Fast learning rate of neural tangent kernel learning and nonconvex optimization by infinite dimensional Langevin dynamics in RKHS

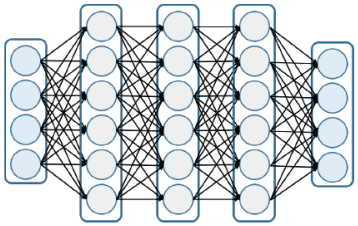
Taiji Suzuki
University of Tokyo
Deep Learning Theory Team, AIP-RIKEN

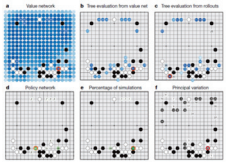
19th/Feb/2020
Functional Inference and Machine Intelligence
@EURECOM

Theory of Deep Learning

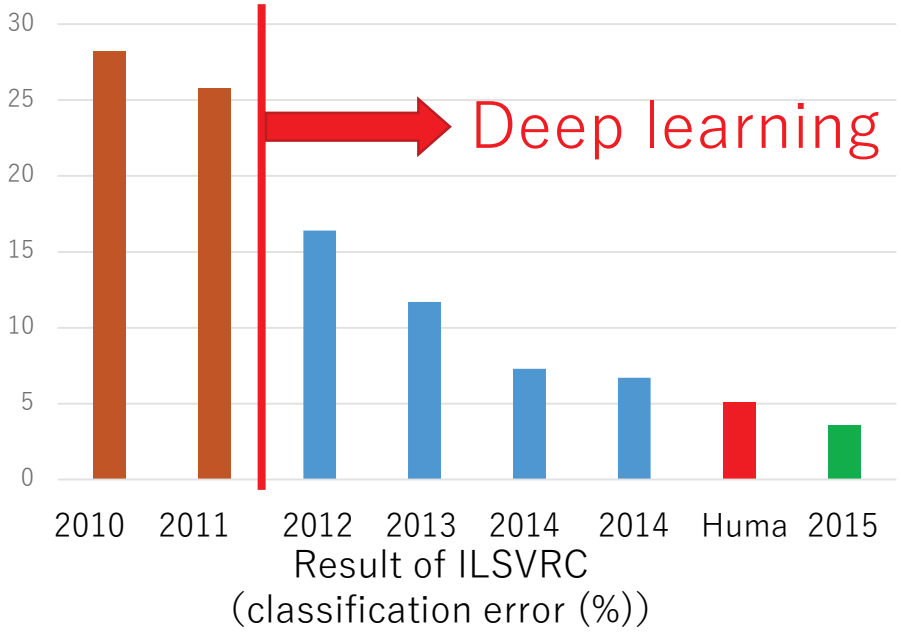
Deep learning

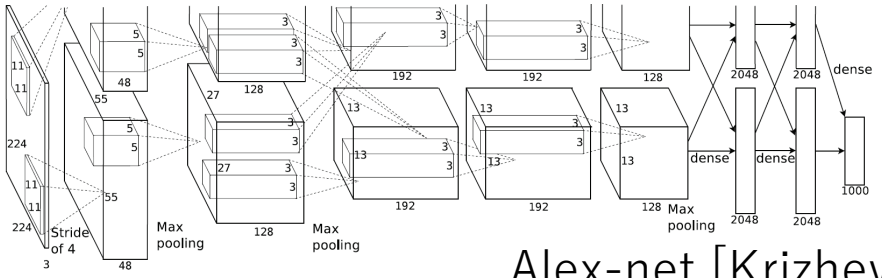
- High performance
- Applied to services in several industries:
  Google Deepmind, Facebook AI Lab., Baidu, …





- **High performance in several applications**
- **We need more theory.**





Alex-net [Krizhevsky, Sutskever + Hinton, 2012]

Theoretical issues

Contribution from our group

Representation ability

How large is the DNN model?

- Approximation theory on Besov spaces
- Integral representation
- Power of graph-CNN

Generalization ability

How well can DL generalize from finite observations?

- Gen error bound on Besov space
- Degree of freedom based on kernel analysis
- Compression based bound
- Function space with sparsity

Optimization

How fast can we find the optimal parameter?

- Particle gradient descent
- Convergence guarantee of gradient descent
- Parallel computation

Outline

- Optimization of shallow neural networks
  - Neural tangent kernel
  - Fast learning rate: regression&classification

  Joint work with Atsushi Nitanda.

- Nonconvex optimization on a Hilbert space
  - Infinite dimensional stochastic gradient Langevin dynamics

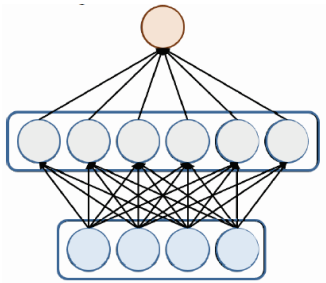
  Joint work with Boris Muzellec (CREST, ENSAE,), Kanji Sato (U-Tokyo), Mathurin Massias (INRIA).

Neural tangent kernel and fast learning rate

Joint work with Atsushi Nitanda.

Universal Approximator

$$f(x) = \sum_{j=1}^{M} a_j \eta(w_j^\top x + b_j)$$

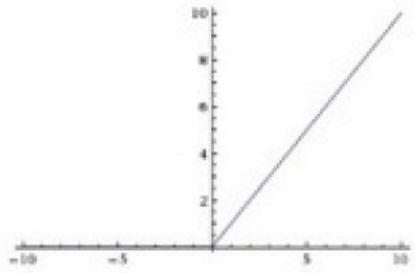


Taking $M \to \infty$, we can approximate “any function” with “any precision.”

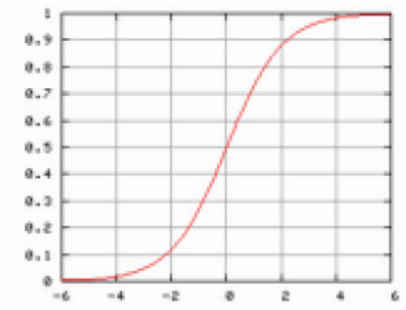
M can be sigmoidal or ReLU.

Activation functions:

ReLU: $\eta(u) = \max\{u, 0\}$



Sigmoid: $\eta(u) = \frac{1}{1+\exp(-u)}$



K is any compact set.

Gradient descent

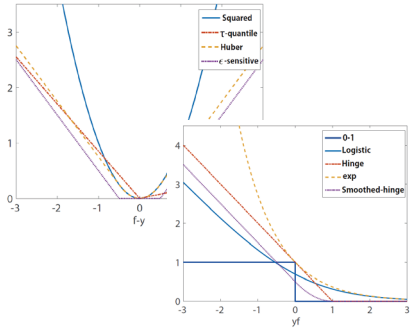
$$f_W(x) = \sum_{j=1}^{M} a_j \eta(w_j^\top x)$$

fix　　optimize

$$\min_W \hat{L}(W) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(f_W(x_i))$$

ℓ_i: loss function

- Regression: squared loss ($\ell_i(f_W(x_i)) = (y_i - f_w(x_i))^2$)
- Classification: convex surrogate loss (e.g., logistic, smoothed hinge, ...)



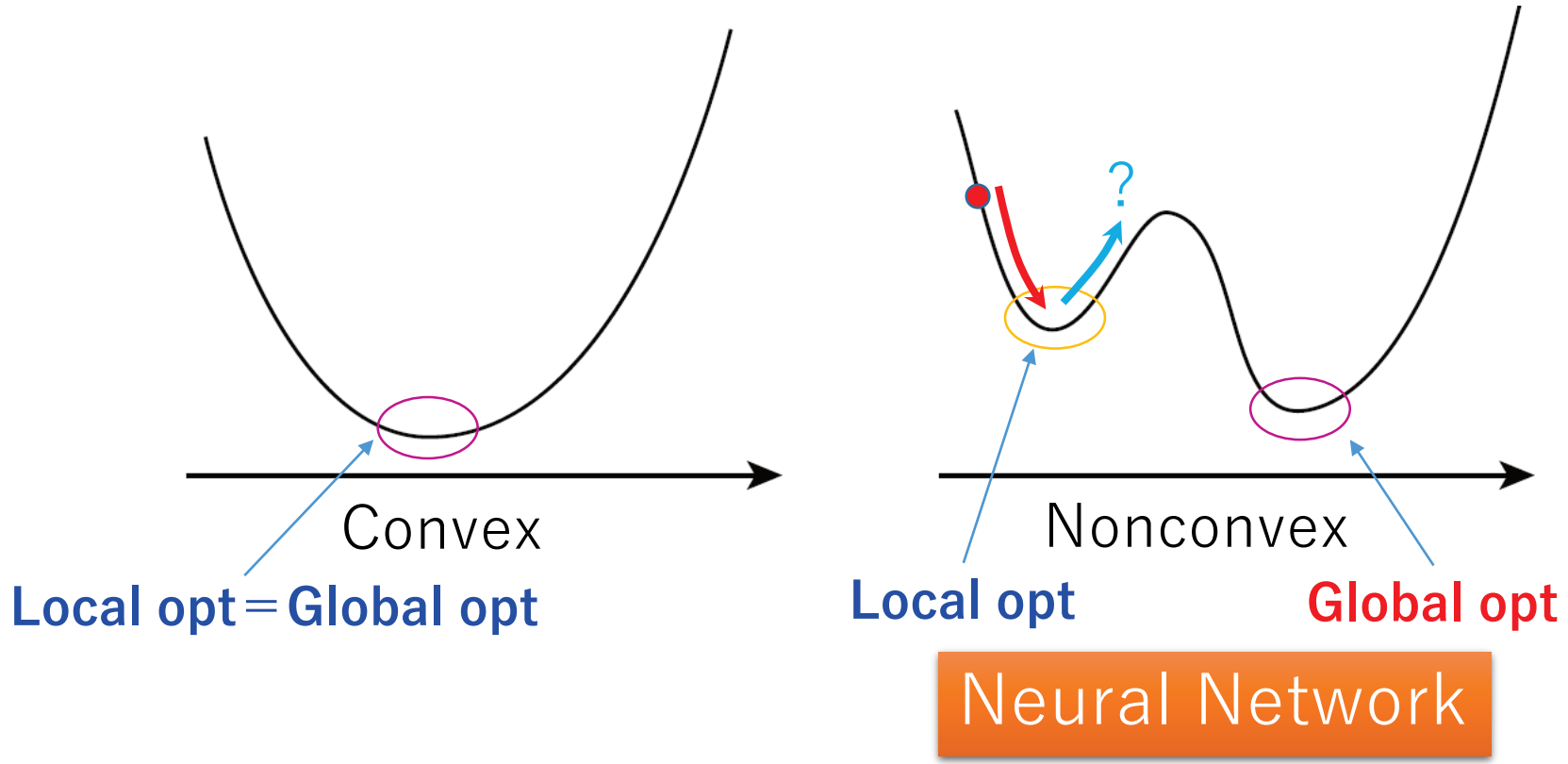
Gradient descent: $W^{(t+1)} = W^{(t)} - \alpha \nabla_W \hat{L}(W)$

