Feature learning theory in multi-task and in-context learning

Taiji Suzuki

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

26th/Mar/2024

FIMI2024

Deep learning theory

Why does deep learning work well?

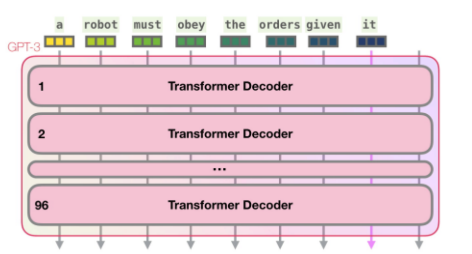
- **Several theoretical work has been conducted.**
- **<u>There are still many things that should be explored.</u>**

- Clarify the principle of deep learning
- What is essential to realize a "good" learning system?

In this presentation:

Feature learning

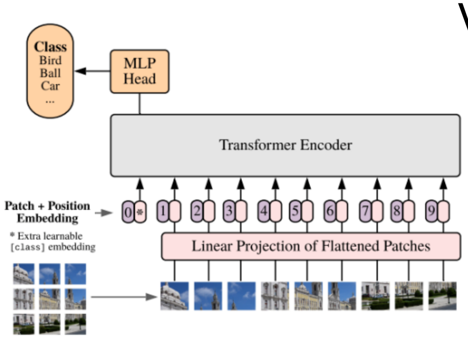
GPT



[Alammar: How GPT3 Works - Visualizations and Animations, https://jalammar.github.io/how-gpt3-works-visualizations-animations/]

[Brown et al. "Language Models are Few-Shot Learners", NeurIPS2020]

ViT



[Dosovitskiy et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929. ICLR2021]

Outline of this talk

2-layer NN

Linear $$f(z) = \beta^\top W x$$

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

Multitask learning/In-context learning

1. **Statistical analysis** for high dimensional regression
2. **Optimization guarantee** for in-context feature learning of Transformer

Effect of feature learning in interpolation regime

[Keita Suzuki, Taiji Suzuki: Optimal criterion for feature learning of two-layer linear neural network in high dimensional interpolation regime. ICLR2024]

High dimensional regression

High dimensional linear regression: $\beta_* \in \mathbb{R}^d$

$$y_i = \beta_*^\top x_i + \epsilon_i \qquad (i = 1, \ldots, n)$$

where $\mathbb{E}[x_i] = 0, \mathbb{E}\left[x_i x_i^\top\right] = \Sigma_X, \mathbb{E}[\epsilon_i] = 0, \mathbb{E}[\epsilon_i] = \sigma^2.$

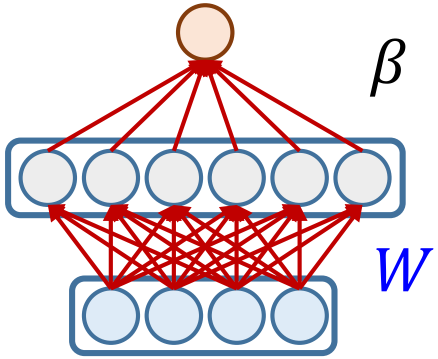
High dimensional setting: $d > n$

Ridge regression: $Y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n, \; X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times d}$

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{n} \|Y - X\beta\|^2 + \lambda \|\beta\|^2$$

Q: How can the predictive error be improved by using a two layer network?

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{n} \|Y - XW^\top \beta\|^2 + \lambda \|\beta\|^2$$



Predictive error

Predictive error: $R(\hat{\beta}) = \mathbb{E}_x[(x^\top \beta_* - \hat{\beta}^\top x)^2]$

$\lambda_i = \mu_i(\Sigma_X)$

Proposition (Tsigler and Bartlett (2020))

When Σ_X is diagonal, then the predictive error can be evaluated as follows:

$$R(\hat{\beta}) \simeq B + V \quad \text{(Bias-Variance trade-off)}$$

$$B = \sum_{j=1}^{k} \beta_{*,j}^2 \frac{(n\lambda + \sum_{j>k} \lambda_j)^2}{n^2 \lambda_j} + \sum_{j=k+1}^{d} \beta_{*,j}^2 \lambda_j$$

$$V = \frac{k}{n} + n \frac{\sum_{j=k+1}^{d} \lambda_j^2}{(n\lambda + \sum_{j>k} \lambda_j)^2}$$

The tail of eigenvalues of covariance matrix Σ_X plays important role.

- Fast decay of λ_j does not generalize when $\lambda = 0$: **Kernel regime**
- Slow decay of λ_j plays regularization
  → Generalize even if $\lambda = 0$: **Benign overfitting**
- Slow decay of λ_j and large d does not generalize: **Harmful overfitting**

Eigenvalue decay and generalization

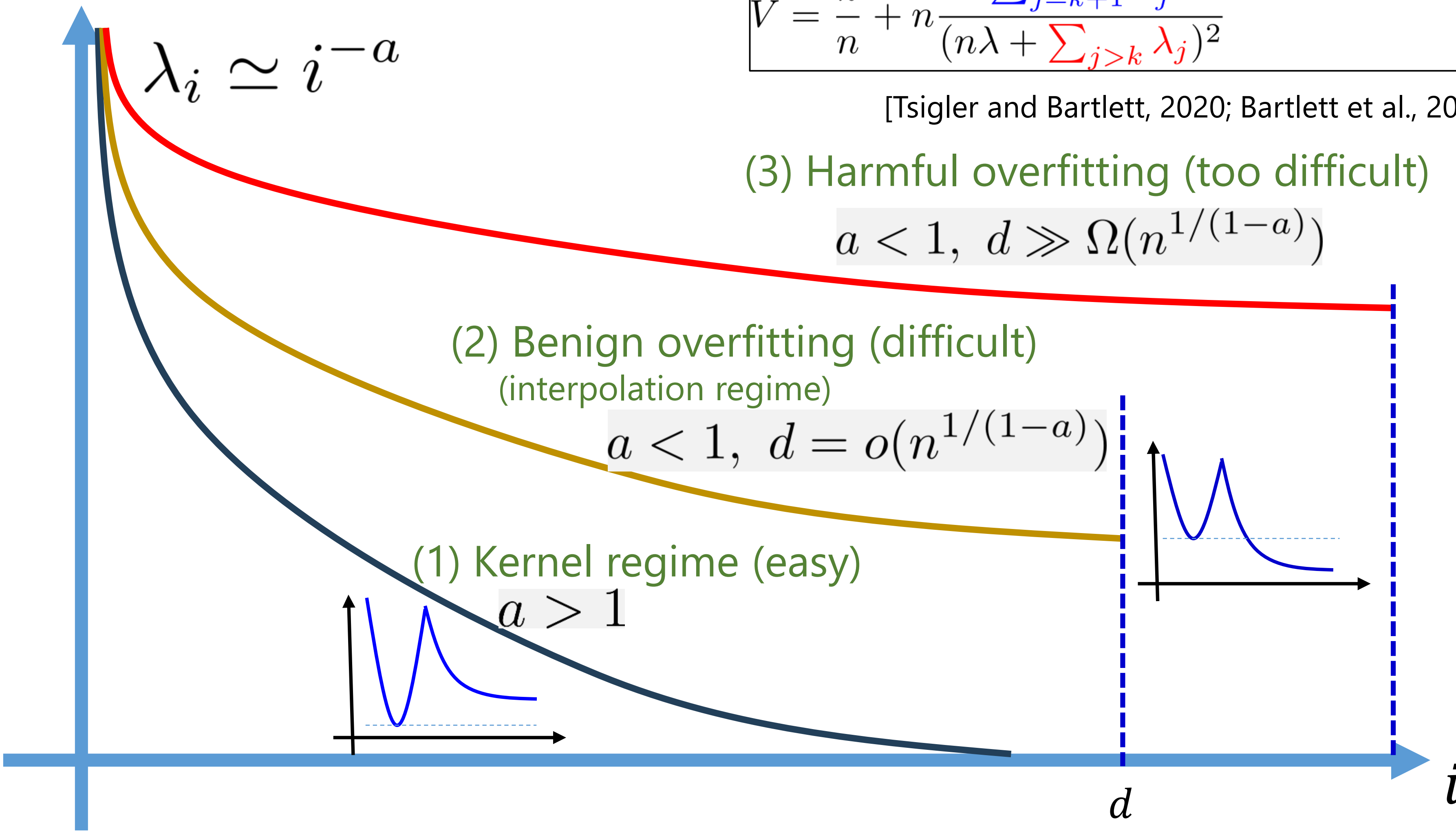
$$\lambda_i = \mu_i(\Sigma_X)$$

$$(\|\beta_*\|^2 < \infty)$$

$$B = \sum_{j=1}^{k} \beta_{*,j}^2 \frac{(n\lambda + \sum_{j>k} \lambda_j)^2}{n^2 \lambda_j} + \sum_{j=k+1}^{d} \beta_{*,j}^2 \lambda_j$$

$$V = \frac{k}{n} + n \frac{\sum_{j=k+1}^{d} \lambda_j^2}{(n\lambda + \sum_{j>k} \lambda_j)^2}$$

