Convergence of mean field Langevin dynamics: Duality viewpoint and neural network optimization

Taiji Suzuki

The University of Tokyo / AIP-RIKEN

(Deep learning theory team)

29th/July/2023

Duality Principles for Modern Machine Learning

ICML2023 workshop@Hawaii

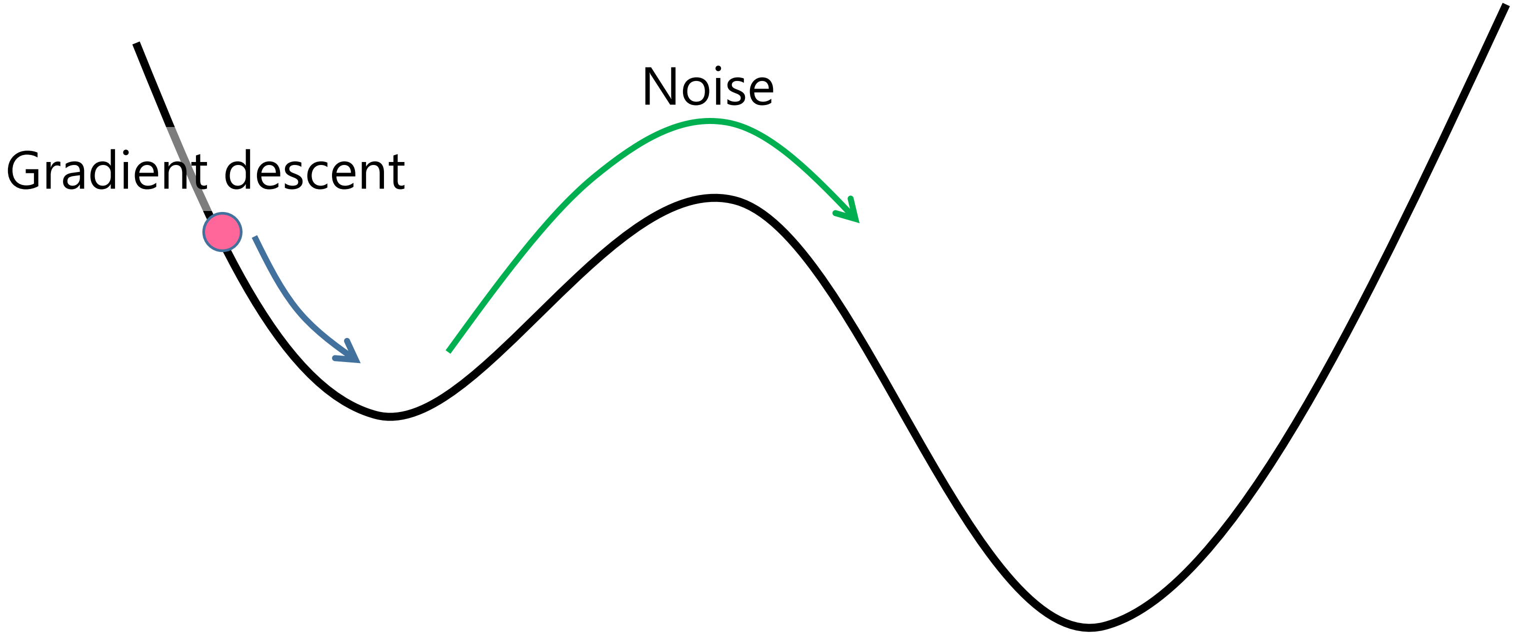
Atsushi Nitanda
(Kyusyu Institute of Technology/
RIKEN-AIP)

Denny Wu
(University of Toronto/
Vector Institute)

Noisy gradient descent

Optimization of neural network is basically non-convex.

- ➢ Noisy gradient descent (e.g, SGD) is effective for non-convex optimization.



Noisy perturbation is helpful to escape a local minimum.

- ➢ Likely converges to a flat global minimum.

Gradient Langevin Dynamics (GLD)

$$\min_{x\in\mathbb{R}^d} L(x) = \min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell_i(x)$$

(Non-convex)

β: inverse temperature

$$\mathrm{d}X_t = -\nabla L(X_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t$$

(Gradient Langevin dynamics)

Discretization

[Gelfand and Mitter (1991); Borkar and Mitter (1999); Welling and Teh (2011)]

(Euler-Maruyama scheme)

$$X_{t+1} = X_t - \eta\nabla L(X_t) + \sqrt{2\eta\beta^{-1}}\xi_t \qquad \xi_t \sim \mathrm{N}(0, I)$$

Stationary distribution : $\mu^* \propto \exp(-\beta L(X))$

Can stay around the global minimum of $L(x)$.

GLD as a Wasserstein gradient flow

$$\mathrm{d}X_t = -\nabla L(X_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t$$

μ_t : Distribution of X_t (we can assume it has a density)

PDE that describes μ_t's dynamics [Fokker-Planck equation]:

$$\begin{aligned}\partial_t \mu_t &= \nabla \cdot [\mu_t \nabla L] + \frac{1}{\beta}\Delta_x \mu_t \\ &= \nabla \cdot \left[\mu_t \left(\nabla L + \frac{1}{\beta}\nabla \log(\mu_t)\right)\right]\end{aligned}$$

This is the Wasserstein gradient flow to minimize the following objective:

$$\mu^* = \mathop{\arg\min}_{\mu \in \mathcal{P}} \underbrace{\int L(x)\mathrm{d}\mu(x)}_{} + \frac{1}{\beta}\mathrm{Ent}(\mu) =: \mathcal{L}(\mu)$$

[linear w.r.t. μ]

$(\mathrm{Ent}(\mu) = \int \log(\mu)\mathrm{d}\mu)$

➡ $\mu_t \rightsquigarrow \mu^*(x) \propto \exp(-\beta L(x))$ = Stationary distribution

c.f., Donsker-Varadan duality formula

Objective of mean field NN

Vanilla GLD

$$\mathcal{L}(\mu) = \underbrace{\int L(x)\mathrm{d}\mu(x)}_{\text{Linear}} + \lambda_2 \mathrm{Ent}(\mu)$$

Nonlinear extension!

Convex objective

$$\boxed{\mathcal{L}(\mu) := \underbrace{F(\mu)}_{\text{convex}} + \lambda_2 \mathrm{Ent}(\mu)}$$

convex + strongly convex = strongly convex

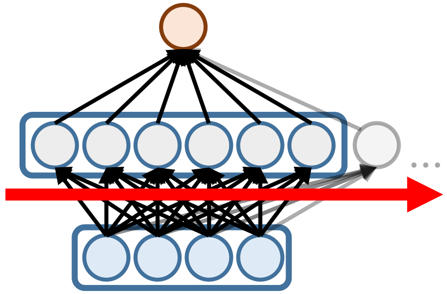
$$(\mathrm{Ent}(\mu) = \textstyle\int \log(\mu)\mathrm{d}\mu)$$

Application:

- Optimization of 2-layer neural network in mean field regime
- Variational inference

Example of loss function

$$f(z) = \frac{1}{M}\sum_{j=1}^{M} r_j \sigma(w_j^\top z)$$



Non-linear with respect to the parameters $(r_j, w_j)_{j=1}^{M}$.

★Mean field limit:

$$f(z) = \frac{1}{M}\sum_{j=1}^{M} r_j \sigma(w_j^\top z) \xrightarrow{M \to \infty} \boxed{f_\mu(z) = \int r\sigma(w^\top z)\mathrm{d}\mu(r, w)}$$

Linear with respect to μ.

[Nitanda&Suzuki, 2017][Chizat&Bach, 2018][Mei, Montanari&Nguyen, 2018][Rotskoff&Vanden-Eijnden, 2018]

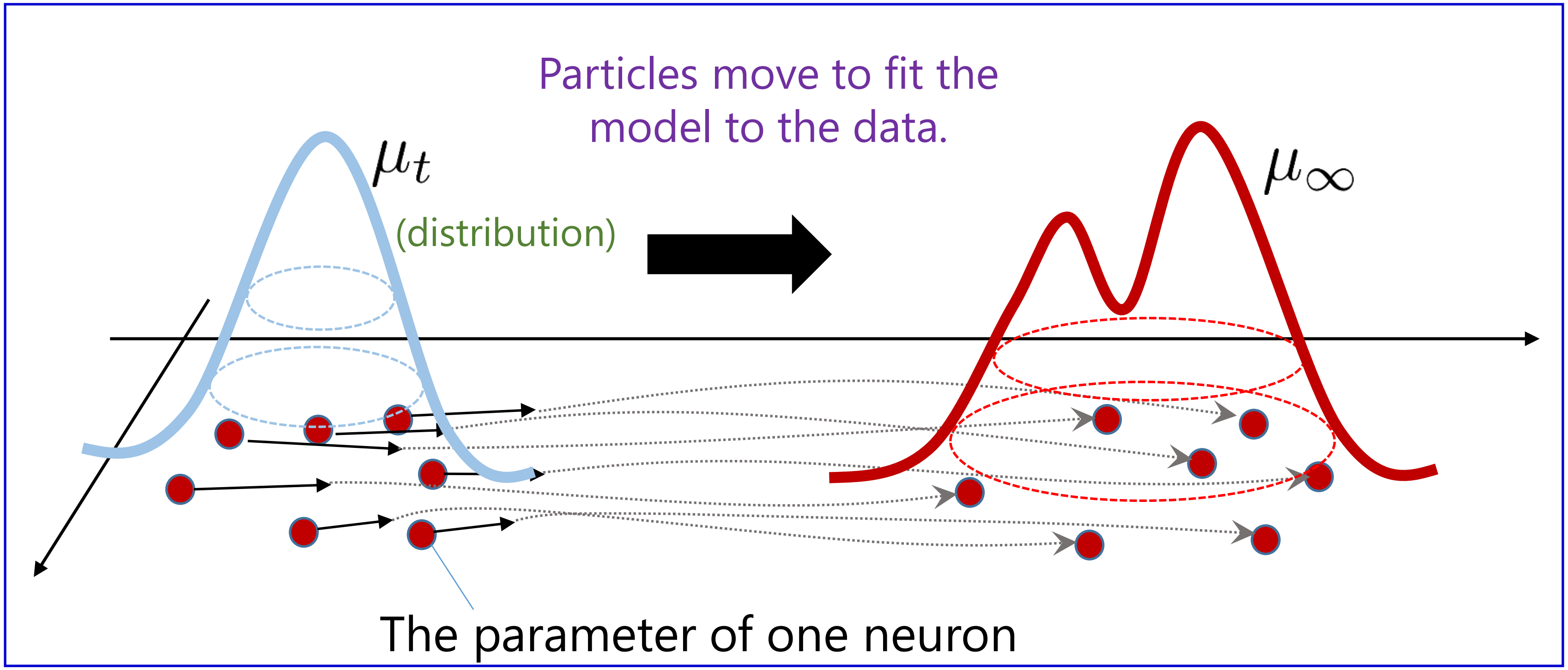
Loss function (empirical risk + regularization):

$$F(\mu) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(f_\mu(z_i)) + \lambda_1 \mathbb{E}_\mu[\|x\|^2]$$

