

# Benefit of deep learning with non-convex noisy gradient descent

**Taiji Suzuki**<sup>1,2</sup> and **Shunta Akiyama**<sup>1</sup>

<sup>1</sup>The University of Tokyo, <sup>2</sup>AIP-RIKEN



THE UNIVERSITY OF TOKYO

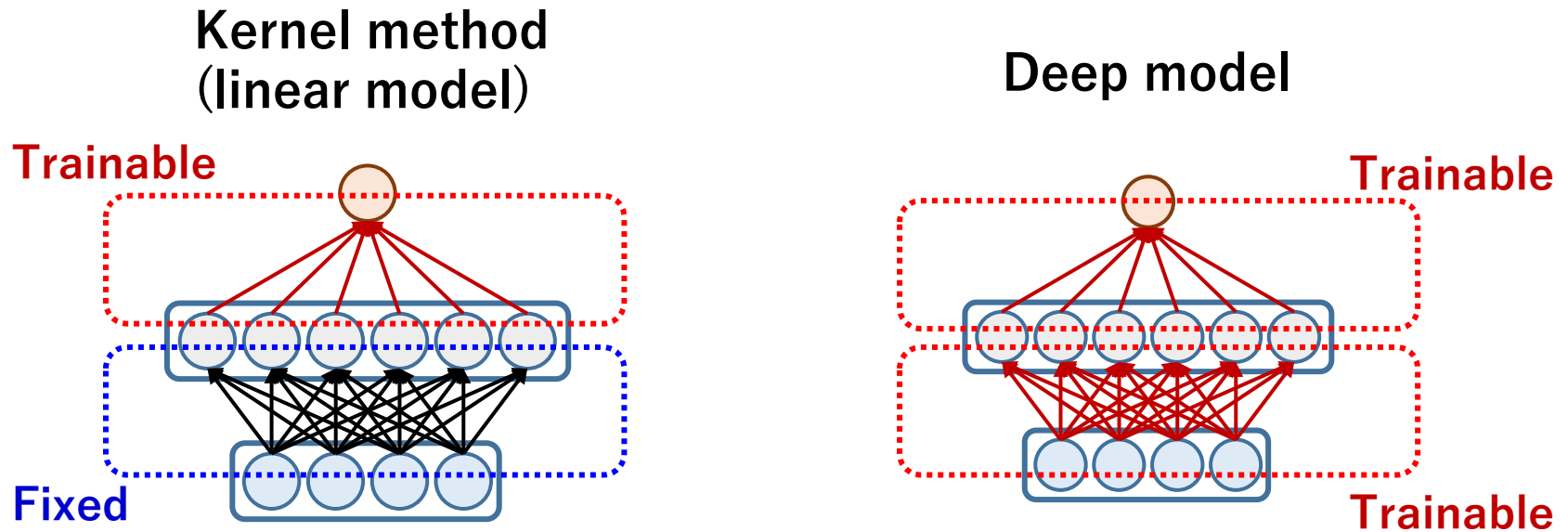


ICLR2021

(spotlight)

# Background

Benefit of neural network with optimization guarantee.



- Statistical efficiency: Curse of dimensionality
- Optimization efficiency: Non-convexity

## We show

- separation between deep and shallow in terms of excess risk,
- global optimality of noisy gradient descent.

# Problem setting (teacher-student model) <sup>3</sup>

Teacher and student model :

$$f_W(x) = \sum_{m=1}^{\infty} a_m w_{2,m} \sigma(b_m^{-1} w_{1,m}^\top x)$$

$W = (w_{1,m}, w_{2,m})_{m=1}^{\infty}$  : trainable parameter

$(a_m, b_m)_{m=1}^{\infty}$  : fixed parameter

$$\mathcal{H}_\gamma := \left\{ W = (w_{1,m}, w_{2,m})_{m=1}^{\infty} \mid \|W\|_{\mathcal{H}_\gamma}^2 = \sum_{m=1}^{\infty} (w_{1,m}^2 + \|w_{2,m}\|^2) / \mu_m^\gamma < \infty \right\}$$

$$\mathcal{F}_\gamma := \{f_W \mid W \in \mathcal{H}_\gamma, \|W\|_{\mathcal{H}_\gamma} \leq 1\}$$

Observation model :  $f^\circ \in \mathcal{F}_\gamma$  (true function),

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

From  $D_n = (x_i, y_i)_{i=1}^n$  (observed data), we estimate  $f^\circ$ .

