Representation theory and optimization of neural networks

Taiji Suzuki
University of Tokyo / AIP-RIKEN
(Deep learning theory team)

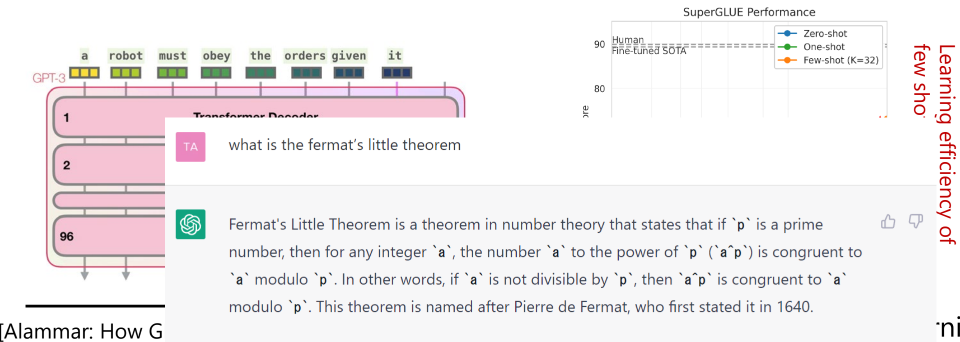
21th/Mar/2023

RIKEN-AIP & PRAIRIE Joint Workshop 2023

Success of deep learning

Deep learning has shown great performances in the AI research field.
→ Why?

Large language model



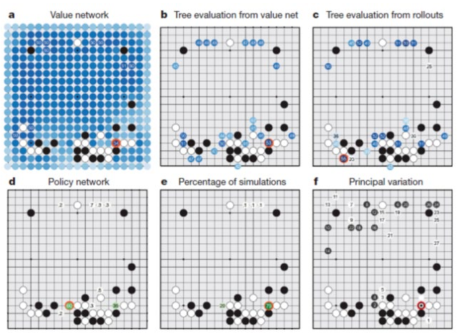
[Alammar: How G https://jalammar.g animations/]

[Brown et al. "Language Models are Fe

rning

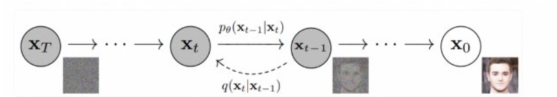
[ChatGPT. OpenAI2022]

AlphaGo/Zero



[Silver et al. (Google Deep Mind): Mastering the game of Go with deep neural networks and tree search, Nature, 529, 484—489, 2016]

Generative models (diffusion models)



[Ho, Jain, Abbeel: Denoising Diffusion Probabilistic Models. 2020]

Image recognition

[He, Gkioxari, Dollár, Girshick: Mask R-CNN, ICCV2017]

Stable diffusion, 2022.

Jason Allen "Théâtre D' Spatial" generated by M Colorado State Fai 1st prize in digital a

Generated by NovelAI

Outline

[Representation theory]

- Minimax optimality of diffusion model
  - ➢Total variation distance and Wasserstein distance
  - ➢Avoids curse of dimensionality

[Kazusato Oko, Shunta Akiyama, Taiji Suzuki: Diffusion Models are Minimax Optimal Distribution Estimators. arXiv:2303.01861, 2023]

[Optimization]

- Mean field Langevin dynamics
  - ➢Unifying frame-work
  - ➢(1) Time discretization, (2) Space discretization, (3) Stochastic gradient

[Taiji Suzuki, Atsushi Nitanda, Denny Wu: Convergence of mean-field Langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction. 2023]

Minimax optimality of diffusion model

[Kazusato Oko, Shunta Akiyama, Taiji Suzuki: Diffusion Models are Minimax Optimal Distribution Estimators. arXiv:2303.01861, 2023]

Kazusato Oko
(The University of Tokyo /RIKEN-AIP)

Shunta Akiyama
(The University of Tokyo)

Diffusion model

「An astronaut riding a horse in a photorealistic style」 ➡

DALL·E: [Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever: Zero-Shot Text-to-Image Generation. ICML2021.]
DALL·E2:[Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen: Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125]

Stable diffusion, 2022.

Jason Allen "Théâtre D'opéra Spatial" generated by **Midjourney**. Colorado State Fair's fine art competition, 1st prize in digital art category

Generated by NovelAI

Decoder : Diffusion model

[Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Song et al., 2020; Ho et al., 2020; Vahdat et al., 2021]

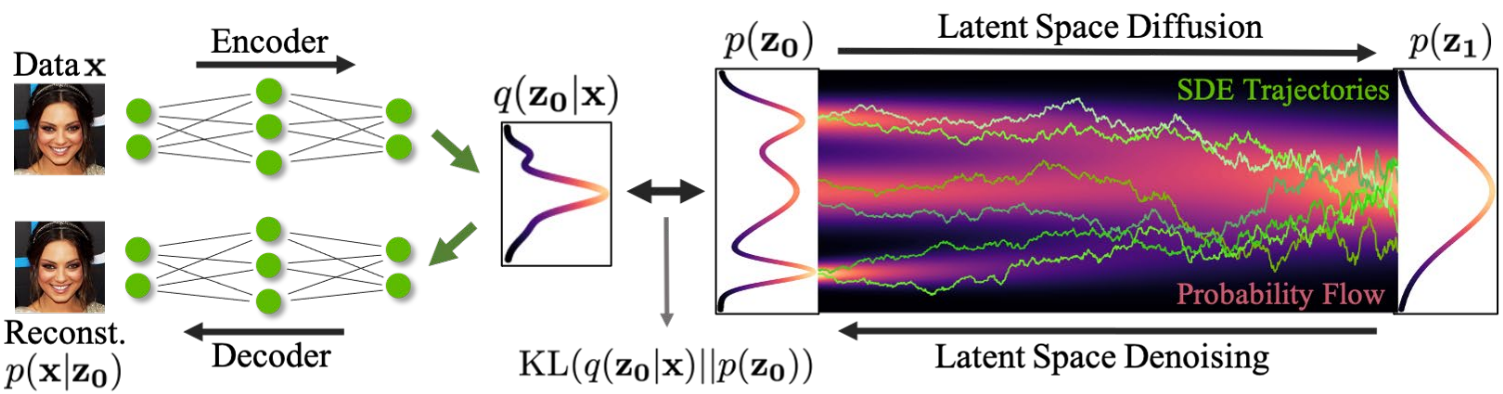
Forward process：Convert the target distribution to a noise distribution (e.g., Gaussian)

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

$$\mathrm{d}Y_t = (Y_t + 2\nabla \log(p_{\overline{T}-t}(Y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

$(Y_t \sim X_{\overline{T}-t})$

Reverse process：Convert the noise distribution to the target distribution



[Vahdat, Kreis, Kautz: Score-based Generative Modeling in Latent Space. arXiv:2106.05931]

Forward process

Forward process:

$$\mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

OU process

- Marginal distribution

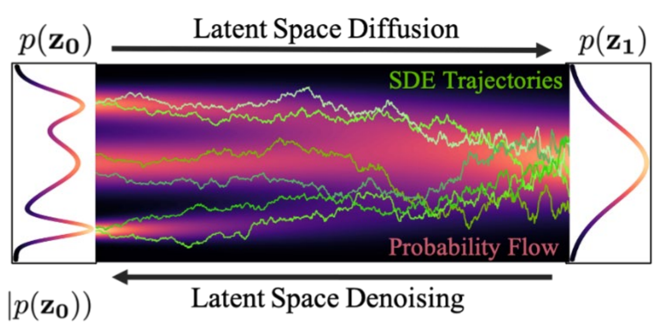
$$p_t = \mathrm{Law}(X_t) \Rightarrow p_t(x) = \int p_0(y) \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\|x - \mu_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y$$

$$p_t = N(\mu_t X_0, \sigma_t^2 I) * p_0$$

where $\mu_t = \exp(-t)$, $\sigma_t^2 = 1 - \exp(-2t)$.

The forward process converges to the noise distribution (standard normal) exponentially:

$$\mathrm{KL}(p_t || N(0, I)) \leq \mathrm{O}(\exp(-2t))$$



[Vahdat, Kreis, Kautz: Score-based Generative Modeling in Latent Space. arXiv:2106.05931]

Reverse process

Reverse process:

(unknown)

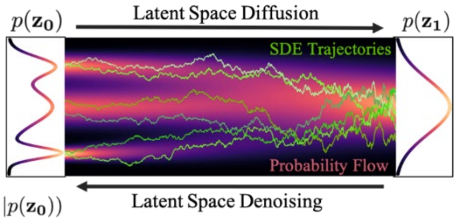
$$Y_0 \sim p_{\overline{T}}$$

(unknown)

$$\mathrm{d}Y_t = (Y_t + 2\nabla \log(p_{\overline{T}-t}(Y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

$$\Rightarrow Y_t \sim p_{\overline{T}-t}$$

[Haussmann & Pardoux, 1986]



Approximated process (generative model):

$$\hat{Y}_0 \sim N(0, I)$$

($N(0,I)$ is close to $p_{\overline{T}}$)

$$\mathrm{d}\hat{Y}_t = (\hat{Y}_t + 2\hat{s}(\hat{Y}_t, \overline{T} - t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

Theorem (Girsanov's theorem; Chen et al. (2023))

If $\hat{Y}_0 \sim p_{\overline{T}}$, then

$$\mathrm{KL}(p_0 \| p_{\hat{Y}_{\overline{T}}}) \leq \frac{1}{4}\int_0^{\overline{T}} \mathbb{E}_{Y_t}[\|\nabla \log(p_{\overline{T}-t}(Y_t)) - \hat{s}(Y_t, \overline{T} - t)\|^2]\mathrm{d}t$$

⇒ It suffices to estimate the *score function* $\nabla \log(p_t)$ as accurate as possible.