Convex w.r.t. μ if the loss ℓ_i is convex (e.g., squared / logistic loss).

$$F(\mu) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(f_\mu(z_i)) + \lambda_1 \mathbb{E}_\mu[\|x\|^2]$$

$$f_{\color{red}\mu}(z) = \int r\sigma(w^\top z)\mathrm{d}{\color{red}\mu(r,w)}$$


Particles move to fit the model to the data.

The parameter of one neuron

General form of mean field LD

Mean field Langevin dynamics:

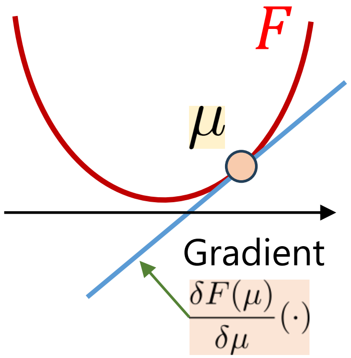
$$\mathcal{L}(\mu) = F(\mu) + \lambda_2 \mathrm{Ent}(\mu)$$

convex

$$(\mathrm{Ent}(\mu) = \int \log(\mu) \mathrm{d}\mu)$$

- SDE the Fokker-Planck equation of which corresponds to the Wasserstein GF:

$$\mathrm{d}X_t = -\nabla \frac{\delta F(\mu_t)}{\delta \mu}(X_t)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t$$
$$\mu_t = \mathrm{Law}(X_t)$$



Distribution dependent SDE

GLD: $\mathrm{d}X_t = -\nabla L(X_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t$, $\frac{\delta F(\mu)}{\delta \mu}(\cdot) = L(\cdot)$
$F(\mu) = \int L(x)\mathrm{d}\mu$

Definition (first variation)

The first variation $\frac{\delta F}{\delta \mu}: \mathcal{P} \times \mathbb{R}^d \to \mathbb{R}$ is defined as a continuous functional such as

$$\lim_{\epsilon \to 0} \frac{F(\epsilon\nu + (1-\epsilon)\mu) - F(\mu)}{\epsilon} = \int \frac{\delta F(\mu)}{\delta \mu}(x)\mathrm{d}(\nu - \mu)(x)$$

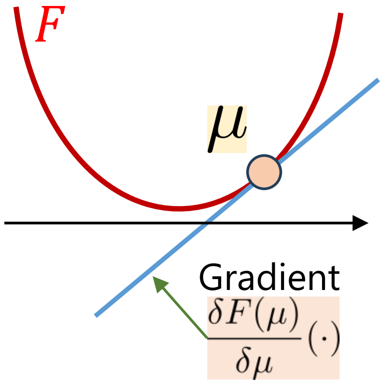
Proximal Gibbs measure

$$\mathcal{L}(\mu) = \underline{F(\mu)} + \lambda_2 \mathrm{Ent}(\mu)$$

$$(\mathrm{Ent}(\mu) = \int \log(\mu) \mathrm{d}\mu)$$

Linearized objective at μ:

$$\bar{\mathcal{L}}_\mu(\nu) = \underline{\int \frac{\delta F(\mu)}{\delta \mu}(x) \mathrm{d}\nu(x)} + \lambda_2 \mathrm{Ent}(\nu)$$



Minimizer

$$\boxed{p_\mu(x) \propto \exp\left(-\frac{1}{\lambda_2}\frac{\delta F(\mu)}{\delta \mu}(x)\right)}$$

$$F(\mu) = \int L(x) \mathrm{d}\mu$$
$$\Rightarrow p_\mu \propto \exp(-\lambda_2^{-1} L(x))$$

Proximal Gibbs measure

- The proximal Gibbs measure is a kind of “tentative” target.
- It plays important role in the convergence analysis.

Dual objective (informal)

[Nitanda, Oko, Wu, Suzuki (ICML2023); Nitanda, Wu, Suzuki (AISTATS2022); Oko, Suzuki, Nitanda, Wu (ICLR2022)]

$$(\mathrm{Ent}(\mu) = \int \log(\mu)\mathrm{d}\mu)$$

Primal $$\min_{\mu\in\mathcal{P}} \mathcal{L}(\mu) = F(\mu) + \lambda_2 \mathrm{Ent}(\mu)$$

||

$$\min_{x\in\mathcal{X}} f(Ax) + g(x) = -\min_{g\in\mathcal{Y}^*} f^*(g) + g^*(-A^*g)$$

(Fenchel's duality theorem)

$A : \mathcal{X} \to \mathcal{Y}$ (bounded linear)

Dual $$\max_{g:\mathbb{R}^d\to\mathbb{R}} \mathcal{D}(g) = -F^*(g) - \lambda_2 \log\left(\int \exp\left(-\frac{g(x)}{\lambda_2}\right)\mathrm{d}x\right)$$

$$F^*(g) := \sup_{\mu\in\mathcal{P}}\{\int g(x)\mathrm{d}\mu(x) - F(\mu)\}$$

Primal-Dual variable correspondence:

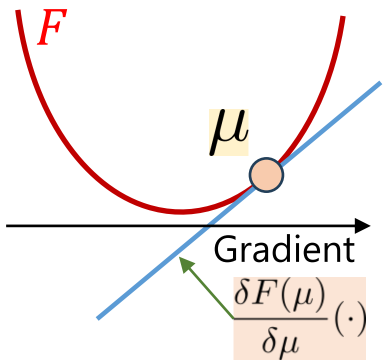
$$\overset{(P)}{\mu} \longrightarrow \overset{(D)}{g_\mu(x) := \frac{\delta F(\mu)}{\delta\mu}(x)} \longrightarrow \overset{(P)}{p_\mu(x) \propto \exp\left(-\frac{1}{\lambda_2}\frac{\delta F(\mu)}{\delta\mu}(x)\right)}$$

Duality gap and divergence:

$\mu^* = \arg\min_{\mu\in\mathcal{P}} \mathcal{L}(\mu)$

- $\mathcal{L}(\mu) - \mathcal{D}(g_\mu) = \lambda_2 \mathrm{KL}(\mu || p_\mu) \geq 0$
- $\mathcal{L}(\mu^*) = \mathcal{D}(g_{\mu^*}) \Rightarrow \mu^* = p_{\mu^*}$

(optimality condition)



Entropy sandwich

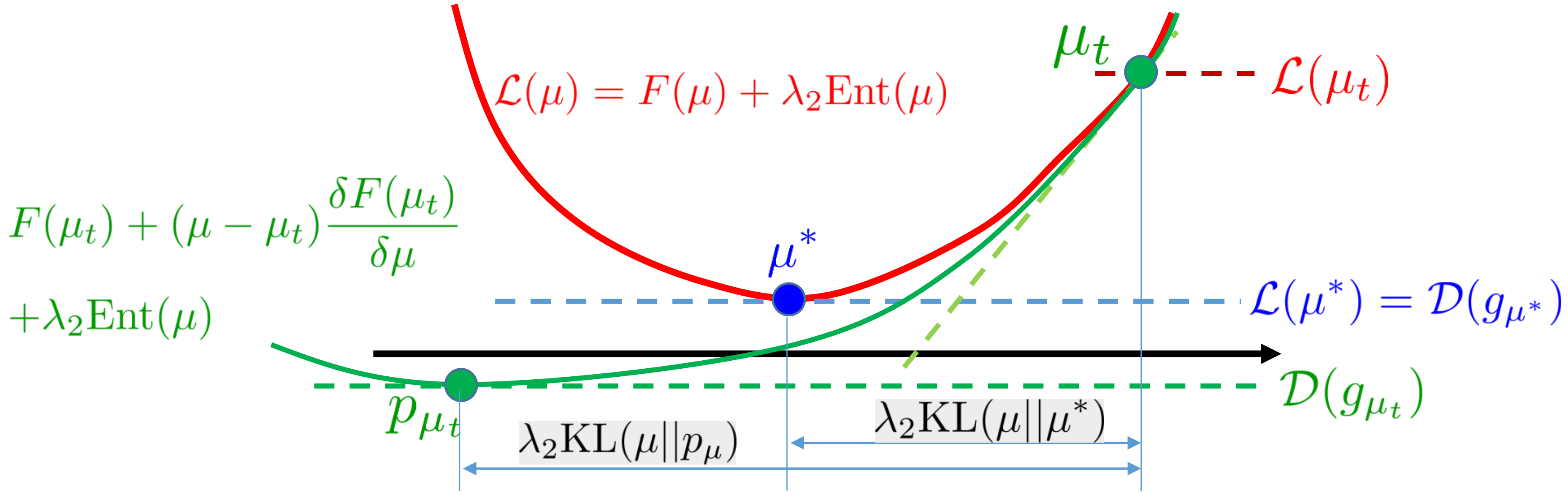
Proximal Gibbs measure:

$$p_\mu(x) \propto \exp\left(-\frac{1}{\lambda_2}\frac{\delta F(\mu)}{\delta\mu}(x)\right) \qquad p_\mu = \arg\min_{\nu\in\mathcal{P}}(\nu-\mu)\frac{\delta F(\mu)}{\delta\mu} + \lambda_2\mathrm{Ent}(\nu)$$