Excess risk (mean squared error):  $\mathbb{E}_{D_n} [\|\hat{f} - f^\circ\|_{L_2(P_X)}^2]$  ➤ Convergence rate?  
➤ Deep vs shallow?

# Condition on parameters

$$f_W(x) = \sum_{m=1}^{\infty} a_m w_{2,m} \sigma(b_m^{-1} w_{1,m}^\top x)$$

## Condition

- $\mu_m \propto m^{-2}$
- $a_m \propto \mu_m^{\alpha_1}$  for  $\alpha_1 > 1/2$
- $b_m \propto \mu_m^{\alpha_2}$  for  $\alpha_2 > \gamma/2$
- Activation function  $\sigma$  is sufficiently smooth.

$$\mathcal{H}_\gamma := \left\{ W = (w_{1,m}, w_{2,m})_{m=1}^{\infty} \mid \|W\|_{\mathcal{H}_\gamma}^2 = \sum_{m=1}^{\infty} (w_{1,m}^2 + \|w_{2,m}\|^2) / \mu_m^\gamma < \infty \right\}$$

$$\mathcal{F}_\gamma := \{ f_W \mid W \in \mathcal{H}_\gamma, \|W\|_{\mathcal{H}_\gamma} \leq 1 \}$$

$f^\circ \in \mathcal{F}_\gamma$  : true function

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

# Related work

## Separation between deep and shallow.

- **Generalization error comparison via Rademacher complexity analysis.**

[Allen-Zhu & Li (2019; 2020); Li et al. (2020); Bai & Lee (2020); Chen et al. (2020)]

- They do not give tight comparison for excess risk.
- Every derived rate is  $O(1/\sqrt{n})$ : difference of rate of conv is not shown.

- **Approximation ability**

[E, Ma & Wu (2018); Ghorbani et al. (2020); Yehudai & Shamir (2019)]

- Estimation error with optimization guarantee is not compared.
- Some of them require  $d \rightarrow \infty$ . What happens for fixed  $d$ ?

- **Sparse regularization and Frank-Wolfe algorithm**

[Barron (1993); Chizat & Bach (2020); Chizat (2019); Gunasekar et al. (2018); Woodworth et al. (2020); Klusowski & Barron (2016)]

- Models with sparsity inducing regularization.
- Frank-Wolfe type method is analyzed. What happens for GD?

**Question:** Convergence of excess risk + Optimization guarantee by GD?

# Approach

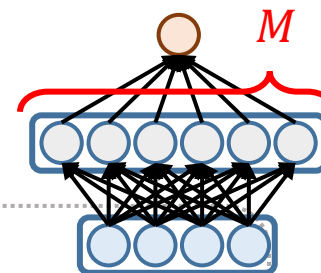
**Loss function (squared loss):**  $W = (w_{1,m}, w_{2,m})_{m=1}^{\infty}$  (element in  $\mathcal{H}_1$ )

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

**Regularized empirical risk minimization:**

$$\min_W \widehat{L}(f_W) + \frac{\lambda}{2} \|W\|_{\mathcal{H}_1}^2$$

**Infinite dimensional non-convex optimization problem**



- Finite-dim truncation**

$$f_{W^{(M)}} = \sum_{m=1}^M a_m w_{2,m} \sigma(b_m^{-1} w_{1,m}^\top x)$$

$$W^{(M)} = (w_{1,m}, w_{2,m})_{m=1}^M$$

- Langevin dynamics for the truncated parameter** (noisy gradient descent)

$$W_{k+1}^{(M)} = W_k^{(M)} - \eta \nabla_{W^{(M)}} \left( \widehat{L}(f_{W_k^{(M)}}) + \frac{\lambda}{2} \|W_k^{(M)}\|_{\mathcal{H}_1}^2 \right) + \sqrt{2 \frac{\eta}{\beta}} \xi_k^{(M)}$$