[Tsigler and Bartlett, 2020; Bartlett et al., 2019]



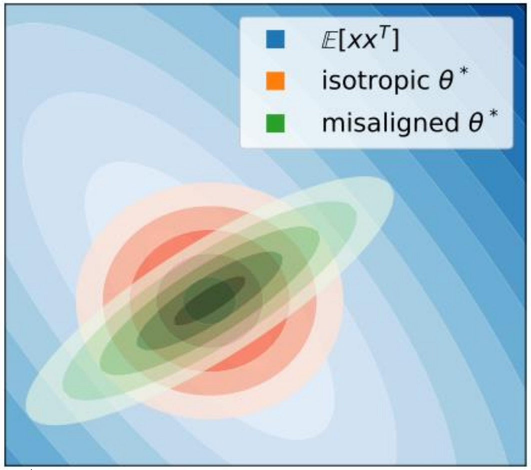
Optimal regularization

$$B=\sum_{j=1}^{k}\beta_{*,j}^2\frac{(n\lambda+\sum_{j>k}\lambda_j)^2}{n^2\lambda_j}+\sum_{j=k+1}^{d}\beta_{*,j}^2\lambda_j,\quad V=\frac{k}{n}+n\frac{\sum_{j=k+1}^{d}\lambda_j^2}{(n\lambda+\sum_{j>k}\lambda_j)^2}$$

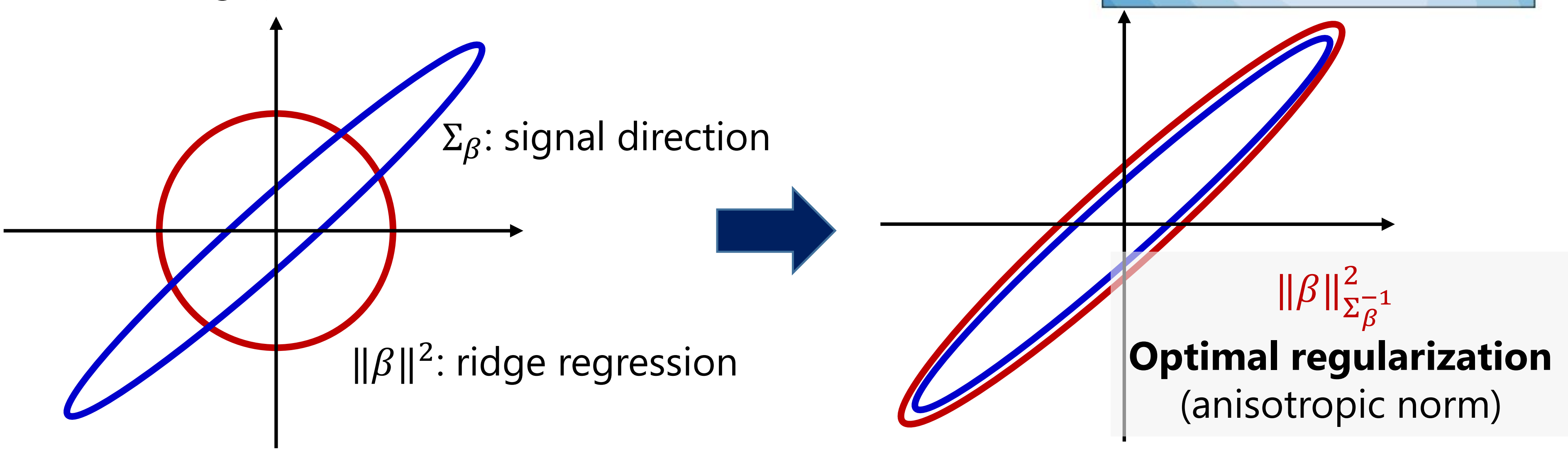
Suppose that $\beta_* \sim \Sigma_\beta$.

- (1) **Slow decay** of eigenvalue λ_j
- (2) **Misalignment** between β and x

→ Bad predictive error.
(Predictive error does not go to 0)



Misalignment:
β has large value toward non-principle components of x (large j)

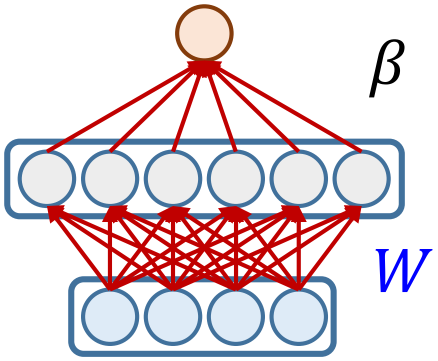


2 layer NN model

Student model
(2 layer linear NN)

$$f(x) = \beta^\top W x$$

$(W \in \mathbb{R}^{d \times d})$



$$\min_{\beta} \frac{1}{n} \|Y - X W^\top \beta\|^2 + \lambda \|\beta\|^2$$
$$\Leftrightarrow \min_{\tilde{\beta}} \frac{1}{n} \|Y - X\tilde{\beta}\|^2 + \lambda \|\tilde{\beta}\|^2_{(W^\top W)^{-1}}$$

$\tilde{\beta} = W^\top \beta$

Feature learning = Metric learning

We want to find the optimal W such that $WW^\top = \Sigma_\beta$.

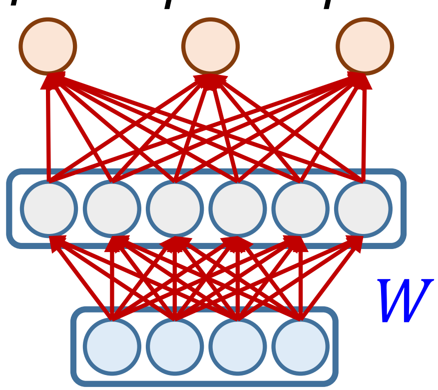
→ We need information of $\beta's$ distribution (i.e., Σ_β).

Multi-task learning (pre-training):

$$\min_{W \in \mathbb{R}^{d\times d}, \beta^{(j)} \in \mathbb{R}^d} \sum_{j=1}^{m} \frac{1}{n} \|Y^{(j)} - X W^\top \beta^{(j)}\|^2 + \lambda \|\beta^{(j)}\|^2$$

$$y_i^{(j)} = \beta_*^{(j)\top} W x_i + \epsilon_i^{(j)}$$

Each task t has the true coefficient $\beta_*^{(j)}$.



How W affects the result

- **Vanilla ridge regression:**

$$f(x) = \beta^\top x$$

Eigenvalues of Σ_X

characterizes the predictive risk.

$$B = \sum_{j=1}^{k} \beta_j^2 \frac{(n\lambda + \sum_{j>k} \lambda_j)^2}{n^2 \lambda_j} + \sum_{j=k+1}^{d} \beta_j^2 \lambda_j$$

$$V = \frac{k}{n} + n \frac{\sum_{j=k+1}^{d} \lambda_j^2}{(n\lambda + \sum_{j>k} \lambda_j)^2}$$

- **Feature learning:**

$$f(x) = \beta^\top W x$$

Eigenvalues of $W \Sigma_X W^\top$

characterizes the predictive risk.

