Optimization theories of neural networks with its statistical perspective

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

20th/Oct/2021
EPFL-AIP joint seminar
Many collaborators and intern students, and graduate students in The University of Tokyo.
Overview of this presentation

• Optimization theory of deep learning
  • SGD in neural tangent kernel regime
  • Infinite dimensional gradient Langevin dynamics
  • Particle gradient descent in mean field regime
  • Optimization theory in double descent

• Its connection to generalization performance of deep learning.
Overparameterization

Wide neural network does not have spurious local minima.

• Two types of analysis
  ➢ Neural Tangent Kernel (NTK)
  ➢ Mean-field analysis

Since the model complexity is increased, the initial solution is already close to the global optimal.

e.g., Venturi, Bandeira and Bruna (2019).
Two regimes

\[ f_W(x) = \sum_{j=1}^{M} a_j \eta(w_j^T x) \]

• Neural Tangent Kernel regime (lazy learning)
  \[ a_j = O\left(\frac{1}{\sqrt{M}}\right) \]
  [Jacot+ 2018][Du+ 2019][Arora+ 2019]
  (Xavier initialization/He initialization)

• Mean field regime
  \[ a_j = O\left(\frac{1}{M}\right) \]

Different scaling of initial solution yields different behavior.
 Neural Tangent Kernel

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

Since the initial scale is large, a linear approximation around the initial solution can fit the data.

\[ f_W(x) \simeq (W - W^{(0)})^\top \nabla_W f_W^{(0)}(x) \] (linear approximation)

Taylor expansion

Feature map

Inner product between feature maps: kernel

\[ k_W(x, x') = \langle \nabla_W f_W^{(0)}(x), \nabla_W f_W^{(0)}(x') \rangle \]

\[ = \sum_{j=1}^{M} a_j^2 \left( x^\top x' \right) \eta'(w_j^\top x) \eta'(w_j^\top x') \]

Neural Tangent Kernel

Optimization dynamics and generalization errors can be analyzed through the linear approximation.
Convergence in NTK regime


- SGD can achieve the best learning error rate.
- The frequency spectrum specific to the initial network determines the learning efficiency.

Theorem:

$f_T$: solution after $T$-updates

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

Decay rate of spectrum of NTK (Neural Tangent Kernel)

- Decreases to 0 as width $M \to \infty$
- Fast learning rate (faster than $O(1/\sqrt{T})$)

First, low frequency components are captured. Afterward, high frequency components are captured.
Lower bound of linear estimator

Non-parametric regression

\[ y_i = f^o(x_i) + \xi_i \quad (i = 1, \ldots, n) \]

where \( \xi_i \sim N(0, \sigma^2) \) and \( x_i \in [0,1]^d \sim P_X(X) \) (i.i.d.).

Ex. Piecewise constant function with 3 jumps.

\[ \mathbb{E}[\|\hat{f} - f^o\|_{L^2(P)}^2] < ? \]

- Deep learning: \( 1/n \)
- Kernel ridge regression:

\[ \sup_{f^o \in \mathcal{F}} \mathbb{E}[\|\hat{f} - f^o\|_{L^2(P)}^2] \geq 1/\sqrt{n} \]

[Donoho & Johnstone, 1994] [Hayakawa & Suzuki: 2020]
• Suzuki&Akiyama: Benefit of deep learning with non-convex noisy gradient descent: Provable excess risk bound and superiority to kernel methods. network training: Transportation map estimation by infinite dimensional Langevin dynamics. ICLR2021, spotlight.

Optimization in non-NTK regime
Optimization beyond NTK regime

The model is not linearly approximated. We need to solve “non-convex” optimization.

SGD is a noisy gradient descent. Noisy perturbation is helpful to escape local minimum.
We can show optimality of noisy gradient descent.

➢ It can achieve the global optimal solution.
➢ DL can avoid the curse of dimensionality.


\[ X_{n+1} = X_n - \eta \left( \nabla L(X_n) + \frac{\lambda}{2} \nabla \|X_{n+1}\|_{\mathcal{H}_K}^2 \right) + \sqrt{\frac{2\eta}{\beta}} \xi_n \]

\[ \int \hat{L}(W_k) d\pi_{(k)}(W_k) - \int \hat{L}(W) d\pi_{\infty}(W) \lesssim \exp\left(-\Lambda_\eta^* k \eta\right) + \frac{\sqrt{\beta}}{\Lambda_0^*} \eta^{1/2-\kappa} \]

We showed noisy gradient descent can achieve the global optimal solution even if there are infinitely many variables.

Optimization of NN

Loss function (squared loss):

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

Regularized empirical risk minimization:

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

Infinite dimensional non-convex optimization problem

Model examples:

• 2-layer NN

$$W = (W_m)_{m=1}^{\infty}$$

$$f_W(x) = \sum_{m=1}^{\infty} a_m \sigma(W_m^T x)$$

(infinite width is allowed)

• ResNet

$$W = ((a_{m,t}, w_{m,t})_{m=1}^{\infty})_{t=1}^{T}$$

$$f_W(x) = u^T \left( \mathbb{I} + \sum_{m=1}^{\infty} a_{m,T} \sigma(w_{m,T}^\top \cdot) \right) \circ \cdots \circ \left( \mathbb{I} + \sum_{m=1}^{\infty} a_{m,1} \sigma(w_{m,1}^\top x) \right)$$
Infinite-dim. Gradient Langevin dynamics

\[
\min_W \left\{ \hat{L}(W) + \frac{\lambda}{2} \| W \|^2_{\mathcal{H}_1} \right\}
\]

\[
\hat{L}(W) := \hat{L}(f_W)
\]