Score matching

$$\int_0^{\overline{T}} \mathbb{E}_{Y_t}[\|\nabla \log(p_{\overline{T}-t}(Y_t)) - \hat{s}(Y_t, \overline{T} - t)\|^2]\mathrm{d}t$$

$$= \int_0^{\overline{T}} \mathbb{E}_{X_t}[\|\nabla \log(p_t(X_t)) - \hat{s}(X_t, t)\|^2]\mathrm{d}t$$

$$= \int_0^{\overline{T}} \mathbb{E}_{X_t, X_0}[\|\nabla \log(p_t(X_t|X_0)) - \hat{s}(Y_t, t)\|^2]\mathrm{d}t + (\mathrm{const})$$

Observation (n data points $D_n = \{x_i\}_{i=1}^n$):

$$x_i \sim p_0 \quad (i = 1, \ldots, n)$$

Empirical score matching loss:

$$\min_{s \in \mathrm{DNN}} \frac{1}{n} \sum_{i=1}^n \int_{t=\underline{T}}^{\overline{T}} \mathbb{E}_{X_t|X_0=x_i}\left[\|s(X_t, t) - \underline{\nabla \log p_t(X_t|x_i)}\|^2\right] \mathrm{d}t$$

Can be sampled via OU process

Explicit form is available

Existing work on error analysis

- Reverse SDE characterization: Song et al. (2021)

[Approximation error analysis]

- KL-divergence bound via Girsanov's theorem: Chen et al. (2022)

- Error bound with LSI: Lee et al. (2022a)
  - With smoothness: Chen et al. (2022) and Lee et al. (2022b)

- Error propagation with manifold assumption: Pidstrigach (2022)

[Generalization analysis]

- Wasserstein dist bound ($n^{-1/d}$) with manifold assumption: De Bortoli (2022)

Problem setting

Assumption 1

The true distribution p_0 is supported on $[-1,1]^d$ and

$$p_0 \in B^s_{p,q}$$

with $s > (1/p - 1/2)_+$ as a density function on $[-1,1]^d$.

Assumption2

p_0 is sufficiently smooth on the edge of the support $[-1,1]^d \setminus [-1+n^{-\frac{1-\delta}{d}}, 1-n^{-\frac{1-\delta}{d}}]^d$.

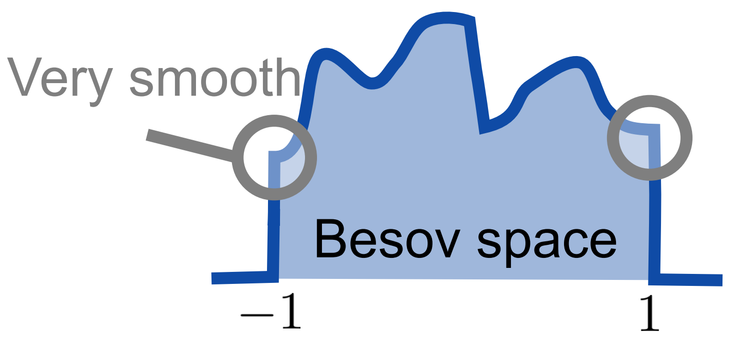
Besov space ($B^s_{p,q}(\Omega)$)

$$\omega_m(f,t)_p := \sup_{\|h\|\le t} \left\| \sum_{j=1}^{m} (-1)^{m-j} \binom{m}{j} f(\cdot + jh) \right\|_{L^p(\Omega)},$$

$$\|f\|_{B^s_{p,q}(\Omega)} = \|f\|_{L^p(\Omega)} + \left(\int_0^\infty [t^{-s}\omega_m(f,t)_p]^q \frac{\mathrm{d}t}{t} \right)^{1/q}.$$

Smoothness

Spatial inhomogeneity



Problem setting

Assumption 1

The true distribution p_0 is supported on $[-1,1]^d$ and

$$p_0 \in B^s_{p,q}$$

with $s > (1/p - 1/2)_+$ as a density function on $[-1,1]^d$.

Assumption2

p_0 is sufficiently smooth on the edge of the support $[-1,1]^d \setminus [-1+n^{-\frac{1-\delta}{d}}, 1-n^{-\frac{1-\delta}{d}}]^d$.

Besov space ($B^s_{p,q}(\Omega)$)

$\omega_m(f,t)_p$

Intuition

$$\|f\|_{B^s_{p,q}(\Omega)} = \|f\|_{L^p(\Omega)} + \|D^s f\|_{L^p(\Omega)}$$

Smoothness

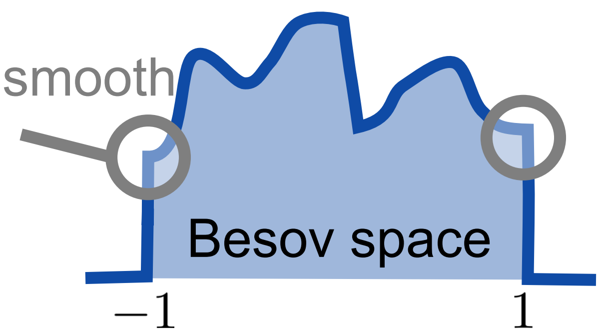
Uniformity of smoothness

$$\|f\|_{B^s_{p,q}(\Omega)} = \|f\|_{L^p(\Omega)} + \left(\int_0^\infty [t^{-s}\omega_m(f,t)_p]^q \frac{\mathrm{d}t}{t}\right)^{1/q}.$$

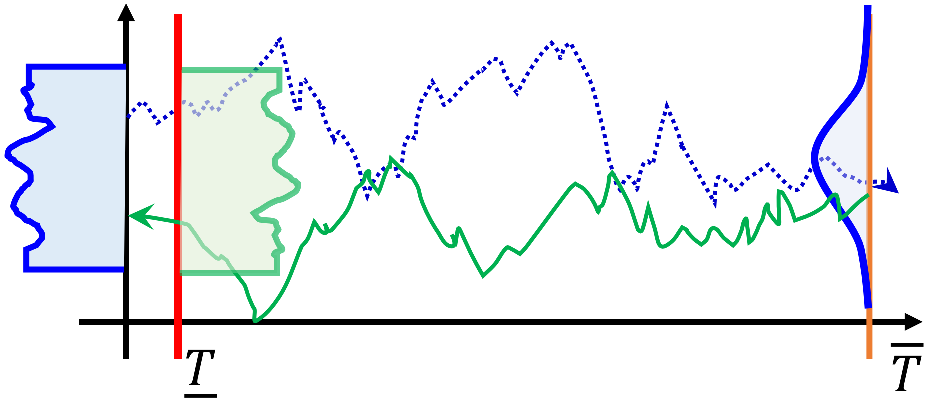
Smoothness

Spatial inhomogeneity

ery smooth



Convergence rate result



Theorem (Estimation error in TV-distance)

Let $\underline{T} = n^{-O(1)}, \overline{T} = O(\log(n))$. Then, the empirical risk minimizer $\hat{s}$ in DNN satisfies

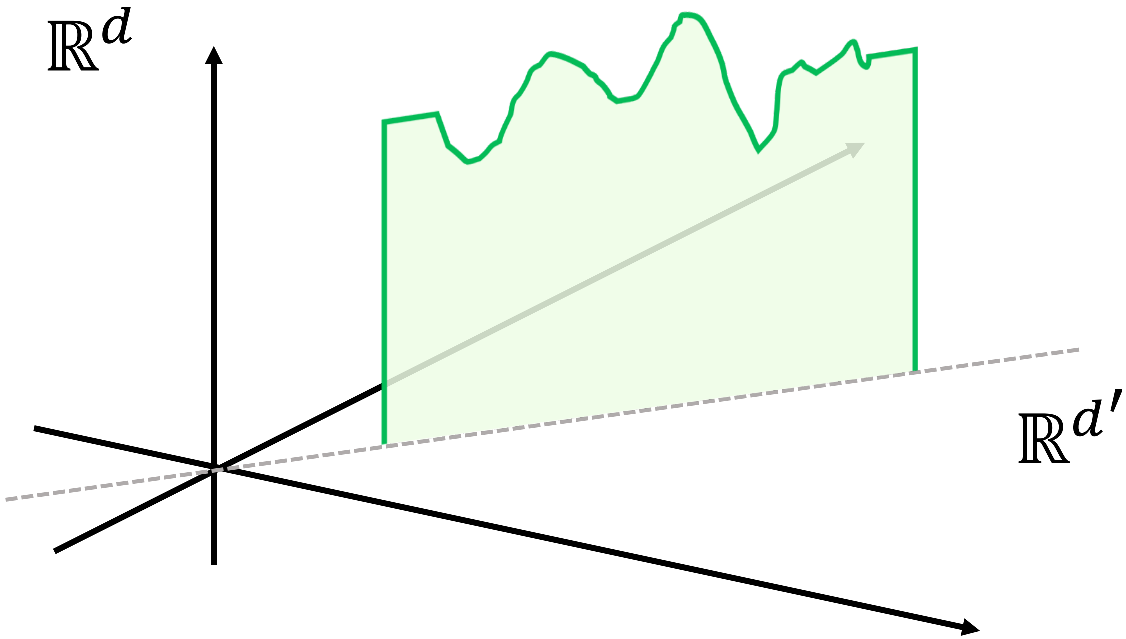
$$\mathbb{E}_{D_n}\left[\mathrm{TV}(\hat{Y}_{\overline{T}-\underline{T}}, X_0)\right] \lesssim n^{-\frac{s}{2s+d}} \log^9(n).$$

This is minimax optimal, that is, it holds

$$n^{-\frac{s}{2s+d}} \lesssim \inf_{\hat{\mu}:\text{estimator}} \sup_{p_0} \mathbb{E}_{D_n}\left[\mathrm{TV}(\hat{\mu}, X_0)\right]$$

Although $\hat{s}(x,t)$ is a function with $d+1$-dimensional input, there appears "d" in the bound instead of $d+1$. This is because Gaussian convolution induces smoothness.

