What causes the test error? Going beyond bias-variance via ANOVA

Edgar Dobriban
joint work with Licong Lin

Wharton, UPenn

May 5, 2021

Collaborator

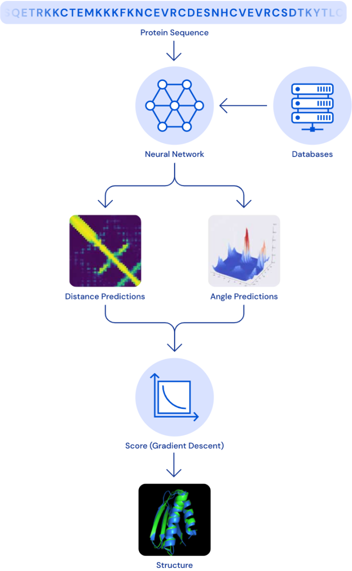
Licong Lin, Peking University undergraduate '21 $\rightarrow$ Berkeley Stats PhD

Overview

Outline

The scientific frontier?

50 years ago



Today

An ideal theory of machine learning?

- Based on problem components (data distribution, sample size, learning algorithm, ...)

An ideal theory of machine learning?

- Based on problem components (data distribution, sample size, learning algorithm, ...)
- Predict:
  - test error, training dynamics

An ideal theory of machine learning?

- Based on problem components (data distribution, sample size, learning algorithm, ...)
- Predict:
  - test error, training dynamics
  - fine-grained characteristics: bias, variance

An ideal theory of machine learning?

- Based on problem components (data distribution, sample size, learning algorithm, ...)
- Predict:
  - test error, training dynamics
  - fine-grained characteristics: bias, variance
  - impact of changing each component

Current works

- Complexity-based generalization bounds

Current works

- Complexity-based generalization bounds
- Distribution-dependent bounds/asymptotics

Bias-variance decomposition

Choose $\hat{f}$ based on the training set, and decompose the test error into bias and variance ($\mathbb{E}_{x,y} = \mathbb{E}_{(x,y)\sim\text{test}}$):

$$\begin{aligned}\mathbb{E}_{x,y}\mathbb{E}\|y - \hat{f}(x)\|^2 &= \mathbb{E}_{x,y}\mathbb{E}\|y - \mathbb{E}\hat{f}(x)\|^2 + \mathbb{E}_{x,y}\mathrm{Var}(\hat{f}(x)) \\ &= \mathrm{Bias}^2 + \mathrm{Variance}.\end{aligned}$$

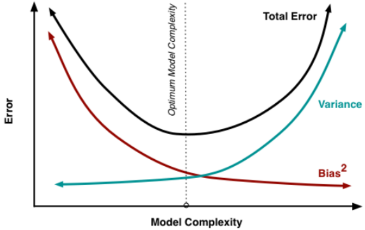


Figure: Bias and variance contributing to total error.[1]

[1]Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

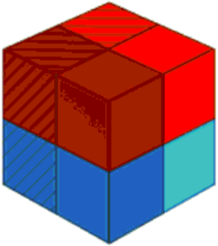
Our approach

- Variance depends on randomness in: initialization, input features, labels...
  and other aspects: randomness in optimization algorithm, ...

[2] Figure: www.graphpad.com/guides/prism/latest/statistics

Our approach

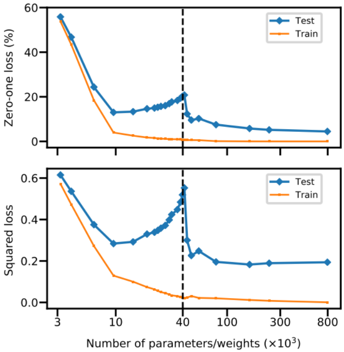
- Variance depends on randomness in: initialization, input features, labels...
  and other aspects: randomness in optimization algorithm, ...
- Decompose the variance into its ANOVA components (R.A. Fisher, 1918)



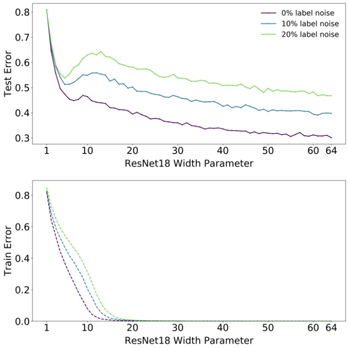
Three-way ANOVA: how is a response affected by three factors?[2]

[2] Figure: www.graphpad.com/guides/prism/latest/statistics

Double descent



[Belkin et al., 2018]



[Nakkiran et al., 2019]

Outline

Setting

- **Data:** n datapoints $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. from $y = f^*(x) + \varepsilon = x^\top \theta + \varepsilon$, where ε is label noise:

$$Y = X\theta + \mathcal{E}.$$

Setting

- **Data:** n datapoints $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. from $y = f^*(x) + \varepsilon = x^\top \theta + \varepsilon$, where ε is label noise:

$$Y = X\theta + \mathcal{E}.$$

- **Training**: Fit a two-layer linear (later nonlinear) "neural net"/random features model

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

Setting

- **Data:** n datapoints $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. from $y = f^*(x) + \varepsilon = x^\top \theta + \varepsilon$, where ε is label noise:

$$Y = X\theta + \mathcal{E}.$$

- **Training**: Fit a two-layer linear (later nonlinear) "neural net"/random features model

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

- Weights $W \in \mathbb{R}^{p \times d}$, $p \leq d$ drawn uniformly from partial orthonormal matrices, $WW^\top = I_p$. Train β with L_2 loss/regularization.

Bias-variance decomposition

- Decompose the test error into irreducible noise, bias and variance:

$$
\begin{aligned}
\mathbb{E}\|y - \hat{f}(x)\|^2 =& \mathbb{E}\|y - \mathbb{E}y\|^2 \\
& + \mathbb{E}\|\mathbb{E}y - \mathbb{E}\hat{f}(x)\|^2 + \mathbb{E}\|\mathbb{E}\hat{f}(x) - \hat{f}(x)\|^2 \\
=& \sigma^2 + \mathrm{Bias}^2 + \mathrm{Variance},
\end{aligned}
$$

where $\mathbb{E} = \mathbb{E}_{x,y,X,W,\varepsilon}$.

Bias-variance decomposition

- Decompose the test error into irreducible noise, bias and variance:

$$\begin{aligned}\mathbb{E}\|y-\hat{f}(x)\|^2 =&\mathbb{E}\|y-\mathbb{E}y\|^2 \\ &+\mathbb{E}\|\mathbb{E}y-\mathbb{E}\hat{f}(x)\|^2+\mathbb{E}\|\mathbb{E}\hat{f}(x)-\hat{f}(x)\|^2 \\ =&\sigma^2+\mathrm{Bias}^2+\mathrm{Variance},\end{aligned}$$

  where $\mathbb{E}=\mathbb{E}_{x,y,X,W,\mathcal{E}}$.
- The randomness of $\hat{f}$ is due to $X, W, \mathcal{E}$. What are their contributions?

