[\|\theta\|_2^2] + \lambda_2 \mathbb{E}_q[\log(q(\theta))]} \right\}.$$

Kullback-Leibler divergence to zero-mean Gaussian

$\mathcal{P}_2$: the set of smooth positive densities with well-defined second moment and entropy.

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

→ Develop new methods with the convergence rate analysis by exploiting the convexity of the loss function w.r.t. the probability density.

→ Quantitative convergence guarantees in discrete-time setting.

PDA Method

- Gradient Descent

$$\theta_r^{(k+1)} = (1 - 2\eta\lambda_1)\theta_r^{(k)} - \frac{\eta}{n}\sum_{i=1}^{n} \partial_z l(g_{\Theta^{(k)}}(x_i), y_i)\partial_\theta h(\theta_r^{(k)}, x_i).$$

PDA Method

- Gradient Descent

$$\theta_r^{(k+1)} = (1 - 2\eta\lambda_1)\theta_r^{(k)} - \frac{\eta}{n}\sum_{i=1}^{n} \partial_z l(g_{\Theta^{(k)}}(x_i), y_i)\partial_\theta h(\theta_r^{(k)}, x_i).$$

Major differences from GD.

- Particle Dual Averaging (a variant of noisy gradient descent)

$$\theta_r^{(k+1)} = \left(1 - \frac{2\eta\lambda_1 t}{\lambda_2(t+2)}\right)\theta_r^{(k)} - \frac{\eta}{n\lambda_2(t+2)(t+1)}\sum_{i=1}^{n} \underline{w_i}\partial_\theta h(\theta_r^{(k)}, x_i) + \underline{\sqrt{2\eta}\zeta_r^{(k)}}.$$

$$(\zeta_r^{(k)} \sim \mathcal{N}(0, I))$$

PDA Method

- Gradient Descent

$$\theta_r^{(k+1)} = (1 - 2\eta\lambda_1)\theta_r^{(k)} - \frac{\eta}{n}\sum_{i=1}^{n} \partial_z l(g_{\Theta^{(k)}}(x_i), y_i)\partial_\theta h(\theta_r^{(k)}, x_i).$$

Major differences from GD.

- Particle Dual Averaging (a variant of noisy gradient descent)

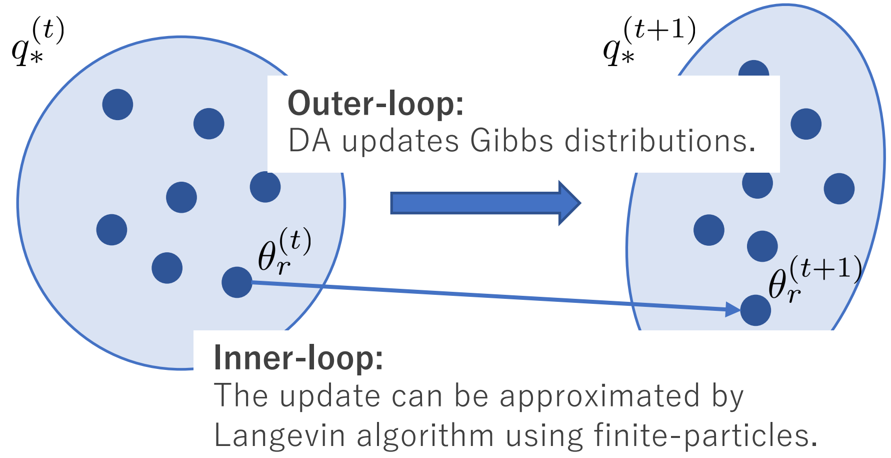
$$\theta_r^{(k+1)} = \left(1 - \frac{2\eta\lambda_1 t}{\lambda_2(t+2)}\right)\theta_r^{(k)} - \frac{\eta}{n\lambda_2(t+2)(t+1)}\sum_{i=1}^{n} w_i \partial_\theta h(\theta_r^{(k)}, x_i) + \sqrt{2\eta}\zeta_r^{(k)}.$$

$$(\zeta_r^{(k)} \sim \mathcal{N}(0, I))$$

Double loops algorithm

(Inner-loop) Run Langevin Monte Carlo to approximate Gibbs distribution $q_*^{(t+1)}$ defined by $\{w_i\}_{i=1}^{n}$.