Low dimensional structure



The support of the target distribution is in a low dimensional subspace.

The estimated distribution is never absolutely continuous to the target distribution.
→ **Wasserstein distance**

W_1-distance convergence rate

Theorem (Estimation error in W1-distance)

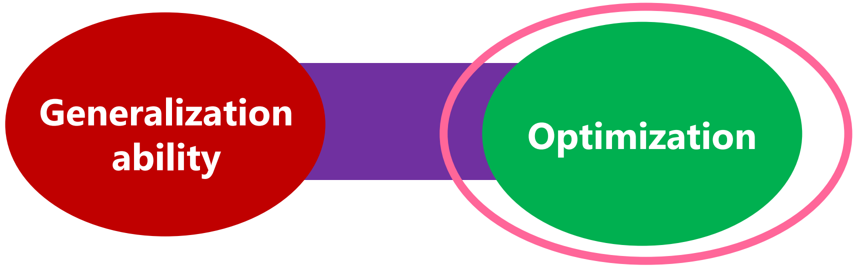
For any fixed $\delta > 0$, by slightly changing the estimator, the empirical risk minimizer $\hat{s}$ in DNN satisfies

$$\mathbb{E}_{D_n}\left[\mathrm{W}_1(\hat{Y}_{\overline{T}-\underline{T}}, X_0)\right] \lesssim n^{-\frac{s+1-\delta}{2s+d'}}.$$

This is also known as minimax optimal (up to δ) [Niles-Weed & Berthet (2022)].

- d' appears instead of d: **Diffusion model can avoid curse of dimensionality**.
- The minimax rate of Wasserstein distance is faster than that of TV distance, which makes it difficult to establish the bound.
  - ➢ We need more precise estimate of the score around $t = 0$.

(TV) $n^{-\frac{s}{2s+d}}$ ➡ $n^{-\frac{s+1}{2s+d}}$ (W1)



Mean field Langevin dynamics to optimize two-layer NN

[Suzuki, Nitanda, Wu: Convergence of mean-field Langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction. 2023]

Atsushi Nitanda
(Kyusyu Institute of Technology)

Denny Wu
(University of Toronto)

Objective of mean field NN

Mean field Langevin dynamics:

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

convex

- Wasserstein gradient flow to minimize $\mathcal{L}$:

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

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

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

$$\mu_t = \mathrm{Law}(X_t)$$

Vanilla GLD: $\mathrm{d}X_t = -\nabla L(X_t)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t$

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

$$\Rightarrow \frac{\delta F}{\delta \mu} = L$$

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

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

MFLD for mean field NN

$$f_\mu(z) = \int h_x(z)\mathrm{d}\mu(x)$$

$$h_x(z) = r\sigma(w^\top z) \text{ for } x = (r, w)$$

Loss function:

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

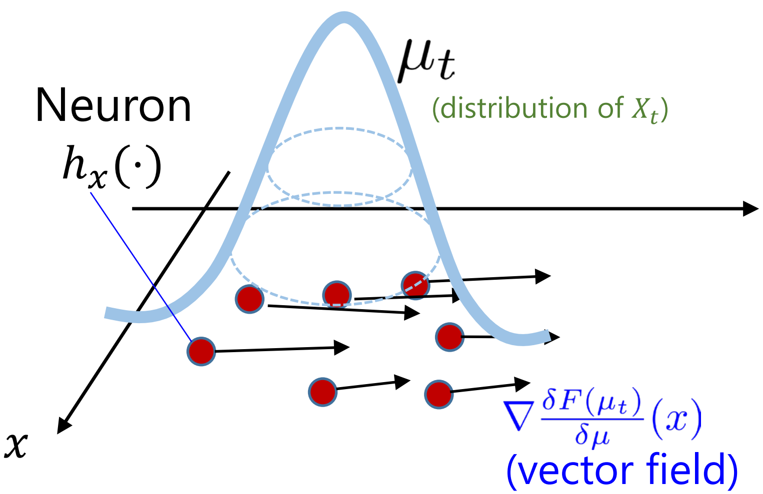
$$\mathrm{d}X_t = -\underbrace{\nabla_{X_t}\left(\frac{1}{n}\sum_{i=1}^{n}\ell_i'(f_{\mu_t})h_{X_t}(z_i) + \lambda_1\|X_t\|^2\right)}_{\nabla\frac{\delta F(\mu_t)}{\delta\mu}(X_t)}\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t$$

(escape from local min.)

[Noise perturbation]

$$\mu_t = \mathrm{Law}(X_t)$$

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



Finite particle approximation:

$$\mathrm{d}\hat{X}_t^i = -\nabla\frac{\delta F\left(\frac{1}{N}\sum_{j=1}^{N}\delta_{\hat{X}_t^j}\right)}{\delta\mu}(\hat{X}_t^i)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t^i$$

(GLD to optimize the finite width neural network)

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)

Summary of our 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, Nitanda, Wu (2023)]

Single loop method

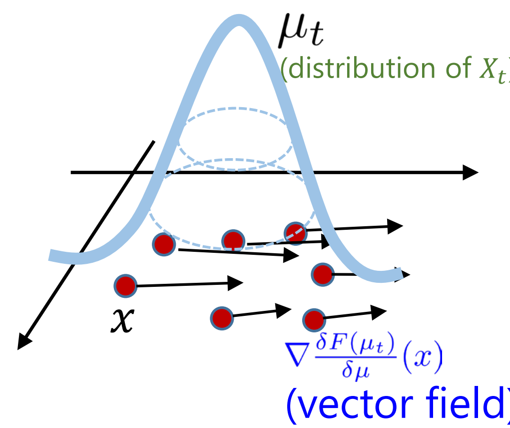
$$\mathrm{d}X_t = -\nabla \frac{\delta F(\mu_t)}{\delta \mu}(X_t)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t$$

(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 <u>finite width</u>.

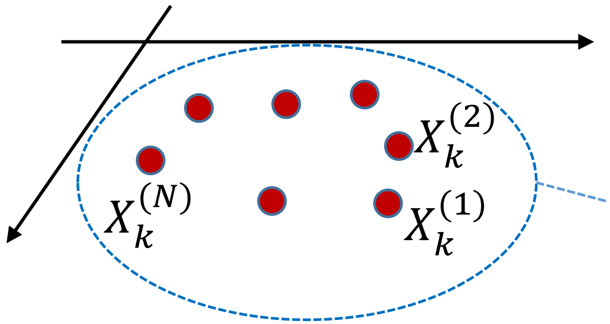
- **Time discretization:** $t \to k\eta$
- **Space discretization:** μ_t is approximatd by N particles

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

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

Uniform log-Sobolev inequality

Reference



$\mathscr{X}_k = \left(X_k^{(i)}\right)_{i=1}^N \sim \mu_k^{(N)}$: 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 $\hat{\mu}_{\mathscr{X}} = \frac{1}{N}\sum_{i=1}^N \delta_{X^{(i)}} \qquad (\mathscr{X} = (X^{(i)})_{i=1}^N)$

- 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,

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

(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.]

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, Nitanda, Wu (2023)]

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

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

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

Assumption:

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

Space discretization is shown through the uniform-log-Sobolev inequality shown by Chen et al. 2022.

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

Computational complexity

- SG-MFLD

$F(\mu) = \frac{1}{n}\sum_{j=1}^{n} f_j(\mu)$ (finite sum),

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

(Mini-batch size = B)

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

Iteration complexity:

By setting $\eta = O\left(\frac{\lambda_2\epsilon}{\alpha}\lambda_2^{-1} \wedge \left(\frac{\lambda_2\epsilon}{\alpha}\right)^2 \frac{B^2(n-1)^2}{(n-B)^2\lambda_2} \wedge \sqrt{\frac{\lambda_2\epsilon}{\alpha}}\right)$,
the iteration complexity becomes

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

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

- ➢ $B = n \wedge \sqrt{\alpha/(\lambda_2\epsilon)}$ is the optimal mini-batch size.

Summary

Deep learning theory
Representation ability + Optimization

[Representation theory]

- Minimax optimality of diffusion model
  - Total variation distance and Wasserstein distance
  - Avoids curse of dimensionality

[Kazusato Oko, Shunta Akiyama, Taiji Suzuki: Diffusion Models are Minimax Optimal Distribution Estimators. arXiv:2303.01861, 2023]

[Optimization]

- Mean field Langevin dynamics
  - Unifying frame-work
  - (1) Time discretization, (2) Space discretization, (3) Stochastic gradient

[Taiji Suzuki, Atsushi Nitanda, Denny Wu: Convergence of mean-field Langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction. 2023]

We are still at a primitive stage.
Hope to have collaborations!