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

**Cylindrical Brownian motion**

Time discretization

(Euler-Maruyama scheme)

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

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

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

\[
W_{k+1} = S_\eta \left( W_k - \eta \nabla \hat{L}(W_k) + \sqrt{\frac{2\eta}{\beta}} \xi_k \right)
\]

\[
(S_\eta := (I + \eta \lambda A)^{-1})
\]

where \( x^* A x = \| x \|^2_{\mathcal{H}_1} \)
Optimization error bound

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

$$\pi_{\infty}(W) \propto \exp \left( - \beta \hat{L}(W) - \frac{\beta \lambda}{2} \|W\|_2^2 \right)$$

invariant measure of continuous dynamics

Likelihood Prior

Analogous to Bayes posterior

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

Suppose that $\|W\|_{H_1}^2 = \sum_{m=1}^{\infty} m^2 W_m$,

$$\kappa > 0: \text{arbitrary small positive real}$$

$$\int \hat{L}(W_k) d\pi_{(k)}(W_k) - \int \hat{L}(W) d\pi_{\infty}(W)$$

$$\leq \exp \left( - \Lambda^* \kappa \eta \right) + \frac{\sqrt{\beta}}{\Lambda^*_0} \eta^{1/2 - \kappa}$$

Geometric ergodicity Time discretization

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

Hilbert space

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

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

RKHS structure

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

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

Assumption (eigenvalue decay)

\[ \mu_k \simeq k^{-2} \]

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

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

**Dissipativity:**

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

$$\langle Cx - \nabla \hat{L}(x), x \rangle \leq -m\|x\|^2 + c.$$
**Assumption (2)**

- **Smoothness:**
  \[
  \| \nabla \hat{L}(x) - \nabla \hat{L}(y) \| \leq M \| x - y \|
  \]

- **Strong smoothness condition:**
  For \( \alpha \in (1/4, 1) \), (without this, the rate becomes slow)
  \[
  \| \nabla \hat{L}(x) - \nabla \hat{L}(y) \|_{-\alpha} \leq M \| x - y \|
  \]
  where
  \[
  \| x \|_{\varepsilon} = \left( \sum_{k \geq 0} (\mu_k)^{2\varepsilon} \| \langle x, f_k \rangle \|^2 \right)^{1/2}.
  \]
  (This is not standard, but, is satisfied in the previous examples)

- **Third order smoothness:**
  Let \( L_N = \hat{L}(P_N x) \). There exists \( \alpha' \in [0, 1) \) such that
  \[
  \| D^3 L_N(x) \cdot (h, k) \|_{\alpha'} \leq C_{\alpha'} \| h \|_0 \| k \|_0,
  \]
  \[
  \| D^3 L_N(x) \cdot (h, k) \|_0 \leq C_{\alpha'} \| h \|_{-\alpha'} \| k \|_0.
  \]
Risk bounds

\[ f_W(x) := \int_{\mathbb{R}^d} W_2(w)\sigma(W_1(w)^\top x) d\rho_0(w) \]

\[ L(W) := \mathbb{E}[\ell(Y, f_W(X))] \quad \hat{L}(W) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_W(x_i)) \]

Gen. error: (Gen. gap)

\[ L(\hat{W}) - \hat{L}(\hat{W}) \]

Excess risk:

\[ L(\hat{W}) - \inf_{f: \text{measurable}} L(f) \]

Optimization method (Infinite dimensional GLD):

\[ dW_t = -\nabla \left( \hat{L}(W_t) + \frac{\lambda}{2} \|W_t\|^2_{\mathcal{H}_K} \right) dt + \sqrt{\frac{2}{\beta}} d\xi_t \]

Time discretization

\[ W_{k+1} = S_\eta \left( W_k - \eta \nabla \hat{L}(W_k) + \sqrt{2\frac{\eta}{\beta}} \xi_k \right) \]

\[ S_\eta := (I + \frac{\lambda}{2} \nabla \cdot (\mathcal{H}_K)^{-1}) \]
Generalization error bound

\[ L(W) := \mathbb{E}[\ell(Y, f_W(X))] \quad \hat{L}(W) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_W(x_i)) \]

**Opt. error:**

\[ \hat{L}(W_k) - \int \hat{L}(w) d\pi_\infty(w) \leq \exp\left(-\frac{\Lambda^*_\eta k \eta}{\Lambda^*_0 \eta^{1/2-\kappa}}\right) + \frac{c\beta}{\Lambda^*_0 \eta^{1/2-\kappa}} \]

\[ \Xi_k \]

**Thm (Generalization error bound)**

\[ \mathbb{E}_{W_k}[L(W_k)] \leq \mathbb{E}_{W_k}[\hat{L}(W_k)] + \frac{R^2}{\sqrt{n}} \left[ 2 \left( 1 + \frac{2\beta}{\sqrt{n}} \right) + \log\left( \frac{1 + e^{R^2/2}}{\delta} \right) \right] + \Xi_k \]

with probability \(1 - \delta\).