Hierarchical decomposition: d'Ascoli et al., 2020

d'Ascoli et al., 2020 decompose the variance of $\hat{f}$ in the order of $\mathcal{E}, W, X$.

$$
\begin{aligned}
\mathbb{E}\|\hat{f}(x) - \mathbb{E}\hat{f}\|^2 &= \mathbb{E}\|\hat{f}(x) - \mathbb{E}\hat{f}(x|W,X)\|^2 \\
&\quad + \mathbb{E}\|\mathbb{E}\hat{f}(x|W,X) - \mathbb{E}\hat{f}(x|X)\|^2 \\
&\quad + \mathbb{E}\|\mathbb{E}\hat{f}(x|X) - \mathbb{E}\hat{f}\|^2 \\
&:= \Sigma_{label} + \Sigma_{init} + \Sigma_{sample}.
\end{aligned}
$$

ANOVA: Symmetric variance decomposition

Denote $(X, W, \mathcal{E})$ by (s, i, l) respectively. We decompose the variance of $\hat{f}$ in a symmetric way via the analysis of variance (ANOVA):

$$\mathrm{Var}[\hat{f}(x)] = V_s + V_l + V_i + V_{sl} + V_{si} + V_{li} + V_{sli},$$

where

$$\begin{aligned}
V_a &= \mathbb{E}_{\theta,x}\mathrm{Var}_a[\mathbb{E}_{-a}(\hat{f}(x)|a)], && a \in \{s,l,i\} \\
V_{ab} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{ab}[\mathbb{E}_{-ab}(\hat{f}(x)|a,b)] - V_a - V_b, && a,b \in \{s,l,i\}, a \neq b. \\
V_{abc} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{abc}[\mathbb{E}_{-abc}(\hat{f}(x)|a,b,c)] - V_a - V_b - V_c - V_{ab} - V_{ac} - V_{bc} && \\
&= \mathrm{Var}[\hat{f}(x)] - V_s - V_l - V_i - V_{sl} - V_{si} - V_{li}, && \{a,b,c\} = \{s,l,i\}.
\end{aligned}$$

ANOVA: Symmetric variance decomposition

Denote $(X, W, \mathcal{E})$ by (s, i, l) respectively. We decompose the variance of $\hat{f}$ in a symmetric way via the analysis of variance (ANOVA):

$$\mathrm{Var}[\hat{f}(x)] = V_s + V_l + V_i + V_{sl} + V_{si} + V_{li} + V_{sli},$$

where

$$\begin{aligned}
V_a &= \mathbb{E}_{\theta,x}\mathrm{Var}_a[\mathbb{E}_{-a}(\hat{f}(x)|a)], && a \in \{s,l,i\}\\
V_{ab} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{ab}[\mathbb{E}_{-ab}(\hat{f}(x)|a,b)] - V_a - V_b, && a,b \in \{s,l,i\}, a \neq b.\\
V_{abc} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{abc}[\mathbb{E}_{-abc}(\hat{f}(x)|a,b,c)] - V_a - V_b - V_c - V_{ab} - V_{ac} - V_{bc} &&\\
&= \mathrm{Var}[\hat{f}(x)] - V_s - V_l - V_i - V_{sl} - V_{si} - V_{li}, && \{a,b,c\} = \{s,l,i\}.
\end{aligned}$$

- V_a: the effect of varying a alone *(main effect)*.

ANOVA: Symmetric variance decomposition

Denote $(X, W, \mathcal{E})$ by (s, i, l) respectively. We decompose the variance of $\hat{f}$ in a symmetric way via the analysis of variance (ANOVA):

$$\mathrm{Var}[\hat{f}(x)] = V_s + V_l + V_i + V_{sl} + V_{si} + V_{li} + V_{sli},$$

where

$$\begin{aligned}
V_a &= \mathbb{E}_{\theta,x}\mathrm{Var}_a[\mathbb{E}_{-a}(\hat{f}(x)|a)], && a \in \{s,l,i\} \\
V_{ab} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{ab}[\mathbb{E}_{-ab}(\hat{f}(x)|a,b)] - V_a - V_b, && a,b \in \{s,l,i\}, a \neq b. \\
V_{abc} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{abc}[\mathbb{E}_{-abc}(\hat{f}(x)|a,b,c)] - V_a - V_b - V_c - V_{ab} - V_{ac} - V_{bc} && \\
&= \mathrm{Var}[\hat{f}(x)] - V_s - V_l - V_i - V_{sl} - V_{si} - V_{li}, && \{a,b,c\} = \{s,l,i\}.
\end{aligned}$$

- V_a: the effect of varying a alone *(main effect)*.
- V_{ab}: the second-order *interaction effect* between a and b beyond their main effects.

ANOVA: Symmetric variance decomposition

Denote $(X, W, \mathcal{E})$ by (s, i, l) respectively. We decompose the variance of $\hat{f}$ in a symmetric way via the analysis of variance (ANOVA):

$$\mathrm{Var}[\hat{f}(x)] = V_s + V_l + V_i + V_{sl} + V_{si} + V_{li} + V_{sli},$$

where

$$\begin{aligned}
V_a &= \mathbb{E}_{\theta,x}\mathrm{Var}_a[\mathbb{E}_{-a}(\hat{f}(x)|a)], && a \in \{s,l,i\} \\
V_{ab} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{ab}[\mathbb{E}_{-ab}(\hat{f}(x)|a,b)] - V_a - V_b, && a,b \in \{s,l,i\}, a \neq b. \\
V_{abc} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{abc}[\mathbb{E}_{-abc}(\hat{f}(x)|a,b,c)] - V_a - V_b - V_c - V_{ab} - V_{ac} - V_{bc} \\
&= \mathrm{Var}[\hat{f}(x)] - V_s - V_l - V_i - V_{sl} - V_{si} - V_{li}, && \{a,b,c\} = \{s,l,i\}.
\end{aligned}$$

- V_a: the effect of varying a alone *(main effect)*.
- V_{ab}: the second-order *interaction effect* between a and b beyond their main effects.
- V_{abc}: interaction effect among a, b, c beyond their pairwise interactions.

Consequence of symmetric variance decomposition

We can recover various orders of variance decompositions $(\{a, b, c\} = \{s, l, i\})$.

$$
\begin{aligned}
\Sigma^a_{abc} &:= \mathbb{E}_{\theta,x}\mathbb{E}_{a,b,c}[\hat{f}(x) - \mathbb{E}_a\hat{f}(x)]^2 & \Sigma^a_{abc} &= V_a + V_{ab} + V_{ac} + V_{abc} \\
\Sigma^b_{abc} &:= \mathbb{E}_{\theta,x}\mathbb{E}_{b,c}[\mathbb{E}_a\hat{f}(x) - \mathbb{E}_{a,b}\hat{f}(x)]^2 & \Sigma^b_{abc} &= V_{bc} + V_b \\
\Sigma^c_{abc} &:= \mathbb{E}_{\theta,x}\mathbb{E}_c[\mathbb{E}_{a,b}\hat{f}(x) - \mathbb{E}_{a,b,c}\hat{f}(x)]^2. & \Sigma^c_{abc} &= V_c.
\end{aligned}
$$

How to interpret these terms?

Consequence of symmetric variance decomposition

We can recover various orders of variance decompositions ($\{a, b, c\} = \{s, l, i\}$).

$$
\begin{aligned}
\Sigma^a_{abc} &:= \mathbb{E}_{\theta,x}\mathbb{E}_{a,b,c}[\hat{f}(x) - \mathbb{E}_a\hat{f}(x)]^2 & \Sigma^a_{abc} &= V_a + V_{ab} + V_{ac} + V_{abc} \\
\Sigma^b_{abc} &:= \mathbb{E}_{\theta,x}\mathbb{E}_{b,c}[\mathbb{E}_a\hat{f}(x) - \mathbb{E}_{a,b}\hat{f}(x)]^2 & \Sigma^b_{abc} &= V_{bc} + V_b \\
\Sigma^c_{abc} &:= \mathbb{E}_{\theta,x}\mathbb{E}_c[\mathbb{E}_{a,b}\hat{f}(x) - \mathbb{E}_{a,b,c}\hat{f}(x)]^2. & \Sigma^c_{abc} &= V_c.
\end{aligned}
$$

How to interpret these terms?

- Σ^a_{abc} is all the variance related to a.
- Σ^b_{abc} is all the variance related to b after subtracting all the variance related to a in the total variance.
- Σ^c_{abc} is the part of the variance that depends only on c.