Appendix

B-spline basis decomposition

$$\nabla \log(p_t(x)) = \frac{\boxed{\nabla p_t(x)}}{\boxed{p_t(x)}}$$

Approximate each term by DNNs

- B-spline decomposition of a Besov function p_0

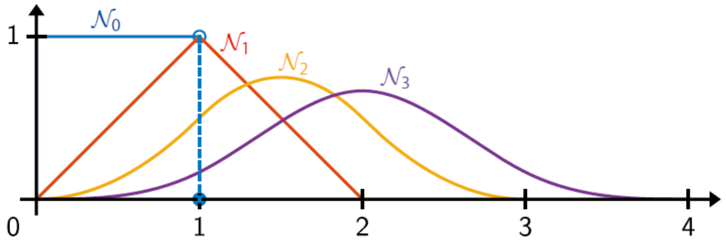
$$p_0(x) \approx \sum_{j=1}^{N} \alpha_j M^d_{a^j, b^j}(x)$$

$$\mathcal{N}(x) = \begin{cases} 1 & (x \in [0,1]), \\ 0 & (\text{otherwise}) \end{cases}$$

Cardinal B-spline of order m:

$$\mathcal{N}_m(x) = (\underbrace{\mathcal{N} * \mathcal{N} * \cdots * \mathcal{N}}_{m+1 \text{ times}})(x)$$

→ Piece-wise polynomial of order m.



Tensor product B-spline:

$$M^d_{a,b}(x) = \prod_{j=1}^{d} \mathcal{N}_m(2^{a_j} - b_j)$$

Cardinal B-spline interpolation (DeVore & Popov, 1988)

- Atomic decomposition:

$$\mathcal{N}_{k,j}^{(d)}(x_1,\dots,x_d) = \prod_{i=1}^{d} \mathcal{N}_m(2^k x_i - j_i)$$

$f \in B^s_{p,q}$ can be decomposed into

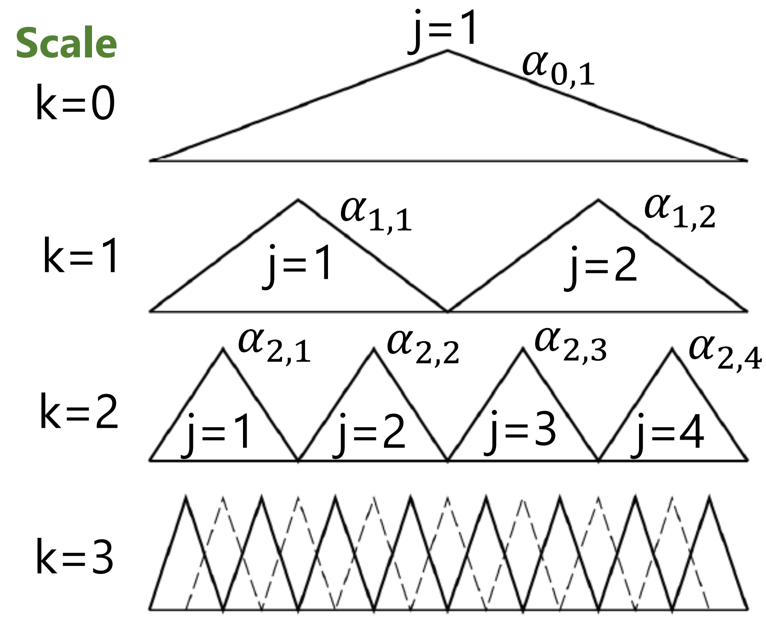
$$f = \sum_{k\in\mathbb{N}+} \sum_{j\in J(k)} \alpha_{k,j}\mathcal{N}_{k,j}^{(d)}$$

such that
(where $J(k) = \{j \in \mathbb{Z}^d \mid -m < j_i < 2^{k_i+1} + m\}$)

$$N(f) = \left[\sum_{k=0}^{\infty}\{2^{sk}(2^{-kd}\sum_{j\in J(k)}|\alpha_{k,j}|^p)^{1/p}\}^q\right]^{1/q} < \infty$$

$$\|f\|_{B^s_{p,q}} \simeq N(f) \quad \text{(Norm equivalence)}$$

Wavelet/multi-resolution expansion



DNN can approximate each B-spline basis efficiently.

$$f = \underbrace{\sum_{k,j\in I_N}\alpha_{k,j}\mathcal{N}_{k,j}^{(d)}}_{} + O(N^{-s/d})$$

N terms (should be appropriately chosen depending on f)

(see also Bolcskei, Grohs, Kutyniok, Petersen: Optimal Approximation with Sparsely Connected Deep Neural Networks. 2018)

Proof outline (1)

$$\nabla \log(p_t(x)) = \frac{\boxed{\nabla p_t(x)}}{\boxed{p_t(x)}}$$

Approximate each term by DNNs

- B-spline decomposition of a Besov function p_0

$$p_0(x) \approx \sum_{j=1}^{N} \alpha_j M^d_{a^j, b^j}(x)$$

- Diffused B-spline basis expansion of p_t

$$p_t(x) = \int \boxed{p_0(y)} \underbrace{\frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \exp\left(- \frac{\|x - \mu_t y\|^2}{2\sigma_t^2} \right)}_{=: K_t(x|y)} \mathrm{d}y$$

Decompose

$$p_t(x) \approx \sum_{j=1}^{N} \alpha_j \underbrace{\int M^d_{a^j, b^j}(y) K_t(x|y) \mathrm{d}y}_{=: E_{a^j, b^j}(x, t)}$$

Diffused B-spline

➢ We approximate *Diffused B-splines* by DNNs.

Approximation error of Diffused B-spline

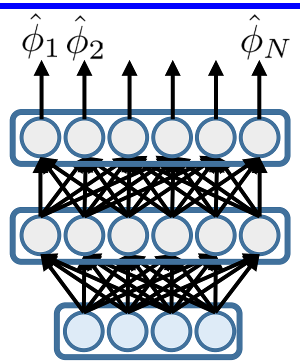
Lemma (Approximation error of diffused B-spline)

There exists a deep neural network $\hat{\phi}: \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d$ such that

$$\left\| \hat{\phi}(x,t) - E_{a^j,b^j}(x,t) \right\|_\infty \leq \epsilon$$

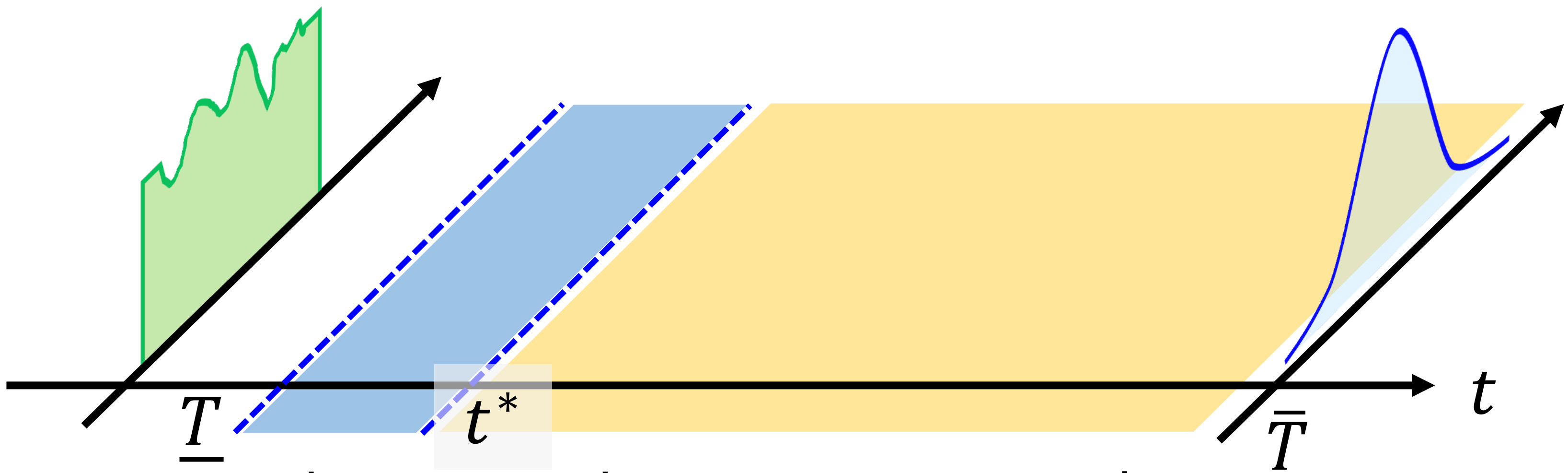
with depth $L = O\left(\log^4(\epsilon^{-1})\right)$, width $W_i = O(\log^6(\epsilon^{-1}))$, sparsity (# of non-zero parameters) $S = O(\log(\epsilon^{-1}))$, and ℓ^∞-norm bound $B = O(\exp(O(\log^2(\epsilon^{-1}))))$ on parameters.