Stochastic gradient descent: $W^{(t+1)} = W^{(t)} - \alpha \frac{1}{B} \sum_{i \in I_t} \nabla_W \ell_i(f_W(x_i))$

Difficulty of DL optimization

Non-convex objective

Convex $\theta f(x) + (1-\theta) f(y) \geq f(\theta x + (1-\theta) y) \quad (\forall x, y \in \mathbb{R}^p, \theta \in [0,1])$

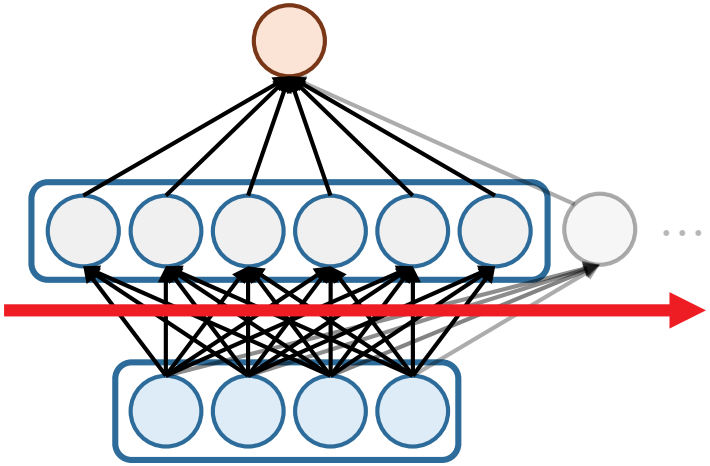


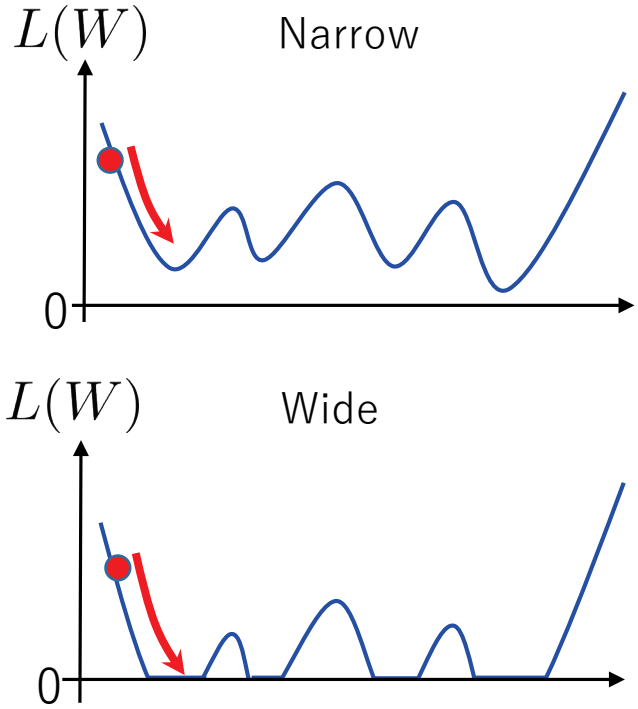
Training "**narrow**" networks is NP-complete:

- Judd (1988), Neural Network Design and the Complexity of Learning.
- Blum&Rivest (1992), Training a 3-node neural network is np-complete.

Over-parametrization

For a very wide network, local optimal solutions become global optimal.





Since the freedom is increased, the initial solution can be near the global opt (complete fit).

- How fast does it converge?
- How well does it generalize?

(sub-level set is connected [Bandeira,Venturi&Bruna,2019])

Two scaling regimes

$$f_W(x) = \sum_{j=1}^{M} a_j \eta(w_j^\top x)$$

- Neural Tangent Kernel regime (lazy learning)
  - $a_j = O(1/\sqrt{M})$

  [Jacot+ 2018][Du+ 2019][Arora+ 2019]
- Mean field regime
  - $a_j = O(1/M)$

  [Nitanda &Suzuki 2017], [Chizat & Bach 2018], [Mei+ 2018]
- w_j is randomly initialized.

The dynamics changes depending on the scaling.
→ Affect convergence and generalization error.

Neural Tangent Kernel

Let us consider a continuous time dynamics.

Model : $f_W(x) = \sum_{j=1}^{M} a_j \eta(w_j^\top x)$

- a_j is fixed
- w_j is trained

[Jacot, Gabriel&Hongler, NeurIPS2018]

$$\frac{\mathrm{d}w_j}{\mathrm{d}t} = -\nabla_{w_j} \hat{L}(f_W) \qquad \text{(Gradient descent, GD)}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \ell_i'(f_W(x_i)) a_j \nabla_{w_j} \eta(w_j^\top x_i)$$

$$\frac{\mathrm{d}f_W(x)}{\mathrm{d}t} = \sum_{j=1}^{M} a_j \nabla_{w_j}^\top \eta(w_j^\top x) \frac{\mathrm{d}w_j}{\mathrm{d}t}$$

$O(1/M)$

$$= -\frac{1}{n} \sum_{i=1}^{n} \underbrace{\left(\sum_{j=1}^{M} a_j^2 \nabla_{w_j}^\top \eta(w_j^\top x) \nabla_{w_j} \eta(w_j^\top x_i) \right)}_{k_W(x, x_i)} \ell_i'(f_W(x_i))$$

residual

(Functional gradient)

Neural Tangent Kernel

Decrease of objective

$$\frac{\mathrm{d}\hat{L}(f_W)}{\mathrm{d}t} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}f_W(x_i)}{\mathrm{d}t} \ell_i'(f_W(x_i))$$

$$= -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \ell_i'(f_W(x_i)) \underbrace{k_W(x_i, x_j)}_{(K_W)_{i,j}} \ell_j'(f_W(x_j))$$

$$\left(= -\frac{1}{n^2} \|\nabla_f \hat{L}(f_W)\|_{K_W}^2 \right)$$

$$\leq 0$$

Unless the gram matrix K_W degenerates, it can find a descent direction.

Fact [Du et al., 2018; Allen-Zhu, Li & Song, 2018]

- For random initialization, $K_{W(0)} \succ \epsilon I$ holds w.h.p.
- During optimization, the gram matrix keeps positive definiteness.

→ **Linear convergence** $(\exp(-t))$

Optimization in NTK regime

We randomly initialize as:

- $a_j \sim (\pm 1)\frac{1}{\sqrt{M}}$ ($+, -$ is generated evenly)
- $w_j \sim N(0, I)$

$$f_W(x) = \sum_{j=1}^{M} a_j \eta(w_j^\top x)$$

Theorem [Arora et al., 2019]

For regression problem (squared loss),
if $M = \Omega(n^2 \log(n)/\lambda_{\min})$, GD can find a solution with gen. error $\sqrt{\boldsymbol{y}^\top (H^\infty)^{-1} \boldsymbol{y}/n}$.

See also[Du et al., 2018; Allen-Zhu, Li & Song, 2018; Li & Liang, 2018]

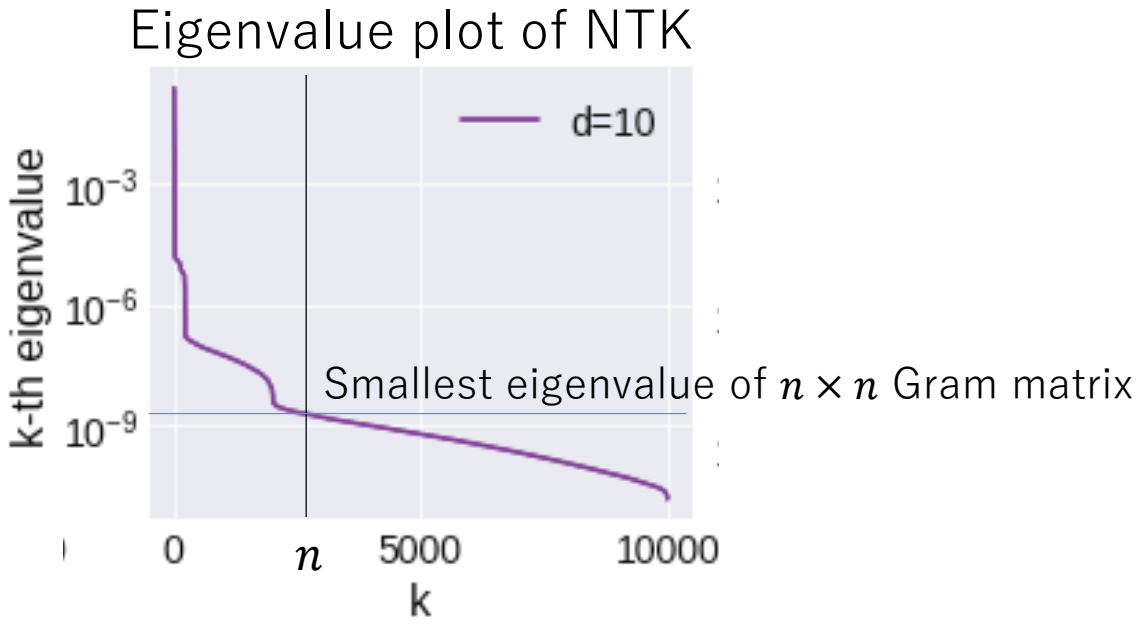
- Linear convergence to a perfect-fit solution.
- Bounded Rademacher complexity.

Issue: inverse of smallest eigenvalue

- **Slow learning rate**.
- Perfect fit is a too strong requirement for **noisy case**.

Spectral bias

- Overparameterization is good for optimization.
- How about generalization?
  The number of parameters is huge $(p \gg n)$...



- Smallest eigenvalue of gram matrix is very small.
- Eigenvalues decay in polynomial order.
  → Spectral bias: beneficial for generalization.