Outline

Details of setup

- **Data:** n datapoints $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. from $y = f^*(x) + \varepsilon = x^\top \theta + \varepsilon$, where x has i.i.d. standardized entries, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is label noise:

$$Y = X\theta + \mathcal{E}.$$

Details of setup

- **Data:** n datapoints $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. from $y = f^*(x) + \varepsilon = x^\top \theta + \varepsilon$, where x has i.i.d. standardized entries, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is label noise:

$$Y = X\theta + \mathcal{E}.$$

- **Training**: Fit a two-layer linear (later nonlinear) neural net

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

Train β with L_2 loss, L_2 regularization λ to get predictor:

$$f(x) = (Wx)^\top \hat{\beta}_{\lambda,\mathcal{T},W} = x^\top W^\top \left(\frac{WX^\top XW^\top}{n} + \lambda I_p \right)^{-1} \frac{WX^\top Y}{n}.$$

Details of setup

- **Data:** n datapoints $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. from $y = f^*(x) + \varepsilon = x^\top \theta + \varepsilon$, where x has i.i.d. standardized entries, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is label noise:

$$Y = X\theta + \mathcal{E}.$$

- **Training**: Fit a two-layer linear (later nonlinear) neural net

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

Train β with L_2 loss, L_2 regularization λ to get predictor:

$$f(x) = (Wx)^\top \hat{\beta}_{\lambda,\mathcal{T},W} = x^\top W^\top \left(\frac{WX^\top XW^\top}{n} + \lambda I_p \right)^{-1} \frac{WX^\top Y}{n}.$$

- Weights $W \in \mathbb{R}^{p \times d}$, $p \leq d$ drawn uniformly from partial orthonormal matrices, $WW^\top = I_p$. Assume $\theta \sim \mathcal{N}(0, \alpha^2 I_d / d)$.

Setup ctd

- Asymptotic regime: data dimension d, number of random features p, sample size n

$$d \to \infty, \qquad \frac{p}{d} \to \pi \in (0,1], \qquad \frac{d}{n} \to \delta.$$

π - parametrization level; δ - data aspect ratio.

Setup ctd

- Asymptotic regime: data dimension d, number of random features p, sample size n

$$d \to \infty, \qquad \frac{p}{d} \to \pi \in (0, 1], \qquad \frac{d}{n} \to \delta.$$

  π - parametrization level; δ - data aspect ratio.

- Let $\gamma := \pi\delta = \lim p/n$ and the resolvent moments:

$$\theta_j(\gamma, \lambda) := \int \frac{1}{(x + \lambda)^j} dF_\gamma(x)$$

  where $F_\gamma(x)$ is the Marchenko-Pastur distribution with parameter γ.

Setup ctd

- Asymptotic regime: data dimension d, number of random features p, sample size n

$$d \to \infty, \qquad \frac{p}{d} \to \pi \in (0,1], \qquad \frac{d}{n} \to \delta.$$

  π - parametrization level; δ - data aspect ratio.

- Let $\gamma := \pi\delta = \lim p/n$ and the resolvent moments:

$$\theta_j(\gamma, \lambda) := \int \frac{1}{(x+\lambda)^j} dF_\gamma(x)$$

  where $F_\gamma(x)$ is the Marchenko-Pastur distribution with parameter γ.

- Let

$$\tilde{\lambda} := \lambda + \frac{1-\pi}{2\pi}\left[\lambda + 1 - \gamma + \sqrt{(\lambda+\gamma-1)^2 + 4\lambda}\right],$$

  and $\tilde{\theta}_1 := \theta_1(\delta, \tilde{\lambda}), \tilde{\theta}_2 := \theta_2(\delta, \tilde{\lambda})$.

Note

θ_1, θ_2 have closed form:

$$\theta_1 = \frac{(-\lambda + \gamma - 1) + \sqrt{(-\lambda + \gamma - 1)^2 + 4\lambda\gamma}}{2\lambda\gamma},$$
$$\theta_2 = -\frac{d}{d\lambda}\theta_1 = \frac{(\gamma - 1)}{2\gamma\lambda^2} + \frac{(\gamma + 1)\cdot\lambda + (\gamma - 1)^2}{2\gamma\lambda^2\sqrt{(-\lambda + \gamma - 1)^2 + 4\lambda\gamma}}.$$

Main result: ANOVA for two-layer linear NN

Theorem. Denoting s: features X; i: initialization W; l: label noise $\mathcal{E}$, we have

$$\lim_{d\to\infty} V_s = \alpha^2[1 - 2\tilde{\lambda}\tilde{\theta}_1 + \tilde{\lambda}^2\tilde{\theta}_2 - \pi^2(1-\lambda\theta_1)^2]$$

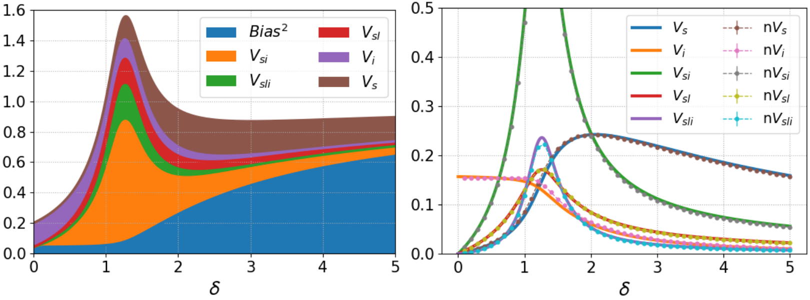
$$\lim_{d\to\infty} V_l = 0$$

$$\lim_{d\to\infty} V_i = \alpha^2\pi(1-\pi)(1-\lambda\theta_1)^2$$

$$\lim_{d\to\infty} V_{sl} = \sigma^2\delta(\tilde{\theta}_1 - \tilde{\lambda}\tilde{\theta}_2)$$

$$\begin{aligned}
\lim_{d\to\infty} V_{li} &= 0 \\
\lim_{d\to\infty} V_{si} &= \alpha^2[\pi(1 - 2\lambda\theta_1 + \lambda^2\theta_2 + (1-\pi)\delta(\theta_1 - \lambda\theta_2)) \\
&\quad - \pi(1-\pi)(1-\lambda\theta_1)^2 - 1 + 2\tilde{\lambda}\tilde{\theta}_1 - \tilde{\lambda}^2\tilde{\theta}_2] \\
\lim_{d\to\infty} V_{sli} &= \sigma^2\delta[\pi(\theta_1 - \lambda\theta_2) - (\tilde{\theta}_1 - \tilde{\lambda}\tilde{\theta}_2)].
\end{aligned}$$

ANOVA for two-layer linear NN



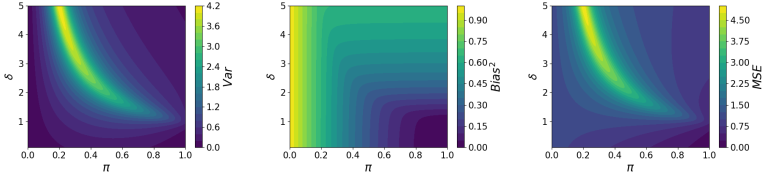
Left: Cumulative figure of the bias and variance components, as fn of $\delta = \lim d/n$.
Right: Variance components with numerical simulations.
Parameters: signal strength $\alpha = 1$, noise level $\sigma = 0.3$, regularization parameter $\lambda = 0.01$, parametrization level $\pi = 0.8$.

Interaction can dominate.

Monotonicity and unimodality

Theorem 2.7 (Bias and variance of ridge models given a fixed λ)**.** *Under the assumptions in our two layer setting, we have*

1. *For any fixed $\lambda > 0$, $\lim_{d\to\infty} \mathbf{Bias}^2(\lambda)$ is monotonically decreasing as a function of π and is monotonically increasing as a function of δ.*
2. *When $\lambda \to 0$, $\lim_{\lambda\to 0}\lim_{d\to\infty} \mathbf{Var}(\lambda) = \infty$ on the curve $\delta = 1/\pi$ (the interpolation threshold where $\lim p/d = 1$).*



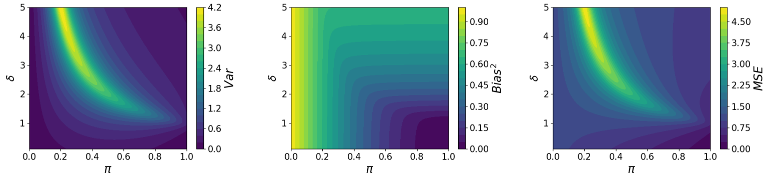
$\lambda = 0.01, \alpha = 1, \sigma = 0.3, \pi = p/d, \delta = d/n$.
d: dimension of x, n: number of samples, p: hidden layer width.