Theorem (Entropy sandwich) [Nitanda, Wu, Suzuki (AISTATS2022)][Chizat (2022)]

$$\mu^* = \arg\min_{\mu\in\mathcal{P}}\mathcal{L}(\mu)$$

$$\lambda_2\mathrm{KL}(\mu\|\mu^*) = \mathcal{L}(\mu) - \underset{\overset{\|}{\mathcal{D}(g_{\mu^*})}}{\underline{\mathcal{L}(\mu^*)}} \le \mathcal{L}(\mu) - \mathcal{D}(g_\mu) = \lambda_2\mathrm{KL}(\mu\|p_\mu)$$



Convergence rate

Proximal Gibbs measure:

$$p_\mu(x) \propto \exp\left(-\frac{1}{\lambda_2}\frac{\delta F(\mu)}{\delta\mu}(x)\right) \qquad p_\mu = \mathop{\arg\min}_{\nu\in\mathcal{P}} (\nu-\mu)\frac{\delta F(\mu)}{\delta\mu} + \lambda_2 \mathrm{Ent}(\nu)$$

Assumption (Log-Sobolev inequality)

c.f., Polyak-Lojasiewicz condition

There exists $\alpha > 0$ such that for any probability measure ν (abs. cont. w.r.t. p_μ),

$$\mathrm{KL}(\nu||p_\mu) \leq \frac{1}{2\alpha} I(\nu||p_\mu)$$

KL-div

$$\mathrm{KL}(\nu||\mu) = \int \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)\mathrm{d}\nu$$

Fisher-div

$$I(\nu||\mu) = \int \left\|\nabla \log\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right\|^2 \mathrm{d}\nu$$

Theorem (Linear convergence) [Nitanda, Wu, Suzuki (AISTATS2022)][Chizat (2022)]

If p_{μ_t} satisfies the LSI condition for any $t \geq 0$, then

$$\mathcal{L}(\mu_t) - \mathcal{L}(\mu^*) \leq \exp(-2\alpha\lambda_2 t)(\mathcal{L}(\mu_0) - \mathcal{L}(\mu^*))$$

This is a non-linear extension of well known GLD convergence analysis.

Example

L2-regularized loss function for mean field 2-layer NN:

$$f_\mu(z) = \int h_x(z)\mathrm{d}\mu(x) \quad \text{where} \quad h_x(z) = r\sigma(w^\top z) \ \text{for}\ x = (r, w)$$

$$F(\mu) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(f_\mu(z_i)) + \lambda_1 \mathbb{E}_\mu[\|X\|^2]$$

Proximal Gibbs: $p_\mu(x) \propto \exp\left(-\frac{1}{\lambda_2}\frac{\delta F(\mu)}{\delta\mu}(x)\right)$

$$= \exp\left[-\frac{1}{\lambda_2}\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n}\ell_i'(f_\mu(z_i))h_x(z_i)}_{\text{Bounded } (\leq B)} + \underbrace{\lambda_1\|x\|^2}_{\text{Strongly convex}}\right)\right]$$

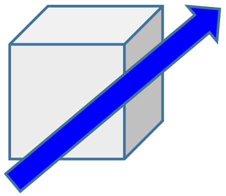
If $\sup_z \left|\ell_i'\big(f_\mu(\cdot)\big)h_x(\cdot)\right| \leq B$, the proximal Gibbs measure p_μ satisfies the LSI with a constant α with

$$\alpha \geq \frac{2\lambda_1}{\lambda_2}\exp(-4B/\lambda_2)$$

∵ **Bakry-Emery criterion** (1085) and **Holley-Strook small perturbation lemma** (1987)

Proof outline of convergence

- MF-LD obeys the following nonlinear Fokker-Planck equation:



$$\partial_t \mu_t = \lambda_2 \Delta_x \mu_t + \nabla \cdot [\mu_t \nabla \frac{\delta F(\mu_t)}{\delta \mu}]$$

$$(= \nabla \cdot \Big[\underbrace{(\lambda_2 \nabla \log(\mu_t) + \nabla \frac{\delta F(\mu_t)}{\delta \mu})}_{=:\ -v_t} \mu_t \Big])$$

$$= -\nabla \cdot [v_t \mu_t]$$

[Continuity equation]

Then,

$$\mathcal{L}(\mu) := F(\mu) + \lambda_2 \mathrm{Ent}(\mu)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(\mu_t) = \int \left\langle v_t, \nabla \frac{\delta \mathcal{L}(\mu_t)}{\delta \mu} \right\rangle \mathrm{d}\mu_t$$

(∵continuity equation)

(Definition of p_{μ_t})

$$p_\mu(x) \propto \exp\left(-\frac{1}{\lambda_2}\frac{\delta F(\mu)}{\delta \mu}(x)\right)$$

$$= \int \langle v_t, \nabla \frac{\delta F(\mu_t)}{\delta \mu} + \lambda_2 \nabla \log(\mu_t) \rangle \mathrm{d}\mu_t$$

$$= -\int \|v_t\|^2 \mathrm{d}\mu_t = -\lambda_2^2 I(\mu_t || p_{\mu_t})$$

LSI & Entropy sandwich

$$\leq -2\alpha\lambda_2^2 \mathrm{KL}(\mu_t || p_{\mu_t}) \leq -2\alpha\lambda_2(\mathcal{L}(\mu_t) - \mathcal{L}(\mu^*))$$

※ Since $\frac{\delta F(\mu_t)}{\delta \mu}$ nonlinearly depends on μ_t, we say “nonlinear Fokker-Planck”.

$$\text{GLD: } F(\mu) = \int L(x)\mathrm{d}\mu \Rightarrow \frac{\delta F(\mu)}{\delta \mu}(\cdot) = L(\cdot)$$

Other applications

Mean field Langevin dynamics can be applied to several problems where a distribution is optimized.

- **Nonparametric density estimation via MMD minimization**

$$F(\mu) = \mathrm{MMD}^2(g * \mu, \hat{\mu}_n) + \lambda_1 \mathbb{E}_\mu[\|x\|^2]$$

k: positive definite kernel

$$\mathrm{MMD}^2(\nu_1, \nu_2) := \|k_{\nu_1} - k_{\nu_2}\|_{\mathcal{H}_k}^2$$

where $k_\mu = \int k(x,\cdot)\mu(\mathrm{d}x)$ (kernel embedding).

- ➢ $g(x) = \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)$
- ➢ $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$: Empirical distribution (training data)

(see also Chizat (2022,TMLR))

- **Variational inference to approximate Bayesian posterior**

$$F(\mu) = \mathrm{KSD}(\mu) + \lambda_1 \mathbb{E}_\mu[\|x\|^2]$$

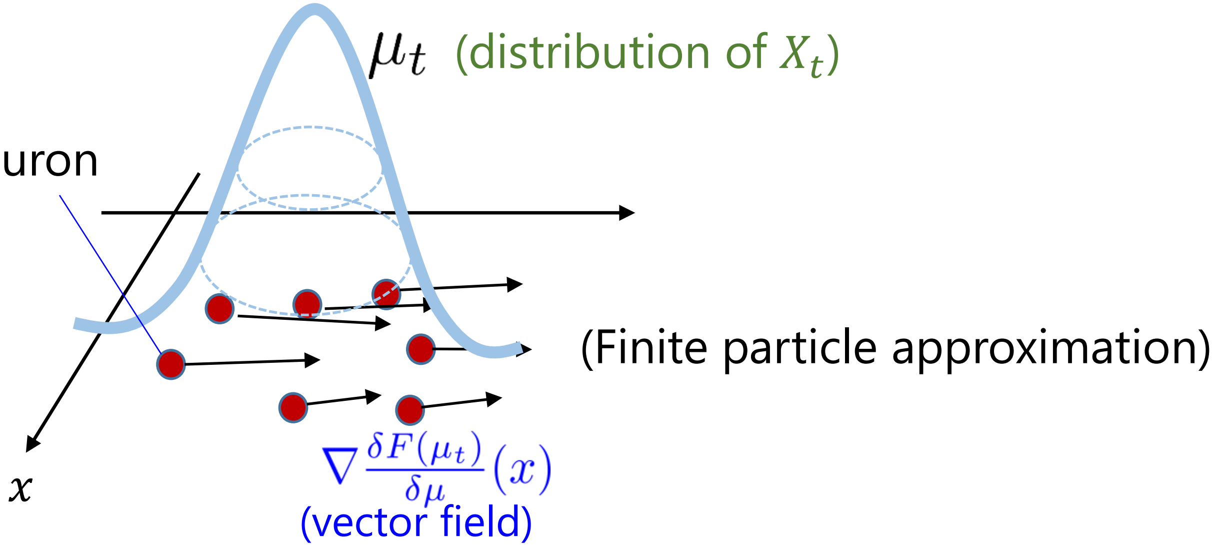
(KSD: Kernel Stein Discrepancy from a posterior distribution)

Finite particle & discrete time algorithm

We have obtained a convergence of infinite width and continuous time dynamics.