Kernel smoothing view point

- Frechet derivative: $\nabla_f \hat{L}(f)$

$$\nabla_f \hat{L}(f) = (\ell_i'(f(x_i)))_{i=1}^n$$

$$\hat{L}(f+h) = \hat{L}(f) + \langle \nabla_f \hat{L}(f), h \rangle_{L_2(P_n)} + o(\|h\|_{L_2(P_n)}^2)$$

- Smoothing integral operator:

$$T_k f(x) := \int k(x, x') f(x') \mathrm{d}P_n(x')$$

$$T_{k_W} \phi_j = \mu_j \phi_j$$

- NTK gradient as a smoothed functional gradient:

$$\frac{\mathrm{d}f_W}{\mathrm{d}t} = -T_{k_W} \nabla_f \hat{L}(f_W)$$

$$\left(= -\frac{1}{n} \sum_{i=1}^n k_W(\cdot, x_i) \ell_i'(f_W(x_i))\right)$$

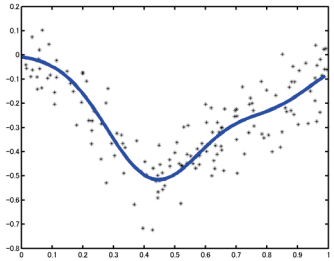
If k_W has a small eigenvalue on high frequency components, then T_{k_W} works as a smoothing operator. → inductive bias.

SGD (our setting)

Model:

$$f_{a,W}(x) = \frac{1}{\sqrt{M}} \sum_{j=1}^{M} a_j \eta({w_j}^\top x)$$

(We train both of first and second layers)



Objective:

$$L(a, W) = \mathbb{E}[(Y - f_{a,W}(X))^2] + \frac{\lambda}{2}(\|a - a^{(0)}\|^2 + \|W - W^{(0)}\|_{\mathrm{F}}^2)$$

Population risk　　Distance from initial value

$$Y = f^*(X) + \epsilon \quad \text{(observation with noise)}$$

Averaged Stochastic Gradient Descent

for $t = 0$ **to** $T - 1$ **do**

Randomly draw a sample $(x_t, y_t) \sim \rho$

Perform SGD update for all $j \in \{1, \ldots, M\}$:

$$a_j^{(t+1)} = a_j^{(t)} - \alpha_t[\nabla_a \ell(y_t, f_{a^{(t)},W^{(t)}}(x_t)) + \lambda(a^{(t)} - a^{(0)})]$$

$$W_j^{(t+1)} = W_j^{(t)} - \alpha_t[\nabla_W \ell(y_t, f_{a^{(t)},W^{(t)}}(x_t)) + \lambda(W^{(t)} - W^{(0)})]$$

end for

Return $\bar{a}^{(T)} = \frac{1}{T}\sum_{t=0}^{T-1} a^{(t)}$, $\bar{W}^{(T)} = \frac{1}{T}\sum_{t=0}^{T-1} W^{(t)}$.

Assumptions

NTK corresponding to 2-layer training:

$$k_\infty(x,x') = \mathbb{E}_{w^{(0)}}[\eta(w^{(0)\top}x)\eta(w^{(0)\top}x')] + \mathbb{E}_{w^{(0)}}[\eta'(w^{(0)\top}x)\eta'(w^{(0)\top}x')x^\top x]$$

Integral operator:

$$T_{k_\infty} f(x) = \int k_\infty(x,x') f(x') \mathrm{d}P_X$$

population

Spectral decomp: $T_{k_\infty}\phi_j = \mu_j\phi_j$, $k_\infty(x,x') = \sum_{j=1}^{\infty} \mu_j\phi_j(x)\phi_j(x')$

- $f^*(x) = \mathrm{E}[Y|X=x]$ can be expressed as
  $$T_{k_\infty}^r h = f^*$$
  for $h \in L_2(P_X)$, and $0 < r$.
- Eigenvalue decay condition:
  $$\mu_j = \mathrm{O}(j^{-\beta}).$$

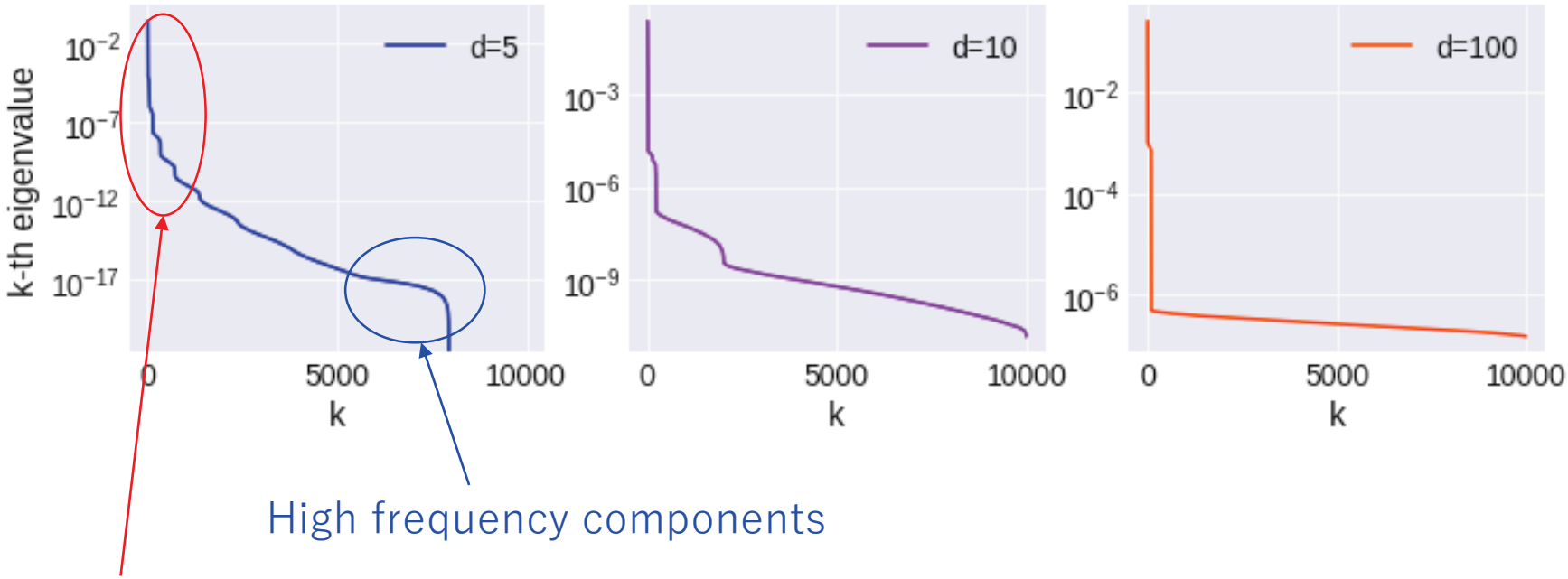
Smoothness of the true function.

"Complexity" of the kernel function.

Standard assumption in analyzing kernel ridge regression; e.g., Dieuleveut et al. (2016); Caponnetto and De Vito (2007).

Inductive bias of NTK

Numerical verification of eigenvalue decay



High frequency components

Low frequency components

See also Cho&Saul (2009), Xie,Liang&Song (2017), Bietti&Marial (2019) for inductive bias.

Fast convergence for regression

[Nitanda&Suzuki: Fast Convergence Rates of Averaged Stochastic Gradient Descent under Neural Tangent Kernel Regime, 2020.]

SGD in NTK regime achieves the “fast learning rate.”

→ Smoothing effect of NTK.

Thm (fast rate)

For $\lambda = T^{-\beta/(2r\beta+1)}$,

Eigen-value decay rate of NTK

$$\mathbb{E}[\|f_T - f^*\|_{L_2}^2] \leq \epsilon_M + O(T^{-\frac{2r\beta}{2r\beta+1}})$$

Converges to 0 as $M \to \infty$

Fast learning rate (existing rate: $O(1/\sqrt{T})$)

→ $T^{-\frac{2r\beta}{2r\beta+1}}$ is the **minimax optimal rate.**

Multiple Kernel Learning

$$k_\infty(x, x') = \mathbb{E}_{w^{(0)}}[\eta(w^{(0)\top}x)\eta(w^{(0)\top}x')] + \mathbb{E}_{w^{(0)}}[\eta'(w^{(0)\top}x)\eta'(w^{(0)\top}x')x^\top x]$$

Kernel for 2nd layer — Kernel for 1st layer

Partly reflects the flexibility of neural networks.