- Model-wise and sample-wise non-monotonicity appear.

Monotonicity and unimodality

Theorem 2.7 (Bias and variance of ridge models given a fixed λ). *Under the assumptions in our two layer setting, we have*

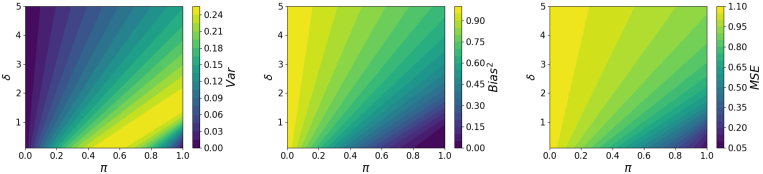
1. *For any fixed $\lambda > 0$, $\lim_{d\to\infty} \mathbf{Bias}^2(\lambda)$ is monotonically decreasing as a function of π and is monotonically increasing as a function of δ.*
2. *When $\lambda \to 0$, $\lim_{\lambda\to 0} \lim_{d\to\infty} \mathbf{Var}(\lambda) = \infty$ on the curve $\delta = 1/\pi$ (the interpolation threshold where $\lim p/d = 1$).*



$\lambda = 0.01, \alpha = 1, \sigma = 0.3, \pi = p/d, \delta = d/n$.
d: dimension of x, n: number of samples, p: hidden layer width.

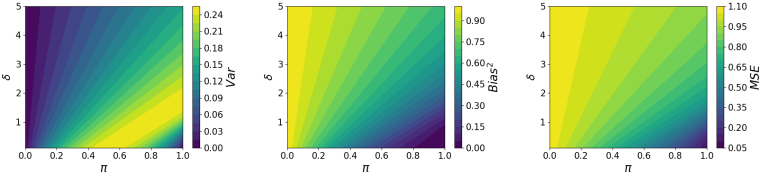
- Model-wise and sample-wise non-monotonicity appear.
- Unimodal variance investigation inspired by Yang, Yu, You, Steinhardt, Ma, 2020.
- The non-monotonicity of MSE comes from the variance.

Monotonicity and unimodality, optimal regularization



Parameters: $\lambda = \lambda^*, \alpha = 1, \sigma = 0.3, \pi = p/d, \delta = d/n$.
d: dimension of x, n: number of samples, p: hidden layer width.

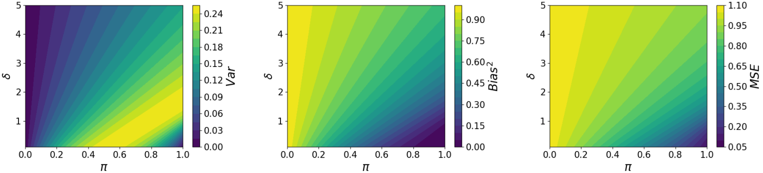
Monotonicity and unimodality, optimal regularization



Parameters: $\lambda = \lambda^*, \alpha = 1, \sigma = 0.3, \pi = p/d, \delta = d/n$.
d: dimension of x, n: number of samples, p: hidden layer width.

- Optimal ridge penalty makes MSE monotonic. (consistent with [Nakkiran et al., 2020])

Monotonicity and unimodality, optimal regularization



Parameters: $\lambda = \lambda^*, \alpha = 1, \sigma = 0.3, \pi = p/d, \delta = d/n$.
d: dimension of x, n: number of samples, p: hidden layer width.

- Optimal ridge penalty makes MSE monotonic. (consistent with [Nakkiran et al., 2020])
- The variance can still be unimodal.

Monotonicity for optimal λ

For optimal $\lambda = \lambda^*$:

Variable / Function	parametrization $\pi = \lim p/d$	aspect ratio $\delta = \lim d/n$
MSE	$\searrow$	$\nearrow$
Bias^2	$\searrow$	$\nearrow$
Var	$\delta < 2\alpha^2/(\alpha^2 + 2\sigma^2)$: $\wedge$, max at $[2 + \delta(1 + 2\sigma^2/\alpha^2)]/4$. $\delta \geq 2\alpha^2/(\alpha^2 + 2\sigma^2)$: $\nearrow$.	$\pi \leq 0.5$: $\searrow$. $\pi > 0.5$: $\wedge$, max at $2(2\pi - 1)/[1 + 2\sigma^2/\alpha^2]$.

Table 1: Monotonicity properties of bias,variance and mse as a function of π or δ, while holding all other parameters fixed. $\nearrow$: non-decreasing. $\searrow$: non-increasing. $\wedge$: unimodal. $\lambda = \lambda^*$ (optimal).

Heatmaps of components

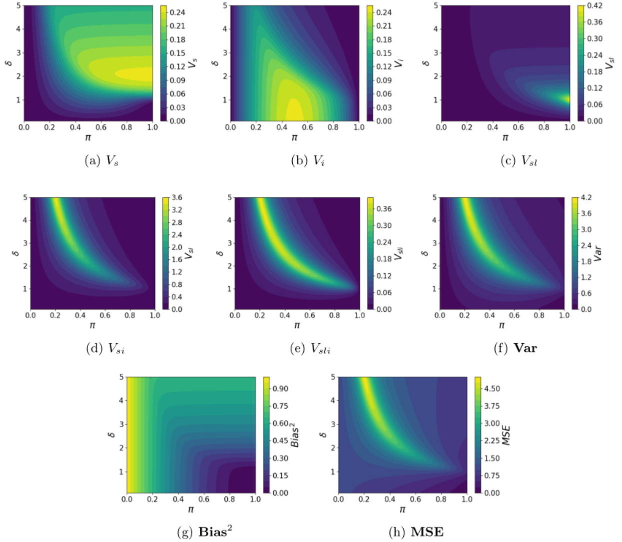
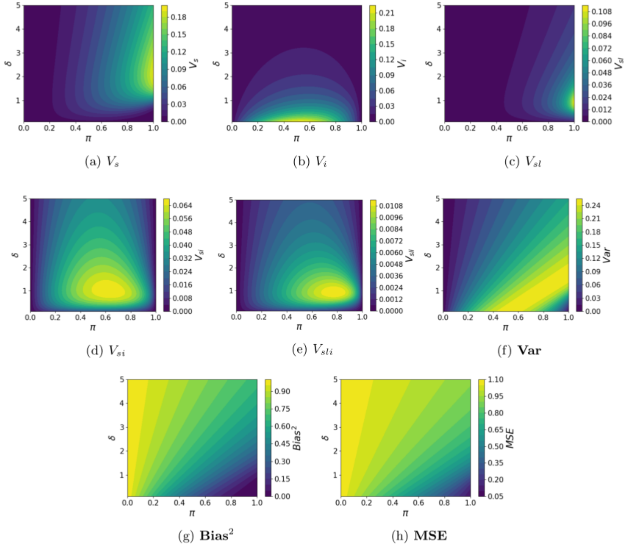


Figure 5: Heatmaps of the performance characteristics for a fixed parameter $\lambda = 0.01$. variance components, variance, bias and the MSE as functions of π and δ when $\alpha = 1, \sigma = 0.3$. ($\mathbf{Var} = V_s + V_i + V_{sl} + V_{si} + V_{sli}$. $\mathbf{MSE} = \mathbf{Bias}^2 + \mathbf{Var} + \sigma^2$.)

s: sample X.
i: initialization W.
l: label noise $\mathcal{E}$.

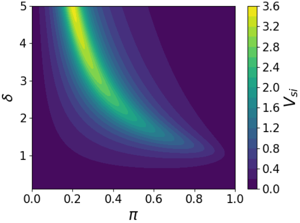
Heatmaps of components, optimal λ^*



s: sample X.
i: initialization W.
l: label noise $\mathcal{E}$.