Gaussian noise

# Infinite-dim. Langevin dynamics

$$\min_W \left\{ \widehat{L}(W) + \frac{\lambda}{2} \|W\|_{\mathcal{H}_1}^2 \right\} \quad \widehat{L}(W) := \widehat{L}(f_W)$$

[Muzellec, Sato, Massias, Suzuki (2020); Suzuki (NeurIPS2020)]

$$dW_t = -\nabla \left( \widehat{L}(W_t) + \frac{\lambda}{2} \|W_t\|_{\mathcal{H}_1}^2 \right) dt + \sqrt{\frac{2}{\beta}} d\xi_t$$

Cylindrical Brownian motion

Time discretization

Gaussian noise

(Euler-Maruyama scheme)

$$W_{k+1} = W_k - \eta \nabla \left( \widehat{L}(W_k) + \frac{\lambda}{2} \|W_k\|_{\mathcal{H}_1}^2 \right) + \sqrt{\frac{2\eta}{\beta}} \xi_k$$

In our theory, we used a bid modified scheme (semi-implicit Euler scheme):

$$W_{k+1} = W_k - \eta \nabla \left( \widehat{L}(W_k) + \frac{\lambda}{2} \|W_{k+1}\|_{\mathcal{H}_1}^2 \right) + \sqrt{\frac{2\eta}{\beta}} \xi_k$$

$$\longleftrightarrow W_{k+1} = S_\eta \left( W_k - \eta \nabla \widehat{L}(W_k) + \sqrt{2\frac{\eta}{\beta}} \xi_k \right) \quad \left( S_\eta := (I + \eta \lambda A)^{-1} \right)$$

where  $x^* A x = \|x\|_{\mathcal{H}_1}^2$

# Optimization error bound

The distribution of  $W_t$  weakly converges to an invariant measure  $\pi_\infty$ :

$$\pi_\infty(W) \propto \exp\left(-\beta \widehat{L}(W) - \frac{\beta\lambda}{2} \|W\|_{\mathcal{H}_1}^2\right)$$

invariant measure  
of continuous dynamics

Likelihood

Prior

Analogous to Bayes posterior

**Thm (informal)** [Muzellec, Sato, Massias, Suzuki (2020); Suzuki (NeurIPS2020)]

$$\widehat{L}(W_k^{(M)}) - \int \widehat{L}(W) d\pi_\infty(W) \stackrel{\mathbb{E}_k}{\lesssim} \exp(-\Lambda_\eta^* k \eta) + \frac{\sqrt{\beta}}{\Lambda_0^*} \eta^{1/2-\kappa} + M^{-2\gamma}$$

Geometric ergodicity

Time discretization

Finite dim truncation

- Convergence to **near global optimal** is guaranteed even though the objective is **non-convex**.
- The rate of convergence is **independent of dimensionality**.



# Excess risk bound for deep learning

## Condition

- $\mu_m \propto m^{-2}$
- $a_m \propto \mu_m^{\alpha_1}$  for  $\alpha_1 > 1/2$
- $b_m \propto \mu_m^{\alpha_2}$  for  $\alpha_2 > \gamma/2$
- Activation function  $\sigma$  is sufficiently smooth.

$$f_W(x) = \sum_{m=1}^{\infty} a_m w_{2,m} \sigma(b_m^{-1} w_{1,m}^\top x)$$

## Neural network training

### Gradient Langevin dynamics (noisy gradient descent):

$$W_{k+1}^{(M)} = W_k^{(M)} - \eta \left( \nabla_{W^{(M)}} \widehat{L}(f_{W_k^{(M)}}) + \frac{\lambda}{2} \nabla_{W^{(M)}} \|W_{k+1}^{(M)}\|_{\mathcal{H}_1}^2 \right) + \sqrt{2\frac{\eta}{\beta}} \xi_k^{(M)}$$