Question:

Can we obtain a finite particle & discrete time (i.e., implementable) algorithm?



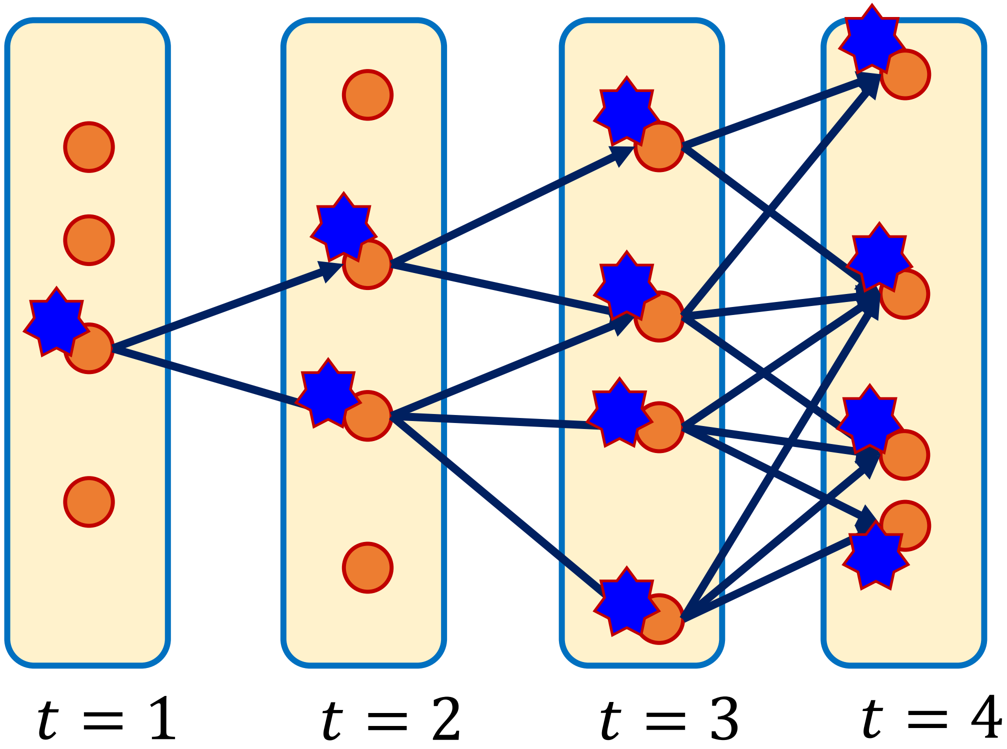
Difficulty

- SDE of interacting particles (McKean, Kac,…, 60’)

Propagation of chaos [Sznitman, 1991; Lacker, 2021]:

The particles behave as if they are independent
as the number of particles increases to infinity.

Finite particle approximation error can be amplified through time.
→ It is difficult to bound the perturbation uniformly over time.



- A naïve evaluation gives exponential growth on time:

$$\exp(t)\,/N$$

[Mei et al. (2018, Theorem 3)]

➢ Weak interaction/Strong regularization in existing work

Outline of research

Infinite particles / Continuous time

Linear convergence of mean field Langevin:
[Nitanda, Wu, Suzuki (AISTATS2022)]
[Chizat (TMLR2022)]

Finite particle / Discrete time

Double loop method:

- PDA [Nitanda, Wu, Suzuki: NeurIPS2021]
- P-SDCA [Oko, Suzuki, Wu, Nitanda: ICLR2022]
- Infinite-dim extension [Nishikawa, Suzuki, Nitanda: NeurIPS2022]

Difficult：
Propagation of chaos (McKean, Kac,…, 60’s)

Finite particle / Continuous time

Uniform-in-time propagation of chaos:

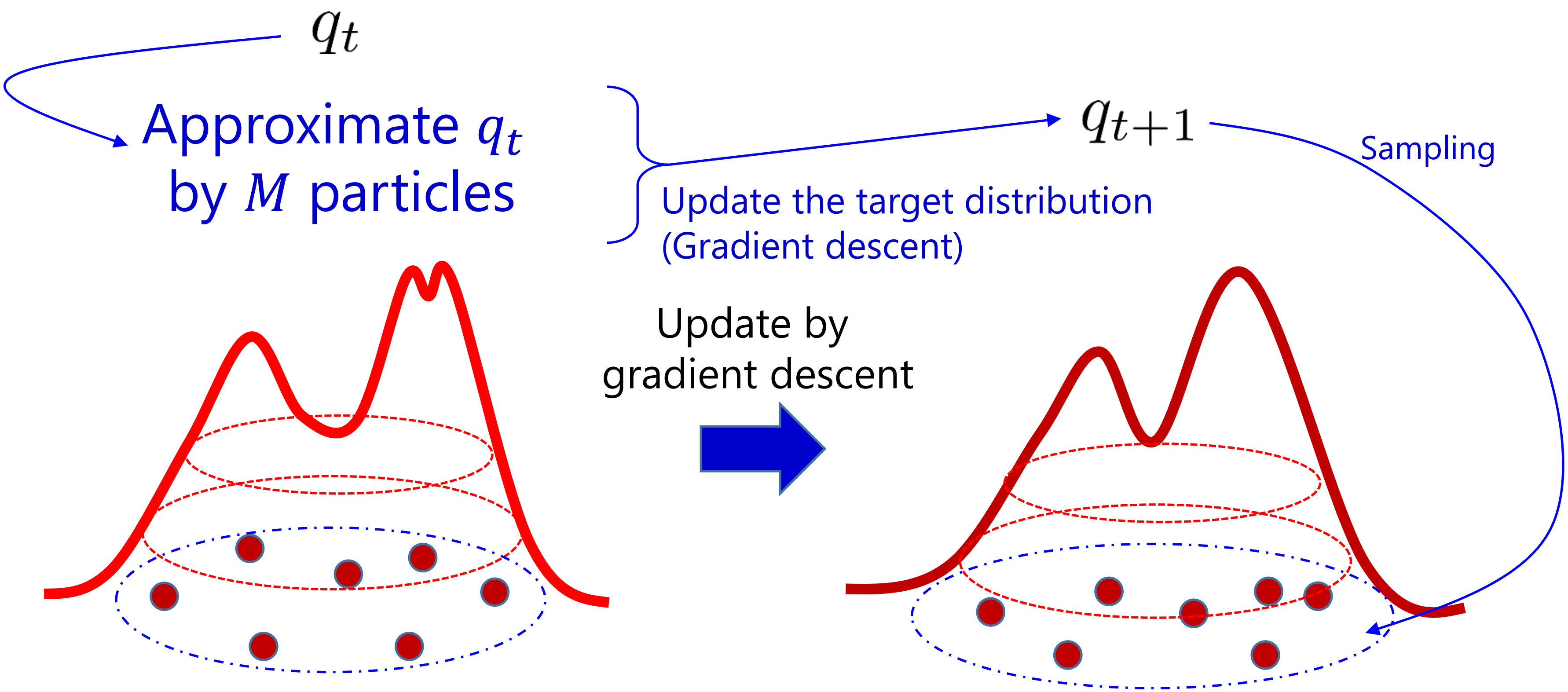
- Super log-Sobolev ineq.
  [Suzuki, Nitanda, Wu (ICLR2023)]
- Leave-one-out type evaluation/Uniform-log-Sobolev
  [Chen, Ren, Wang (arXiv2022)]

Finite particle / Discrete time

Single loop method:
Time-space discretization, stochastic gradient
[Suzuki, Wu, Nitanda (arXiv:2306.07221)]

(1) Double loop algorithm

Apply convex optimization techniques developed in finite dimensional settings.



q_t

Approximate q_t
by M particles

q_{t+1}

Sampling

Update the target distribution
(Gradient descent)

Update by
gradient descent

Double loop algorithms

PDA:
Particle Dual Averaging

[Nitanda, Wu, Suzuki: NeurIPS2021]

$$\min_{q:\text{prob.density}} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\ell\left(\mathbb{E}_q[h_\theta(x_i)], y_i\right) + \lambda_1\mathbb{E}_q[\|\theta\|^2]}\, + \lambda_2\mathbb{E}_q[\log(q)]$$

Approximate this by a linear functional of q.

Approx

$\mathbb{E}_{\theta\sim q}[\bar{g}^{(t)}(\theta)]$ (linear approx)

$\bar{g}^{(t)}$ is determined by <u>dual averaging method</u>.

$$\min_{q:\text{prob.density}} \mathbb{E}_{\theta\sim q}[\bar{g}^{(t)}(\theta)] + \lambda_2\mathbb{E}_q[\log(q)]$$

Solution: $q^{(t+1)}(\theta) \propto \exp(-\bar{g}^{(t)}(\theta)/\lambda_2)$

→ We can sample particles from this distribution by using the following GLD.

$$\mathrm{d}\theta_t = -\nabla(\bar{g}^{(t)}(\theta)/\lambda_2)\mathrm{d}t + \sqrt{2}\mathrm{d}\xi_t.$$

Time discretization

$$\theta_k = \theta_{k-1} - \eta\nabla\bar{g}^{(t)}(\theta)/\lambda_2 + \sqrt{2\eta}\xi_{k-1}$$

Computational complexity :

1. **Inner loop:** $\mathcal{L}(\hat{q}^{(t)}) - \mathcal{L}(q^*) \le O(1/t)$
2. **Outer loop:** $T_t = \tilde{O}\left(t^2\exp(8/\lambda_2)/(\lambda_1\lambda_2)\right)$ (by GLD)

⇒Total: $O(\epsilon^{-3})$ times gradient update

➢ **<u>The first polynomial time method</u>**

P-SDCA:
Particle Stochastic Dual Coordinate Ascent

[Oko, Suzuki, Wu, Nitanda: ICLR2022]

Primal

$$\min_p P(p) = \frac{1}{n}\sum_{i=1}^{n}\ell_i\left(\int p(\theta)h_i(\theta)\right) + \lambda_1\int\|\theta\|^2 p(\theta)\mathrm{d}\theta + \lambda_2\int p(\theta)\log(p(\theta))\mathrm{d}\theta$$

|| by Fenchel duality theorem

Dual

$$\ell_i^*(g) := \sup_{u\in\mathbb{R}}\{ug - \ell_i(u)\}$$

$$-\min_{g\in\mathbb{R}^n} D(g) = \frac{1}{n}\sum_{i=1}^{n}\ell_i^*(g_i) + \lambda_2\log\left(\int q[g](\theta)\mathrm{d}\theta\right)$$

where $q[g](\theta) := \exp\left\{-\frac{1}{\lambda_2}\left(\frac{1}{n}\sum_{i=1}^{n}h_i(\theta)g_i + \lambda_1\|\theta\|^2\right)\right\}$

- Randomly choose a coordinate of the dual variable and optimize the selected coordinate.