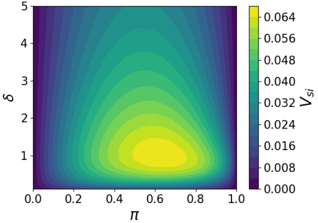
Figure 4: Heatmaps of the performance characteristics for the optimal regularization parameter $\lambda = \lambda^*$. variance components, variance, bias and the MSE as functions of π and δ when $\alpha = 1, \sigma = 0.3$. ($\mathbf{Var} = V_s + V_i + V_{sl} + V_{si} + V_{sli}$. $\mathbf{MSE} = \mathbf{Bias}^2 + \mathbf{Var} + \sigma^2$.)

What is the effect of regularization?



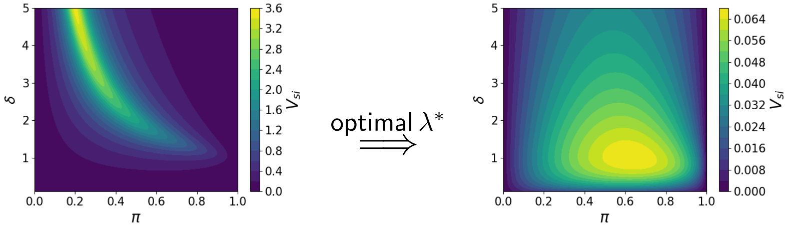
V_{si}, $\lambda = 0.01$

optimal λ^* $\Longrightarrow$



V_{si}, $\lambda = \lambda^*$

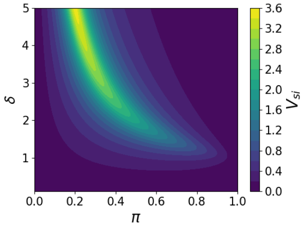
What is the effect of regularization?



V_{si}, $\lambda = 0.01$ $\qquad$ V_{si}, $\lambda = \lambda^*$

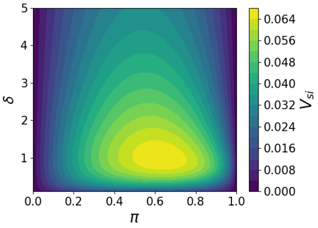
- Large reduction in V_{si}

What is the effect of regularization?



V_{si}, $\lambda = 0.01$

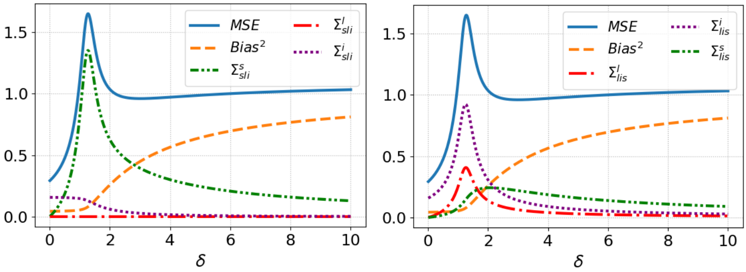
optimal λ^*
$\Longrightarrow$



V_{si}, $\lambda = \lambda^*$

- Large reduction in V_{si}
- V_{si}: The part of variance that can be reduced via ensembling over the sample X or initialization W.

Decomposition order has large effect



where

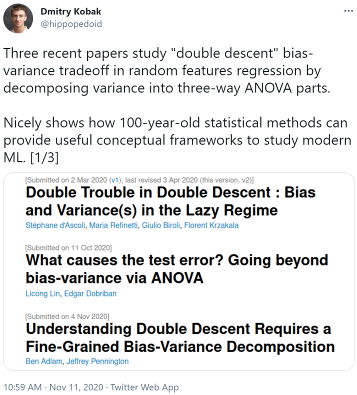
$$
\begin{aligned}
\Sigma^a_{abc} &:= \mathbb{E}_{\theta,x}\mathbb{E}_{a,b,c}[\hat{f}(x) - \mathbb{E}_a \hat{f}(x)]^2 & \Sigma^a_{abc} &= V_a + V_{ab} + V_{ac} + V_{abc} \\
\Sigma^b_{abc} &:= \mathbb{E}_{\theta,x}\mathbb{E}_{b,c}[\mathbb{E}_a \hat{f}(x) - \mathbb{E}_{a,b} \hat{f}(x)]^2 & \Sigma^b_{abc} &= V_{bc} + V_b \\
\Sigma^c_{abc} &:= \mathbb{E}_{\theta,x}\mathbb{E}_{c}[\mathbb{E}_{a,b} \hat{f}(x) - \mathbb{E}_{a,b,c} \hat{f}(x)]^2. & \Sigma^c_{abc} &= V_c.
\end{aligned}
$$

Related Works

- early works in 1980/90s: Hertz et al. [1989], Opper et al. [1990], Hansen [1993], Barber et al. [1995], Duin [1995], Opper [1995], Opper and Kinzel [1996], Raudys and Duin [1998]
- Advani Saxe, 2017, ...
- Belkin, Rakhlin, Tsybakov, 2018, Belkin, Hsu, Xu, 2019
- Liang, Rakhlin, 2018
- Hastie, Montanari, Rosset, Tibshirani, 2019, Bartlett, Long, Lugosi, Tsigler, 2019
- Muthukumar, Vodrahalli, Sahai, 2019
- Mei and Montanari, 2019
- d'Ascoli, Refinetti, Biroli, Krzakala, 2020
- Nakkiran, Venkat, Kakade, Ma, 2020
- Yang, Yu, You, Steinhardt, Ma, 2020
- Many others... see paper for details.

Most closely related works

- Yang, Yu, You, Steinhardt, Ma, 2020: variance unimodality, different theoretical model
- d'Ascoli, Refinetti, Biroli, Krzakala, 2020: hierarchical decomposition, Gaussian, "physics-level" rigor
- Adlam and Pennington [2020]: parallel work, Gaussian initialization, different tools; study ensemble learning, do not focus on properties of the bias/variance/mse.



Outline

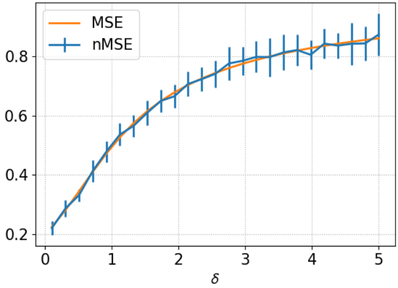
Numerical verification of MSE formula

- Generate $k = 400$ i.i.d. tuples $(x_i, \theta_i, \varepsilon_i, X_i, W_i)$, $1 \le i \le k$, X and x with i.i.d. $\mathcal{N}(0, 1)$ entries.

Numerical verification of MSE formula

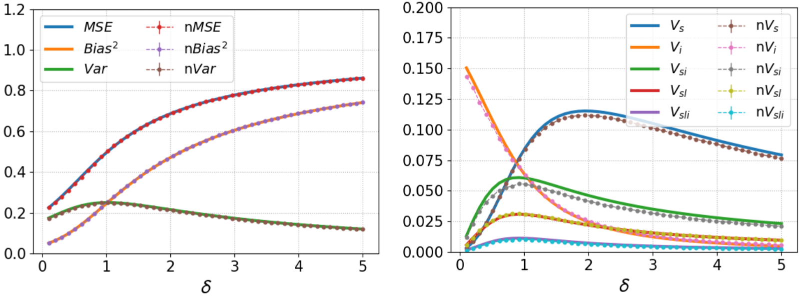
- Generate $k = 400$ i.i.d. tuples $(x_i, \theta_i, \varepsilon_i, X_i, W_i)$, $1 \le i \le k$, X and x with i.i.d. $\mathcal{N}(0,1)$ entries.
- With $\hat{f}_i(x_i) = x_i^\top W_i^\top \left(n^{-1} W_i X_i^\top y_i W_i^\top + \lambda I_p\right)^{-1} n^{-1} W_i X_i^\top y_i$, estimate MSE:

$$\mathrm{nMSE} = k^{-1} \sum_{i=1}^{k} (\hat{f}_i(x_i) - x_i^\top \theta_i)^2.$$



Parameters: $\alpha = 1, \sigma = 0.3, \pi = 0.8, n = 150, d = \lfloor n\delta \rfloor, p = \lfloor d\pi \rfloor$, $\lambda = \lambda^*$. Mean and s.e. over 20 repetitions.