- ➢ **Alignment can be improved.**
- ➢ **Harmful overfitting regime can be turned to kernel regime.**

$\mu_j(\Sigma_X) \geq j^{-1}$ $\qquad$ $\mu_j(W \Sigma_X W^\top) \leq j^{-1}$

Feature learning with DoF reg

$$\min_{W\in\mathbb{R}^{d\times d},\beta^{(j)}\in\mathbb{R}^d} \sum_{j=1}^{m} \frac{1}{n}\|Y^{(j)} - XW^\top\beta^{(j)}\|^2 + \lambda\|\beta^{(j)}\|^2$$

Just minimizing W does not lead to a good generalization.
(It can cause harmful overfitting)
→ Difficulty of feature learning in high dimensional settings.

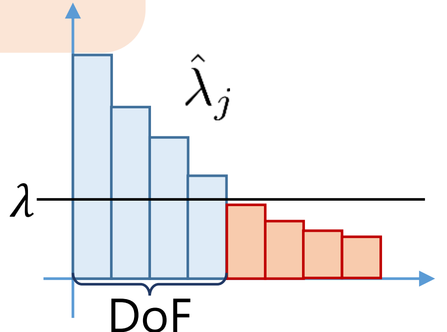
Our proposal: Mallows' C_p type regularization

[Mallows (1973)]

$$R(W) := \min_{\beta^{(j)}} \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\|Y^{(j)} - XW^\top\beta^{(j)}\|^2 + \lambda\|\beta^{(j)}\|^2\right) + \frac{\sigma'^2}{n}\underbrace{\mathrm{Tr}[WX^\top XW^\top(WX^\top XW^\top + n\lambda I)^{-1}]}_{\text{Degrees of freedom (DoF)}}$$

$$\hat{\lambda}_j = \mu_j(W(\tfrac{1}{n}X^\top X)W^\top)$$

$$\text{DoF} = \sum_{j=1}^{d} \frac{\hat{\lambda}_j}{\hat{\lambda}_j + \lambda}$$



Main result

Predictive error: $$\bar{R}(W) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_x[(x^\top \beta_*^{(j)} - x^\top \hat{\beta}^{(j)})^2]$$

$$R(W) := \min_{\beta^{(j)}} \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\|Y^{(j)} - XW^\top \beta^{(j)}\|^2 + \lambda\|\beta^{(j)}\|^2\right) + \frac{{\sigma'}^2}{n}\mathrm{Tr}[WX^\top XW^\top (WX^\top XW^\top + n\lambda I)^{-1}]$$

Theory (Predictive risk bound)

For sufficiently small $\delta > 0$, under some technical conditions, we have that with high probability, the following holds uniformly over W:

$$\bar{R}(W) \lesssim \max\{R(W) - \sigma^2, \delta\}$$

- ***R*(*W*) can be an estimator of the predictive risk.**

→ Minimization of $R(W)$ leads to small predictive risk.

Main result 2

$\Sigma_\beta = \frac{1}{m}\sum_{j=1}^{m} \beta_*^{(j)} \beta_*^{(j)\top}$

Theory (Optimal risk bound)

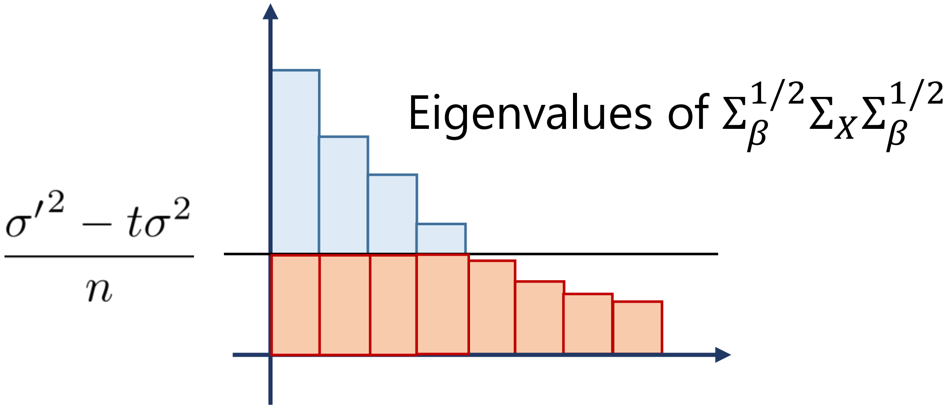
Suppose that $\mathrm{Tr}[\Sigma_\beta^{1/2}\Sigma_X\Sigma_\beta^{1/2}] \leq C < \infty$.

If $t\sigma^2 \leq \sigma'^2$, then with high probability, it holds that:

$$\min_W \{R(W) - \sigma^2\} \lesssim \sum_{j=1}^{d} \min\left\{\frac{\sigma'^2 - t\sigma^2}{n}, \mu_i(\Sigma_\beta^{1/2}\Sigma_X\Sigma_\beta^{1/2})\right\}$$

→ The bound is that of kernel regime for $\tilde{x} \leftarrow \Sigma_\beta^{1/2} x$.

- $\mathrm{Tr}[\Sigma_X] \to \infty$ $(d \to \infty)$: **Interpolation regime** (Benign/harmful overfitting)
- $\mathrm{Tr}\left[\Sigma_\beta^{1/2}\Sigma_X\Sigma_\beta^{1/2}\right] < \infty$: It becomes **kernel regime** by feature learning.



The bound is achieved by
$W^\top W \simeq \Sigma_\beta$.
⇒ The optimal regularization.

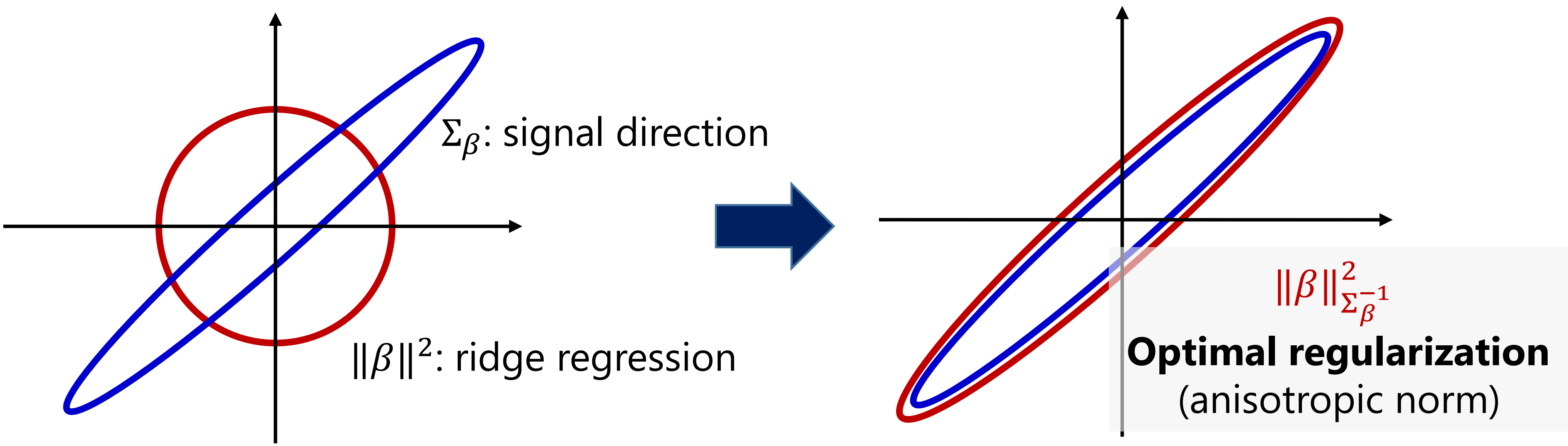
Optimal regularization

$$\min_{\beta} \frac{1}{n}\|Y - XW^\top \beta\|^2 + \lambda\|\beta\|^2 \quad \Leftrightarrow \quad \min_{\tilde\beta} \frac{1}{n}\|Y - X\tilde\beta\|^2 + \lambda\|\tilde\beta\|^2_{(W^\top W)^{-1}}$$

The optimal bound in the theorem is achieved by

$$W^\top W \simeq \Sigma_\beta.$$

⇒ The optimal regularization.