PAC-Bayesian stability bound [Rivasplata, Kuzborskij, Szepesvári, and Shawe-Taylor, 2019]

**Assumption**

- Loss function \(\ell\) is “sufficiently smooth.”
- Loss and its gradients are bounded:
  \[ 0 \leq \ell(f_W, z) \leq R, \quad \|\nabla_W \ell(f_W, z)\|_\mathcal{H} \leq R \quad (\forall W \in \mathcal{H}, \; z \in \text{supp}(P)) \]
Excess risk evaluation

\[ L(f) := \mathbb{E}[\ell(Y, f(X))] \]

Additional assumption:

- \( \exists W^* \in \mathcal{H} \) s.t. \( \inf_f L(f) = L(f_{W^*}) = L(f^*) \)
- \( \exists \gamma > 1/4 \): model complexity
  \[ \hat{L}(W) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_{T_K^{\gamma/2}}^W(x_i)) \]
  \( (T_K^{a} x := \sum_{k=0}^{\infty} \mu_k^a x_k e_k \) where \( x = \sum_{k=0}^{\infty} x_k e_k \) and \( \|x\|_{\mathcal{H}_K}^2 = \sum_{k=0}^{\infty} \mu_k x_k^2 \)

- Bernstein condition [Erven et al., 2015]:
  \[ \mathbb{E}[(\ell(Y, f(X)) - \ell(Y, f^*(X)))^2] \leq B(L(f) - L(f^*))^s \]
  - Squared loss: \( s = 1 \)
  - Logistic loss with bounded \( f, f^* \): \( s = 1 \)

- \[ \mathbb{E} \left[ \exp \left( -\frac{\beta}{n} (\ell(Y, f(X)) + \ell(Y, f^*(X))) \right) \right] \leq 1 \]
  - Loss function needs not be a log likelihood.
  - The true distribution should has a light tail.
Fast rate: general result

Let $T_K^a x = \sum_{k=0}^{\infty} \mu_k^a x_k f_k$ where $x = \sum_{k=0}^{\infty} x_k f_k$ and $\|x\|_{\mathcal{H}_K}^2 = \sum_{k=0}^{\infty} \mu_k x_k^2$. Accordingly, define $\mathcal{H}_{\tilde{K}} = T_K^{(\gamma+1)/2} \mathcal{H}$ and $\mathcal{H}_{\tilde{K}}^\theta = T_K^{\theta(\gamma+1)/2} \mathcal{H}$.

**Thm (Excess risk bound: fast rate)**

Suppose that $W^* \in \mathcal{H}_{\tilde{K}}^\theta$ for $0 < \theta < 1 - \frac{1}{2(\gamma+1)}$. Then, for $\tilde{\alpha} = \frac{1}{2(\gamma+1)}$, it holds that

$$
\mathbb{E}_{D_n} \left[ \mathbb{E}_{W_k} [L(W_k)] - L(f_{W^*}) \right] 
\lesssim \max \left\{ (\lambda \beta)^{\frac{2\tilde{\alpha}/\theta}{2-s(1-\tilde{\alpha}/\theta)}} n^{-\frac{1}{2-s(1-\tilde{\alpha}/\theta)}}, \lambda^{-\tilde{\alpha}} \beta^{-1}, \lambda^\theta \right\} + \Xi_k
$$

Can be faster than $O(1/\sqrt{n})$.
Example: classification & regression

Model: \[ f_W(x) := \int_{\mathbb{R}^d} W_2(a) \sigma(W_1(w)^\top x) d\rho_0(a,w) \]

Classification

Strong low noise condition: \[ |P(Y = 1|X) - 1/2| \geq \delta \quad \text{(a.s.)} \]

For sufficiently large \( n \) and any \( \beta \leq n \),

\[
\mathbb{E}[P_{\pi_k}(\{W_k \in \mathcal{H} \mid P_X[\text{sign}(f_{W_k}(X)) = \text{sign}(f^*(X))] \neq 0\})] \\
\lesssim \exp(-c\beta\delta^{2m/(2m-d)}) + \frac{\Xi_k}{\delta^{2m/(2m-d)}}
\]

Regression

- \( \mathcal{H}: L_2(\rho_0) \)
- \( \mathcal{H}_\overline{K}: W^{a+d/2}(\mathbb{R}^d) \) (Sobolev space)
- \( \theta = \frac{2b}{2a+d} \) for \( b < a \)

If we set \( \lambda^{-1} = \beta = n \),

\[
\mathbb{E}_{D_n} \left[ \mathbb{E}_{W_k}[L(W_k)] - L(W^*) \right] \lesssim n^{-\frac{2\min\{a,b\}}{2a+d}} + \Xi_k
\]
Problem setting (teacher-student model)

Teacher-student model:

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

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

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

\[ W^* : \text{the true parameter satisfies} \]

\[ \|W^*\|^2_{H^\gamma} = \sum_{m=1}^{\infty} (w_{1,m}^2 + \|w_{2,m}\|^2)/m^{-2\gamma} \leq 1 \]

Observation model:

\[ y_i = f_{W^*}(x_i) + \varepsilon_i \quad (i = 1, \ldots, n) \]

From \( D_n = (x_i, y_i)_{i=1}^{n} \) (observed data), we estimate \( f_{W^*} \).

Excess risk (mean squared error):

\[ \mathbb{E}_{D^n} \left[ \| \hat{f} - f^\circ \|_{L_2(P_X)}^2 \right] \]  ➤ Convergence rate?

➤ Deep vs shallow?
Comparison between deep and shallow

Teacher-student model:
\[
\hat{f}_W(x) = \sum_{m=1}^{\infty} a_m w_{2,m} \sigma\left(b_m^{-1} w_{1,m} \top x\right)
\]

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

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

\(W^*\) : the true parameter satisfies
\[
\|W^*\|_{\mathcal{H}_\gamma}^2 = \sum_{m=1}^{\infty} (w_{1,m}^*^2 + \|w_{2,m}^*\|^2)/m^{-2\gamma} \leq 1
\]

[Theorem & Suzuki&Akiyama, ICLR2021]

Estimation error \(\mathbb{E}[\|\hat{f} - f^*\|_{L_2(P)}^2]\) can be bounded by

Deep

\[n \succ \frac{\gamma}{\alpha_1 - 3\alpha_2 + 1}\]