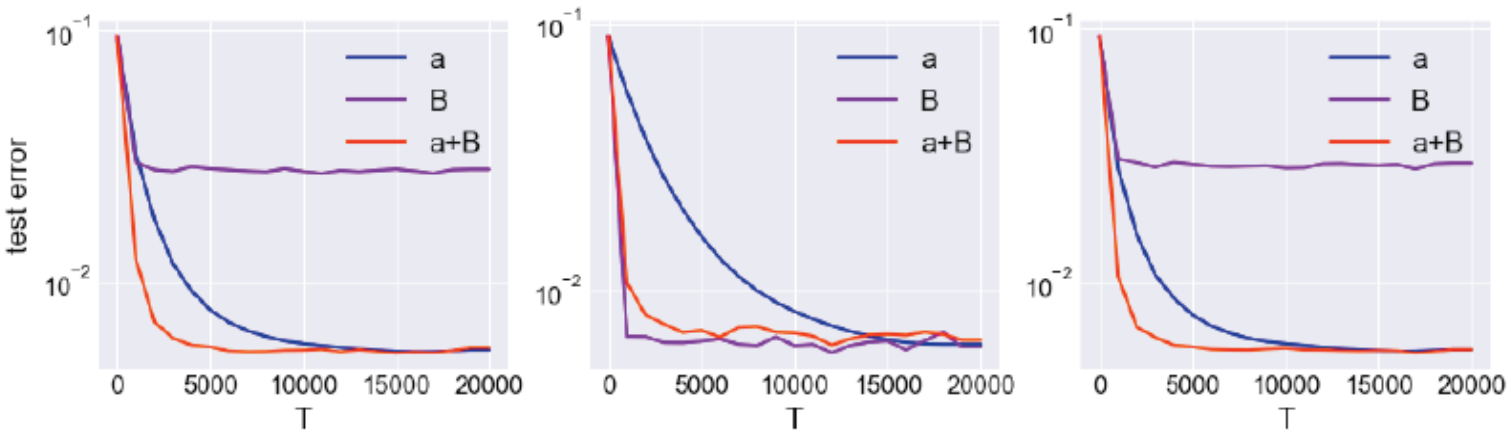
Numerical evaluation of MKL effect

f^* is in

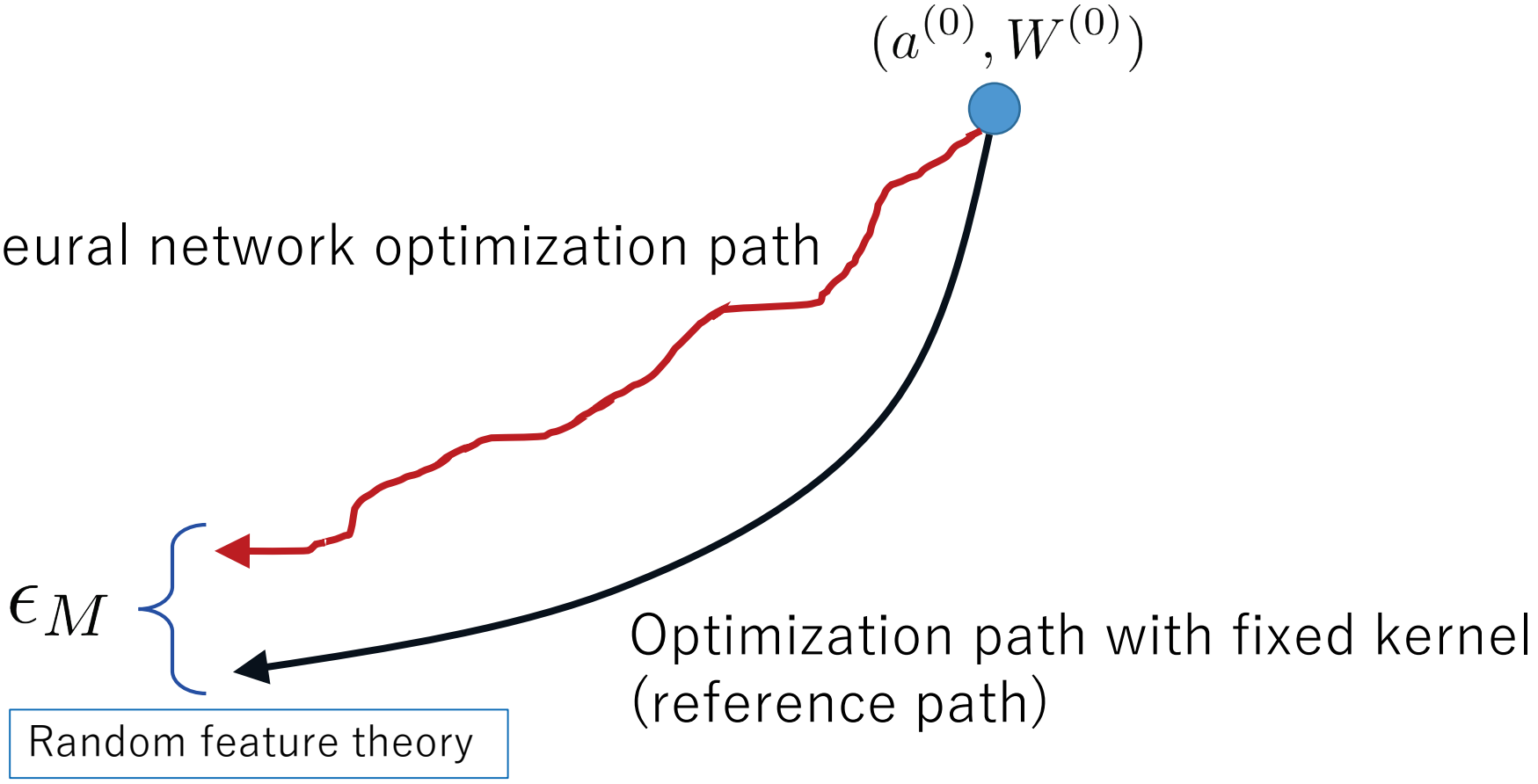
RKHS w.r.t. NTK for the 1st layer.

RKHS w.r.t. NTK for the 2nd layer.

RKHS w.r.t. NTK for the both layer.



Proof idea

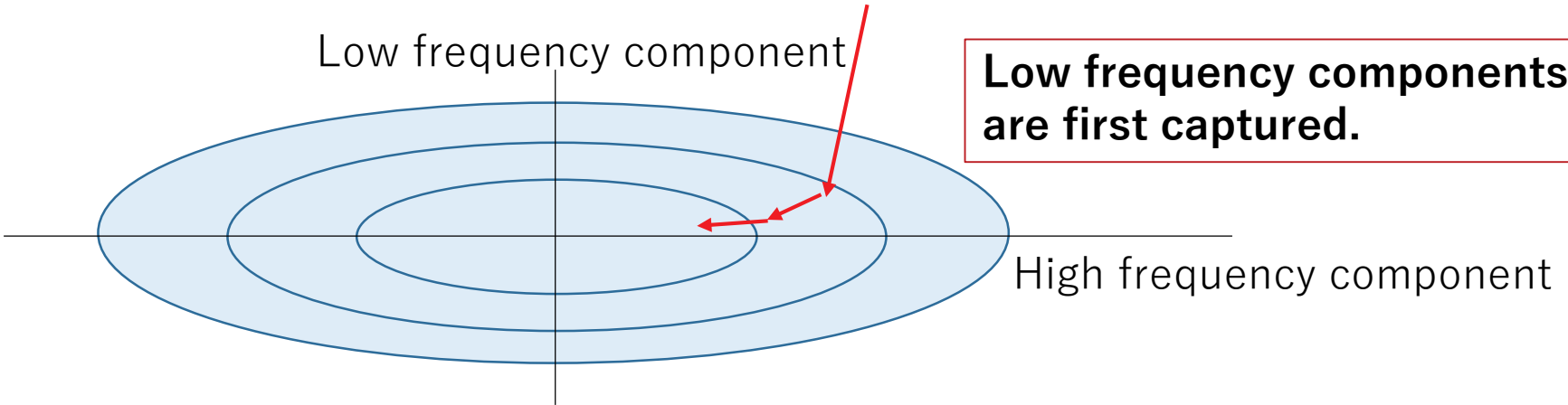


In NTK regime, the optimization path of neural network is close to the optimization path on RKHS with fixed kernel.
[Weinan,Ma&Wu (2019); Arora et al., (2019)]

- Difference from usual NTK analysis:
  - ➢Positive definiteness is not used.
  - ➢Perfect fitting is not used.
  - ➢Linear convergence is not required.

- With a stronger condition on the true, the optimal rate is obtained. $T_{k_\infty}^r h = f^*$

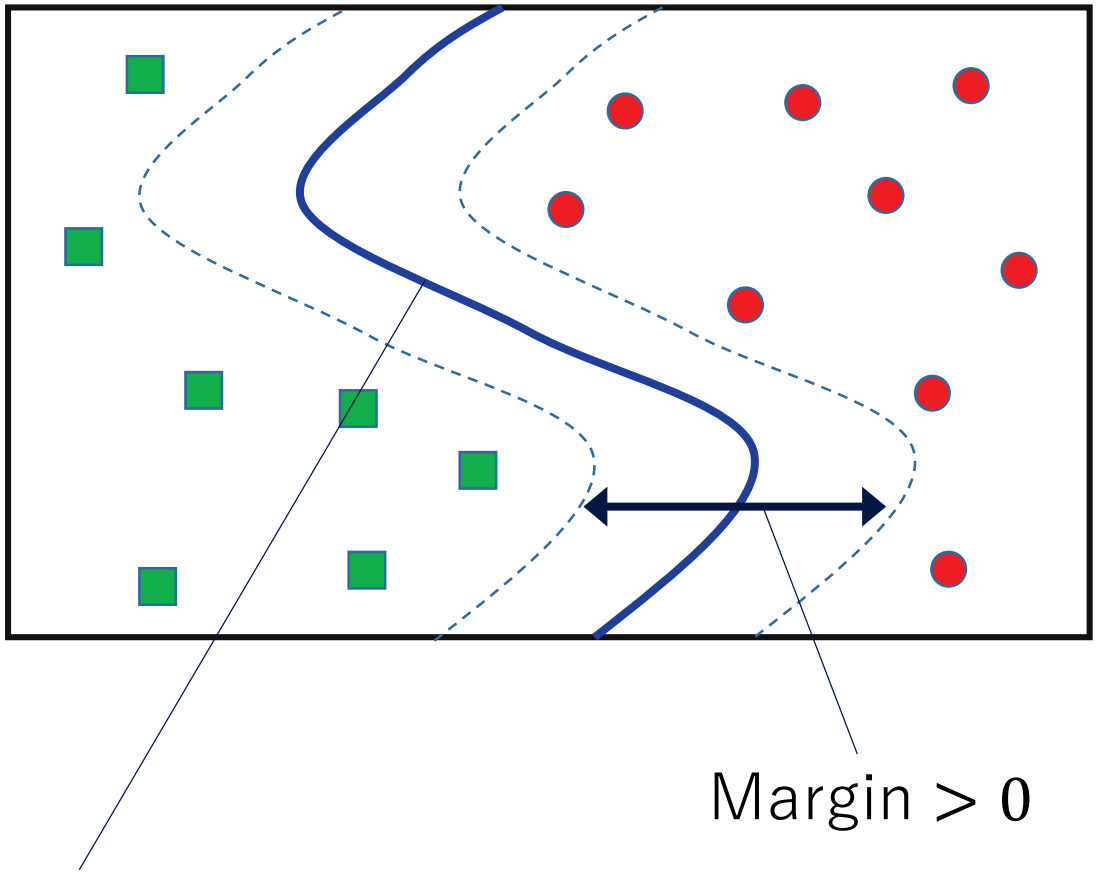
(usual bound) $\frac{1}{\sqrt{T}}$ $\rightarrow$ $T^{-\frac{2r\beta}{2r\beta+1}}$ (optimal rate)



Low frequency components are first captured.

Classification

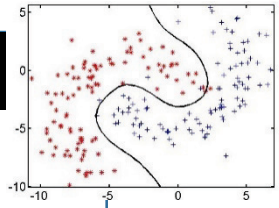
Setting



Bayes optimal classifier is in RKHS corr. to NTK.

Fast convergence for classification

[Atsushi Nitanda, Geoffrey Chinot, and Taiji Suzuki: Refined Generalization Analysis of Gradient Descent for Over-parameterized Two-layer Neural Networks with Smooth Activations on Classification Problems. arXiv:1905.09870.]



Theorem

If the width is $M = \Omega(n^{1/4})$, then $T = \mathrm{O}(n^{1/2})$ iteration of gradient descent achieves,

$$\text{Class. error} \leq n^{-\frac{1}{4}}.$$

Requirement on the width is much improved:

$\Omega(n^2)$ (regression), $\Omega(n^{3.5})$(classification) → $\Omega(n^{1/4})$

Existing work [Cao&Gu:arXiv:1902.01384] **Ours**

Pros: We don't use the positive definiteness of the Gram matrix.
Cons: We should assume "margin condition" – a strong assumption.