→ stochastic coordinate ascent

Computational complexity :

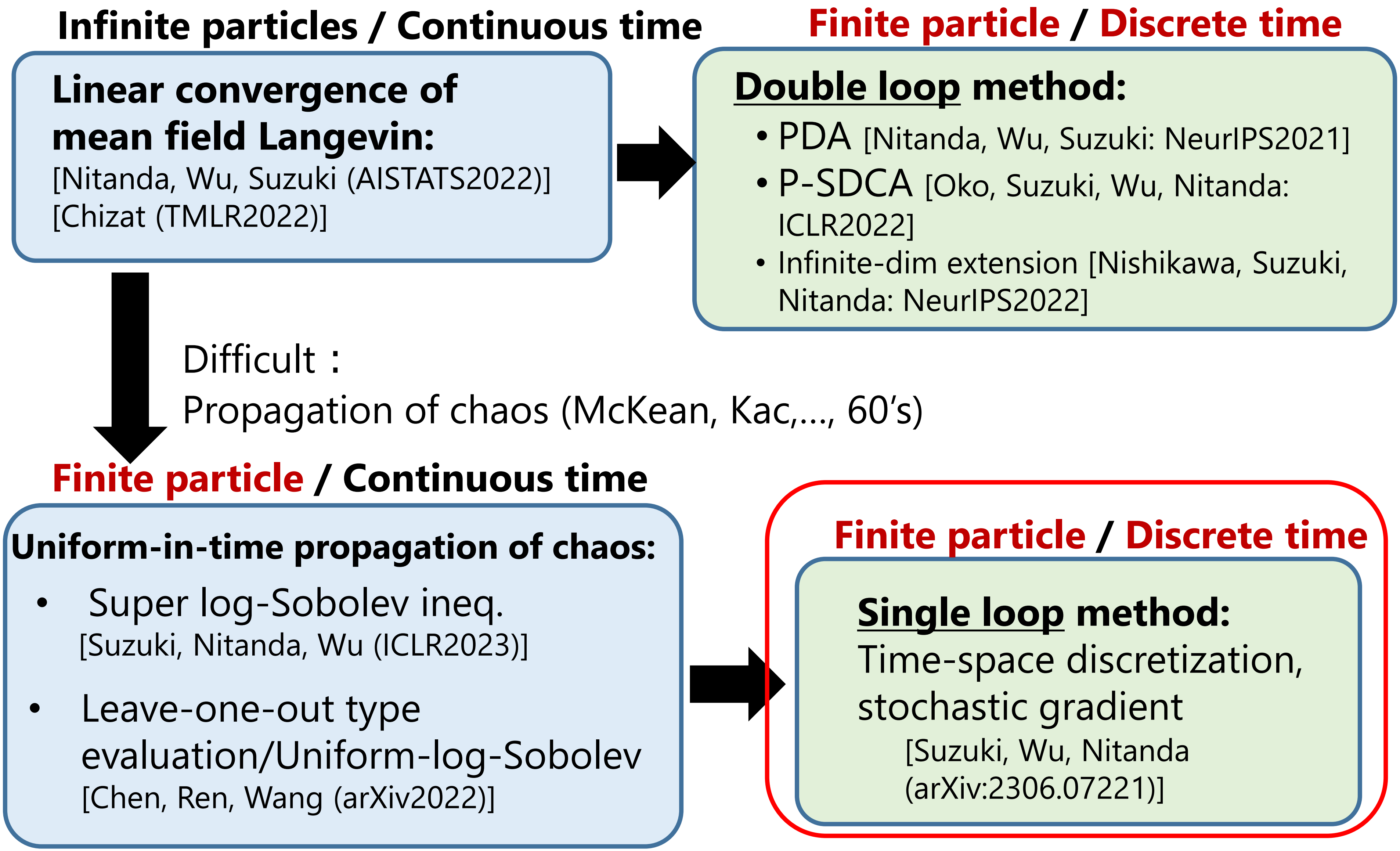
# of outer loops to obtain the duality gap ϵ_P :

$$t_{\text{end}} = 2\left(n + \frac{1}{\lambda_2\gamma}\right)\log\left(\frac{nC}{\epsilon_P}\right)$$

➢ <u>Exponential order convergence</u>

➢ <u>Relax the dependency on sample size</u>

Outline of research



Infinite particles / Continuous time

Linear convergence of mean field Langevin:
[Nitanda, Wu, Suzuki (AISTATS2022)]
[Chizat (TMLR2022)]

Finite particle / Discrete time

Double loop method:

- PDA [Nitanda, Wu, Suzuki: NeurIPS2021]
- P-SDCA [Oko, Suzuki, Wu, Nitanda: ICLR2022]
- Infinite-dim extension [Nishikawa, Suzuki, Nitanda: NeurIPS2022]

Difficult：
Propagation of chaos (McKean, Kac,…, 60’s)

Finite particle / Continuous time

Uniform-in-time propagation of chaos:

- Super log-Sobolev ineq.
  [Suzuki, Nitanda, Wu (ICLR2023)]
- Leave-one-out type evaluation/Uniform-log-Sobolev
  [Chen, Ren, Wang (arXiv2022)]

Finite particle / Discrete time

Single loop method:
Time-space discretization, stochastic gradient
[Suzuki, Wu, Nitanda (arXiv:2306.07221)]

[Suzuki, Wu, Nitanda: Convergence of mean-field Langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction. arXiv:2306.07221]

(2) Single loop method

$$\mathrm{d}X_t = -\nabla \frac{\delta F(\mu_t)}{\delta \mu}(X_t)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t$$

N particles $\left(X_k^{(i)}\right)_{i=1}^{N}$

(time discretization)

$$X_{k+1}^{(i)} = X_k^{(i)} - \eta_k v_k^i + \sqrt{2\eta_k\lambda_2}\xi_k^{(i)}$$

where $\mathbb{E}[v_k^i] = \nabla \frac{\delta F(\hat{\mu}_k)}{\delta \mu}(X_k^i)$ and $\hat{\mu}_k = \frac{1}{N}\sum_{i=1}^{N}\delta_{X_k^{(i)}}$

(stochastic gradient) (space discretization)

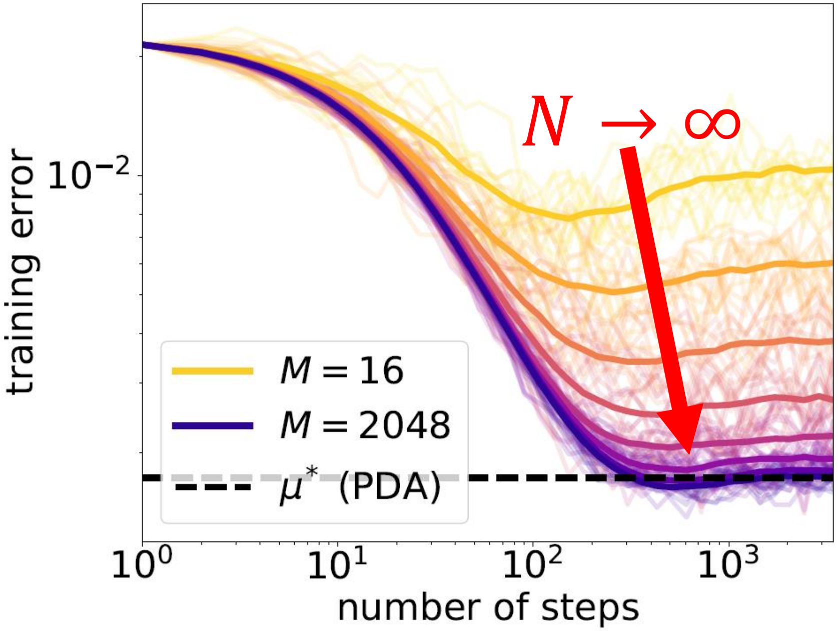
➢ Noisy gradient descent on 2-layer NN with finite width.