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

DL trained by GLD

Linear (kernel)

\[R_{\text{lin}}(\mathcal{F}_\gamma) \succ n - \frac{2\tilde{\beta} + d}{2\tilde{\beta} + 2d}\]

\[\tilde{\beta} = \frac{\alpha_1 + \alpha_2}{\alpha_2 - \gamma/2}\]

Worst case error of kernel method
Comparison between deep and shallow

Teacher-student model:

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

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

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

\[ W^* \text{: the true parameter satisfies} \]

\[ \|W^*\|_{\mathcal{H}}^2 = \sum_{m=1}^{\infty} \left( w_{1,m}^* \right)^2 \leq 1 \]

[Theorem: Suzuki&Akiyama, ICLR2021]

Estimation error \( \mathbb{E}[\|\hat{f} - f^*\|_{L^2(P)}^2] \) can be bounded by

**Deep**

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

\[ \frac{1}{n} \]

DL trained by GLD

**Linear (kernel)**

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

\[ \frac{1}{\sqrt{n}} \]

Worst case error of kernel method

\[ \sqrt{n}\text{-times large!!} \]
Particle optimization method in mean field regime


Mean field limit of 2-layer NN

2-layer neural network:

\[ f(x) = \frac{1}{M} \sum_{j=1}^{M} r_j \sigma(w_j^T x) \]

Non-linear with respect to parameters \((r_j, w_j)^M_{j=1}\).

Overparameterization (Mean field limit):

\[ f(x) = \frac{1}{M} \sum_{j=1}^{M} r_j \eta(w_j^T x) \xrightarrow{M \to \infty} \int r \sigma(w^T x) d\nu(r, w) \]

Linear with respect to the prob. measure \(\nu\).

Particles move to fit the model to the data.
**Objective**

\[
\min_{\nu: \mathcal{P}(\Theta)} \frac{1}{n} \sum_{i=1}^{n} \ell \left( \mathbb{E}_{\theta \sim \nu} [h_{\theta}(x_i)], y_i \right) + \lambda \mathbb{E}_{\nu} [\|\theta\|^2]
\]

- **\ell**: Smooth loss function
- **\(h_{\theta}\)**: Neuron with param. \(\theta\)
- i.e., \(h_{\theta}(x) = r \sigma(w^\top x)\) for \(\theta = (r, w)\)

**Prob meas.**

**L2-regularization**

**Negative entropy regularization**

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

\(\lambda_2 \text{KL}(\nu, N(0, \lambda_2/\lambda_1 I))\)

**KL-div from a Gaussian distribution.**

\(A \text{ convex function with respect to the density function } q.\)

\(\rightarrow\) We can apply a standard convex optimization technique.

**Difficulty:** We don’t have any closed form representations of the expectations.

\(\rightarrow\) **Solution:**
- Particle approximation.
- Sampling from gradient Langevin dynamics.
Outline of the proposed algorithm

\[
\begin{align*}
\min_{q:\text{prob.\ density}} & \quad \frac{1}{n} \sum_{i=1}^{n} \ell \left( \mathbb{E}_q[h_{\theta}(x_i)], y_i \right) + \lambda_1 \mathbb{E}_q[\|\theta\|^2] + \lambda_2 \mathbb{E}_q[\log(q)] \\
\text{Approximate this by a linear functional of } q. \quad \text{e.g., } & \quad \mathbb{E}_{\theta \sim q}[\tilde{g}^{(t)}(\theta)] \quad \text{(which is something like a gradient w.r.t. } q) \\
\text{Dual averaging} & \quad \text{(Nesterov, 2005; 2009; Xiao, 2009)} \\
\min_{q:\text{prob.\ density}} & \quad \mathbb{E}_{\theta \sim q}[\tilde{g}^{(t)}(\theta)] + \lambda_2 \mathbb{E}_q[\log(q)] \\
\text{Solution: } & \quad q^{(t+1)}(\theta) \propto \exp(-\tilde{g}^{(t)}(\theta)/\lambda_2) \quad \text{Explicit form} \\
\rightarrow & \quad \text{This is the stationary distribution of the gradient Langevin dynamics:} \\
\frac{d\theta_t}{dt} & = -\nabla(\tilde{g}^{(t)}(\theta)/\lambda_2) dt + \sqrt{2} d\xi_t. \\
\theta_k & = \theta_{k-1} - \eta \nabla\tilde{g}^{(t)}(\theta)/\lambda_2 + \sqrt{2\eta} \xi_{k-1} \\
\text{Gradient Langevin dynamics} & \\
\text{The dual averaging method employs} & \\
\tilde{g}^{(t)}(\theta) & \leftarrow \ell'(\mathbb{E}_{\theta \sim q^{(t)}}[h_{\theta'}(x_{i_t})], y_{i_t}) h_{\theta}(x_{i_t}) + \lambda_1 \|\theta\|^2 \\
\bar{g}^{(t)} & \leftarrow \frac{2}{(t+2)(t+1)} \sum_{s=1}^{t} s g^{(s)} \\
\mathbb{E}_{q^{(t)}}[h_{\theta}(x)] & \simeq \frac{1}{M} \sum_{r=1}^{M} h_{\theta(r)}(x) & \text{Can be approximated by GLD.}
\end{align*}
\]
Algorithm description

Algorithm 1 Particle Dual Averaging (PDA)