Numerical verification of formulas for variance components



Simulations verifying accuracy of the bias, variance, ANOVA components. $\star$: theory, $n\star$: numerical. Parameters: $\alpha = 1, \sigma = 0.3, \pi = 0.8, n = 150, d = \lfloor n\delta \rfloor, p = \lfloor d\pi \rfloor$.

Experiments on empirical data

- Superconductivity data set.[3]
  - Goal: predict critical temperature T_c below which material is superconductive.

[3] Hamidieh, 2018; archive.ics.uci.edu/ml/datasets/superconductivty+data

Experiments on empirical data

- Superconductivity data set.[3]
  - Goal: predict critical temperature T_c below which material is superconductive.
  - $d = 81$ features: mean/entropy/SD of material properties: atomic mass, radius, thermal conductivity, ...

[3] Hamidieh, 2018; archive.ics.uci.edu/ml/datasets/superconductivty+data

Experiments on empirical data

- Superconductivity data set.[3]
  - Goal: predict critical temperature T_c below which material is superconductive.
  - $d = 81$ features: mean/entropy/SD of material properties: atomic mass, radius, thermal conductivity, ...
  - $N = 21,263$ materials, e.g., $\mathrm{RbAsO_2}$: Rubidium arsenic dioxide

[3] Hamidieh, 2018; archive.ics.uci.edu/ml/datasets/superconductivty+data

Experiments on empirical data

- Superconductivity data set.[3]
  - Goal: predict critical temperature T_c below which material is superconductive.
  - $d = 81$ features: mean/entropy/SD of material properties: atomic mass, radius, thermal conductivity, ...
  - $N = 21,263$ materials, e.g., $\mathrm{RbAsO_2}$: Rubidium arsenic dioxide
- Standard preprocessing: random $90 - 10\%$ train-test split, feature standardization

[3] Hamidieh, 2018; archive.ics.uci.edu/ml/datasets/superconductivty+data

Experiments on empirical data

- Superconductivity data set.[3]
  - Goal: predict critical temperature T_c below which material is superconductive.
  - $d = 81$ features: mean/entropy/SD of material properties: atomic mass, radius, thermal conductivity, ...
  - $N = 21,263$ materials, e.g., $\mathrm{RbAsO_2}$: Rubidium arsenic dioxide
- Standard preprocessing: random $90 - 10\%$ train-test split, feature standardization
- Fitting: Randomly select n samples: X; map into random p-subspace with W; do ridge.

[3] Hamidieh, 2018; archive.ics.uci.edu/ml/datasets/superconductivty+data

Experiments on empirical data: estimating components

- Generate i.i.d. X_i, $1 \leq i \leq n_s$, W_j, $1 \leq j \leq n_i$, form (X_i, W_j). Let

$$\hat{f}_{ij}(x) = x^\top \left(\frac{W_j X_i^\top X_i W_j^\top}{n} + \lambda I_p \right)^{-1} \frac{W_j X_i^\top y_i}{n}, \quad 1 \leq i, j \leq 50.$$

Experiments on empirical data: estimating components

- Generate i.i.d. X_i, $1 \le i \le n_s$, W_j, $1 \le j \le n_i$, form (X_i, W_j). Let

$$\hat{f}_{ij}(x) = x^\top \left(\frac{W_j X_i^\top X_i W_j^\top}{n} + \lambda I_p \right)^{-1} \frac{W_j X_i^\top y_i}{n}, \quad 1 \le i, j \le 50.$$

- With $n_i = n_s = 50$, test set size L, test data x_k, y_k, estimate

$$\widehat{\mathrm{MSE}} = \frac{1}{L} \sum_{k=1}^{L} \hat{\mathbb{E}}(\hat{f}_{ij}(x_k) - y_k)^2,$$

$$\widehat{\mathrm{Var}} = \frac{1}{L} \sum_{k=1}^{L} \hat{\mathbb{E}}(\hat{f}_{ij}(x_k) - \hat{\mathbb{E}}\hat{f}_{ij}(x_k))^2, \ \ \widehat{\mathrm{Bias}}^2 = \frac{1}{L} \sum_{k=1}^{L} (\hat{\mathbb{E}}\hat{f}_{ij}(x_k) - y_k)^2,$$

$$\widehat{V_s} = \frac{1}{L n_s} \sum_{k=1}^{L} \sum_{i=1}^{n_s} (\hat{\mathbb{E}}_j \hat{f}_{ij}(x_k) - \hat{\mathbb{E}}\hat{f}_{ij}(x_k))^2,$$

$$\widehat{V_i} = \frac{1}{L n_i} \sum_{k=1}^{L} \sum_{j=1}^{n_i} (\hat{\mathbb{E}}_i \hat{f}_{ij}(x_k) - \hat{\mathbb{E}}\hat{f}_{ij}(x_k))^2.$$

Experiments on empirical data

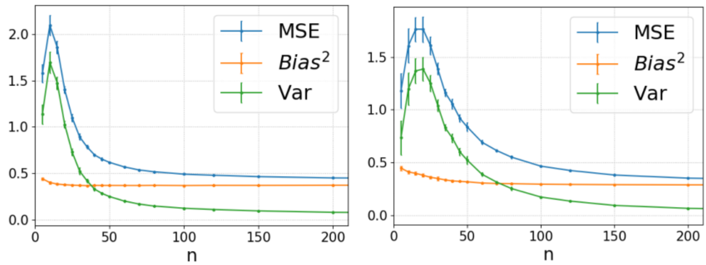
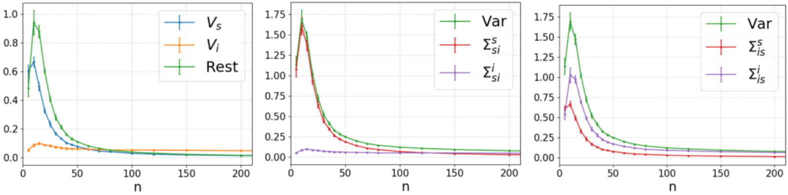


Figure 8: Empirically estimated MSE, variance and bias as functions of number of samples n. We display the mean and one standard deviation of the numerical results over 10 repetitions. Left: $\pi = 0.2, \lambda = 0.01$. Right: $\pi = 0.9, \lambda = 0.01$.

Experiments: Decomposition order has large effect



Outline

Nonlinear activation: Setup

- **Data**: n datapoints $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ drawn i.i.d. from $y = f^*(x) + \varepsilon = x^\top \theta + \varepsilon$, $\theta \in \mathbb{R}^d$, where x has i.i.d. $\mathcal{N}(0, 1)$ entries, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is the label noise independent of x. In matrix form, $Y = X\theta + \mathcal{E}$.

Nonlinear activation: Setup

- **Data**: n datapoints $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ drawn i.i.d. from $y = f^*(x) + \varepsilon = x^\top \theta + \varepsilon$, $\theta \in \mathbb{R}^d$, where x has i.i.d. $\mathcal{N}(0, 1)$ entries, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is the label noise independent of x. In matrix form, $Y = X\theta + \mathcal{E}$.
- **Model**: Learn $f^*(x) = x^\top \theta$ using a two-layer neural network,

$$f(x) = \sigma(Wx)^\top \beta.$$

Assume that the parameters θ are random: $\theta \sim \mathcal{N}(0, \alpha^2 I_d / d)$.