- **Time discretization:** $t \to k\eta$ (η: step size, k: # of steps)
- **Space discretization:** μ_t is approximatd by N particles

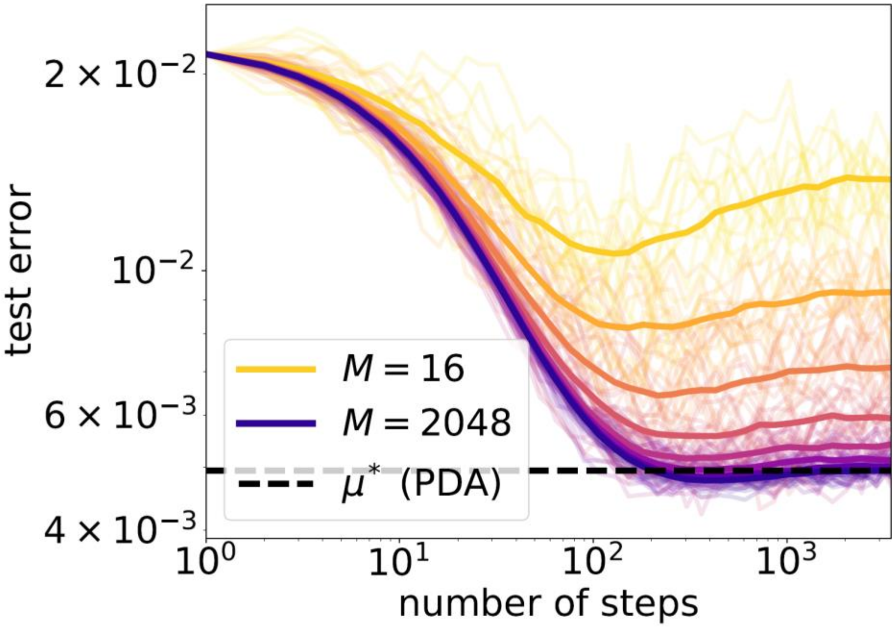
$$\mu_t \to \hat{\mu}_k = \frac{1}{N}\Sigma\delta_{X_k^{(i)}}$$

- **Stochastic gradient:** $\nabla \frac{\delta F(\mu)}{\delta \mu} \to v_k^i$

Numerical experiment



Training error with $r(x) = \|x\|^4$



Test error with $r(x) = \|x\|^2$

Convergence analysis

$p_\mu(x) \propto \exp\left(-\frac{1}{\lambda_2}\frac{\delta F(\mu)}{\delta \mu}(x)\right)$: proximal Gibbs measure

Theorem (One-step update) [Suzuki, Wu, Nitanda (2023)]

Suppose that p_μ satisfies log-Sobolev inequality with a constant α.

Under smoothness and boundedness of the loss function, it holds that

$$\begin{aligned}
&\mathscr{L}^{(N)}(\hat{\mu}_{k+1}) - \mathcal{L}(\mu^*) \\
&\leq \exp(-\lambda_2\eta_k\alpha)\left(\mathscr{L}^{(N)}(\hat{\mu}_k) - \mathcal{L}(\mu^*)\right) \\
&\quad + C\Big(\underbrace{\eta_k^3 + \lambda_2\eta_k^2}_{\text{Time discr.}} + \underbrace{\frac{\eta_k}{N}}_{\text{Space discr.}} + \underbrace{\eta_k^{\frac{3}{2}}\lambda_2^{\frac{1}{2}}\sigma_k\tilde{\sigma}_k}_{\text{Stochastic approx.}}\Big)
\end{aligned}$$

Naïve bound:
$\eta_k\sigma_k^2$

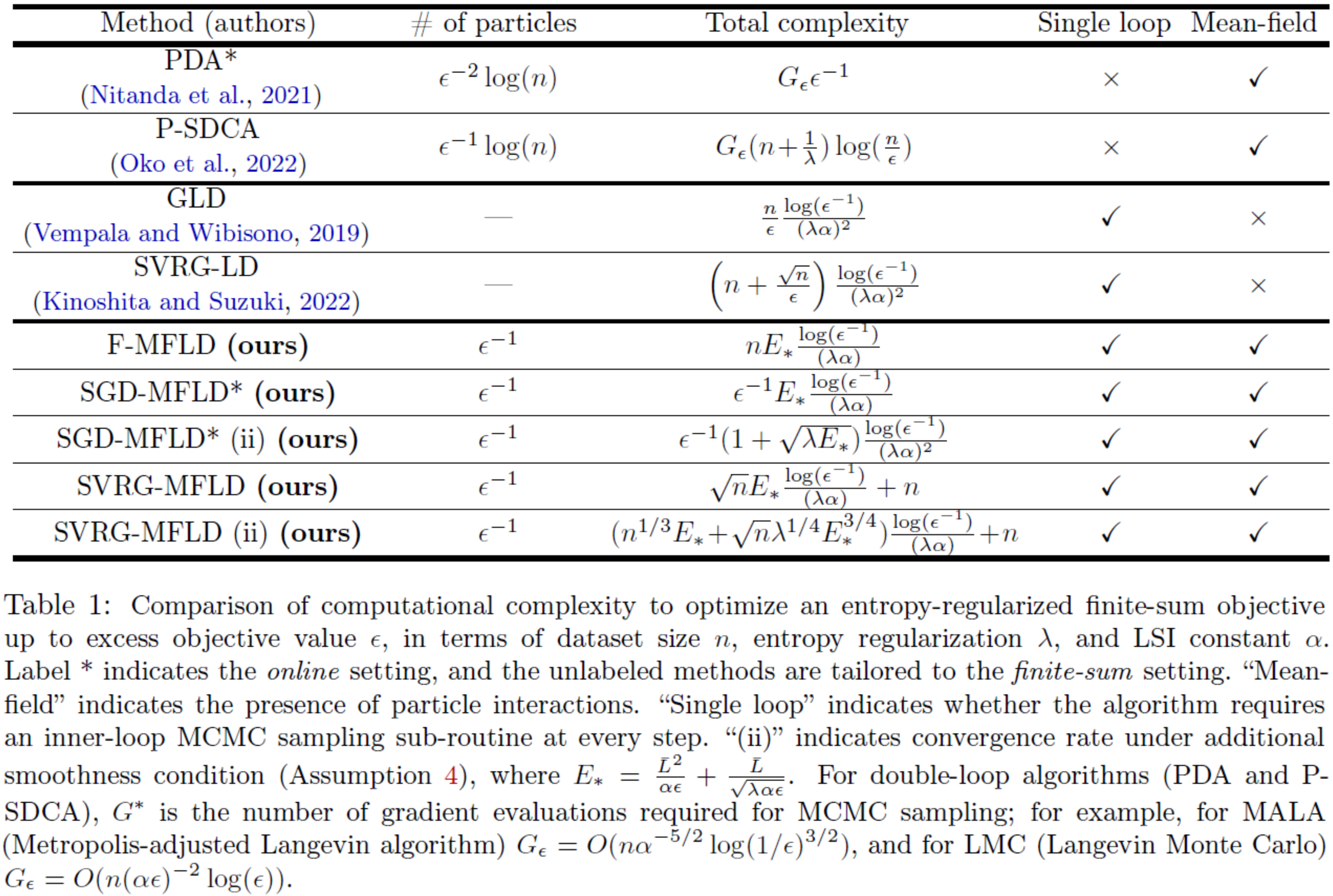
$$\begin{cases}
\sigma_k^2 = \max_i \mathbb{E}\left[\|v_k^i - \mathbb{E}[v_k^i]\|^2\right] \\
\tilde{\sigma}_k^2 = \max_i \mathbb{E}\left[\left\|\nabla v_k^{i\top}(\mathscr{X}) - \nabla\nabla^\top \frac{\delta F(\mu_{\mathscr{X}})}{\delta \mu}(X^i)\right\|_{\mathrm{op}}^2\right]
\end{cases}$$

Assumption:

1. $F\colon \mathcal{P} \to \mathbb{R}$ is convex and has a form of $\boldsymbol{F(\mu) = L(\mu) + \lambda_1 \mathbb{E}_\mu[\|x\|^2]}$.
2. (smoothness) $\left\|\nabla\frac{\delta L(\mu)}{\delta\mu}(x) - \nabla\frac{\delta L(\nu)}{\delta\mu}(y)\right\| \leq C(W_2(\mu,\nu) + \|x-y\|)$ and

   (boundedness) $\left\|\nabla\frac{\delta L(\mu)}{\delta\mu}(x)\right\| \leq R$. (+ second order differentiability)

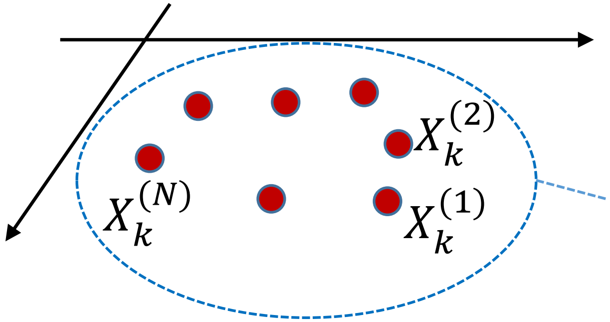
[Suzuki, Wu, Nitanda: Convergence of mean-field Langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction. arXiv:2306.07221]



Method (authors)	# of particles	Total complexity	Single loop	Mean-field
PDA* (Nitanda et al., 2021)	$\epsilon^{-2}\log(n)$	$G_\epsilon \epsilon^{-1}$	×	✓
P-SDCA (Oko et al., 2022)	$\epsilon^{-1}\log(n)$	$G_\epsilon(n+\frac{1}{\lambda})\log(\frac{n}{\epsilon})$	×	✓
GLD (Vempala and Wibisono, 2019)	—	$\frac{n}{\epsilon}\frac{\log(\epsilon^{-1})}{(\lambda\alpha)^2}$	✓	×
SVRG-LD (Kinoshita and Suzuki, 2022)	—	$\left(n+\frac{\sqrt{n}}{\epsilon}\right)\frac{\log(\epsilon^{-1})}{(\lambda\alpha)^2}$	✓	×
F-MFLD **(ours)**	ϵ^{-1}	$nE_*\frac{\log(\epsilon^{-1})}{(\lambda\alpha)}$	✓	✓
SGD-MFLD* **(ours)**	ϵ^{-1}	$\epsilon^{-1}E_*\frac{\log(\epsilon^{-1})}{(\lambda\alpha)}$	✓	✓
SGD-MFLD* (ii) **(ours)**	ϵ^{-1}	$\epsilon^{-1}(1+\sqrt{\lambda E_*})\frac{\log(\epsilon^{-1})}{(\lambda\alpha)^2}$	✓	✓
SVRG-MFLD **(ours)**	ϵ^{-1}	$\sqrt{n}E_*\frac{\log(\epsilon^{-1})}{(\lambda\alpha)}+n$	✓	✓
SVRG-MFLD (ii) **(ours)**	ϵ^{-1}	$(n^{1/3}E_*+\sqrt{n}\lambda^{1/4}E_*^{3/4})\frac{\log(\epsilon^{-1})}{(\lambda\alpha)}+n$	✓	✓