Randomly draw i.i.d. initial parameters \( \tilde{\Theta}^{(1)} \sim q^{(1)}(\theta) d\theta \) \( (r \in \{1, 2, \ldots, M\}) \)

\[ \tilde{\Theta}^{(1)} \leftarrow \left\{ \tilde{\theta}^{(1)}_r \right\}_{r=1}^M \]

for \( t = 1 \) to \( T \) do

Randomly draw a data index \( i_t \) from \( \{1, 2, \ldots, n\} \)

\[ g^{(t)} \leftarrow \partial_z \ell(h_{\tilde{\Theta}^{(t)}}(x_{i_t}, y_{i_t}) h(\cdot, x_{i_t}) + \lambda_1 \| \cdot \|_2^2 \]

\[ \bar{g}^{(t)} \leftarrow \frac{2}{(t+2)(t+1)} \sum_{s=1}^t s g^{(s)} \]

Obtain \( q^{(t+1)} \) by running the Langevin algorithm to approximate the following density function:

\[ q^{(t+1)} \propto \exp \left( -\bar{g}^{(t)}/\lambda_2 \right) . \]

\[ \tilde{\Theta}^{(t+1)} \leftarrow \left\{ \tilde{\theta}^{(t+1)}_r \right\}_{r=1}^M \text{ where } \tilde{\theta}^{(t+1)}_r \sim q^{(t+1)} \]

end for

Randomly pick up \( t \) from \( \{2, 3, \ldots, T + 1\} \) following the probability \( P[t] = \frac{2^t}{T(T+3)} \) and return \( h_{\tilde{\Theta}^{(t)}} \)

In each iteration, the potential for updating each particle is given by \( \bar{g}^{(t)} \).
Convergence analysis

**Theorem (informal)**

1. **Outer loop:**
   \[ \mathcal{L}(\hat{q}) - \mathcal{L}(q^*) \leq O(1/T) \]

2. **Inner loop:**
   By setting the step size at the \( t \)-th iteration as \( \eta_t = O\left(\frac{\lambda_1 \lambda_2}{t^2 \exp(8/\lambda_2)}\right) \),
   \[ T_t = \tilde{O}\left(\eta_t^{-1}\right) = \tilde{O}\left(t^2 \exp(8/\lambda_2)/(\lambda_1 \lambda_2)\right) \]
   is sufficient for the number of inner iterations (GLD updates).

**Total complexity:**

\[ O(\varepsilon^{-3}) \] GLD updates to obtain \( \varepsilon \)-optimal solution.

The network width (\# of particles) \( M = \varepsilon^{-2} \text{poly}(n, d) \) is sufficient to obtain the iteration complexity described above.

- **Polynomial order**
- **Simple analysis**
Modification to SDCA

[Oko, Suzuki, Nitanda, Wu (2021)]

• **Motivation:**
  - We want to improve the outer-iteration complexity for **finite sample ERM setting**.
  - SDCA (Stochastic Dual Coordinate Ascent) achieves linear convergence:
    \[
    \left( n + \frac{L}{\mu} \right) \log(1/\epsilon).
    \]

    ※ DA: \(1/\epsilon\)

• **Difficulty:**
  - How to combine gradient Langevin sampling and SDCA?
  - We want to skip the number of exact sampling as many as possible.
    (One iteration of GLD requires \(O(n)\) computation!)
**Fenchel dual**

**Primal**

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

\[
\min_{x \in X} f(Ax) + g(x) = -\min_{g \in Y^*} f^*(g) + g^*(-A^* g)
\]  
(Fenchel’s duality theorem)

\[
A : X \rightarrow Y \text{ (bounded linear)}
\]

**Dual**

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

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

where

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

\[
p[g](\theta) := \frac{q[g](\theta)}{\int q[g](\theta') d\theta'}
\]

**Strategy:**

- We randomly pick-up one coordinate \( i \in [n] \). (sampling one data point)
- Update \( g_i \) by minimizing the dual problem: coordinate descent.
\[
\min_{g_i \in \mathbb{R}} D(g) = \frac{1}{n} \sum_{i=1}^{n} \ell_i^*(g_i) + \lambda_2 \log \left( \int q[g](\theta) d\theta \right)
\]

We update just one coordinate \( g_i \) per iteration.

(ideal update)

\[
\begin{cases}
\tilde{g}_{i}^{(t+1)} := \arg \min_{g_i \in \mathbb{R}} \left\{ \ell_i^*(g_i) - \int p^{(t)}(\theta) h_i(\theta) d\theta (g_i - \bar{g}_i^{(t)}) + \frac{1}{2n\lambda_2} (g_i - \bar{g}_i^{(t)})^2 \right\} \\
\bar{g}_j^{(t+1)} = \bar{g}_j^{(t)} \quad (j \neq i) \\
p^{(t+1)}(\theta) := p[\bar{g}^{(t+1)}](\theta)
\end{cases}
\]

(proximal gradient descent (2\(^{nd}\) term is linearized))

(requires integration)

\[
\int p^{(t)}(\theta) h_i(\theta) d\theta \approx \sum_{m=1}^{M} r^{(t)}_m h_i(\theta_m)
\]

\[
r^{(0)}_m = \frac{1}{M}, \quad \delta \tilde{g}_i^{(t+1)} := \tilde{g}_i^{(t+1)} - \bar{g}_i^{(t)}
\]

\[
\begin{cases}
\tilde{r}^{(t+1)}_m = r^{(t)}_m \exp \left( -\frac{1}{n} h_i(\theta_m) \delta \tilde{g}_i^{(t+1)} \right) \\
r^{(t+1)}_m = \frac{\tilde{r}^{(t+1)}_m}{\sum_{m=1}^{M} \tilde{r}^{(t+1)}_m} \quad (m \in [M])
\end{cases}
\]

\[
p^{(t)}(\theta) \propto \exp \left\{ -\frac{1}{\lambda_2} \left( \frac{1}{n} \sum_{i=1}^{n} h_i(\theta) g_i^{(t)} + \lambda_1 \|\theta\|^2 \right) \right\}
\]

\[
\rightarrow \text{We can sample particles via GLD.} \quad \theta_m \sim p^{(t)} \quad (m = 1, \ldots, M)
\]

We “refresh” particles each \( \tilde{n} \) iteration.
Algorithm 2 Dual Coordinate Descent with the particle method

Require: training data \(\{(x_i, y_i)\}_{i=1}^n\) and numbers of inner-loop iterations \(\tilde{n}\) and outer-loop iterations \(T_{end}\).