(Outer-loop) Update $\{w_i\}_{i=1}^{n}$ based on dual averaging method so that Gibbs distributions $\{q_*^{(t)}\}_t$ converges to the solution.



(Remark: PDA can be also applied to expected risk minimization.)

Idea behind Mean field Limit of PDA

- The problem we want to solve is an entropic regularized nonlinear functional:

$$\min_q \left\{ \frac{1}{n} \sum_{i=1}^n \underbrace{l(\mathbb{E}_q[h(\cdot, x_i)], y_i)}_{\text{nonlinear w.r.t. } q} + \lambda_1 \underbrace{\mathbb{E}_q[\|\theta\|_2^2]}_{\text{linear w.r.t. } q} + \lambda_2 \mathbb{E}_q[\log(q(\theta))] \right\}.$$

Idea behind Mean field Limit of PDA

- The problem we want to solve is an entropic regularized nonlinear functional:

$$\min_q \left\{ \frac{1}{n} \sum_{i=1}^n \underbrace{l(\mathbb{E}_q[h(\cdot, x_i)], y_i)}_{\text{nonlinear w.r.t. } q} + \lambda_1 \underbrace{\mathbb{E}_q[\|\theta\|_2^2]}_{\text{linear w.r.t. } q} + \lambda_2 \mathbb{E}_q[\log(q(\theta))] \right\}.$$

- Linearize this based on DA method and obtain an entropic regularized linear functional:

$$\min_q \{ \underbrace{\mathbb{E}_q[f]}_{\text{linear w.r.t. } q} + \mathbb{E}_q[\log(q)] \}.$$

Idea behind Mean field Limit of PDA

- The problem we want to solve is an entropic regularized nonlinear functional:

$$\min_q \left\{ \underbrace{\frac{1}{n}\sum_{i=1}^n l(\mathbb{E}_q[h(\cdot, x_i)], y_i)}_{\text{nonlinear w.r.t. } q} + \lambda_1 \underbrace{\mathbb{E}_q[\|\theta\|_2^2]}_{\text{linear w.r.t. } q} + \lambda_2 \mathbb{E}_q[\log(q(\theta))] \right\}.$$

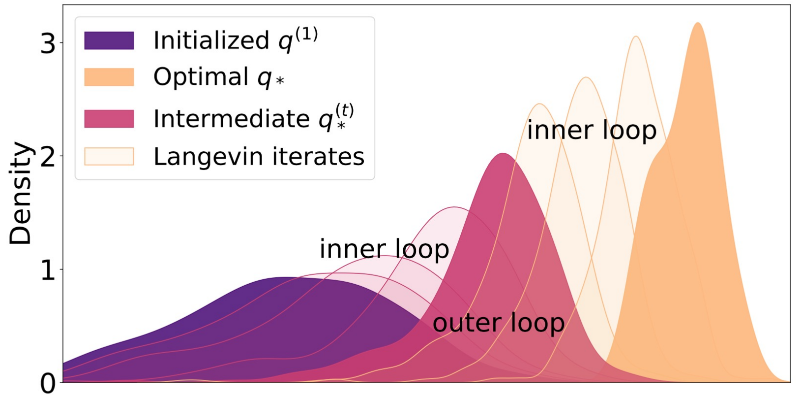
- Linearize this based on DA method and obtain an entropic regularized linear functional:

$$\min_q \{ \underbrace{\mathbb{E}_q[f]}_{\text{linear w.r.t. } q} + \mathbb{E}_q[\log(q)] \}.$$

The minimizer is the Gibbs distribution $\propto \exp(-f)$.

LMC converges to this distribution up to $O(\eta)$-error.

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



Convergence Analysis

Theorem. Under appropriate assumptions:

(Outer loop complexity)

$$\min_{t \in \{1,\ldots,T\}} \mathcal{L}(q^{(t)}) - \mathcal{L}(q^*) = \tilde{O}(1/T).$$ (We ignore λ_1, λ_2 for simplicity)

Convergence Analysis

Theorem. Under appropriate assumptions:

(Outer loop complexity)

$$\min_{t\in\{1,\ldots,T\}} \mathcal{L}(q^{(t)}) - \mathcal{L}(q^*) = \tilde{O}(1/T).$$ (We ignore λ_1, λ_2 for simplicity)

(Inner loop complexity) $k_t = \tilde{O}\left(t^2 \exp(16/\lambda_2)/\lambda_1^2\right)$ iteration is sufficient at t-th outer loop to guarantee the above convergence.