Table 1: Comparison of computational complexity to optimize an entropy-regularized finite-sum objective up to excess objective value ϵ, in terms of dataset size n, entropy regularization λ, and LSI constant α. Label * indicates the *online* setting, and the unlabeled methods are tailored to the *finite-sum* setting. "Mean-field" indicates the presence of particle interactions. "Single loop" indicates whether the algorithm requires an inner-loop MCMC sampling sub-routine at every step. "(ii)" indicates convergence rate under additional smoothness condition (Assumption 4), where $E_* = \frac{\bar{L}^2}{\alpha\epsilon} + \frac{\bar{L}}{\sqrt{\lambda\alpha\epsilon}}$. For double-loop algorithms (PDA and P-SDCA), G^* is the number of gradient evaluations required for MCMC sampling; for example, for MALA (Metropolis-adjusted Langevin algorithm) $G_\epsilon = O(n\alpha^{-5/2}\log(1/\epsilon)^{3/2})$, and for LMC (Langevin Monte Carlo) $G_\epsilon = O(n(\alpha\epsilon)^{-2}\log(\epsilon))$.

Uniform log-Sobolev inequality



$\mathscr{X}_k = \left(X_k^{(i)}\right)_{i=1}^{N} \sim \mu_k^{(N)}$: Joint distribution of N particles.

Potential of the joint distribution $\mu_k^{(N)}$ on $\mathbb{R}^{d\times N}$:

$$\mathscr{L}^N(\mu_k^{(N)}) = N\mathbb{E}_{\mathscr{X}\sim\mu_k^{(N)}}[F(\hat{\mu}_{\mathscr{X}})] + \lambda_2 \mathrm{Ent}(\mu_k^{(N)}).$$

where $\hat{\mu}_{\mathscr{X}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{X^{(i)}}$ $\quad (\mathscr{X} = (X^{(i)})_{i=1}^{N})$

- ➢ The finite particle dynamics is the Wasserstein gradient flow that minimizes $\mathscr{L}^N$.

(Approximate) Uniform log-Sobolev inequality [Chen et al. 2022]

For any N,

$$\frac{1}{N}\mathscr{L}^N(\mu_k^{(N)}) - \mathcal{L}(\mu^*) \le \frac{\lambda_2}{2\alpha}\left(\frac{1}{N}I(\mu_k^{(N)}||p^{(N)})\right) + \frac{C_{\alpha,\lambda_2}}{N}$$

(Fisher divergence)

where $p^{(N)}(\mathscr{X}) \propto \exp(-\frac{N}{\lambda_2}F(\hat{\mu}_{\mathscr{X}}))$

Recall $\mathcal{L}(\mu) = F(\mu) + \lambda_2\mathrm{Ent}(\mu)$

[Chen, Ren, Wang. Uniform-in-time propagation of chaos for mean field langevin dynamics. arXiv:2212.03050, 2022.]

Log Sobolev for Lipschitz cont obj

Proximal Gibbs measure:

$$p_\mu(x) \propto \exp\left(-\frac{1}{\lambda_2}\frac{\delta F(\mu)}{\delta\mu}(x)\right) \qquad p_\mu = \mathop{\arg\min}_{\nu\in\mathcal{P}}(\nu-\mu)\frac{\delta F(\mu)}{\delta\mu} + \lambda_2\mathrm{Ent}(\nu)$$

Assumption: $F(\mu) = L(\mu) + \lambda_1\mathbb{E}_\mu[\|x\|^2]$

μ satisfies the LSI if there exits $\alpha > 0$ such that for any ϕ s.t. $\mu(\phi^2) = 1$, it holds that

$$\mu(\phi^2\log(\phi^2)) \leq \frac{2}{\alpha}\int\|\nabla\phi\|^2\mathrm{d}\mu$$

1. Holley—Strook argument: [Bakry & Emery, 1985; Holley & Stroock, 1987]

$$\left\|\frac{\delta L(\mu)}{\delta\mu}\right\|_\infty \leq R \qquad \Rightarrow \qquad \alpha \geq \frac{\lambda_1}{\lambda_2}\exp\left(-\frac{4R}{\lambda_2}\right)$$

(New)

2. Lipschitz perturbation argument + Miclo's trick:

[Cattiaux & Guillin, 2022; Bardet et al., 2018]

$\sup_x\left\|\nabla\frac{\delta L(\mu)}{\delta\mu}(x)\right\| \leq R$ (Lipschitz continuous)

$$\Rightarrow \qquad \alpha \geq \frac{\lambda_1}{2\lambda_2}\exp\left(-\frac{4R^2}{\lambda_1\lambda_2}\sqrt{2d/\pi}\right)\vee$$

$$\left\{\frac{4\lambda_2}{\lambda_1} + e^{\frac{R^2}{2\lambda_1\lambda_2}}\left(\frac{R}{\lambda_1}+\sqrt{\frac{2\lambda_2}{\lambda_1}}\right)^2\left[2+d+\frac{d}{2}\log\left(\frac{\lambda_2}{\lambda_1}\right)+4\frac{R^2}{\lambda_1\lambda_2}\right]\right\}^{-1}$$

Computational complexity

SG-MFLD

$$F(\mu) = \frac{1}{n}\sum_{j=1}^{n} \ell_j(\mu) + \lambda_1 \mathbb{E}[\|x\|^2]$$ (finite sum),

$$v_k^i = \frac{1}{B}\sum_{j\in I_k} \nabla \frac{\delta \ell_j(\hat{\mu}_k)}{\delta \mu}(X_k^{(i)}) + \lambda_1 x$$ (stochastic gradient)

(Mini-batch size = B)

$$\boxed{\mathscr{L}^{(N)}(\hat{\mu}_k) - \mathcal{L}(\mu^*) \lesssim \exp(-\lambda_2 \eta k \alpha) + \frac{1}{\alpha\lambda_2}\left(\underbrace{\eta^2 + \lambda_2\eta}_{\text{Time discr.}} + \underbrace{\frac{1}{N}}_{\text{Space discr.}} + \underbrace{\frac{\eta + \sqrt{\eta\lambda_2}}{B}}_{\text{Stochastic approx.}}\right)}$$

Iteration complexity:

By setting $\eta = O\left(\epsilon\alpha \wedge \sqrt{\lambda_2\epsilon\alpha} \wedge (\lambda_2\epsilon\alpha)^2 \frac{B^2}{\lambda_2} \wedge (\epsilon\alpha B\lambda_2)\right)$,
the iteration complexity becomes

$$k = O\left(\frac{1}{\epsilon\alpha} + \sqrt{\frac{1}{\lambda_2\epsilon\alpha}} + \left(\frac{1}{\lambda_2\epsilon\alpha}\right)^2 \frac{\lambda_2}{B^2} + \frac{1}{\lambda_2\epsilon\alpha B}\right)\frac{1}{\lambda_2\alpha}\log(\epsilon^{-1})$$

to achieve $\epsilon + O(1/(\lambda_2\alpha N))$ accuracy.

- ➢ $B = \sqrt{1/(\lambda_2\alpha\epsilon)}$ is the optimal mini-batch size. → $k = O(\log(\epsilon^{-1})/\epsilon)$.

Variance reduction

SVRG-MFLD:

$$F(\mu) = \frac{1}{n}\sum_{j=1}^{n} f_j(\mu) = \frac{1}{n}\sum_{j=1}^{n} \ell_j(\mu) + \lambda\mathbb{E}[\|x\|^2]$$

(finite sum),

$$\boxed{v_k^i = \frac{1}{B}\sum_{j\in I_k}\nabla\frac{\delta f_j(\hat{\mu}_k)}{\delta\mu}(X_k^{(i)}) - \frac{1}{B}\sum_{j\in I_k}\nabla\frac{\delta f_j(\dot{\mu})}{\delta\mu}(\dot{X}^{(i)}) + \nabla\frac{\delta F(\dot{\mu})}{\delta\mu}(\dot{X}^{(i)})}$$

Variance reduction

($\dot{X}$ is updated once at every m steps)

$$\boxed{\begin{aligned}&\mathscr{L}^{(N)}(\hat{\mu}_k) - \mathcal{L}(\mu^*)\\ &\lesssim \exp(-\lambda_2\eta k\alpha)\\ &\quad + \frac{1}{\lambda_2\alpha}\left(\underbrace{\eta^2 + \lambda_2\eta}_{\text{Time discr.}} + \underbrace{\frac{1}{N}}_{\text{Space discr.}} + \underbrace{\frac{n-B}{B(n-1)}\lambda_2^{1/2}\eta\sqrt{m(\eta+\lambda_2)}}_{\text{Stochastic approx.}}\right)\end{aligned}}$$

Tighter than the analysis in linear GLD [Kinoshita, Suzuki: NeurIPS2022] ($m = B = \sqrt{n}$)

$$\frac{n-B}{B(n-1)}m(\eta^2 + \lambda_2\eta)$$

\# of update : $\eta = \epsilon\alpha \wedge \sqrt{\lambda_2\alpha\epsilon}$,

$$k = \frac{1}{\lambda_2\alpha\eta}\log(1/\epsilon) = O\left(\frac{1}{\epsilon\alpha} + \sqrt{\frac{1}{\lambda_2\alpha\epsilon}}\right)\frac{1}{\lambda_2\alpha}\log(\epsilon^{-1})$$

where $B = \sqrt{m} = n^{1/3}$.