1: Choose \(g^{(0)}_i\) s.t. \(|\ell^*_i(g^{(0)}_i)| \leq 1\) for \(i = 1, \ldots, n\) and \(\ell^*_i(g^{(0)}_i) \leq \ell^*(0)\).
2: \(g^{(0)} \leftarrow 0\).
3: for \(T = 0, 1, \ldots, T_{end} - 1\) do
4: Randomly (approximately) draw i.i.d. parameters \(\theta_m\) for \(m = 1, \ldots, M(\tilde{n}T)\) from \(p(\tilde{n}T)(\theta)\) that satisfies \(\text{TV}\left(p(\tilde{n}T) \| p[g(\tilde{n}T)]\right) \leq \epsilon(\tilde{n}T)\).
5: \(r^{(\tilde{n}T)}_m \leftarrow \frac{1}{M(\tilde{n}T)}\) for \(m = 1, \ldots, M(\tilde{n}T)\).
6: for \(t = \tilde{n}T, \tilde{n}T + 1, \ldots, \tilde{n}T + \tilde{n} - 1\) do
7: Randomly choose \(i_t\) from \(\{1, 2, \ldots, n\}\).
8: \(g^{(t+1)}_{i_t} \leftarrow \arg\max_{g_{i_t} \in \mathbb{R}} \left\{ -\ell^*_i(g_{i_t}) + \sum_{m=1}^{M(\tilde{n}T)} r^{(t)}_m h_{i_t}(\theta_m)(g_{i_t} - g^{(t)}_{i_t}) - \frac{1}{2n\lambda_2}(g_{i_t} - g^{(t)}_{i_t})^2 \right\} \).
9: \(r^{(t+1)}_m \leftarrow r^{(t)}_m \exp\left(-\frac{1}{n\lambda_2} h_{i_t}(\theta_m)(g^{(t+1)}_{i_t} - g^{(t)}_{i_t})\right)\) for \(m = 1, \ldots, M(\tilde{n}T)\).
10: end for
11: end for
12: return Option (A): \(g^{(A)}_{out} = g^{(\tilde{n}T_{end})}\); Option (B): \(g^{(B)}_{out} = g^{(t_{end}')}\) for \(t_{end}'\) that is randomly chosen from \(\{\tilde{n}T_{end} - n + 1, \ldots, \tilde{n}T_{end}\}\).

At every \(\tilde{n}\) iteration, we refresh particles.

Particle weight update
Dual coordinate ascent
(A1) $\ell_i$ is $\gamma$-smooth.

(A2) $|h_i(\theta)| \leq 1$ for all $\theta$.

(A3) Other technical conditions.

$$g_i^{(t+1)} := \arg\min_{g_i \in \mathbb{R}} \left\{ \ell_i^*(g_i) - \sum_{m=1}^{M} r_m^{(t)} h_i(\theta_m)(g_i - g_i^{(t)}) + \frac{1}{2n\lambda_2} (g_i - g_i^{(t)})^2 \right\}$$

**Lemma (informal)**

It holds that

$$|g_i^{(t)} - \bar{g}_i^{(t)}| \lesssim \sqrt{\frac{1}{M} \log(n/\delta)}$$

uniformly over $i \in [n], t \in [n]$ with probability $1 - \delta$.

If $t > n$, the error can exponentially diverge.

$\Rightarrow$ We re-sample $(\theta_m)_{m=1}^{M}$ by GLD at each $t = \tilde{n}$ updates.
Convergence rate

(A1) $\ell_i$ is $1/\gamma$-smooth.

(A2) $|h_i(\theta)| \leq 1$ for all $\theta$.

(A3) Other technical conditions.

**Theorem (convergence rate, informal)**

Suppose that $\frac{\bar{n}}{n\lambda_2} = o(1)$ and the number of particles satisfies

$$M^* \gtrsim \frac{1}{\epsilon_P \lambda_2}.$$  

Then,

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

iterations are sufficient to achieve $\epsilon_P$ duality gap:

(Duality gap) $\mathbb{E}[P(p^{(t_{\text{end}})}) - D(g^{(t_{\text{end}})})] \leq \epsilon_P$

**Total complexity:**

$$M^* \left(1 + \frac{K^*}{\bar{n}}\right) \left(n + \frac{1}{\lambda_2 \gamma}\right) \log(n/\epsilon_P)$$

If deterministic optimization is used, the number of gradient evaluations become

$$t_{\text{end}} = O\left(\frac{n}{\lambda_2 \gamma} \log(1/\epsilon_P)\right)$$
Sampling algorithm

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

\[ U(\theta) := \frac{1}{\lambda_2} \left( \frac{1}{n} \sum_{i=1}^{n} h_i(\theta)g_i + \lambda_1 \|\theta\|^2 \right) \]

- **ULA (Unadjusted Langevin algorithm)**

\[ \theta^{k+1} = \theta^k - \eta \nabla U(\theta^k) + \sqrt{2\eta} \xi_k \]

\[ \xi_k \sim N(0, I) \]

- **MALA (Metropolis adjusted Langevin algorithm)**

\[ \tilde{\theta}^{k+1} = \theta^k - \eta \nabla U(\theta^k) + \sqrt{2\eta} \xi_k \]