$$\check{f}_N(x,t) = \sum_{i=1}^{N} \alpha_i \hat{\phi}_i(x,t)$$: Deep neural network



$$\|p_t(\cdot) - \check{f}_N(\cdot,t)\|_{L^r} \leq \sum_{i=1}^N |\alpha_i| \underbrace{\|\phi_i(\cdot,t) - \hat{\phi}_i(\cdot,t)\|_{L^r}}_{\leq O(e^{-L})} + \underbrace{\| \textstyle\sum_{i=N+1}^\infty \alpha_i \phi_i(\cdot,t)\|_{L^r}}_{\leq N^{-s/d}}$$

Error bound of score



- Bound by diffused B-spline approximation

$$p_t(x) \approx \sum_{j=1}^{N} \alpha_j E_{a^j, b^j}(x, t)$$

$$\|p_t - \check{f}_N(\cdot, t)\|_{L^r(X_t)} \lesssim N^{-s/d} \|p_0\|_{B^s_{p,q}}$$

  - Similar argument is applied to ∇p_t: $\|\nabla \log p_t - \dot{f}_N(\cdot, t)\|_{L^2}^2 \lesssim \dfrac{N^{-2s/d} \log(N)}{\sigma_t^2}$

- A tighter bound on the smooth part $(t > t_*)$

- Useful for W1 bound.
- Smoothness around the edge (A2) is not requires.

$$\|p_t\|_{W_p^k} = \sum_{|\alpha| \le k} \left\| \frac{\partial^\alpha p_t}{\partial x^\alpha} \right\|_{L^p} \lesssim \sigma_t^{-k} \left(\le t_*^{-\frac{k}{2}}\right)$$

$$\Rightarrow \quad \|p_t - \check{f}_{N'}\|_{L^2(X_t)}^2 \lesssim N'^{-2k/d} t_*^{-k} \qquad \text{(take } k = s + 1\text{)}$$

Error decomposition

Score matching loss

$$\mathrm{TV}(X_0, \hat{Y}_{\overline{T}-\underline{T}}) \lesssim \left[\int_{t=\underline{T}}^{\overline{T}} \mathbb{E}_{X_t \sim p_t}[\|\hat{s}(X_t,t) - \nabla \log p_t(X_t)\|^2]\mathrm{d}t\right]^{\frac{1}{2}}$$

$$+ n^{O(1)}\sqrt{\underline{T}} + \exp(-O(\overline{T})) \lesssim n^{-\frac{s}{d+2s}} \log^9 n$$

Truncation loss at $\underline{T}$.

Truncation loss at $\overline{T}$.

$$t_* = N^{-(2-\delta)/d}$$

$$\int_{t=\underline{T}}^{\overline{T}} \mathbb{E}_{X_t}[\|\nabla \log p_t - \hat{s}(\cdot,t)\|^2]\mathrm{d}t$$

$$\frac{\log(\text{covering num})}{n}$$

Bias

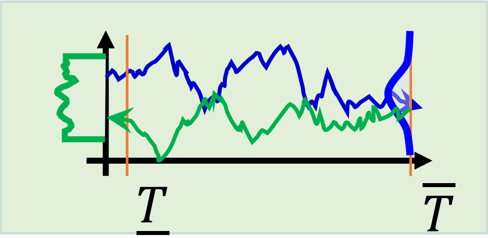
Variance

$$\lesssim \int_{\underline{T}}^{\overline{T}} \frac{N^{-2s/d}}{\sigma_t^2} \log(N)\mathrm{d}t \quad + \quad \frac{N\mathrm{polylog}(N)}{n}$$

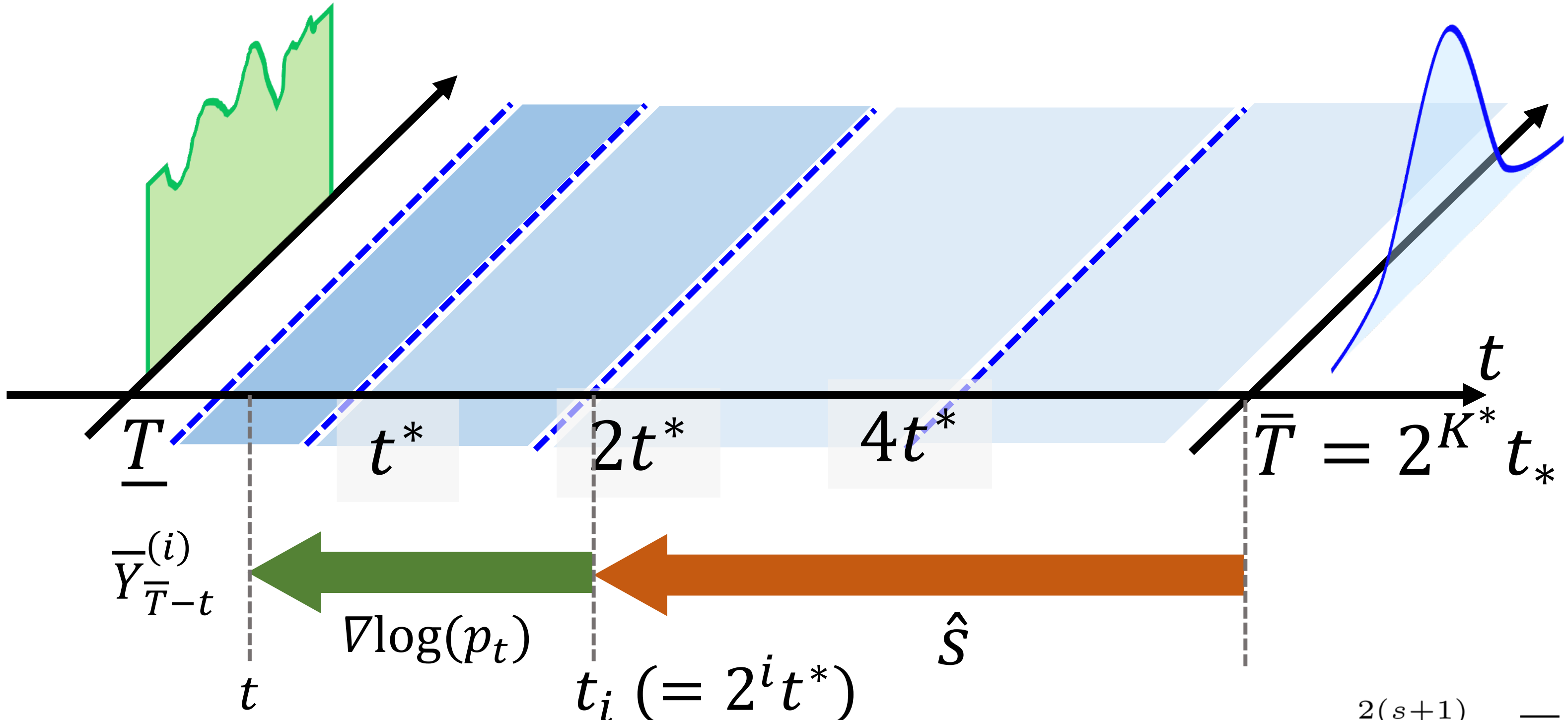
$$\lesssim \left(N^{-2s/d} + \frac{N}{n}\right) \mathrm{polylog}(N)$$

$$N \simeq n^{d/(2s+d)}$$

$$\lesssim n^{-2s/(2s+d)} \mathrm{polylog}(n)$$



Bound for W1 distance



$$\underline{T} = n^{-\frac{2(s+1)}{2s+d}}, \ \overline{T} \simeq \log(n)$$

$$t_* = n^{-\frac{2-\delta}{d+2s}}, \quad t_k = t_* 2^k$$

$$\mathrm{W}_1(X_0, \hat{Y}_{\overline{T}-\underline{T}}) \le \underset{\text{(negligible)}}{\mathrm{W}_1(X_0, X_{\overline{T}})} + \underset{(\exp(-\overline{T}))}{\mathrm{W}_1(X_{\overline{T}}, \bar{Y}^{(K^*)}_{\overline{T}-\underline{T}})} + \sum_{i=1}^{K^*} \underbrace{\mathrm{W}_1(\bar{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \bar{Y}^{(i)}_{\overline{T}-\underline{T}})}$$

$$\boxed{\sqrt{t_{i-1} \int_{t_{i-1}}^{t_i} \mathbb{E}_{X_t}[\| \hat{s}(X_t, t) - \nabla \log p_t(X_t) \|^2] \mathrm{d}t}} \lesssim n^{-\frac{s+1-\delta}{2s+d}}$$

Implementable discretization

$$\min_{s \in \mathrm{DNN}} \frac{1}{n} \sum_{i=1}^{n} \int_{t=\underline{T}}^{\overline{T}} \mathbb{E}_{X_t|X_0=x_i}\left[\|s(X_t,t) - \nabla \log p_t(X_t|x_i)\|^2\right] \mathrm{d}t$$

Finite sample approximation

$$\min_{s \in \mathrm{DNN}} \frac{1}{M} \sum_{j=1}^{M} \|s(x_{t_j,j}, t_j) - \nabla \log p_{t_j}(x_{t_j,j}|x_{i_j})\|^2$$

- $i_j \sim \mathrm{Unif}(\{1,\ldots,\mathrm{n}\})$
- $t_j \sim \mathrm{Unif}([\underline{T},\overline{T}])$
- $x_{t_j,j} \sim p_{t_j}(\cdot\,|x_{i_j})$

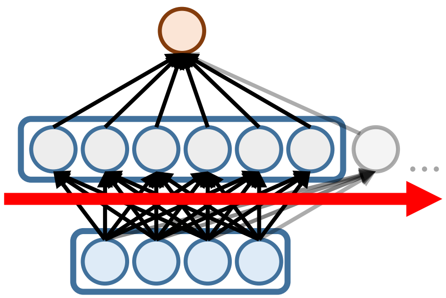
Prop

$$M \gtrsim n \cdot \underline{T}^{-1} = n^{1+\frac{2(s+1)}{2s+d}}$$

is sufficient to attain the same convergence rate.