- Alignment is improved
- Fast decay of $\mu_j(\Sigma_\beta)$ turns the problem into a kernel regime.

Bayes optimal regularization

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_x\left[\left(x^\top\beta_{*i} - x^\top W^\top\hat{\beta}_i(W)\right)^2\right] \approx \text{Bias} + \text{Variance} = \mathbb{E}_{\beta_*\sim\mathcal{N}(0,\Sigma_\beta),Y\sim\mathcal{N}(X\beta_*,\sigma^2 I)}\left[\|\beta_* - \hat{\beta}(W)\|_{\Sigma_X}^2\right]$$

$:= R(X,\sigma,\hat{\beta}(W))$ (Bayes risk)

This transformation is merit of multi-output setting

$$\Sigma_\beta = \frac{1}{m}\sum_{i=1}^{m}\beta_{*i}\beta_{*i}^\top$$

Lemma

Suppose Σ_β is positive, then the minimizer of Bayes risk is given by

$$\hat{\beta}_B := \mathrm{argmin}_\beta R(X,\sigma,\beta) = \left(X^\top X + \sigma^2\Sigma_\beta^{-1}\right)^{-1}X^\top y.$$

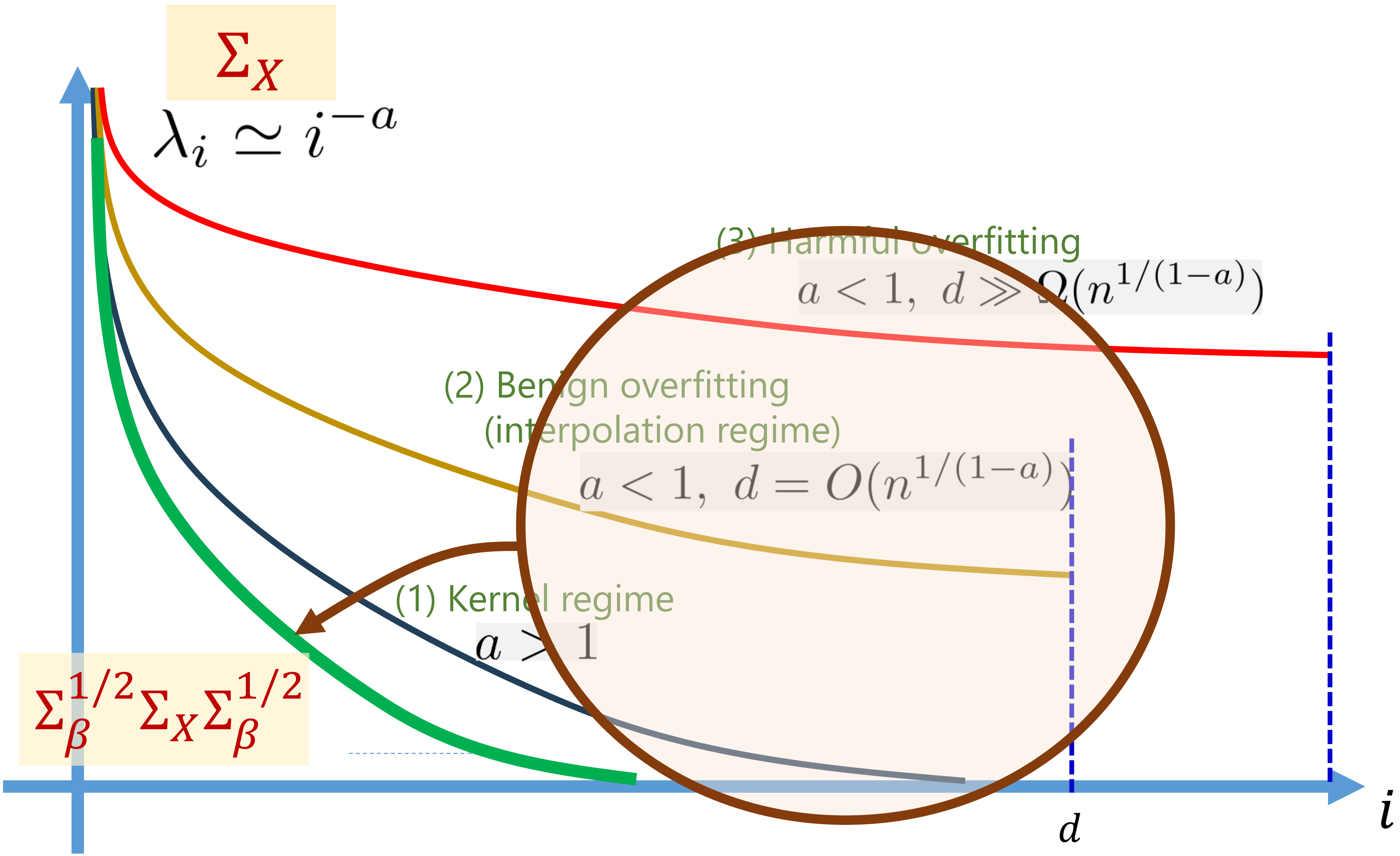
(Bayes estimator)

Optimal regularization

Remark

Since we don't know Σ_β, we have to obtain good regularization by feature learning

Effect of feature learning



Feature learning makes the problem easy one.

Case study

In some concrete situations, **the feature learning method can provably outperform the vanilla ridge regression**.

- Ridge regression: Predictive error=$\Omega(1)$
- Feature learning: Predictive error= $o(1)$

Here, we give two examples:

1. Harmful overfitting setting
2. Misaligned setting

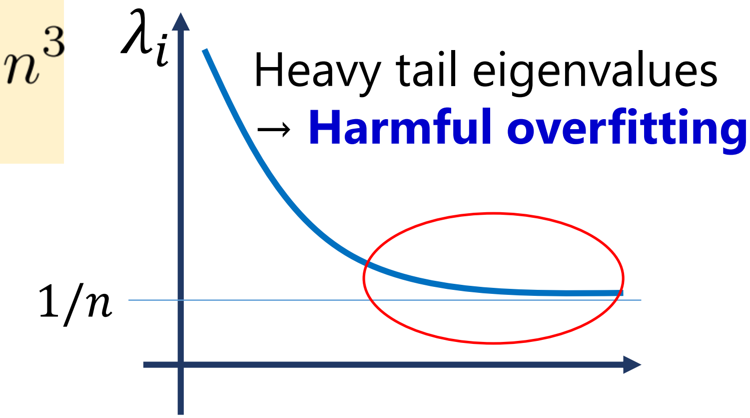
(※ These are just typical situations. There are uncountable situations where 2 layer NN with DoF regularization can outperform ridge regression)

(1) Harmful overfitting

$$\lambda_i = \mu_i(\Sigma_X)$$
$$\nu_i = \mu_i(\Sigma_\beta)$$

Σ_X and Σ_β share the same eigen vectors.

$$\lambda_i = \begin{cases} i^{-1} & (i \le n) \\ \frac{1}{n} & (i > n) \end{cases}, \quad \nu_i = i^{-2}, \quad d = n^3$$