The proposal is accepted with prob. \( \alpha \) and rejected otherwise:

\[ \alpha = \min \left\{ 1, \frac{U(\tilde{\theta}_{k+1})q(\theta^k|\tilde{\theta}^{k+1})}{U(\theta_k)q(\tilde{\theta}^{k+1}|\theta^k)} \right\} \]

\[ q(\theta'|\theta) \propto \exp(-\|\theta' - \theta - \eta \nabla U(\theta)\|^2 / 4\eta) \]
Log-Sobolev inequality

\[ p(\theta) \propto \exp \left\{ -\frac{1}{\lambda_2} \left( \frac{1}{n} \sum_{i=1}^{n} h_i(\theta) g_i + \lambda_1 \|\theta\|^2 \right) \right\} : \pi \]

Log-Sobolev inequality with a constant \( c_{\text{LS}} \)

\[
d\nu(\theta) = f(\theta) d\pi(\theta) \quad \text{(probability measure)}
\]

\[
\int f \log(f) d\pi \leq 2c_{\text{LS}} \int \frac{\|\nabla f\|^2}{f} d\pi \quad \text{for} \quad (D(\nu||\pi_\infty) \leq 2c_{\text{LS}} I(\nu||\pi_\infty))
\]

Lemma

\[
\|h_i\| \leq 1, \quad |g_i| \leq B
\]

\[
c_{\text{LS}} = \frac{2\lambda_1}{\lambda_2} \exp(-4B/\lambda_2)
\]

Convergence analysis

TV($p||p_k$) : TV-distance between the target $p$ and the marginal distribution of the $k$-th step sample.

TV($p||p_{K^*}$) $\leq \epsilon_C$

- **ULA (Unadjusted Langevin algorithm)**

\[ K^* = O\left(\frac{L^2}{c_{LS}^2} \frac{d}{\epsilon_C} \log(1/(\lambda_2 \epsilon_C))\right) \]

[Vempala and Wibisono, 2019]

- **MALA (Metropolis adjusted Langevin algorithm)**

\[ K^* = O\left(\frac{1}{c_{LS}^{5/2}} \left(\frac{L}{\lambda_2} + \log(1/\epsilon_C)\right)^{3/2} d\right) \]

Large conductance $\leftarrow$ log-Sobolev


$y = \sigma(w^*_x + b^*) + \epsilon$

Experiments

- $\lambda_1 = 10^{-2}$: fixed

- (a) $\lambda_2 = 0.01$

- (b) $\lambda_2 = 0.001$

- (c) $\lambda_2 = 0.0001$
Convergence in teacher-student setting

[Shunta Akiyama, Taiji Suzuki: On Learnability via Gradient Method for Two-Layer ReLU Neural Networks in Teacher-Student Setting. ICML2021]

Shunta Akiyama
Problem setting

Noiseless observation:

\[ y_i = f^\circ (x_i) \quad (i = 1, \ldots, n) \]

where \( x_i \sim \text{Unif}(\mathbb{S}^{d-1}) \).

**Teacher-student model with ReLU activation:**

**Teacher**

\[ f^\circ (x) = \sum_{j=1}^{m} a^\circ_j \sigma(\langle w^\circ_j, x \rangle) \]

\( \sigma : \text{ReLU} \)

**Student (overparameterization)**

\[ f_\Theta (x) = \sum_{j=1}^{M} a_j \sigma(\langle w_j, x \rangle) \]

- **Overparameterized setting:** \( M \gg m \).
- Can the student model estimate the teacher model by GD?
Sparse regularized learning:

\[
\min_{\Theta=(a_j, w_j)_{j=1}^M} F(\Theta) = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{M} a_j \sigma(\langle w_j, x \rangle) \right)^2 + \lambda \sum_{j=1}^{M} |a_j| \|w_j\|
\]

\[
\sum_{j=1}^{M} |a_j| \|w_j\| \leq \frac{1}{2} \sum_{j=1}^{M} (a_j^2 + \|w_j\|^2)
\]

**Sparse regularization**

Weight decay yields sparse regularization.

**Norm-dependent step size for gradient descent:**

**Initialization**

- **Mean field setting**
  
  \[
  a_j^{(0)} = \begin{cases} 
  2/M & (1 \leq j \leq M/2) \\
  -2/M & (M/2 + 1 \leq j \leq M)
  \end{cases}
  \]
  
  \[
  w_j^{(0)} \sim \text{Unif}(S^{d-1}) & (1 \leq j \leq M)
  \]

**Gradient descent**

\[
a_j^{(k+1)} = a_j^{(k)} - \eta_{j,k} \partial_{a_j} F(\Theta^{(k)})
\]

\[
w_j^{(k+1)} = w_j^{(k)} - \eta_{j,k} \partial_{w_j} F(\Theta^{(k)})
\]

\[
\eta_{j,k} = \alpha \frac{|a_j^{(k)}| \|w_j^{(k)}\|}{a_j^{(k)} + \|w_j^{(k)}\|^2}
\]

Norm dependent step-size
• **Result 2:** Convergence of GD

\[ J^* = \inf_{\nu} J_{\lambda}(\nu) \]

**Theorem (informal)**

There exists \( J_0 \) such that \( J^* < J_0 < J_{\lambda}(\nu_0) \) and sufficiently large \( M \) such that

**Stage 1 (Global exploration):** \( \exists k_0 \geq \Omega(\sqrt{J_0 - J^*}) \) such that

\[ J_{\lambda}(\nu_{k_0}) - J^* \leq J_0 - J^*. \]

**Stage 2 (Local convergence):** \( \exists \zeta > 0 \) such that

\[ J_{\lambda}(\nu_k) - J^* \leq (1 - \zeta)^{k-k_0} (J_{\lambda}(\nu_{k_0}) - J^*) \quad (\forall k \geq k_0). \]

Dual certificate + convergence guarantee by Chizat (2019) + some technical modifications for ReLU.