Nonlinear activation: Setup

- **Data**: n datapoints $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ drawn i.i.d. from $y = f^*(x) + \varepsilon = x^\top \theta + \varepsilon$, $\theta \in \mathbb{R}^d$, where x has i.i.d. $\mathcal{N}(0,1)$ entries, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is the label noise independent of x. In matrix form, $Y = X\theta + \mathcal{E}$.
- **Model**: Learn $f^*(x) = x^\top \theta$ using a two-layer neural network,

  $$f(x) = \sigma(Wx)^\top \beta.$$

  Assume that the parameters θ are random: $\theta \sim \mathcal{N}(0, \alpha^2 I_d / d)$.
- **Orthogonality**: The first-layer weight matrix W is drawn uniformly from matirces with orthonormal rows, i.e., $p \leq d$, $WW^\top = I_p$.

Nonlinear activation: Setup

- **Data**: n datapoints $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ drawn i.i.d. from $y = f^*(x) + \varepsilon = x^\top \theta + \varepsilon$, $\theta \in \mathbb{R}^d$, where x has i.i.d. $\mathcal{N}(0,1)$ entries, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is the label noise independent of x. In matrix form, $Y = X\theta + \mathcal{E}$.
- **Model**: Learn $f^*(x) = x^\top \theta$ using a two-layer neural network,

  $$f(x) = \sigma(Wx)^\top \beta.$$

  Assume that the parameters θ are random: $\theta \sim \mathcal{N}(0, \alpha^2 I_d/d)$.
- **Orthogonality**: The first-layer weight matrix W is drawn uniformly from matirces with orthonormal rows, i.e., $p \leq d$, $WW^\top = I_p$.
- **Training**: Train the second layer weight β with L_2 loss+penalty. Corresponds to random feature model.

Nonlinear activation: Setup

- Suppose that σ, σ' grows at most exponentially, i.e., there exist $c_1, c_2 > 0$ such that $|\sigma(x)|, |\sigma'(x)| \leq c_1 e^{c_2|x|}$. Assume $\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\sigma(Z) = 0$. Define the moments

$$\mu := \mathbb{E}_{Z\sim\mathcal{N}(0,1)} Z\sigma(Z), \quad v := \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\sigma^2(Z).$$

Nonlinear activation: Setup

- Suppose that σ, σ' grows at most exponentially, i.e., there exist $c_1, c_2 > 0$ such that $|\sigma(x)|, |\sigma'(x)| \le c_1 e^{c_2|x|}$. Assume $\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\sigma(Z) = 0$. Define the moments

$$\mu := \mathbb{E}_{Z\sim\mathcal{N}(0,1)} Z\sigma(Z), \quad v := \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\sigma^2(Z).$$

- Training the second layer gives us the ridge estimator:

$$\hat{f}(x) := \sigma(Wx)^\top \hat{\beta} = \sigma(x^\top W^\top)\left(\frac{\sigma(WX^\top)\sigma(XW^\top)}{n} + \lambda I_p\right)^{-1} \frac{\sigma(WX^\top)Y}{n}.$$

Main result #2: ANOVA for two-layer NN, non-linear activation

Theorem. As $d, p, n \to \infty$ proportionally:

$$\lim_{d\to\infty} \mathbf{MSE}(\lambda) = \alpha^2\pi\left[\frac{1}{\pi} - 1 + \delta(1-\pi)\theta_1 + \frac{\lambda}{v}\left(\frac{\lambda\mu^2}{v^2} - \delta(1-\pi)\right)\theta_2 \right.$$
$$\left. +(v-\mu^2)\left(\frac{\gamma}{v}\theta_1 + \frac{1}{v} - \frac{\lambda\gamma}{v^2}\theta_2\right)\right] + \sigma^2\gamma\left(\theta_1 - \frac{\lambda}{v}\theta_2\right) + \sigma^2, \quad (16)$$

$$\lim_{d\to\infty} \mathbf{Bias}^2(\lambda) = \alpha^2\left[\pi\frac{\mu^2}{v}\left(1 - \frac{\lambda}{v}\theta_1\right) - 1\right]^2, \quad (17)$$

$$\lim_{d\to\infty} \mathbf{Var}(\lambda) = \alpha^2\pi\left[\frac{2\mu^2}{v} - 1 + \left(-\frac{2\lambda\mu^2}{v^2} + \delta(1-\pi)\right)\theta_1 + \frac{\lambda}{v}\left(\frac{\lambda\mu^2}{v^2} - \delta(1-\pi)\right)\theta_2 \right.$$
$$\left. -\frac{\pi\mu^4}{v^2}\left(1 - \frac{\lambda}{v}\theta_1\right)^2 + (v-\mu^2)\left(\frac{\gamma}{v}\theta_1 + \frac{1}{v} - \frac{\lambda\gamma}{v^2}\theta_2\right)\right] + \sigma^2\gamma\left(\theta_1 - \frac{\lambda}{v}\theta_2\right), \quad (18)$$

where $\theta_1 := \theta_1(\gamma, \lambda/v)$, $\theta_2 := \theta_2(\gamma, \lambda/v)$, $\gamma = \pi\delta$. *Similar to the linear case, the limiting MSE has a unique minimum at* $\lambda^* := \frac{v^2}{\mu^2}\left[\delta(1 - \pi + \sigma^2/\alpha^2) + \frac{(v-\mu^2)\gamma}{v}\right]$.

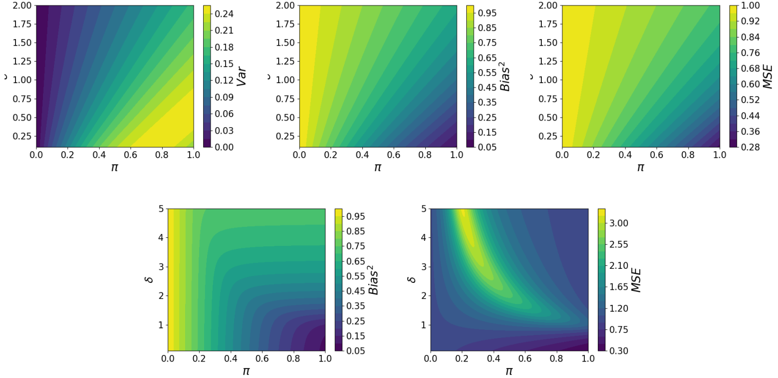
Monotonicity properties

For the optimal penalty $\lambda = \lambda^*$, we have the same monotonicity properties as in the linear case:

Function \ Variable	parametrization $\pi = \lim p/d$	aspect ratio $\delta = \lim d/n$
MSE	$\searrow$	$\nearrow$
Bias^2	$\searrow$	$\nearrow$
Var	$\delta < 2\frac{\mu^2}{v}\left(2\frac{\mu^2}{v}-1\right) / (1+2\sigma^2/\alpha^2)$: $\wedge$, max at $\frac{v}{\mu^2}\left[2+\frac{\delta v}{\mu^2}\left(1+\frac{2\sigma^2}{\alpha^2}\right)\right]/4$. $\delta \geq 2\frac{\mu^2}{v}\left(2\frac{\mu^2}{v}-1\right) / (1+2\sigma^2/\alpha^2)$: $\nearrow$.	$\pi \leq \frac{v}{2\mu^2}$: $\searrow$. $\pi > \frac{v}{2\mu^2}$: $\wedge$, max at $\frac{2\mu^2(2\pi\mu^2/v-1)}{v(1+2\sigma^2/\alpha^2)}$.

Table 2: Monotonicity properties of various components of the risk for a two-layer network with nonlinear activation, as a function of π or δ, while holding all other parameters fixed. $\nearrow$: non-decreasing. $\searrow$: non-increasing. $\wedge$: unimodal. Thus, e.g., the MSE is non-increasing as a function of the parameterization level π, while holding δ fixed.

Monotonicity properties



Parameters:

$\sigma(\cdot) = \mathrm{ReLU}(\cdot) - \mathbb{E}\mathrm{ReLU}, \lambda = 0.01, \alpha = 1, \sigma = 0.3, \pi = p/d, \delta = d/n.$

d: dimension of x, n: number of samples, p: hidden layer width.