Mean field limit of 2-layer NN

- 2-layer neural network:



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

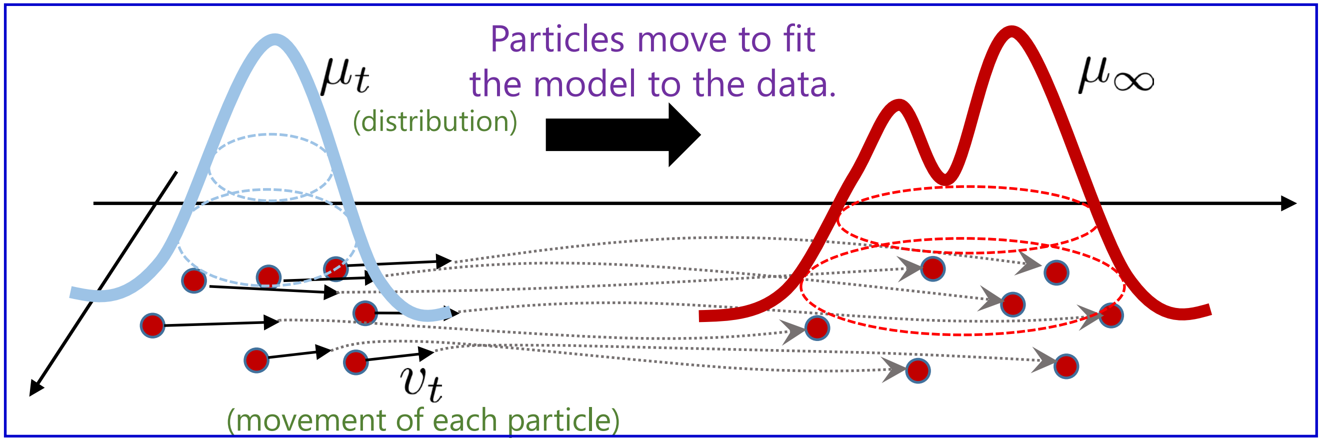
<u>Non-linear</u> with respect to parameters $(r_j, w_j)_{j=1}^{M}$.

- Overparameterization (Mean field limit):

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

<u>Linear</u> with respect to the prob. measure μ .

Optimization of f $\Leftrightarrow$ Optimization of μ



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]:

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

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

$$\mu^* = \mathop{\arg\min}_{\mu \in \mathcal{P}} \underbrace{\int L(x) \mathrm{d}\mu(x)}_{\text{[linear w.r.t. } \mu\text{]}} + \frac{1}{\beta} \mathrm{Ent}(\mu) =: \mathcal{L}(\mu)$$

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

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

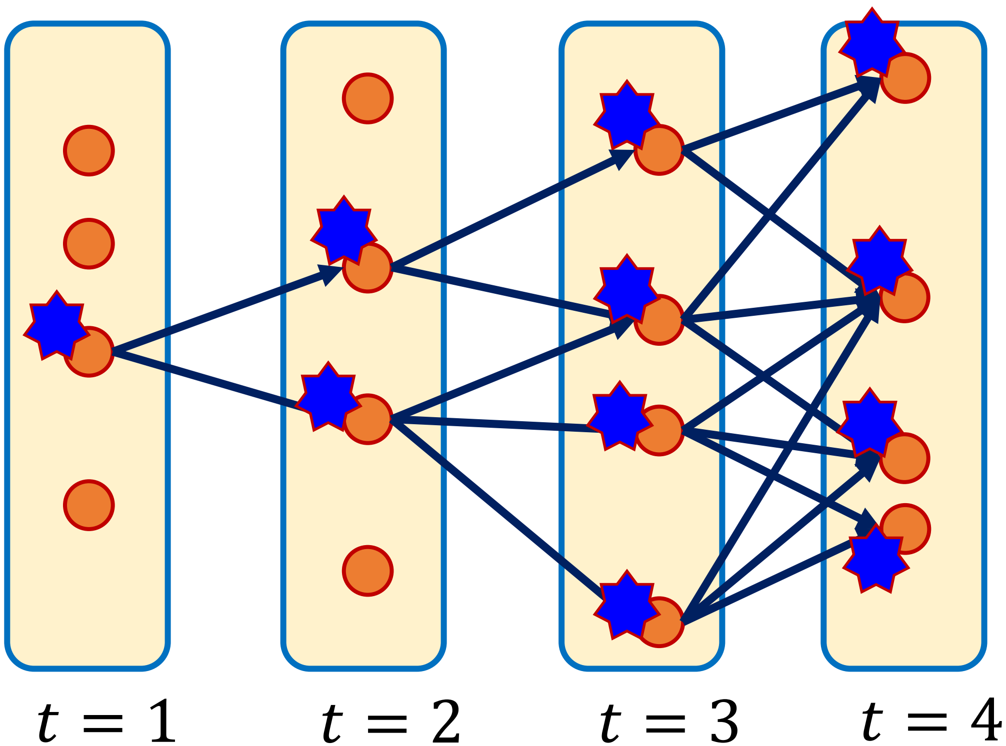
Difficulty

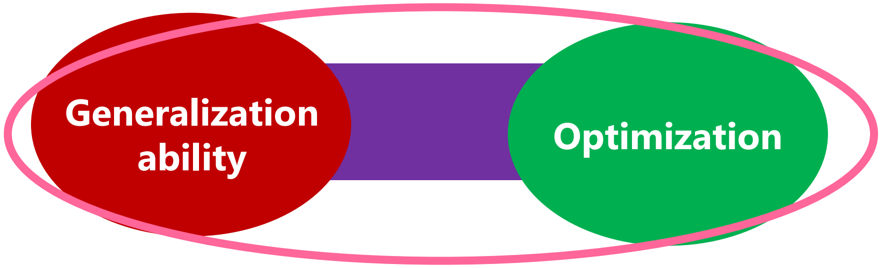
- SDE of interacting particles (McKean, Kac,..., 60′)

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

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

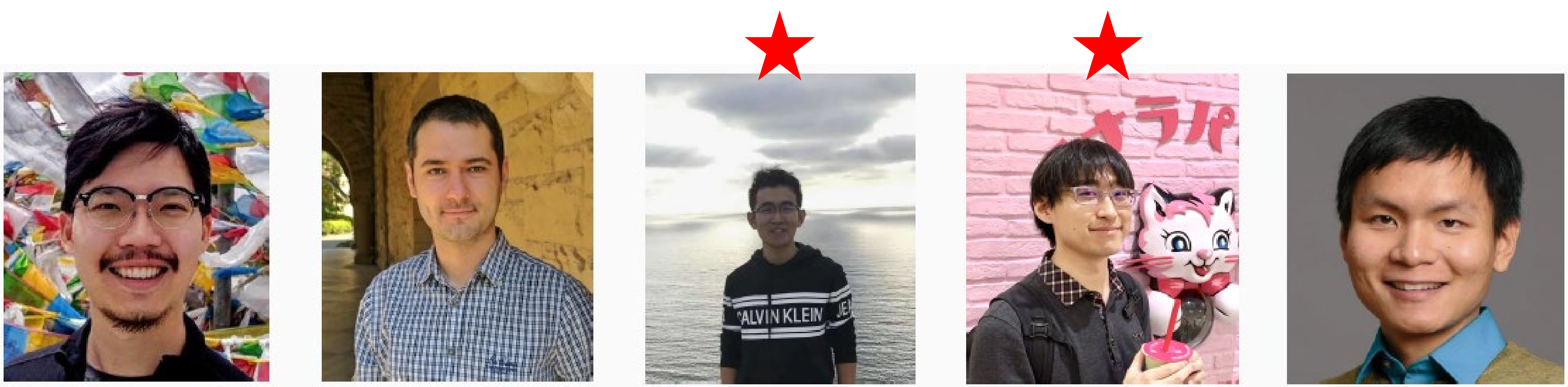
Finite particle approximation error can propagate through time.
→ It is difficult to bound the perturbation uniformly over time.





Feature learning with one-step gradient descent

[Ba, Erdogdu, Suzuki, Wang, Wu, Yang: High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation. NeurIPS2022]



Jimmy Ba   Murat A. Erdogdu   Zhichao Wang   Denny Wu   Greg Yang

Problem setting

Observation model:

$$y_i = f^*(x_i) + \epsilon_i \quad (i = 1, \dots, n)$$

where $x_i \sim N(0, I), \epsilon_i \sim N(0,1)$, and $x_i \in \mathbf{R}^d$.

- We fit 2-layer NN of mean field scaling:

$(\because a_i = O_p(1/\sqrt{N}))$

Mean field regime $O(1/N)$

$$f_{\mathrm{NN}}(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i \sigma(\langle x, w_i \rangle) = \overbrace{\frac{1}{\sqrt{N}}}^{} a^\top \sigma(W^\top x)$$

where $a_i \sim N(0,1/\underbrace{N}_{\text{var}})$ and $W_{ij} \sim N(0,1/\underbrace{d}_{\text{var}})$.