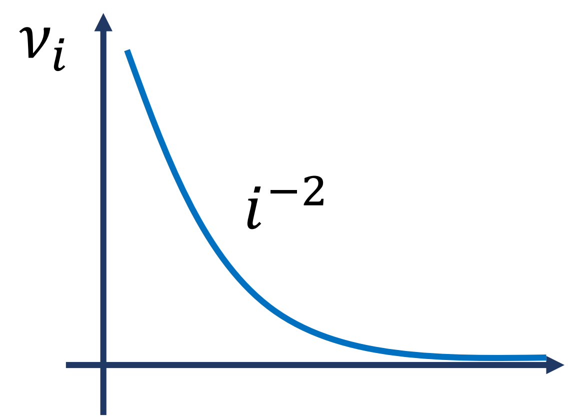


- **Two layer NN**

  Predictive error = $\dfrac{1}{n^{2/3}}$

- **Ridge regression**

  Predictive error = $\Omega(1)$

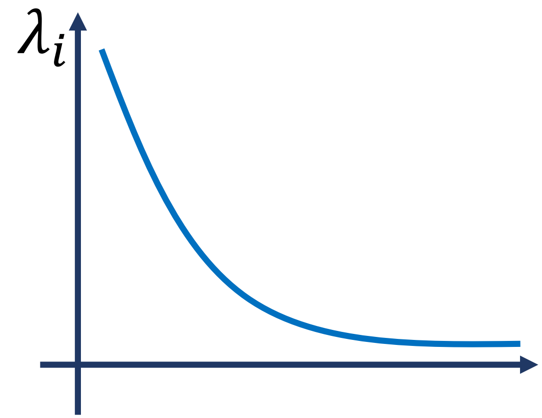


(2) Misaligned setting

$$\lambda_i = \mu_i(\Sigma_X)$$
$$\nu_i = \mu_i(\Sigma_\beta)$$

Σ_X and Σ_β share the same eigen vectors.

$$\lambda_i = i^{-1}, \quad \nu_i = \begin{cases} n & (i = n) \\ i^{-1} & (\text{otherwise}) \end{cases}, \quad d \gg n$$

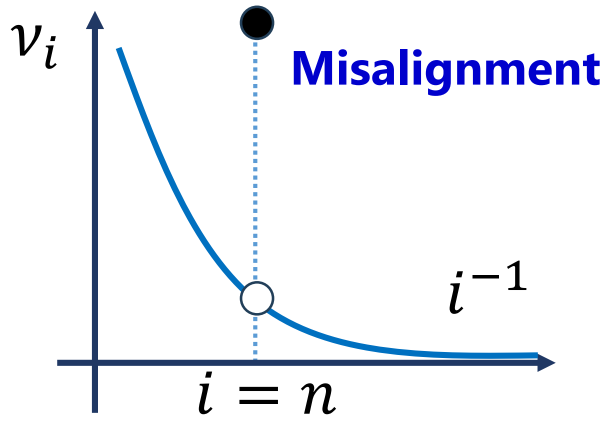


- **Two layer NN**

Predictive error = $\dfrac{1}{\sqrt{n}}$

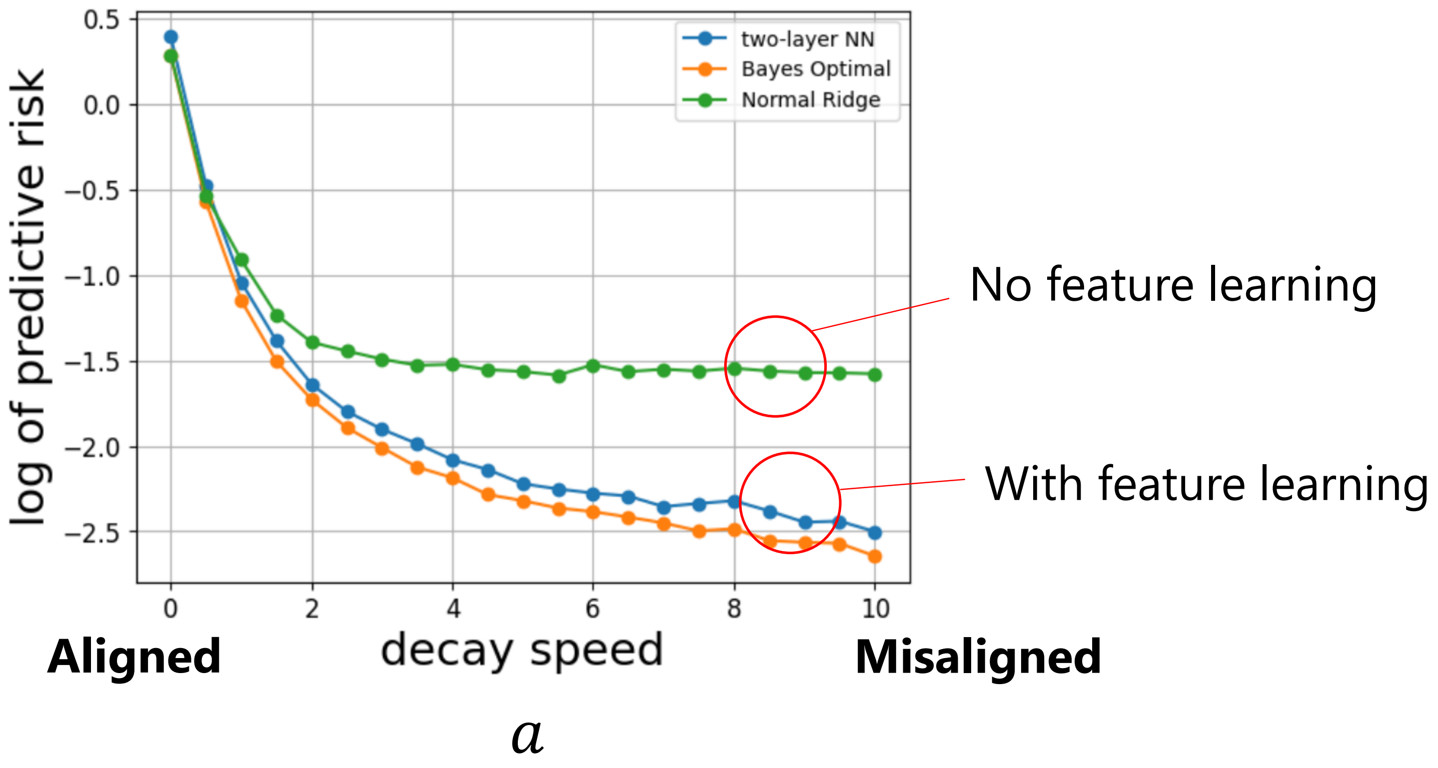
- **Ridge regression**

Predictive error = $\Omega(1)$



Experiment

- $\mu_j(\Sigma_X) = j^{-1}$
- $\Sigma_\beta = \text{diag}(1, 2^{-a}, \ldots, j^{-a}, \ldots, 1000^{-a})$ with $a \in \{0, 0.5, 1, \ldots, 10\}$



Computational issue

$$R(W) := \min_{\beta^{(j)}} \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\|Y^{(j)} - XW^\top\beta^{(j)}\|^2 + \lambda\|\beta^{(j)}\|^2\right)$$
$$+ \frac{\sigma'^2}{n}\mathrm{Tr}[WX^\top XW^\top(WX^\top XW^\top + n\lambda I)^{-1}]$$

How to minimize $R(W)$?

→ Global optimality of noisy gradient descent

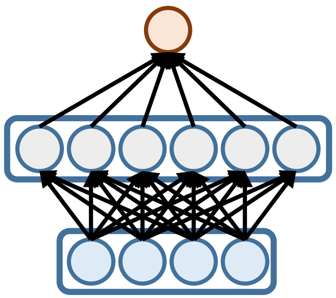
The DoF regularization is not a standard technique in the deep learning literature.

→ Label noise acts as the DoF regularizer

2-layer NN in mean-field scaling

- Extension to 2-layer nonlinear neural network:

$$f(x) = \beta^\top W x = \sum_{j=1}^{d} \beta_j W_{j,:} x$$



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

σ is a nonlinear activation such as sigmoid function.