Convergence Analysis

Theorem. Under appropriate assumptions:

(Outer loop complexity)

$$\min_{t\in\{1,\dots,T\}} \mathcal{L}(q^{(t)}) - \mathcal{L}(q^*) = \tilde{O}(1/T).$$

(We ignore λ_1, λ_2 for simplicity)

(Inner loop complexity) $k_t = \tilde{O}\left(t^2 \exp(16/\lambda_2)/\lambda_1^2\right)$ iteration is sufficient at t-th outer loop to guarantee the above convergence.

(Total) To obtain ϵ-accurate solution,

Iteration complexity: $\tilde{O}(\epsilon^{-3})$, Particle complexity: $\tilde{O}(\epsilon^{-2})$.

Remark

- We use restarting scheme to guarantee the particle complexity.
- Inner and total complexities can be reduced by using more efficient sampling than Langevin MC.

Convergence Analysis

Theorem. Under appropriate assumptions:

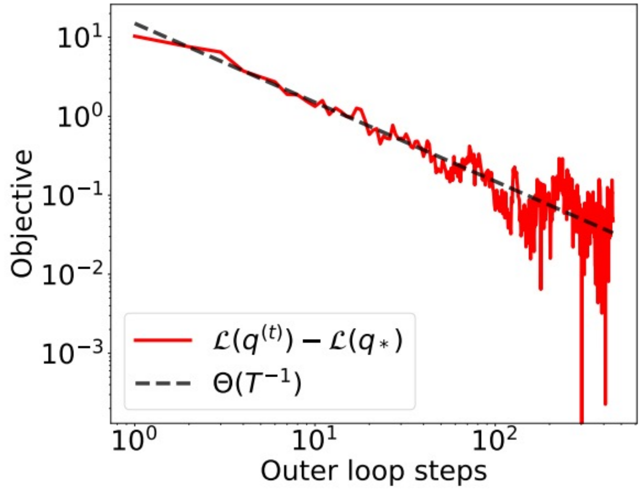
(Outer loop complexity)

$$\min_{t\in\{1,\dots,T\}} \mathcal{L}(q^{(t)}) - \mathcal{L}(q^*) = \tilde{O}(1/T).$$ (We ignore λ_1, λ_2 for simplicity)

(Inner loop complexity) $k_t = \tilde{O}\left(t^2 \exp(16/\lambda_2)/\lambda_1^2\right)$ iteration is sufficient at t-th outer loop to guarantee the above convergence.

(Total) To obtain ϵ-accurate solution,

Iteration complexity: $\tilde{O}(\epsilon^{-3})$, Particle complexity: $\tilde{O}(\epsilon^{-2})$.

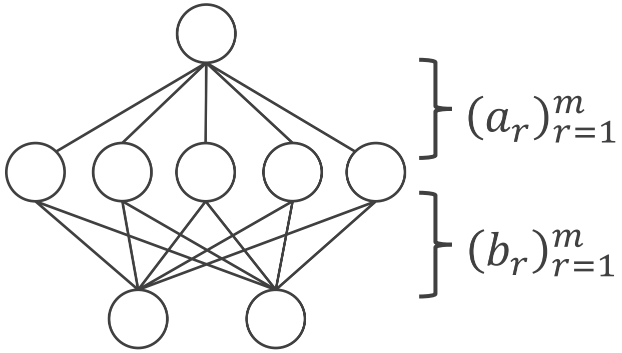


(a) objective value (regression).

Remark

- We use restarting scheme to guarantee the particle complexity.
- Inner and total complexities can be reduced by using more efficient sampling than Langevin MC.

Summary



- We study the optimization of mean field neural networks for KL-regularized problems over the space of distributions.

$$\min_{q \in \mathcal{P}_2} \left\{ \frac{1}{n} \sum_{i=1}^{n} l(\mathbb{E}_q[h(\cdot, x_i)], y_i) + \lambda_1 \mathbb{E}_q[\|\theta\|_2^2] + \lambda_2 \mathbb{E}_q[\log(q(\theta))] \right\}.$$

- Utilizing the convexity, we give the **quantitative convergence guarantees**:

  Iteration complexity: $\tilde{O}(\epsilon^{-3})$, Particle complexity: $\tilde{O}(\epsilon^{-2})$.

Future work:

More efficient optimization methods inspired by finite-dimensional optimization.