	Activation	Separability	M (width)	T	Gen. Error
Allen-Zhu et al. (2018)	ReLU	Smooth target	$n^{2.5}$	$n^{1/2}$	$n^{-1/4}$
Cao&Gu (2019)	ReLU	ReLU NN	$n^{3.5}$	$n^{1/2}$	$n^{-1/4}$
Ours	Smooth	RKHS of NTK	$n^{1/4}$	$n^{1/2}$	$n^{-1/4}$

Assumptions

1. Smoothness: the activation function is in C^2-class:

$$\|\eta'\|_\infty,\ \|\eta''\|_\infty < \infty$$

2. Symmetric initialization:

$$a_j = -a_{j+M/2},\ w_j = w_{j+M/2}\quad (j = 1, \dots, M/2)$$

3. **Separability** (most essential):

$$\exists \rho > 0, \exists v : \mathbb{R}^d \to \mathbb{R}^d \ \text{ s.t. } \ \|v(w)\| \le 1,$$

$$\mathbb{E}_{w \sim N(0,I)}[y \eta'(w^\top x) x^\top v(w)] \ge \rho$$

$$\forall (x, y) \in \mathrm{supp}(P_{XY})$$

The RKHS corresponding to the neural tangent kernel can classify perfectly with margin ρ.

⇒ We don't need positive definiteness of the gram matrix!

Remark

- Recently, Ji & Telgarsky (2019) showed stronger results based on our analysis:

  With $M = \text{poly} - \log(\text{n})$, ReLU activation,

  expected class. Error $\leq \frac{1}{\sqrt{n}}$,

  after $T = \mathrm{O}(\sqrt{n})$ iterations.

Infinite dim non-convex optimization by Langevin dynamics

Joint work with Boris Muzellec (CREST, ENSAE,), Kanji Sato (U-Tokyo), Mathurin Massias (INRIA).

Non-convex optimization

- NTK is essentially convex optimization on RKHS.
- We need more non-convex optimization.

→ Langevin dynamics on a Hilbert space

Non-convex optimization is important not only for deep learning but also for robust classification, Bayesian optimization (and so on).

GLD/SGLD

- Stochastic Gradient Langevin Dynamics

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

β: Inverse temperature

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

Stationary distribution : $\pi \propto \exp(-\beta L(X))$

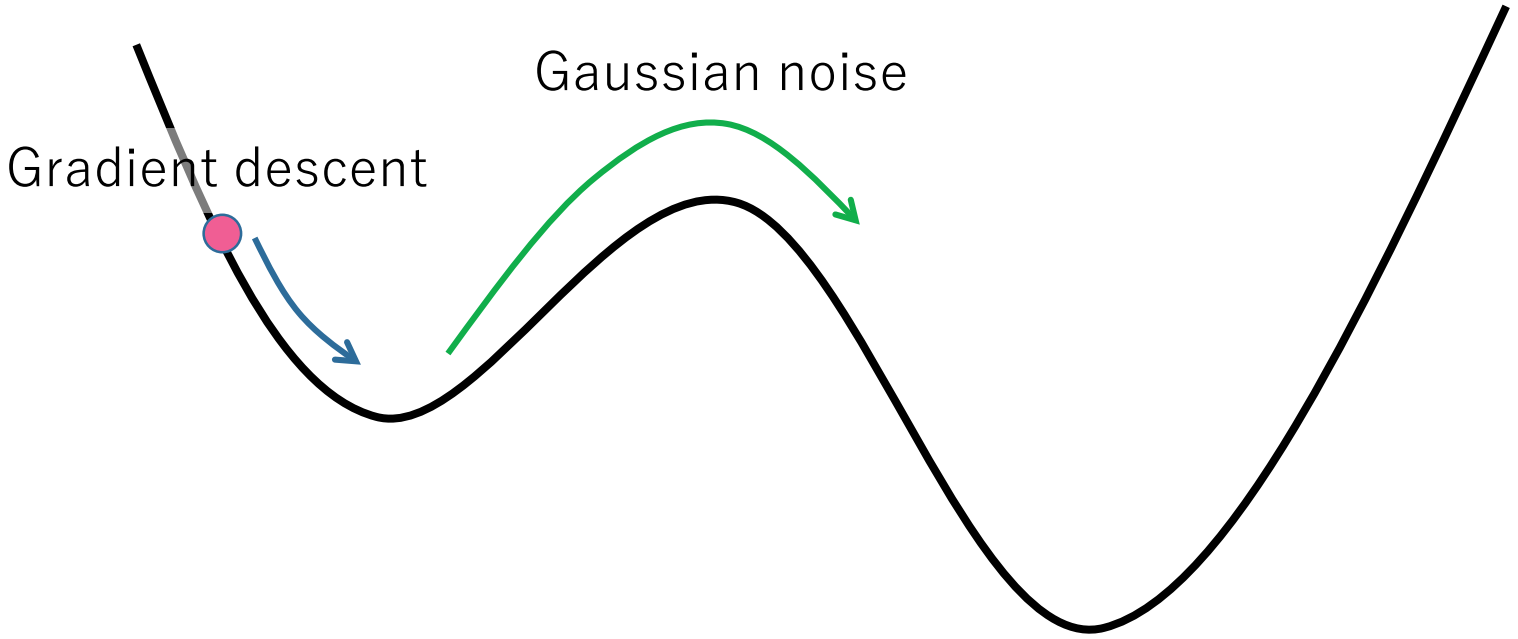
Discretization [Gelfand and Mitter (1991); Borkar and Mitter (1999); Welling and Teh (2011)]

GLD: $X_{t+1} = X_t - \eta\nabla L(X_t) + \sqrt{2\eta\beta^{-1}}\xi_t$ (Euler-Maruyama scheme)

$\xi_t \sim \mathrm{N}(0, I)$

SGLD: $X_{t+1} = X_t - \eta\frac{1}{|I_B|}\sum_{i\in I_B}\nabla\ell_i(X_t) + \sqrt{2\eta\beta^{-1}}\xi_t$

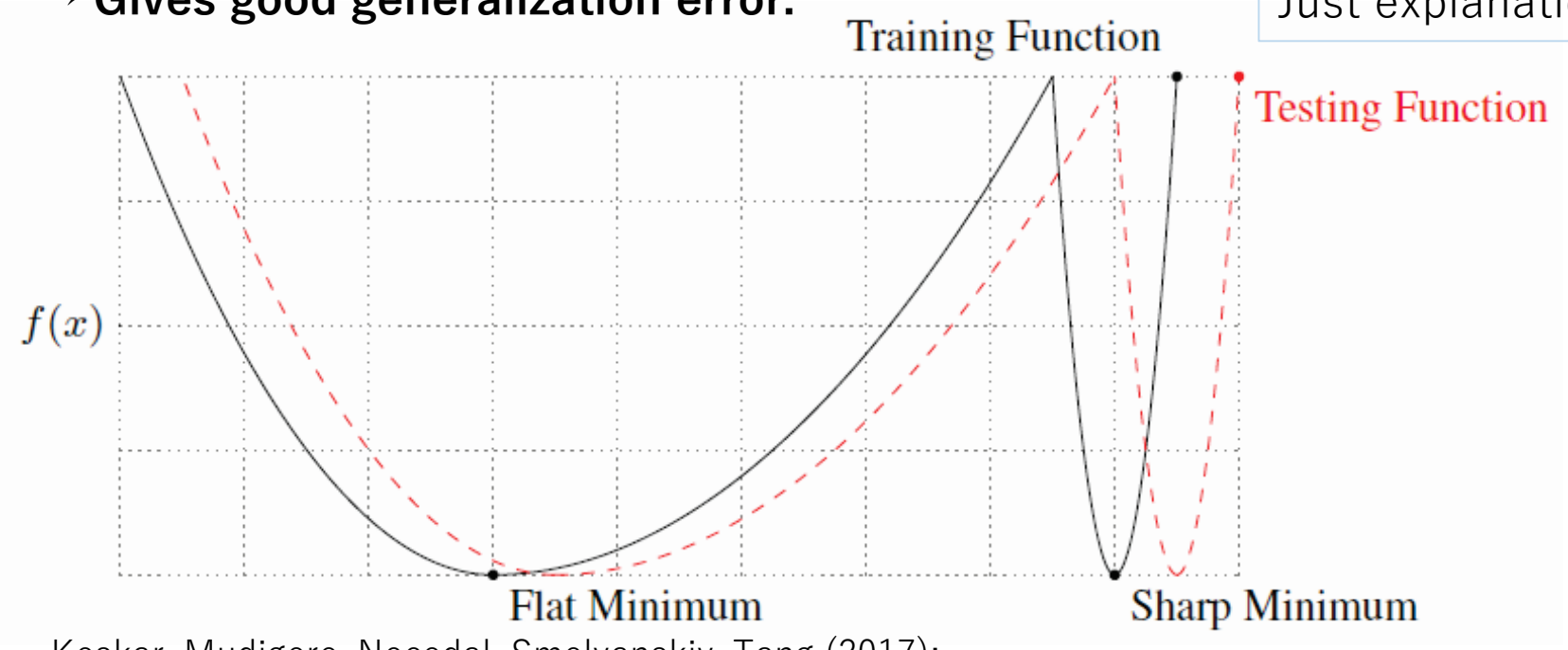
Stochastic


Gaussian noise
Gradient descent

Sharp minima vs flat minima

It is said that SGD likely stay in "flat local minimum"
→ Gives good generalization error.

This is not theory,
Just explanation.



Keskar, Mudigere, Nocedal, Smelyanskiy, Tang (2017):
On large-batch training for deep learning: generalization gap and sharp minima.

$$\theta_t = \theta_{t-1} - \alpha_b \underbrace{\left(\frac{1}{b} \sum_{j=1}^{b} \nabla_\theta \ell(z_{i_j}; \theta) \right)}$$

≅Normal distribution

→ Random walk is likely to be captured in a flat region.