Total complexity : $Bk + \frac{nk}{m} \lesssim n^{1/3}\left(\frac{1}{\alpha\epsilon} + \sqrt{\frac{1}{\lambda_2\alpha\epsilon}}\right)\frac{1}{\lambda_2\alpha}\log(\epsilon^{-1})$.

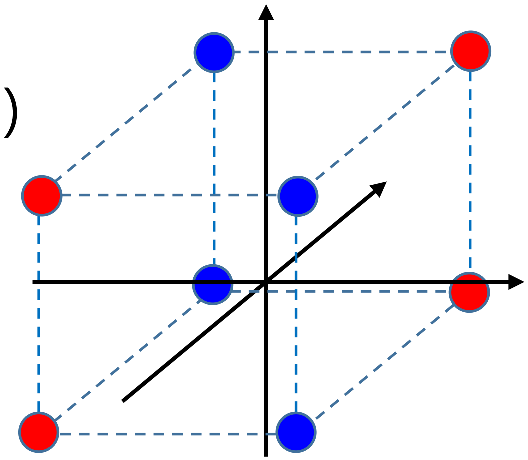
$\sqrt{n}$ by Kinoshita&Suzuki (2022)

Generalization analysis

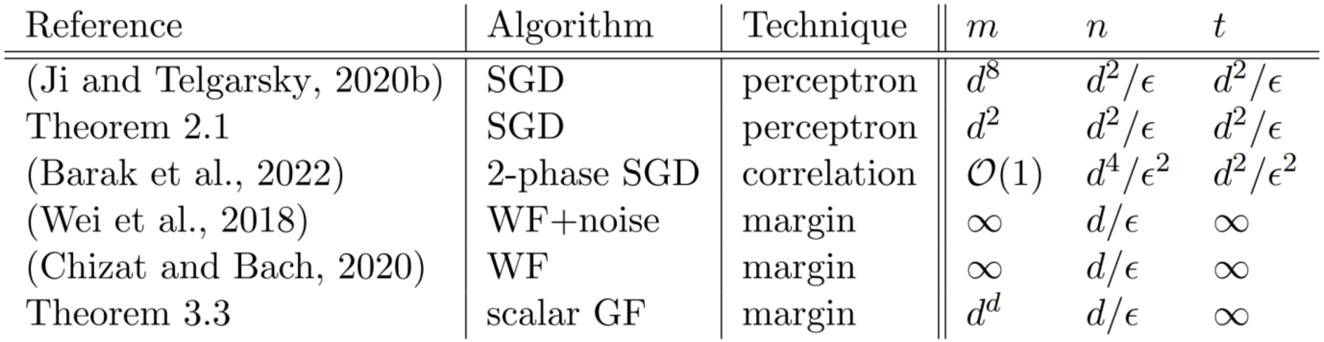
Reference

- ℓ_i: logistic loss
- $h_z(x) = \bar{R} \cdot \tanh(\langle x_1, z\rangle + x_2)/2$

- Learning XOR function on high dimensional data.
  - ➢ $X \sim \mathrm{Unif}(\{-1,1\}^d)$ (up to freedom of rotation)
  - ➢ $Y = X_k X_l$ for $k, l \in [d]$ with $k \neq l$.



Q: Can we learn XOR function with GD?
How large is the computational cost?

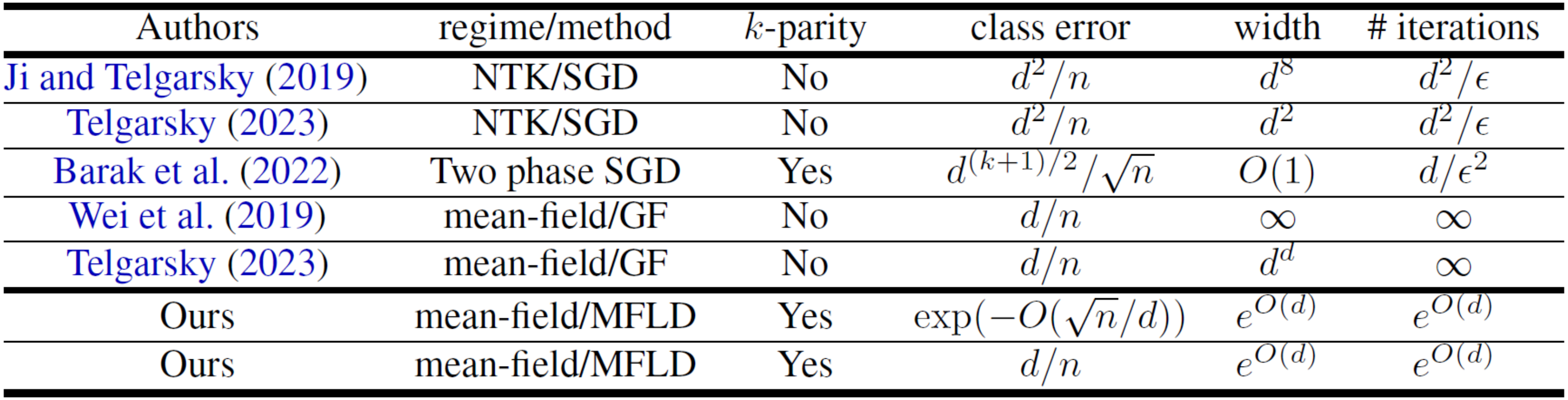


Reference	Algorithm	Technique	m	n	t
(Ji and Telgarsky, 2020b)	SGD	perceptron	d^8	d^2/ϵ	d^2/ϵ
Theorem 2.1	SGD	perceptron	d^2	d^2/ϵ	d^2/ϵ
(Barak et al., 2022)	2-phase SGD	correlation	$\mathcal{O}(1)$	d^4/ϵ^2	d^2/ϵ^2
(Wei et al., 2018)	WF+noise	margin	∞	d/ϵ	∞
(Chizat and Bach, 2020)	WF	margin	∞	d/ϵ	∞
Theorem 3.3	scalar GF	margin	d^d	d/ϵ	∞

Table 1 of [Telgarsky: Feature selection and low test error in shallow low-rotation ReLu networks, ICLR2023].

Reference

- Setting 1: $n > d^2$
  - Comp complexity: $\exp(O(d))$
  - Test error (classification error) = $\boldsymbol{O}(\exp(-\sqrt{\boldsymbol{n}}/\boldsymbol{d}))$
- Setting 2: $n > d$
  - Comp complexity: $\exp(O(d))$
  - Test error (classification error) = $\boldsymbol{O}(\boldsymbol{d}/\boldsymbol{n})$



Authors	regime/method	k-parity	class error	width	# iterations
Ji and Telgarsky (2019)	NTK/SGD	No	d^2/n	d^8	d^2/ϵ
Telgarsky (2023)	NTK/SGD	No	d^2/n	d^2	d^2/ϵ
Barak et al. (2022)	Two phase SGD	Yes	$d^{(k+1)/2}/\sqrt{n}$	$O(1)$	d/ϵ^2
Wei et al. (2019)	mean-field/GF	No	d/n	∞	∞
Telgarsky (2023)	mean-field/GF	No	d/n	d^d	∞
Ours	mean-field/MFLD	Yes	$\exp(-O(\sqrt{n}/d))$	$e^{O(d)}$	$e^{O(d)}$
Ours	mean-field/MFLD	Yes	d/n	$e^{O(d)}$	$e^{O(d)}$

Conclusion

Mean field Langevin dynamics

- Entropy sandwich
  - ➢ Connecting **duality gap** with KL-div between the current solution and its proximal Gibbs measure.
  - ➢ Exponential convergence

$$\lambda_2 \mathrm{KL}(\mu||\mu^*) = \mathcal{L}(\mu) - \mathcal{L}(\mu^*) \leq \mathcal{L}(\mu) - \mathcal{D}(g_\mu) = \lambda_2 \mathrm{KL}(\mu||p_\mu)$$

$$\mathcal{L}(\mu_t) - \mathcal{L}(\mu^*) \leq \exp(-2\alpha\lambda_2 t)(\mathcal{L}(\mu_0) - \mathcal{L}(\mu^*))$$

- Finite particle approximation
  - ➢ Uniform-in-time propagation of chaos

$$\begin{aligned}&\mathscr{L}^{(N)}(\hat{\mu}_{k+1}) - \mathcal{L}(\mu^*)\\ &\leq \exp(-\lambda_2\eta_k\alpha)\left(\mathscr{L}^{(N)}(\hat{\mu}_k) - \mathcal{L}(\mu^*)\right) + C\left(\eta_k^3 + \lambda_2\eta_k^2 + \frac{\eta_k}{N} + \eta_k^{\frac{3}{2}}\lambda_2^{\frac{1}{2}}\sigma_k\tilde{\sigma}_k\right)\end{aligned}$$

Many other interesting topics:

- Entropic fictious play [Chen, Ren, Wang (2022); Nitanda et al. (ICML2023)]
- Learning theory, better sample complexity than NTK [Suzuki et al. (2023)]
- Application to Reinforcement Learning: Policy-Gradient [Yamamoto et al. (2023)]
- Infinite dimensional mean field Langevin [Nishikawa, Suzuki, Nitanda (NeurIPS2022)]