Empirical risk:

$$\mathcal{L}(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Predictive risk:

$$\mathcal{R}(f) = \mathbb{E}[(f^*(X) - f(X))^2]$$

Question: Can we provably improve the predictive risk by gradient descent?

We analyze the risk especially for the single index model:

$$f^*(x) = \sigma^*(\langle x, \beta^* \rangle)$$

Feature learning with optimization guarantee

$$f_{\mathrm{NN}}(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i \sigma(\langle x, w_i \rangle) = \frac{1}{\sqrt{N}} a^\top \sigma(W^\top x)$$

$$W_{k+1} = W_k - \eta \sqrt{N} \nabla_W L(f_{\mathrm{NN}})$$

We consider the **proportional limit** $(n, d, N \to \infty$ with $n/d \to \psi_1, N/d \to \psi_2)$.

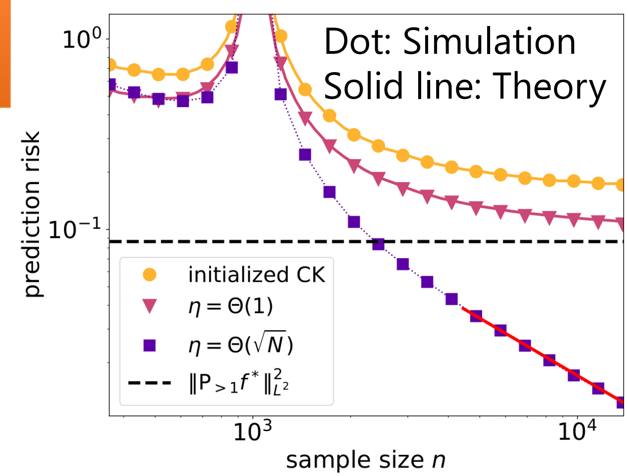
It allows to derive precise risk.

We evaluate predictive risk of **one-step GD.**

Take home message:
GD with Large step-size can outperform **any** random feature model by only one-step update.

[Outline of our result]

- $\boldsymbol{\eta = \Theta(\sqrt{N})}$ can get out of NTK regime and outperform random feature models.
- $\boldsymbol{\eta = \Theta(1)}$ can outperform the initial setting of W.
- $\boldsymbol{\eta = o(1)}$ does not improve the performance.



Ridge regression with RF

Feature learning vs Random feature

Random features (without feature learning):

- **Conjugate kernel at initialization**:

$$\phi_{\mathrm{CK}}(x) = \frac{1}{\sqrt{N}}\sigma(W_0^\top x)$$

Precise asymptotics has been extensively studied. (e.g., [Louart, Liao, and Couillet, 2018; Mei and Montanari, 2019])

- **NTK** (Neural tangent kernel):

$$\phi_{\mathrm{NTK}}(x) = \frac{1}{\sqrt{Nd}}\mathrm{Vec}(\sigma'(W_0^\top x)x^\top)$$

$$\hat{a}_{\mathrm{RF}} = \mathop{\arg\min}_{a \in \mathbb{R}^N} \left\{ \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle a, \phi_{\mathrm{RF}}(x_i)\rangle)^2 + \frac{\lambda}{N}\|a\|^2 \right\}$$

RF ∈ {CK,NTK}

Trained feature:

$$\phi_{\mathrm{CK}^{(t)}}(x) = \frac{1}{\sqrt{N}}\sigma(W_t^\top x)$$

Rank 1 property of first gradient step

Reference

- The gradient G_t can be approximated by rank one matrix.
  $\Rightarrow$ There appears "spike" in the spectral distribution of W_1.

$$G_t = -\frac{1}{n} X^\top \left[\left(\frac{1}{\sqrt{N}} \left(\frac{1}{\sqrt{N}} \sigma(X W_t) a - y \right) a^\top \right) \odot \sigma'(X W_t) \right]$$

(generally, this is not low rank due to the nonlinearlity of σ')

Theorem (Rank one approximation of gradient)

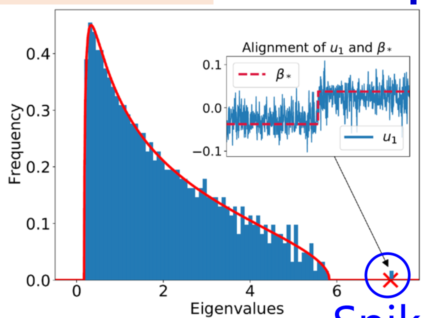
Remember that $G_0 = \frac{1}{\eta\sqrt{N}}(W_1 - W_0) \qquad (\because W_1 = W_0 + \eta\sqrt{N} G_0)$

Let $\mu_1 = \mathbb{E}[z\sigma(z)], \quad \mu_2 = \sqrt{\mathbb{E}[\sigma(z)^2] - \mu_1^2}, \quad \text{where } z \sim \mathcal{N}(0,1).$

Define $A := \frac{\mu_1}{n\sqrt{N}} X^\top y a^\top$ (rank one matrix), then we have

$$\|G_0 - A\| \lesssim \frac{\log^2(n)}{\sqrt{n}} \cdot \|G_0\|$$

with high probability for sufficiently large n, d, N.



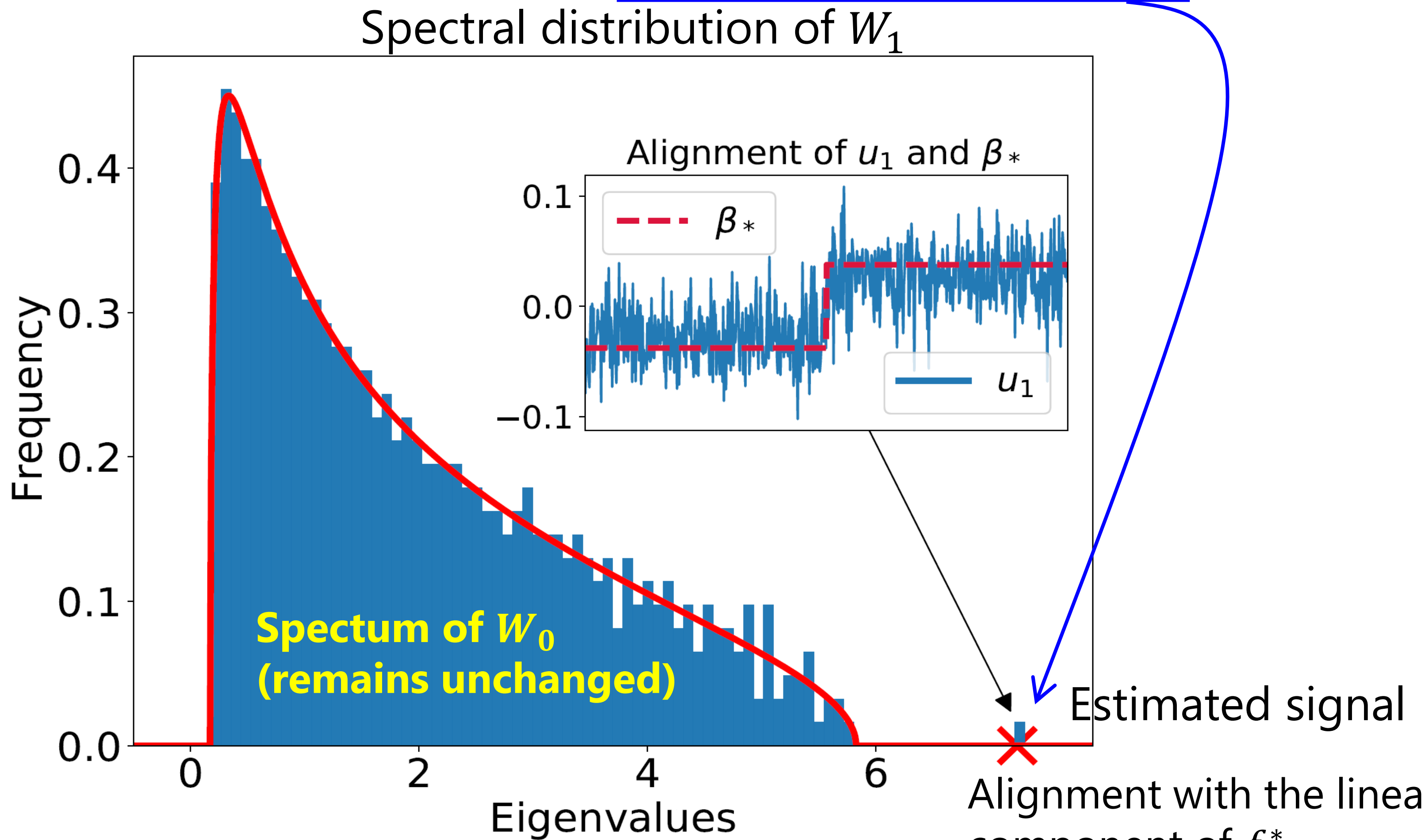
Spike

$W_1 = W_0 + \eta \times$ (rank one matrix).
$\Rightarrow$ For large step size η, spike becomes more dominant.