- (criticism) The concept of "flat" depends on the choice of coordinate system. (Dinh et al., 2017)
- PAC-Bayesian analysis (Dziugaite, Roy, 2017)

Convergence theorem (finite dim)

- f_i is bounded and Lipschitz continuous, and has smooth gradient.

$$\|\ell_i\|_\infty \le A,\ \|\nabla \ell_i\|_\infty \le B,\ \|\nabla \ell_i(x) - \nabla \ell_i(y)\| \le M\|x-y\|$$

- Dissipative condition:

$$\langle \nabla L, w \rangle \ge m\|w\|^2 - b \quad (\forall w \in \mathbb{R}^d)$$

(with other technical conditions)

Thm [Raginsky, Rakhlin and Telgarsky, COLT2017]

$$\mathrm{E}[L(X_k)] - L(X^*) \le O\left(\frac{\beta(\beta+d)^2}{\lambda_*}\epsilon + \frac{d\log(\beta+1)}{\beta}\right)$$

for $k = \Omega\left(\frac{\beta(d+\beta)}{\lambda_*\epsilon^4}\log^5(1/\epsilon)\right)$ and any $\epsilon > 0$

Can dependent on d exponentially!

(λ_* is the spectral gap of the Langevin dynamics with β)

Thm [Xu et al., NeurIPS2018] (proof and results contain some flaws)

$$\mathrm{E}[L(X_k)] - L(X^*) \le O\left(\frac{\epsilon}{\beta} + \frac{d\log(\beta+1)}{\beta}\right) \quad \text{for } k = \Omega\left(\frac{d}{\epsilon\lambda_*}\log(1/\epsilon)\right)$$

and any $\epsilon > 0$.

Infinite dimensional setting

Hilbert space

$$\mathcal{H} = \left\{ \sum_{k=0}^{\infty} \alpha_k f_k \mid \sum_{k=0}^{\infty} \alpha_k^2 < \infty \right\}$$

$$\langle x, y \rangle = \sum_{k=0}^{\infty} \alpha_k \beta_k \quad \text{for } x = \sum_k \alpha_k f_k,\ y = \sum_k \beta_k f_k.$$

RKHS structure

$$\mathcal{H}_K = \left\{ \sum_{k=0}^{\infty} \alpha_k f_k \mid \sum_{k=0}^{\infty} \alpha_k^2 / \mu_k < \infty \right\}$$

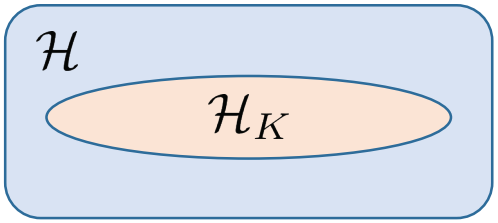
$$\langle x, y \rangle_{\mathcal{H}_K} = \sum_{k=0}^{\infty} \alpha_k \beta_k / \mu_k \quad \text{for } x = \sum_k \alpha_k f_k,\ y = \sum_k \beta_k f_k.$$

Assumption (eigenvalue decay)

$$\mu_k \simeq k^{-2}$$

(not essential, can be relaxed to $\mu_k \sim k^{-\beta}$ for $\beta > 1$)

$$\min_{x \in \mathcal{H}} L(x) = \min_{x \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell_i(x) + \left(\frac{\lambda_0}{2} \|x\|^2 \right)$$



Motivative situation

$$\min_{x\in\mathcal{H}} L(x) = \min_{x\in\mathcal{H}} \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \ell_i(x) + \left(\frac{\lambda_0}{2}\|x\|^2\right)$$

$$\ell_i(x) = \tilde{\ell}_i(T_K^\gamma x) \begin{cases} \tilde{\ell}_i\text{: non-convex loss} \\ T_K^\gamma x = \sum_{k=0}^{\infty} \mu_k^\gamma \alpha_k f_k \quad \text{for } x = \sum_{k=0}^{\infty} \alpha_k f_k \end{cases}$$

$$\mathcal{H}_{K^\gamma} := T_K^\gamma(\mathcal{H})$$

$$\langle x, y\rangle_{\mathcal{H}_{K^\gamma}} := \sum_{k=0}^{\infty} \alpha_k \beta_k / \mu_k^{2\gamma}$$

If $f = T_K^\gamma x$, then $\|f\|_{\mathcal{H}_{K^\gamma}} = \|x\|$.

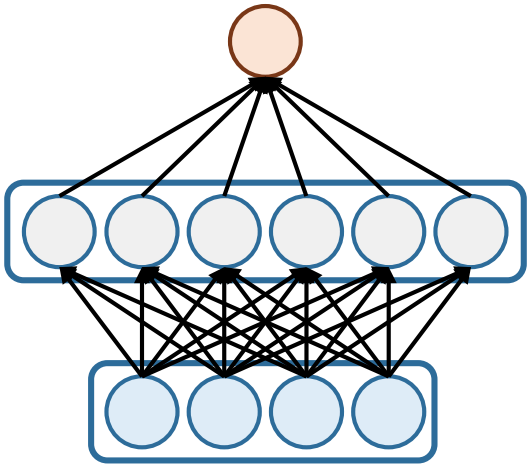
$$\frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \ell_i(x) + \frac{\lambda_0}{2}\|x\|^2 = \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \tilde{\ell}_i(f) + \frac{\lambda_0}{2}\|f\|^2_{\mathcal{H}_{K^\gamma}}$$

2-layer neural network

$$L(W) = \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \ell_i(f_W(x_i)) + \frac{\lambda_0}{2} \|W\|_{\mathrm{F}}^2$$

$$f_W(x) = \sum_{j=1}^{\infty} a_j \eta(w_j^\top x)$$

- $a_j \leq j^{-\gamma}$ for $\gamma > 1/2$
- η is a smooth activation, e.g., sigmoid.



Unlike NTK, a_j is fixed with respect to the width and the sample size.

Extension to infinite dim

$$x = \sum_{j=1}^{\infty} x_j \phi_j \in \mathcal{H}$$

$$\min_{x \in \mathcal{H}} L(x) \Rightarrow \min_{x \in \mathcal{H}} \left\{ L(x) + \frac{\lambda}{2} \|x\|_{\mathcal{H}_K}^2 \right\}$$

$\mathcal{H}_K$: RKHS with kernel K.

$\mathcal{H}_K \hookrightarrow \mathcal{H}$

$$\mathrm{d}X_t = -\nabla \left(L(X_t) + \frac{\lambda}{2} \|X_t\|_{\mathcal{H}_K}^2 \right) \mathrm{d}t + \sqrt{\frac{2}{\beta}} \mathrm{d}W_t$$

Norm: For $x = \sum_{j=1}^{\infty} x_j \phi_j \in \mathcal{H}$, we let $\|x\|_{\mathcal{H}_K}^2 = \sum_{j=1}^{\infty} \mu_j x_j^2$ where $\mu_j \sim j^2$.

Cylindrical Brownian motion: $W_t = \sum_{j=1}^{\infty} W_{j,t} \phi_j$

Time discretization:

$$X_{n+1} = S_\eta \left(X_n - \eta \nabla L(X_n) + \sqrt{2 \frac{\eta}{\beta}} \xi_n \right) \qquad \left(S_\eta := (I + \eta \frac{\lambda}{2} \nabla \| \cdot \|_{\mathcal{H}_K})^{-1} \right)$$

(Semi-implicit Euler scheme)

$\xi_n = \sum_{j=1}^{\infty} \gamma_{n,j} \phi_j$ where $\gamma_{n,j} \sim N(0,1)$ (i.i.d.).

Galerkin & Stochastic approx

- Spectral Galerkin approximation:

  The infinite dimensional dynamics cannot be computed.
  → We approximate it by N-dimensional subspace.

$$H_N := \mathrm{span}\{f_0, \ldots, f_N\} \quad P_N\text{: Projection to } \mathcal{H}_N$$

$$X_{n+1}^N = S_\eta \left(X_n^N - \eta P_N \nabla L(X_n^N) + \sqrt{2\frac{\eta}{\beta}} P_N \xi_n \right)$$

- Stochastic gradient approximation:

$$X_{n+1}^N = S_\eta \left(X_n^N - \eta P_N \frac{1}{|I_n|} \sum_{i \in I_n} \nabla \ell_i(X_n^N) + \sqrt{2\frac{\eta}{\beta}} P^N \xi_n \right)$$

Invariant measure

$$\mathrm{d}X_t = -\nabla \left(L(X_t) + \frac{\lambda}{2} \|X_t\|^2_{\mathcal{H}_K} \right) \mathrm{d}t + \sqrt{\frac{2}{\beta}} \mathrm{d}W_t$$

$$\frac{\mathrm{d}\pi}{\mathrm{d}\mu_*}(x) \quad \propto \quad \exp\left(-\beta L(x)\right)$$

$$\mu_* = N(0, C)$$

$$\text{where } C = \beta \mathrm{diag}(\mu_0, \mu_1, \dots).$$

Related work

- **Finite dimensional Langevin dynamics:**

➢ Convergence in low (convex case): Dalalyan and Tsybakov, 2012; Dalalyan, 2016; Durmus and Moulines, 2015, ..

➢ Non-convex Optimization: Raginsky et al., 2017; Xu et al., 2018; Erdogdu, Mackey and Shamir, 2018, ...

- **Infinite dimensional Langevin dynamics:**

➢ Continuous time:

- Existence & Uniqueness of invariant measure: Da Prato and Zabczyk,1992; Maslowski, 1989; Sowers, 1992.
- Geometric ergodicity: Jacquot and Royer, 1995; Shardlow, 1999; Hairer, 2002, Its explicit rate: Goldys and Maslowski, 2006.

➢ Discrete time:

- Weak approximation rate of discretized scheme: Hausenblas, 2003; Debussche, 2011; Bréhier, 2014; Bréhier and Kopec 2016.

Other topics (MCMC in Hilbert space):

- preconditioned Crank–Nicolson (pCN): Hairer et al., 2014; Eberle, 2014; Vollmer, 2015; Rudolf and Sprungk, 2018.
- Metropolis-Adjusted Langevin Algorithm (MALA): Durmus and Moulines, 2015; Beskos et al., 2017.

Assumption (1)

- Smoothness:

$$\|\nabla L(x) - \nabla L(y)\| \leq M\|x - y\|$$

- Strong smoothness condition:

For $\alpha \in (1/4, 1)$,

$$\|\nabla L(x) - \nabla L(y)\|_{-\alpha} \leq M\|x - y\|$$

where $\|x\|_\varepsilon = \left(\sum_{k\geq 0}(\mu_k)^{2\varepsilon}|\langle x, f_k\rangle|^2\right)^{1/2}$.