Mean field limit $M \to \infty$

$$f_{a,\mu}(x) = \int a(w) \sigma(w^\top x) \mathrm{d}\mu(w)$$

$$w \in \mathbb{R}^d,\ \mu \in \mathcal{P}(\mathbb{R}^d),\ a \in L^2(\mu)$$

Label noise and WG flow

[Takakura&Suzuki: Mean-field Analysis on Two-layer Neural Networks from a Kernel Perspective. 2024]

$$\min_{\mu, a^{(j)}} \frac{1}{m}\sum_{j=1}^{m}\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(j)} - \int a^{(j)}(w)\sigma(w^\top x_i)\mathrm{d}\mu(w)\right)^2 + \lambda\|a^{(j)}\|_{L^2(\mu)}^2\right]$$

2 time scale optimization:

(1) Optimization with respect to a with fixed μ:

$$F(\mu) := \min_{a^{(j)}} \frac{1}{m}\sum_{j=1}^{m}\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(j)} - \int a^{(j)}(w)\sigma(w^\top x_i)\mathrm{d}\mu(w)\right)^2 + \lambda\|a^{(j)}\|_{L^2(\mu)}^2\right]$$

(2) Optimization of F with respect to μ: (Wasserstein gradient flow)

$$\mu_{t+1} \leftarrow \mu_t + \eta\nabla\cdot\left(\nabla\frac{\delta F(\mu_t)}{\delta\mu}\right)$$ (+ Entropy regularization)

More precisely, mean field Langevin dynamics

$$w^{(t+1)} \leftarrow w^{(t)} - \eta\nabla\frac{\delta F(\mu_t)}{\delta\mu}(w^{(t)}) + \sqrt{2\eta\lambda}\xi_t,\ \ \mu_{t+1} = \mathrm{Law}(w^{(t+1)})$$

Theorem (informal)

We have convergence of this algorithm with log-Sobolev assumption.

$$F(\mu_t) - F(\mu^*) \le \exp(-\alpha\lambda t)(F(\mu_0) - F(\mu^*))$$

Label noise as regularization

- **Label noise training**

$$\tilde{a}^{(j)} := \underset{a^{(j)}}{\arg\min} \ \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{y}_i^{(j)} - \int a^{(j)}(w)\sigma(w^\top x_i)\mathrm{d}\mu(w)\right)^2 + \lambda\|a^{(j)}\|_{L^2(\mu)}^2$$

Label noise: $y_i^{(j)} + \tilde{\epsilon}_i^{(j)}$ where $\tilde{\epsilon}_i^{(j)} \sim U([-\tilde{\sigma}, \tilde{\sigma}])$

- **First layer training**

$$G(\mu) := \frac{1}{m}\sum_{j=1}^{m}\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(j)} - \int \tilde{a}^{(j)}(w)\sigma(w^\top x_i)\mathrm{d}\mu(w)\right)^2 + \lambda\|\tilde{a}^{(j)}\|_{L^2(\mu)}^2\right]$$

(no label noise)

Lemma [Takakura&Suzuki, 2024]

Degrees of freedom

$$\mathbb{E}_{\tilde{\epsilon}}[G(\mu)] = F(\mu) + \frac{\lambda\tilde{\sigma}^2}{n}\mathrm{Tr}\left[\hat{\Sigma}_\mu(\hat{\Sigma}_\mu + n\lambda I)^{-1}\right]$$

where $(\hat{\Sigma}_\mu)_{i,j} = \int \sigma(w^\top x_i)\sigma(w^\top x_j)\mathrm{d}\mu(w) \quad (i \in [n],\ j \in [n])$

- **Label noise training acts as Degrees of Freedom regularization.**
- **<u>Mean field Langevin dynamics</u> can optimize the objective.**

In-context learning by Transformer

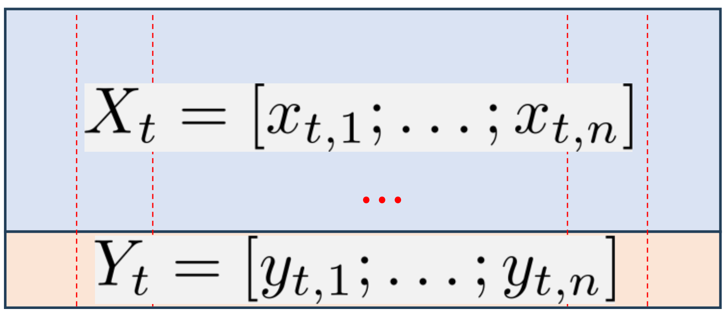
Kim, Suzuki: Transformers Learn Nonlinear Features In Context: Nonconvex Mean-field Dynamics on the Attention Landscape. arXiv:2402.01258.

In-context learning

Model $y_{t,i} = F_t^{\circ}(x_{t,i}) \qquad (i = 1, \ldots, n)$

- The true function F_t is different for different tasks.
- F_t is randomly generated from some distribution.

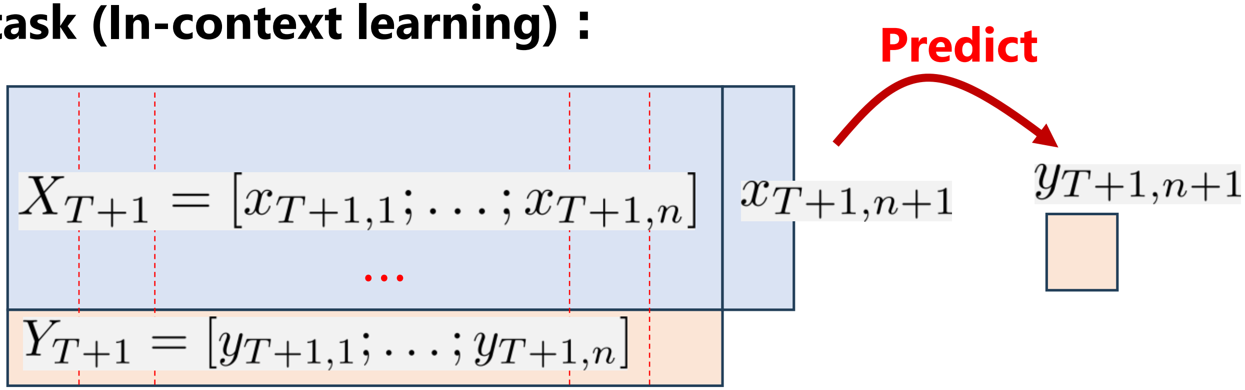
Pretraining (T tasks)：



$\times T$

- ➢ We observe pretraining task data T times.
- ➢ Each task has n data.

Test task (In-context learning)：



Model: Nonlinear feature

Linear model with nonlinear features:

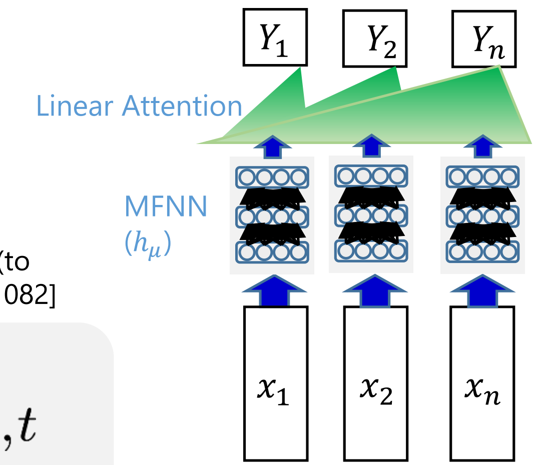
$$F_t^\circ(x) = v_t^\top f^\circ(x) \quad \text{where } v_t \sim N(0, I) \text{ and } f^\circ(x) \in \mathbb{R}^k.$$

We want to estimate the nonlinear feature f° by pretraining.