## Thm (Bound of excess risk for deep learning)

Let  $\lambda = \beta^{-1} = \Theta(1/n)$ , then for  $M = \Omega(n^{1/2(\alpha_1 - 3\alpha_2 + 1)})$ , it holds that

$$\mathbb{E}_{D^n} \left[ \mathbb{E}_{W_k} \left[ \|f_{W_k^{(M)}} - f^\circ\|_{L_2(P_X)}^2 \mid D_n \right] \right] \lesssim n^{-\frac{\gamma}{\alpha_1 - 3\alpha_2 + 1}} + \Xi_k$$

$$\Xi_k := \exp(-\Lambda_\eta^* k \eta) + \frac{\sqrt{\beta}}{\Lambda_0^*} \eta^{1/2 - \kappa} + M^{-2\gamma}$$

Independent of dimension  
(free from curse of dim.)

# Lower bound for linear estimators

We compare the rate of convergence of DL with that of linear estimators.

**Linear estimator:** Any estimator that has the following form:

$$\hat{f}(x) = \sum_{i=1}^n \varphi_i(x; X_n) \underline{y_i} \quad (X_n = (x_1, \dots, x_n))$$

linear

- Kernel ridge estimator
- Sieve estimator
- Nadaraya-Watson estimator
- k-NN estimator

e.g., Kernel ridge regression:

$$\hat{f}(x) = K_{x,X} (K_{X,X} + \lambda I)^{-1} \underline{Y}$$

$R_{\text{lin}}(\mathcal{F}_\gamma) := \inf_{\hat{f}: \text{linear}} \sup_{f^\circ \in \mathcal{F}_\gamma} \mathbb{E}_{D_n} [\|\hat{f} - f^\circ\|_{L_2(P_X)}^2]$  : Minimax risk of linear estimators

Lower bound of worst case error for **any linear estimators**

**Thm** (Lower bound of minimax risk for linear estimators)

For  $\tilde{\beta} = \frac{\alpha_1 + \alpha_2}{\alpha_2 - \gamma/2}$  and any  $\kappa' > 0$ , it holds that

Dependent on dimension  
(curse of dim!!!)

$$R_{\text{lin}}(\mathcal{F}_\gamma) \gtrsim n^{-\frac{2\tilde{\beta} + d}{2\tilde{\beta} + 2d} - \kappa'}$$

# Comparison between deep and shallow <sup>11</sup>

## 1. Excess risk of deep learning

$$n^{-\frac{\gamma}{\alpha_1 - 3\alpha_2 + 1}}$$

## 2. Minimax excess risk of linear estimators

$$R_{\text{lin}}(\mathcal{F}_\gamma) \gtrsim n^{-\frac{2\tilde{\beta} + d}{2\tilde{\beta} + 2d}} \quad \left( \tilde{\beta} = \frac{\alpha_1 + \alpha_2}{\alpha_2 - \gamma/2} \right)$$

This rate depends on  $d$

→ Large error for high dimensional setting

(curse of dimensionality)

Ex.:  $\alpha_1 = \gamma + 3\alpha_2$ ,  $\alpha_2 = 4\alpha_2$

Deep

$$n^{-\left(1 + \frac{1}{\gamma}\right)^{-1}}$$

$$\frac{1}{n} \text{ (large } \gamma)$$

Linear (kernel)

$$n^{-\left(1 + \frac{d}{d+11.3}\right)^{-1}}$$

$$\frac{1}{\sqrt{n}} \text{ (large } d)$$

$\sqrt{n}$ -times large!!

Separation of estimation performance is shown for a realistic optimization.

# Summary

- Analyzed excess risk of deep and shallow methods in a teacher-student model.
- A gradient Langevin dynamics (noisy gradient descent) was applied to train neural network.
- The Langevin dynamics achieves a near global optimal solution.
- We derived excess risk of deep learning and linear estimators:
  - **DL can avoid the curse of dimensionality.**
  - **Any linear estimator suffers from the curse of dimensionality.**

Deep

$$n^{-\frac{\gamma}{\alpha_1 - 3\alpha_2 + 1}}$$

Linear (kernel)

$$n^{-\frac{2\tilde{\beta} + d}{2\tilde{\beta} + 2d}}$$

Separation between deep and shallow was shown in terms of excess risk with optimization guarantee.