(This is non-standard, but, is satisfied in the previous examples)

- Third order smoothness:

Let $L_N = L(P_N x)$. There exists $\alpha' \in [0, 1)$ such that

$$\|D^3 L_N(x)\cdot(h,k)\|_{\alpha'} \leq C_{\alpha'}\|h\|_0\|k\|_0,$$
$$\|D^3 L_N(x)\cdot(h,k)\|_0 \leq C_{\alpha'}\|h\|_{-\alpha'}\|k\|_0.$$

Assumption (2)

- It either holds:
  - (Strict Dissipativity) $\lambda > M\mu_0$, or
  - (Bounded gradients) $\|\nabla L(\cdot)\| \leq B$, for $B > 0$.

Dissipativity:

For $A = -\frac{\lambda}{2}\nabla\|\cdot\|^2_{\mathcal{H}_K}$

$$\langle Ax - \nabla L(x), x\rangle \leq -m\|x\|^2 + c.$$

$$L(x) + \frac{\lambda}{2}\|x\|^2_{\mathcal{H}_K}$$

Error bound analysis

- Assumption: Smoothness of L.

$$x^* := \arg\min_{x \in \mathcal{H}} L(x) \qquad \tilde{x} := \arg\min_{x \in \mathcal{H}} \left\{ L(x) + \frac{\lambda}{2} \|x\|_{\mathcal{H}_K}^2 \right\}$$

Thm [Muzellec, Sato, Massias, Suzuki, 2020]

Under some smoothness assumption on L, we have

$$L(X_n) - L(x^*) \lesssim \exp\left(-\Lambda^* n\eta\right) + \frac{1}{\Lambda^*}\eta^{1/2-\kappa} \quad \text{(geometric ergodicity + time discretization)}$$
$$+ \frac{1}{\beta}\left(\sqrt{\frac{1}{\lambda}} + 1\right) + \lambda\left(\frac{\|\tilde{x}\|_{\mathcal{H}_K}}{\sqrt{\beta}} + \|\tilde{x}\|_{\mathcal{H}_K}^2\right)$$
$$+ L(\tilde{x}) - L(x^*) \quad \text{(bias of invariant measure)}$$

with high probability, where $\kappa > 0$ is an arbitrary small constant.

Note: Λ^* can be dependent on β exponentially, which seems unavoidable without additional assumption.

Our proof depends on the technique by Brehier 2014; Brehier&Kopec 2016; Mattingly et al., 2002; Goldys&Maslowski, 2006.

Decomposition of errors

- It is difficult to show strong convergence.
- Instead, we show weak convergence.

$$|\mathbb{E}[\phi(X_n)] - \phi(x^*)| \leq ?$$

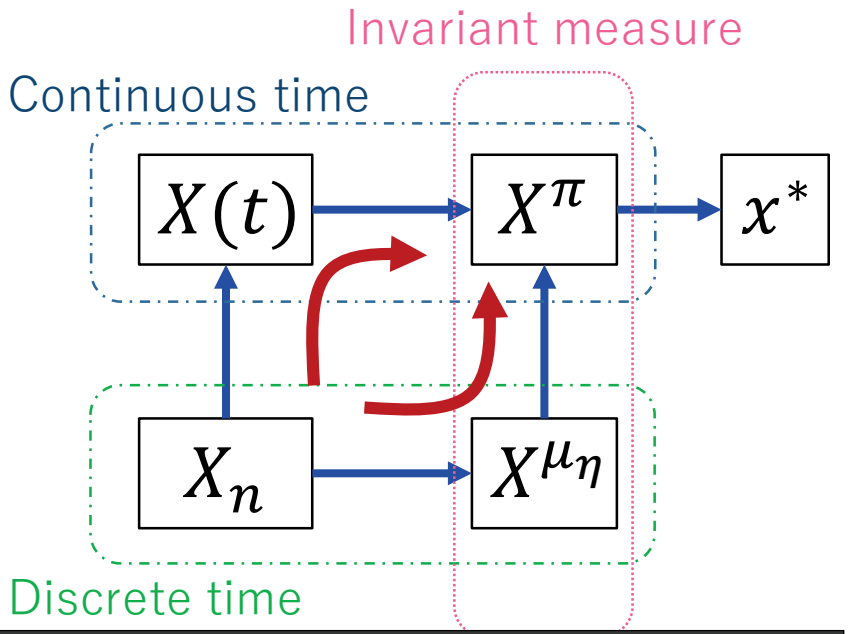
for a smooth function ϕ.

- Raginsky et al. (2017), Bréhier (2014), Bréhier and Kopec (2016):

$$\mathbb{E}[\phi(X_n) - \phi(X(n\eta))] + \mathbb{E}[\phi(X(n\eta)) - \phi(X^\pi)] + \mathbb{E}[\phi(X^\pi) - \phi(x^*)]$$

- Xu et al. (2018):

$$\mathbb{E}[\phi(X_n) - \phi(X^{\mu_\eta})] + \mathbb{E}[\phi(X^{\mu_\eta}) - \phi(X^\pi)] + \mathbb{E}[\phi(X^\pi) - \phi(x^*)]$$



Better rate, but requires stronger condition (Strong smoothness)

First term bound

$$\mathbb{E}[\phi(X_n) - \phi(x^*)] = \mathbb{E}[\phi(X_n) - \phi(X^{\mu_\eta})] + \mathbb{E}[\phi(X^{\mu_\eta}) - \phi(X^{\pi})] + \mathbb{E}[\phi(X^{\pi}) - \phi(x^*)]$$

Lemma (Geometric ergodicity)

There exists a unique invariant measure μ_η, and the geometric ergodicity is satisfied:

$$\mathbb{E}[\phi(X_n) - \phi(X^{\mu_\eta})] \leq C(1 + \|x_0\|) \exp\left(-\Lambda_\eta^* n\eta\right)$$

where the "spectral gap" Λ_η^* is given by

(i) (Strict dissipative)

$$\Lambda_\eta^* = \frac{\frac{\lambda}{\mu_0} - M}{1 + \eta\frac{\lambda}{\mu_0}}$$

(ii) (Bounded gradient)

$$\Lambda_\eta^* = C \min\left(\frac{\lambda}{2\mu_0}, \frac{1}{2}\right)\delta$$

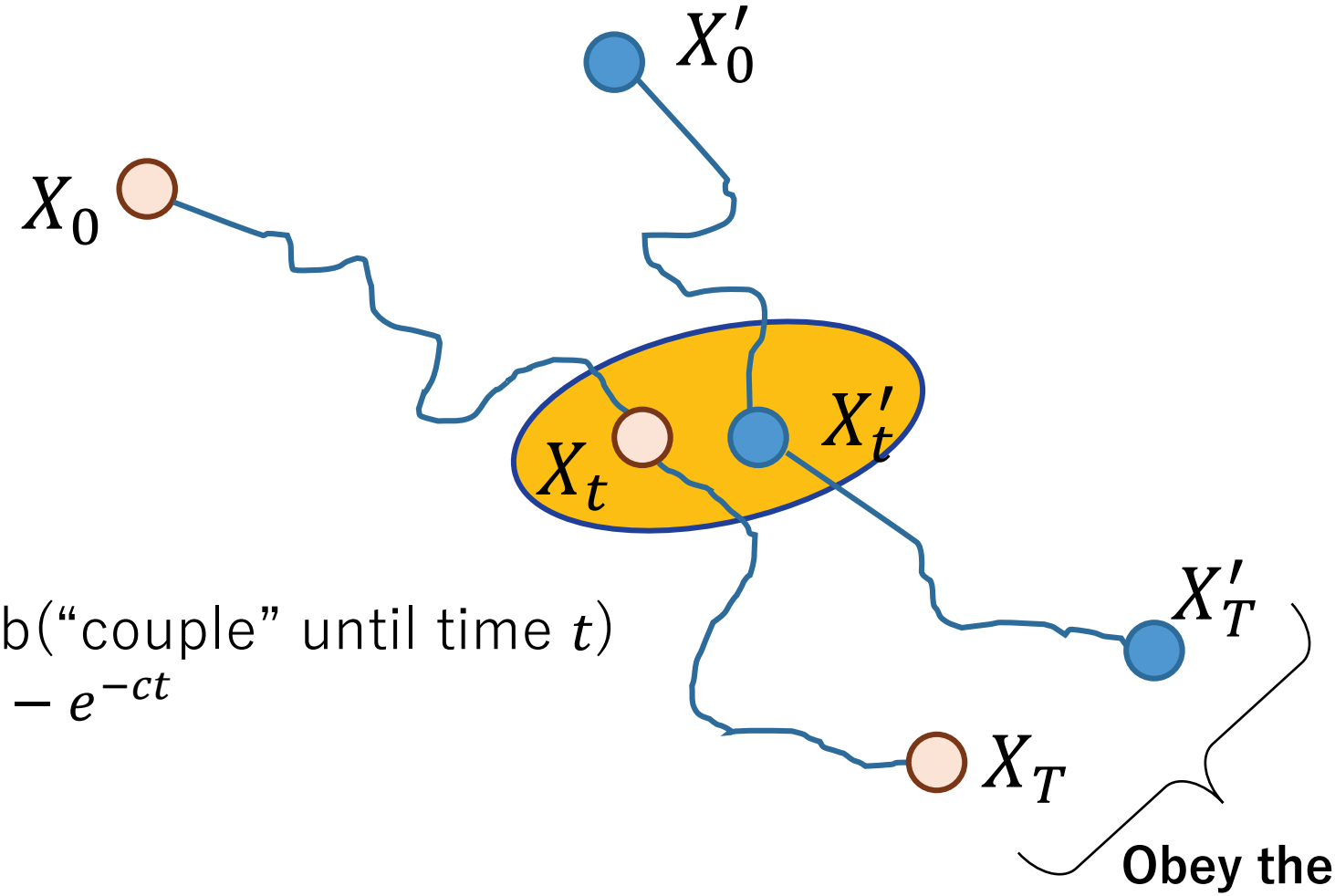
for $\delta = \exp(-O(\beta))$

X^{μ_η}: r.v. obeying μ_η $\quad$ $X_0 = x_0$ (constant)

- Unlike finite dimensional case, this would not hold without an assumption like strong smoothness.
- **Coupling argument**: Lyapunov condition, majorization condition (combining techniques by Mattingly et al. (2002), Goldys&Maslowski (2006))

Geometric ergodicity

- Coupling argument



Prob("couple" until time t)
$\geq 1 - e^{-ct}$

Second term bound

$$\mathbb{E}[\phi(X_n) - \phi(x^*)] = \mathbb{E}[\phi(X_n) - \phi(X^{\mu_\eta})] + \mathbb{E}[\phi(X^{\mu_\eta}) - \phi(X^{\pi})] + \mathbb{E}[\phi(X^{\pi}) - \phi(x^*)]$$

X^{μ_η}: the invariant measure of discrete time dynamics
X^{π}: the invariant measure of continuous time dynamics
(existence and uniqueness are well known)

Lemma (Discrepancy between invariant measures)

For any $0 < \kappa < 1/2$, there exists a constant C such that

$$|\mathbb{E}[\phi(X^{\mu_\eta}) - \phi(X^{\pi})]| \leq C \frac{\|\phi\|_{0,2}}{\Lambda_0^*} \eta^{1/2-\kappa}.$$

$$\|\phi\|_{0,2} = \max\{\|\phi\|_\infty, \|D\phi\|_\infty, \|D^2\phi\|_\infty\}$$

- As the step-size goes to 0, the discrete time dynamics approaches the continuous one.
- β affects the bound through the spectral gap Λ_0^*.