Mean field neural network:

$$h_\mu(x) = \int h_\theta(x)\mathrm{d}\mu(\theta) \in \mathbb{R}^k$$

$$h_\theta(x) = \mathbf{a}\sigma(\mathbf{w}^\top x) \quad (\theta = (\mathbf{a}, \mathbf{w}) \in \mathbb{R}^k \times \mathbb{R}^d)$$



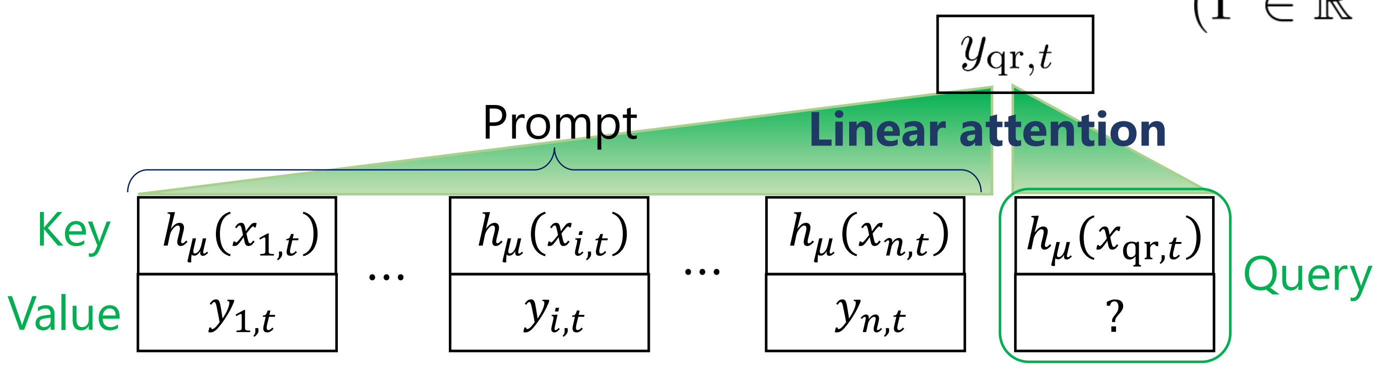
Linear attention model

[Ahn et al.: Linear attention is (maybe) all you need (to understand transformer optimization). arXiv:2310.01082]

$$\frac{1}{n}\sum_{i=1}^{n} y_{i,t} h_\mu(x_{i,t})^\top \Gamma h_\mu(x_{\mathrm{qr},t})$$

Value, Key, Query → Predict $y_{\mathrm{qr},t}$

$(\Gamma \in \mathbb{R}^{k\times k})$



In-Context Learning (ICL) risk

Empirical ICL risk :

$$\widehat{\mathcal{L}}(\mu,\Gamma) := \frac{1}{T}\sum_{t=1}^{T}\left(y_{\mathrm{qr},t} - \frac{1}{n}\sum_{i=1}^{n} y_{i,t} h_\mu(x_{i,t})^\top \Gamma h_\mu(x_{\mathrm{qr},t})\right)^2$$

→ Minimize with respect to μ, Γ.

The expected ICL risk: (Large sample limit: $n \to \infty$ and $T \to \infty$)

$$\mathcal{L}(\mu,\Gamma) := \mathbb{E}_{x_{\mathrm{qr}}}\left[\left\| f^\circ(x_{\mathrm{qr}}) - \mathbb{E}_x[f^\circ(x) h_\mu(x)^\top]\Gamma h_\mu(x_{qr})\right\|^2\right]$$

(note that $y_{i,t} = v_t^\top f^\circ(x_{i,t})$)

Question : Can we optimize μ, Γ by a gradient descent?
(Infinite-dimensional non-convex problem)

There have been many work on optimization guarantee on ICL for **linear model**: Zhang et al., (2023), Mahankali et al. (2023), Guo et al. (2023) to name a few.
Our novelty: Optimization guarantee w.r.t. **nonlinear feature learning (h_μ).**

Two time scale dynamics

Feature covariance $\Sigma_{\mu,\nu} := \mathbb{E}_X[h_\mu(X)h_\nu^\top(X)]$

$h_\mu(x) := \int h_\theta(x)\mathrm{d}\mu(\theta)$

Assumption (realizability of the true feature)

There exists μ° such that $f^\circ = h_{\mu^\circ}$ and $\Sigma_{\mu^\circ,\mu^\circ} \propto I_k$.

Two time-scale dynamics (Γ is optimized first):

$$\mathcal{L}(\mu) := \min_{\Gamma} \mathcal{L}(\mu, \Gamma) = \min_{\Gamma} \mathbb{E}_{x_{\mathrm{qr}}}\left[\left\| f^\circ(x_{\mathrm{qr}}) - \mathbb{E}_x[f^\circ(x)h_\mu(x)^\top]\Gamma h_\mu(x_{qr})\right\|^2\right]$$
$$= \mathbb{E}_{x_{\mathrm{qr}}}\left[\left\| f^\circ(x_{\mathrm{qr}}) - \Sigma_{\mu^\circ,\mu}\Sigma_{\mu,\mu}^{-1} h_\mu(x_{qr})\right\|^2\right]$$

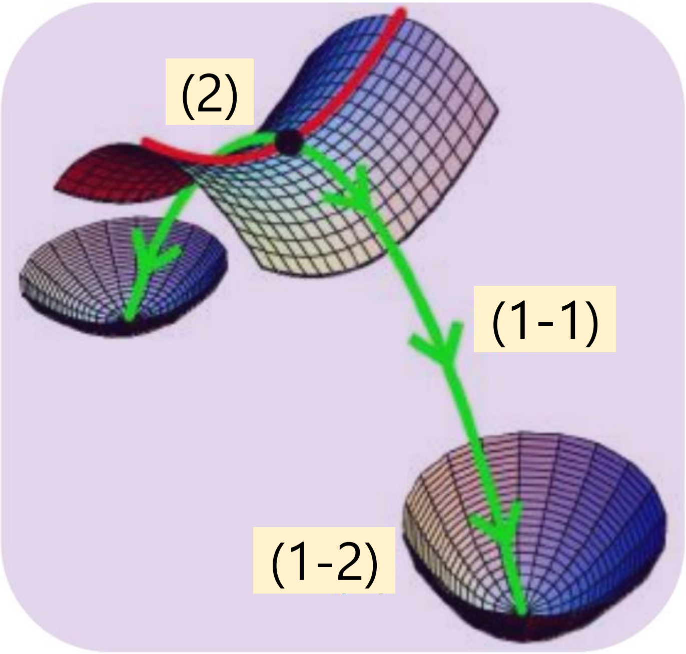
- μ is the minimizer iff $h_\mu = R h_{\mu^\circ}$ for an invertible matrix R

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

- $\partial_t \mu_t = \nabla \cdot \left(\mu_t \nabla \frac{\delta \mathcal{L}(\mu_t)}{\delta \mu}\right)$
- $\frac{\mathrm{d}\theta_t}{\mathrm{d}t} = -\nabla \frac{\delta \mathcal{L}(\mu_t)}{\delta \mu}(\theta_t) \qquad (\mu_t = \mathrm{Law}(\theta_t))$

Strict saddle

Theorem 1 (**Strict saddle** property of the loss landscape)



There exists a **descent direction** or **negative curvature**.

Strict saddle

For an orthogonal matrix $\mathbf{R} \in \mathrm{O}(k)$, define $\boldsymbol{R}\#\mu$ as the push-forward of μ along the rotation $\mathbf{R}$: $(a, w) \mapsto (\mathbf{R}a, w)$, i.e., $h_{\mathbf{R}\#\mu} = \mathbf{R}h_\mu$.

Theorem 1 (Strict saddle property of the loss landscape)

If $\mu \in \mathcal{P}$ is not the global minimum, then one of the followings holds:

(1) (1-1) There exists $\mathbf{R} \in \mathrm{conv}(\mathrm{O}(k))$ such that

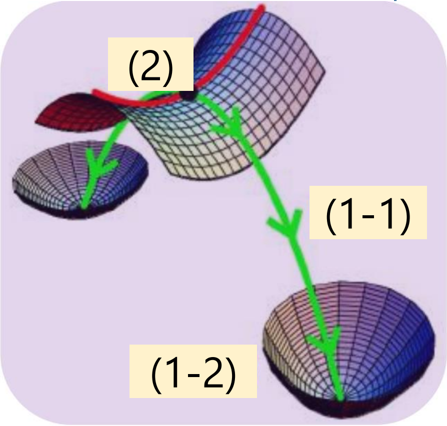
$$\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{L}(\bar{\mu}_s)\Big|_{s=0} < 0 \quad \text{where } \bar{\mu}_s = (1-s)\mu + s\mathbf{R}\sharp\mu^\circ.$$

(1-2) Furthermore, if $0 < \mathcal{L}(\mu) < r^\circ/2$, then

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{L}(\bar{\mu}_s)\Big|_{s=0} \le -\frac{4}{\|\sigma\|_\infty^2}\mathcal{L}(\mu)\left(\frac{r_0}{2} - \mathcal{L}(\mu)\right)$$

(2) Otherwise,

$$\mathcal{L}(\mu) > \frac{r_0}{2} \quad \text{and} \quad \frac{\mathrm{d}^2\mathcal{L}(\bar{\mu}_s)}{\mathrm{d}s^2}\Big|_{s=0} \le -\frac{4}{k\|\sigma\|_\infty^2}\mathcal{L}(\mu)^2.$$



There exists a **descent direction** or **negative curvature**.