Effect of large step-size update

Reference

$$W_1 = W_0 + \eta\sqrt{N}\underbrace{(-\nabla_W \mathcal{L}(f_{\mathrm{NN}}^{(0)})/2)}_{\text{Theorem: almost rank 1}}$$



Alignment with the linear component of f^*.
(feature learning)

Limitation of RF

(1) Random feature models and
(2) GD updates with <u>small learning rate</u>
can learn only <u>linear functions</u> in the proportional

[El Karoui (2010); Ghorbani et al. (2019), Hu and Lu (2020), ...] $\mathcal{R}_{\mathrm{XX}}(f) = \mathbb{E}[(f^*(X) - \hat{f}_{\mathrm{XX}}(X))^2]$

Theorem (Lower bound of predictive risk for RF)

If the step size is not large $\eta = \Theta(1)$, then for any finite number steps t, we have

$$\inf_{\lambda>0} \min\{\mathcal{R}_{\mathrm{CK}}(\lambda), \mathcal{R}_{\mathrm{NTK}}(\lambda), \mathcal{R}_{\mathrm{CK}^{(t)}}(\lambda)\} \geq \|P_{>1} f^*\|_{L^2(P_X)}^2 + o_{p,d}(1)$$

Nonlinear part cannot be trained!

$$P_{>1} f^* := (I - P_{\leq 1}) f^*$$

where $P_{\leq 1}$ is the projection operator in $L^2(P_X)$ to the subspace consisting of linear functions and constants.

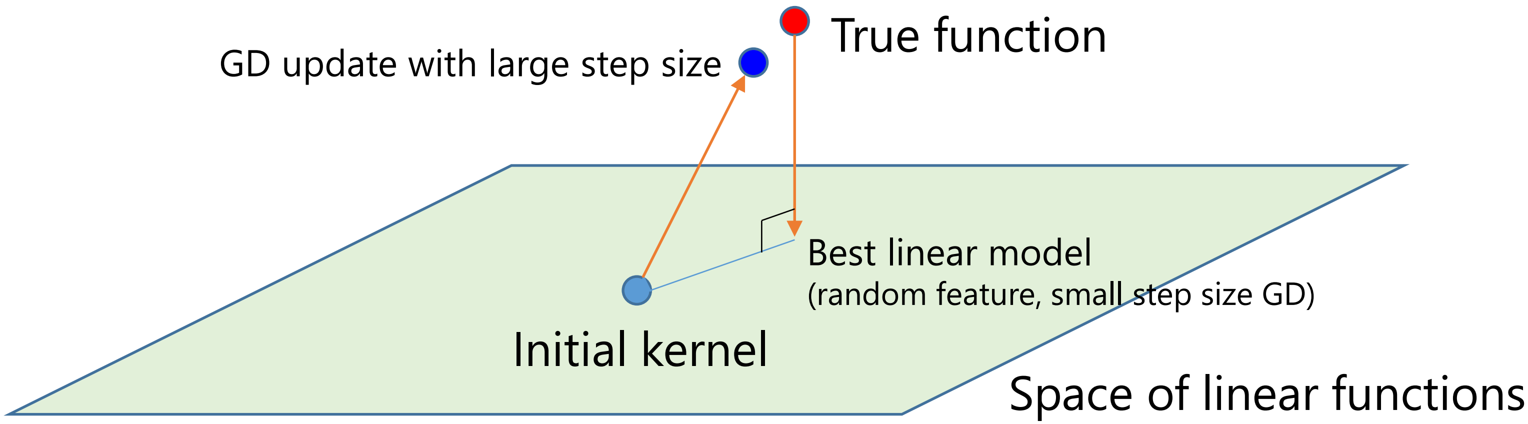
Remark: The same is true for "rotational invariant kernel" [El Karoui (2010)].

This is because in high dimensional setting, a central limit theorem yields

$$a^\top \phi_{\mathrm{CK}}(x) = \frac{1}{\sqrt{N}} a^\top \sigma(W_0^\top x_i) \approx \frac{1}{\sqrt{N}} a^\top (\mu_1 W_0^\top x_i + \mu_2 z)$$

(linear function;
Gaussian equivalence)

Improvement over the Initial CK



- $\eta = \Theta(\sqrt{N})$ (large learning rate):

Known as maximal update parameterization (μP) [Yang and Hu, 2020].

$$\tau^* = \inf_{\eta > 0} \mathbb{E}_{\xi_1 \sim N(0,1)} \left[\sigma^*(\xi_1) - \mathbb{E}_{\xi_2 \sim N(0,1)} [\sigma(\eta \xi_1 + \xi_2)] \right]$$

(measure for model misspecification)

$f^*(x) = \sigma^*(\langle x, \beta^* \rangle)$ is assumed.

- $\tau^* = 0$ if $\sigma = \sigma^* = \mathrm{erf}$.
- $\tau^* \ll 1$ if $\sigma = \sigma^* = \tanh$.

$$\boxed{\mathcal{R}_{W_1}(\lambda) \le 16\tau^* + C(\sqrt{\tau^*}\psi_1^{-1/2} + \psi_1^{-1}) + o_p(1)}$$

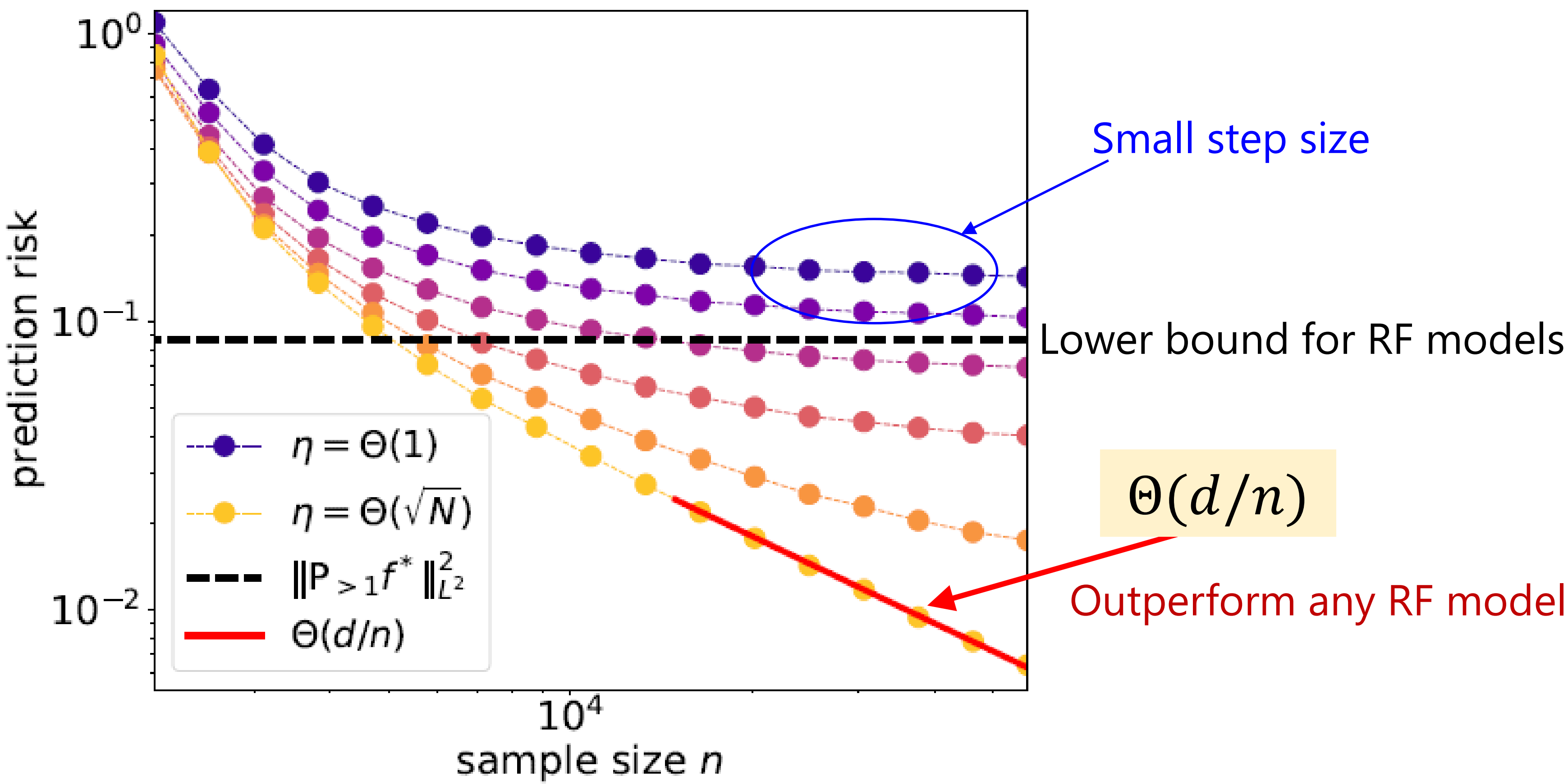
$n/d \to \psi_1,\ N/d \to \psi_2$

Large learning rate yields feature learning and can be better than the small step size regime if $\tau^* \ll \|P_{>1} f^*\|^2$.

Implications

Corollary

If $\sigma = \sigma^* = \mathrm{erf}$, then $\tau^* = 0$.
In particular, we have $R_{W_1}(\lambda) = \Theta(\psi_1^{-1}) = \Theta(d/n)$.



Predictive risk of ridge regression on CK obtained by one step GD (empirical simulation, $d = 1024$): brighter color represents larger step size scaled as $\eta = N^{\alpha}$ for $\alpha \in [0,1/2]$. We chose $\sigma = \sigma^* = \mathrm{erf}$, $\psi_2 = 2, \lambda = 10^{-3}$, and $\sigma_\epsilon = 0.1$.