- \( M \) could be \( \exp(\Omega(d)) \).
- It also holds that \( \widetilde{W}_2^2(\nu_k, \nu^*) \leq O((1 - \zeta)^{k-k_0}) \), but we **don’t** have \( ||\Theta_k - \Theta^*|| \to 0 \).
Numerical experiment

Optimization dynamics

The parameter does not converge to the true one.
The measure representation converges to the true one.

Linear model requires $\epsilon^{-d}$ neurons [Yehudai and Shamir, 2019].
The solution with sparse regularization possesses only $m$ atoms. (resolving the curse of dimensionality)
Double descent and optimization

[Amari, Ba, Grosse, Li, Nitanda, Suzuki, Wu, Xu: When Does Preconditioning Help or Hurt Generalization? ICLR2021]

Denny Wu

[Belkin et al.: Reconciling modern machine learning practice and the bias-variance trade-off. 2018]

- Even if the model size is larger than the sample size, it can generalize.
- The variance decreases as the model complexity increases.
Preconditioned Gradient Descent

\[ y_i = x_i^\top \beta^* + \epsilon_i \]

\[ \min_{\beta \in \mathbb{R}^d} \; \left\| \beta \right\|_P^{-1}^2 \quad \text{s.t.} \quad y_i = x_i^\top \beta \quad (\forall i \in [n]) \]

Q: How does the preconditioner \( P \) affect the predictive accuracy?

\[ \frac{d\beta(t)}{dt} = -PX^\top (Y - X\beta(t))/n \]

\( P = I \): Gradient descent (GD)

\( P = \Sigma_x^{-1} \): Natural Gradient descent (NGD)  
(population Fisher)
Optimal choice of preconditioner

Bias-variance decomposition

\[ \mathbb{E}_{x,D_n}[(x^\top \hat{\beta}^* - x^\top \hat{\beta})^2] = \mathbb{E}_x[(x^\top \beta^* - x^\top \mathbb{E}[\hat{\beta}])^2] + \operatorname{tr}[\operatorname{Cov}(\hat{\beta})\Sigma_x] \]

\(B: \text{Bias}\)
\(V: \text{Variance}\)

Theorem (informal)
[Amari, Ba, Grosse, Li, Nitanda, Suzuki, Wu, Xu: When Does Preconditioning Help or Hurt Generalization? ICLR2021]

We derived an exact form of the asymptotic risk when \(d/n \to \gamma > 1\) as \(n \to \infty\).

1. Variance:

\(P = \Sigma_x^{-1}\) (population cov) minimizes the variance.

NGD is optimal in terms of variance.

2. Bias:

No free-lunch: the optimal \(P\) is not known \(a\ priori\):
- GD generalizes better when the target is \(\text{isotropic} \ \Sigma_{\beta^*} = I\).
- NGD is better when the target is \(\text{misaligned} \ \Sigma_{\beta^*} = \Sigma_x^{-1}\).

(Bayesian setting: Average predictive risk over a random \(\beta^*\) with \(\mathbb{E}[\beta^* \beta^{*\top}] = \Sigma_{\beta^*}\))

Interpolation of GD and NGD is beneficial.

\[\begin{align*}
\text{Additive:} & \quad P = (\alpha \Sigma_x + (1 - \alpha)I_d)^{-1} \\
\text{Geometric:} & \quad P = \Sigma_x^{-\alpha} \\
\alpha & \in [0, 1]
\end{align*}\]
(A2) The spectral distribution of $\Sigma_{XP} := P^{1/2}\Sigma P^{1/2}$ converges weakly to $H_{XP}$.

**self-consistent equation:**

$$\frac{1}{m(z)} = -z + \gamma \int \frac{\tau}{1 + \tau m(z)} dH_{XP}(\tau)$$

$\rightarrow$ Limiting distribution of eigenvalues of $\frac{1}{n}XPX^\top$.

1. Variance:

$$V \xrightarrow{P} \sigma^2 \left( \lim_{\lambda \to +0} m'(-\lambda)m^{-2}(-\lambda) - 1 \right)$$

$$V \geq \sigma^2(\gamma - 1)^{-1} \text{ and equality holds by } P = \Sigma^{-1}$$

(A3) $P$ and $\Sigma$ shares the same eigenvectors $U$.

2. Bias:

$$\mathbb{E}_{\beta^*}[B] \xrightarrow{P} \lim_{\lambda \to +0} m'(-\lambda)m^{-2}(-\lambda)\mathbb{E}[v_x v_\theta (1 + v_{xp}m(-\lambda))^{-2}]$$

where $(e_x, e_\theta, e_{xp})$ are eigenvalues of $\Sigma, \Sigma_{XP}, \text{diag}(U^\top \Sigma_{\beta^*}U)$ and jointly converge weakly to $(v_x, v_\theta, v_{xp})$. 
**Summary**

- **Optimization theory**
  - SGD in Neural Tangent Kernel regime
  - Noisy gradient descent: a near global optimum
    - Estimation error separation between kernel and deep learning
  - Particle gradient method in mean field regime
    - Combination of known 1st order optimization technique and particle sampling
  - Optimization method selection for minimum norm interpolator

In deep learning, optimization and generalization cannot be separated.

More detailed analysis will be required by bridging these two research fields: feature extraction, loss landscape, benign overfitting…