Behavior around the critical point

Let the "Hessian" at μ be

$$H_\mu(\theta, \theta') := \nabla_\theta \nabla_{\theta'} \frac{\delta^2 \mathcal{L}(\mu)}{\delta \mu^2}(\theta, \theta')$$

Lemma

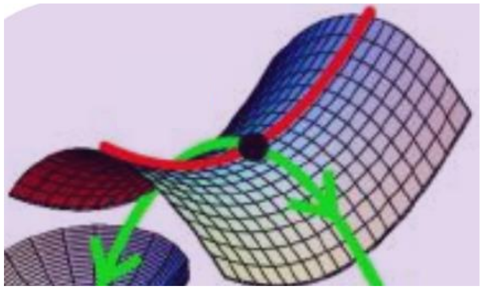
The Wasserstein GF μ_t around a critical point μ^+ can be written as $(\mathrm{id} + \epsilon v_t)\#\mu^+$ where the velocity field v_t follows

$$\partial_t v_t(\theta) = -\int H_{\mu^+}(\theta, \theta') v_t(\theta') \mathrm{d}\mu^+(\theta') + O(\epsilon)$$

(c.f., Otto calculus)

➡ Negative curvature direction exponentially grows up!

➡ μ_t moves away from the critical point.



Theorem (Informal)

The solution is not captured by any critical point *almost surely*.
(The solution converges to the global optimal solution almost surely)

Decay speed of objective

Suppose that $\left\|\frac{\mathrm{d}\mu^\circ}{\mathrm{d}\mu_t}\right\|_\infty \leq R$ (which could be ensured by using birth-death process).

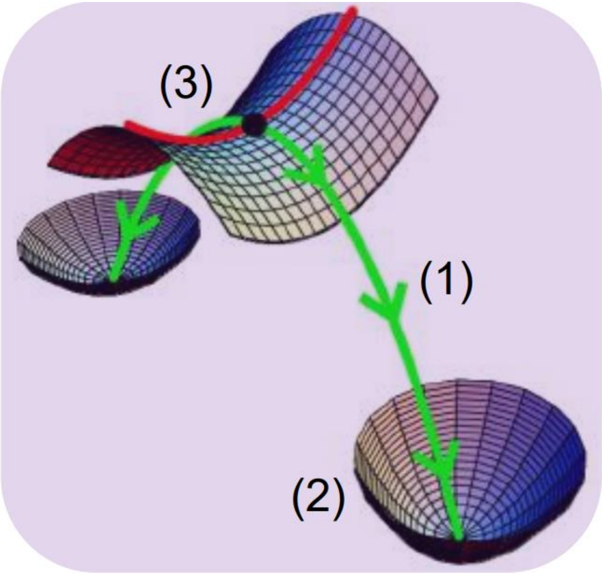
Theorem (GF moves toward a descent direction (1))

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{L}(\bar{\mu}_s)\Big|_{s=0} < -\delta \quad \Rightarrow \quad \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(\mu_t) \leq -R^{-1}\delta^2.$$

Theorem (Accelerated convergence phase (2))

Once $\mathcal{L}(\mu_t) \leq \frac{r^\circ}{2} - \epsilon$ is satisfied,

$$\mathcal{L}(\mu_{t+T}) \leq O\left(\frac{Rk^2}{T}\right)$$



Theorem (Negative curvature around a saddle point (3))

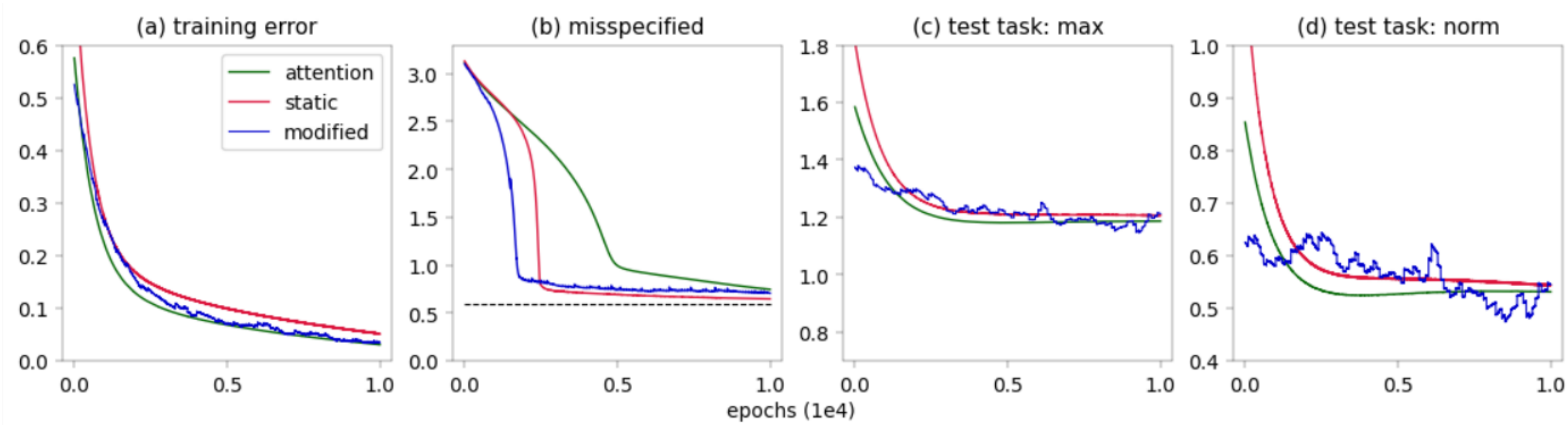
$$\frac{\mathrm{d}^2\mathcal{L}(\bar{\mu}_s)}{\mathrm{d}s^2} \leq -\Lambda \quad \Rightarrow \quad \text{min-eigen-value}(H_{\mu_t}) \leq -\Lambda/R$$

➡ Escape from the critical point exponentially fast.

Numerical experiment

We compare 3 models with $d = 20$, $k = 5$, and 500 neurons with sigmoid act.
All models are pre-trained using SGD on 10K prompts of 1K token pairs.

1. **attention**: jointly optimizes $\mathcal{L}(\mu, \Gamma)$.
2. **static**: directly minimizes $\mathcal{L}(\mu)$.
3. **modified**: static model implementing birth-death & GP



→ verify global convergence as well as improvement for misaligned model ($k_{\text{true}} = 7$) and nonlinear test tasks $g(x) = \max_{j \le k} h_{\mu^\circ}(x)_j$ or $g(x) = \|h_{\mu^\circ}(x)\|^2$.

Conclusion

- Feature learning by 2-layer NN
  - ➢ Statistical analysis in high dimensional regression
  - ➢ Optimization theory of in-context feature learning in Transformer
- [High dimensional regression]
  - ➢ Optimal regularization via Degrees of Freedom reg
  - ➢ Overfitting regime → Kernel regime

$$\bar{R}(W) \lesssim \sum_{j=1}^{d} \min\left\{1/n, \mu_i(\Sigma_\beta^{1/2}\Sigma_X\Sigma_\beta^{1/2})\right\}$$

- [In-context feature learning by Transformer]
  - ➢ The loss landscape is like strict saddle
  - ➢ The solution is hardly captured by a saddle point

(i) $\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{L}(\bar{\mu}_s)\Big|_{s=0} < 0 \quad \text{where } \bar{\mu}_s = (1-s)\mu + s\mathbf{R}\sharp\mu^\circ.$

(ii) Otherwise, $\mathcal{L}(\mu) > \frac{r_0}{2} \quad \text{and} \quad \frac{\mathrm{d}^2\mathcal{L}(\bar{\mu}_s)}{\mathrm{d}s^2}\Big|_{s=0} \leq -\frac{4}{k\|\sigma\|_\infty^2}\mathcal{L}(\mu)^2.$