Comparison to existing bound

- Thanks to geometric ergodicity, the weak convergence rate becomes different.

Brehier 2014:

$$|\mathbb{E}[\phi(X_n) - \phi(X^\pi)]| \leq C \left[\frac{1}{\Lambda_0^*} \left(\frac{\beta}{n} \right)^{1/2-\kappa} + \frac{1}{\Lambda_0^*} \eta^{1/2-\kappa} \right]$$

Ours:

$$|\mathbb{E}[\phi(X_n) - \phi(X^\pi)]| \leq C \left[\exp\left(-\Lambda_\eta^* n\eta\right) + \frac{1}{\Lambda_0^*} \eta^{1/2-\kappa} \right]$$

This is because of additional assumption:

Strong smoothness

$$\|\nabla L(x) - \nabla L(y)\|_{-\alpha} \leq M \|x - y\|$$

(strong, but natural in machine learning setting)

Third term bound

$$\mathbb{E}[\phi(X_n) - \phi(x^*)] = \mathbb{E}[\phi(X_n) - \phi(X^{\mu_\eta})] + \mathbb{E}[\phi(X^{\mu_\eta}) - \phi(X^{\pi})] + \mathbb{E}[\phi(X^{\pi}) - \phi(x^*)]$$

$$\tilde{x} := \arg\min_{x \in \mathcal{H}} \left\{ L(x) + \frac{\lambda}{2} \|x\|^2_{\mathcal{H}_K} \right\} \qquad \frac{\mathrm{d}\pi}{\mathrm{d}\mu_*}(x) \propto \exp\left(-\beta L(x)\right)$$

$$\mu_* = N(0, C) \text{ where } C = \beta \mathrm{diag}(\mu_0, \mu_1, \dots).$$

Lemma (Discrepancy between invariant measures)

$$\int L \mathrm{d}\pi - L(\tilde{x}) \lesssim \frac{1}{\beta} \left(\sqrt{\frac{2M}{\lambda}} + 1 \right) + \lambda \left(\frac{\|\tilde{x}\|_{\mathcal{H}_K}}{\sqrt{\beta}} + \|\tilde{x}\|^2_{\mathcal{H}_K} \right).$$

- A technique analogous to contraction rate analysis of Bayesian nonparametrics is used.
- Gaussian correlation inequality is used to bound the mass of π around $\tilde{x}$.

Galerkin approximation & SGLD

- Assumption: Smoothness of L.

$$x^* := \arg\min_{x \in \mathcal{H}} L(x) \qquad \tilde{x} := \arg\min_{x \in \mathcal{H}} \left\{ L(x) + \frac{\lambda}{2} \|x\|^2_{\mathcal{H}_K} \right\}$$

Thm

Under some smoothness assumption on L, we have

$$L(X_n) - L(x^*) \lesssim \exp\left(-\Lambda^* n\eta\right) + \frac{1}{\Lambda^*}\eta^{1/2-\kappa} \quad \text{(geometric ergodicity + time discretization)}$$

$$+ \frac{1}{\beta}\left(\sqrt{\frac{1}{\lambda}} + 1\right) + \lambda\left(\frac{\|\tilde{x}\|_{\mathcal{H}_K}}{\sqrt{\beta}} + \|\tilde{x}\|^2_{\mathcal{H}_K}\right) \quad \text{(bias of invariant measure)}$$

$$+ L(\tilde{x}) - L(x^*)$$

Galerkin approx.

$$+ \frac{\mu_{N+1}^{1/2-\kappa}}{\Lambda_0^*} + \sqrt{\frac{n\beta\eta(n_{\mathrm{tr}} - n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}} - 1)}}$$

Mini-batch size

with high probability, where $\kappa > 0$ is an arbitrary small constant.

(c.f., Brehier 2014; Goldys&Maslowski, 2006)

Summary

- Fast learning rate in NTK regime.
  - Smoothness matters.

➢ Infinite dimensional Langevin dynamics.
  - Weak approximation error
  - Geometric ergodicity + time discretization error + bias of invariant measure

What is missing?

- Adaptivity of deep learning (minimax-optimality).
  - Hölder class [Schmidt-Hieber, 2017]
  - Besov space [Suzuki, 2019][Hayakawa&Suzuki, 2019]
  - Piece-wise smooth [Imaizumi&Fukumizu, 2018]
  - Anisotropic Besov [Suzuki&Nitanda, 2019]

→ Non-convex optimization is required.

Estimation error in anisotropic Besov space

(kernel ridge) $n^{-\frac{2(s_{\min}-D/p+d/2)}{2(s_{\min}-D/p+d/2)+d}}$ Sub-optimal

$\gg$

(deep learning) $n^{-\frac{2\bar{s}}{2\bar{s}+1}}$ Optimal

$\bar{s} := \left(\frac{1}{s_1}+\cdots+\frac{1}{s_d}\right)^{-1}$

Appendix

Mean field analysis

- We regard neural network optimization as a distribution optimization.

$$f(x) = \frac{1}{M}\sum_{j=1}^{M} a_j \eta(w_j^\top x) \xrightarrow{M \to \infty} \int a\eta(w^\top x)\rho(a,w)\mathrm{d}a\mathrm{d}w$$

➡ Mean w.r.t. prob. density ρ of (a, w):

Optimization of f ⇔ Optimization of ρ

$$\frac{\mathrm{d}\rho_t}{\mathrm{d}t} = -\nabla \cdot (v_t \rho_t)$$ Continuity equation

Continuity equation, Wasserstein gradient flow
(fluid dynamics, probability theory)

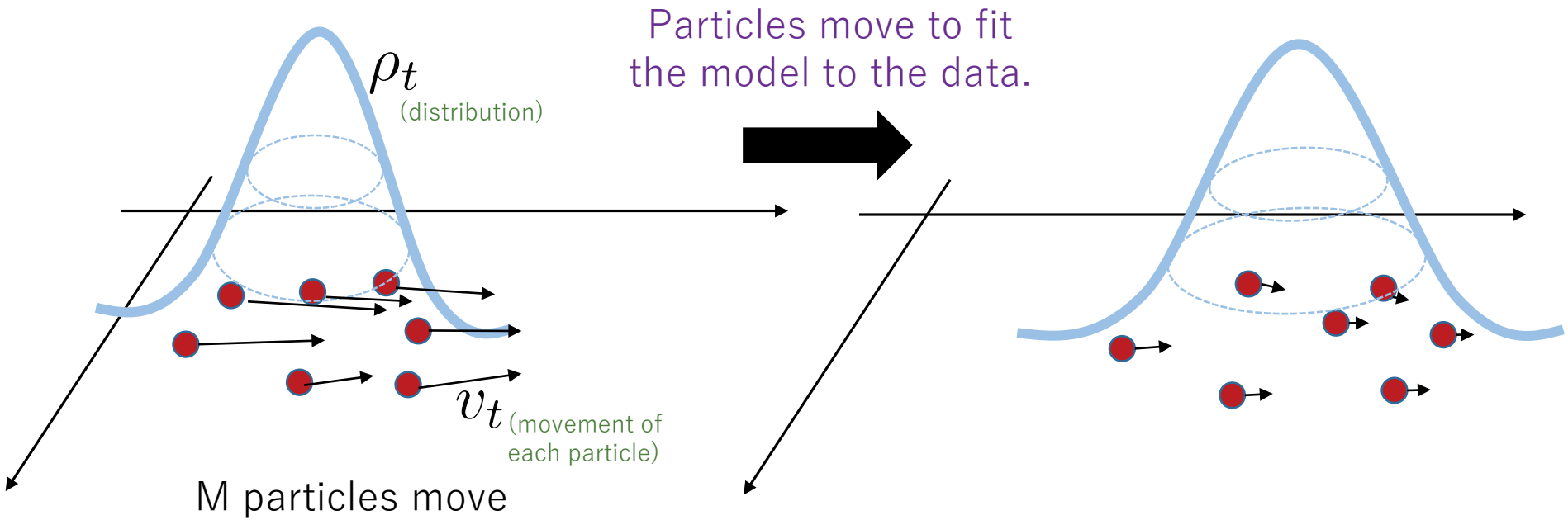
[Atsushi Nitanda and Taiji Suzuki: Stochastic Particle Gradient Descent for Infinite Ensembles. arXiv:1712.05438.]

Particle gradient descent

$$f(x) = \frac{1}{M}\sum_{j=1}^{M} a_j \eta(w_j^\top x)$$

One partilce

- Each neuron corresponds to one particle.
- We optimize the distribution of the all particles.



We can show global convergence in the limit of $M \to \infty$.

[Nitanda&Suzuki, 2017][Chizat&Bach, 2018][Mei, Montanari&Nguyen, 2018]

Assumption (1)

- Smoothness:

$$\|\nabla L(x) - \nabla L(y)\| \leq M\|x - y\|$$

- Second order smoothness:

For $\alpha \in (1/4, 1)$,

$$|D^2 L(x) \cdot (h, k)| \leq C_{\alpha,2}\|h\|_{\mathcal{H}}\|k\|_{\alpha},$$

where $\|x\|_{\varepsilon} = \left(\sum_{k\geq 0}(\mu_k)^{2\varepsilon}|\langle x, f_k\rangle|^2\right)^{1/2}$.

(This is non-standard, rather strong. But, is satisfied in the previous examples)

- Third order smoothness:

Let $L_N = L(P_N x)$. There exists $\alpha' \in [0, 1)$ such that

$$\|D^3 L_N(x) \cdot (h, k)\|_{\alpha'} \leq C_{\alpha'}\|h\|_0\|k\|_0,$$
$$\|D^3 L_N(x) \cdot (h, k)\|_0 \leq C_{\alpha'}\|h\|_{-\alpha'}\|k\|_0.$$