Outline

Proof techniques

- The proof uses techniques from asymptotic random matrix theory (Marchenko & Pastur, 1967, Bai & Silverstein, 2010, Couillet & Debbah, 2011, ...)

Proof techniques

- The proof uses techniques from asymptotic random matrix theory (Marchenko & Pastur, 1967, Bai & Silverstein, 2010, Couillet & Debbah, 2011, ...)
- We leverage deterministic equivalent results for Haar random matrices from [Couillet et al., 2012]. Have not been used in the area before?

Calculation of the variance components

Define

$$\begin{aligned}
\tilde{M}_{X,W}(\lambda) &:= W^\top (n^{-1} W X^\top X W^\top + \lambda I_p)^{-1} W X^\top / n \\
M_{X,W}(\lambda) &:= \tilde{M}_{X,W}(\lambda) X.
\end{aligned}$$

Then we have $f_{\lambda,\mathcal{T},W}(x) = x^\top \tilde{M} Y = x^\top M \theta + x^\top \tilde{M} \mathcal{E}$.

For V_s,

$$\begin{aligned}
V_s &= \mathbb{E}_{\theta,x} \mathrm{Var}_X(\mathbb{E}_{\mathcal{E},W}(\hat{f}(x)|X)) = \mathbb{E}_{\theta,x,X}[x^\top (\mathbb{E}_W M - \mathbb{E} M)\theta]^2 \\
&= \frac{\alpha^2}{d} \mathbb{E}_X \|\mathbb{E}_W M - \mathbb{E} M\|_F^2.
\end{aligned}$$

Calculation of the variance components

Similarly, we can write down all the variance components.

$$\begin{aligned}V_s &= \mathbb{E}_{\theta,x}\mathrm{Var}_X(\mathbb{E}_{\mathcal{E},W}(\hat{f}(x)|X)) = \mathbb{E}_{\theta,x,X}[x^\top(\mathbb{E}_W M - \mathbb{E}M)\theta]^2\\ &= \frac{\alpha^2}{d}\mathbb{E}_X\|\mathbb{E}_W M - \mathbb{E}M\|_F^2.\end{aligned}$$

$$V_l = \mathbb{E}_{\theta,x}\mathrm{Var}_{\mathcal{E}}(\mathbb{E}_{X,W}(\hat{f}(x)|\mathcal{E})) = \sigma^2\|\mathbb{E}\tilde{M}\|_F^2.$$

$$\begin{aligned}V_{li} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{\mathcal{E},W}(\mathbb{E}_X(\hat{f}(x)|\mathcal{E},W)) - V_i - V_l\\ &= \mathbb{E}_{\theta,x,\mathcal{E},W}[x^\top(\mathbb{E}_X M - \mathbb{E}M)\theta + x^\top\mathbb{E}_X\tilde{M}\mathcal{E}]^2 - V_i - V_l\\ &= \sigma^2\mathbb{E}_W\|\mathbb{E}_X\tilde{M} - \mathbb{E}\tilde{M}\|_F^2\end{aligned}$$

$$\begin{aligned}V_i &= \mathbb{E}_{\theta,x}\mathrm{Var}_W(\mathbb{E}_{\mathcal{E},X}(\hat{f}(x)|W)) = \mathbb{E}_{\theta,x,W}[x^\top(\mathbb{E}_X M - \mathbb{E}M)\theta]^2\\ &= \frac{\alpha^2}{d}\mathbb{E}_W\|\mathbb{E}_X M - \mathbb{E}M\|_F^2.\end{aligned}$$

$$\begin{aligned}V_{si} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{X,W}(\mathbb{E}_\tau(\hat{f}(x)|X,W)) - V_s - V_i\\ &= \mathbb{E}_{\theta,x,X,W}[x^\top(M - \mathbb{E}M)\theta]^2 - V_s - V_i\\ &= \frac{\alpha^2}{d}\left(\mathbb{E}\|M\|_F^2 - \mathbb{E}_X\|\mathbb{E}_W M\|_F^2 - \mathbb{E}_W\|\mathbb{E}_X M\|_F^2 + \|\mathbb{E}M\|_F^2\right).\end{aligned}$$

$$\begin{aligned}V_{sl} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{\mathcal{E},X}(\mathbb{E}_W(\hat{f}(x)|\mathcal{E},X)) - V_s - V_l\\ &= \mathbb{E}_{\theta,x,X,\mathcal{E}}[x^\top(\mathbb{E}_W M - \mathbb{E}M)\theta + x^\top\mathbb{E}_W\tilde{M}\mathcal{E}]^2 - V_s - V_l\\ &= \sigma^2\mathbb{E}_X\|\mathbb{E}_W\tilde{M} - \mathbb{E}\tilde{M}\|_F^2.\end{aligned}$$

$$\begin{aligned}V_{sli} &= \mathrm{Var}(\hat{f}(x)) - (V_s + V_l + V_i + V_{sl} + V_{si} + V_{li})\\ &= \sigma^2(\mathbb{E}\|\tilde{M}\|_F^2 - \mathbb{E}_W\|\mathbb{E}_X\tilde{M}\|_F^2 - \mathbb{E}_X\|\mathbb{E}_W\tilde{M}\|_F^2 + \|\mathbb{E}\tilde{M}\|^2).\end{aligned}$$

Calculation of the variance components

It remains to calculate the following terms.

Lemma 6.1 (Behavior of $\mathbb{E}M$).

$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathrm{tr}(M) = \pi(1-\lambda\theta_1), \quad \forall i \geq 1. \qquad \lim_{d\to\infty} \frac{1}{d}\|\mathbb{E}M\|_F^2 = \pi^2(1-\lambda\theta_1)^2.$$

Lemma 6.2 (Behavior of the Frobenius norm of M).

$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}\|M\|_F^2 = \pi\left[1 - 2\lambda\theta_1 + \lambda^2\theta_2 + (1-\pi)\delta(\theta_1 - \lambda\theta_2)\right].$$

Lemma 6.3 (Behavior of the Frobenius norm of $\tilde{M}$).

$$\lim_{d\to\infty} \mathbb{E}\|\tilde{M}\|_F^2 = \pi\delta(\theta_1 - \lambda\theta_2).$$

Lemma 6.4 (Behavior of the Frobenius norm of $\mathbb{E}_X M$).

$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}_W\|\mathbb{E}_X M\|_F^2 = \pi(1-\lambda\theta_1)^2.$$

Lemma 6.5 (Behavior of the Frobenius norm of $\tilde{M}$).

$$\lim_{d\to\infty} \|\mathbb{E}\tilde{M}\|_F^2 = \lim_{d\to\infty} \mathbb{E}_W\|\mathbb{E}_X\tilde{M}\|_F^2 = 0.$$

Lemma 6.6 (Behavior of the Frobenius norm of $\mathbb{E}_W\tilde{M}$).

$$\lim_{d\to\infty} \mathbb{E}_X\|\mathbb{E}_W\tilde{M}\|_F^2 = \delta(\tilde{\theta}_1 - \tilde{\lambda}\tilde{\theta}_2).$$

Lemma 6.7 (Behavior of the Frobenius norm of $\mathbb{E}_W M$).

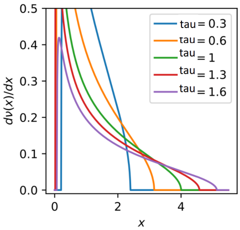
$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}_X\|\mathbb{E}_W M\|_F^2 = (1 - 2\tilde{\lambda}\tilde{\theta}_1 + \tilde{\lambda}^2\tilde{\theta}_2).$$

Marchenko Pastur theorem, 1967

Theorem (MP'67; Bai-Silverstein '95-'10)

Suppose $X \in R^{n\times d}$ has i.i.d. entries with zero mean and unit variance. If $d \to \infty, d/n \to \tau$, then with probability one, the empirical spectral distribution of $X^\top X/n$ weakly converges to the Marchenko Pastur distribution μ_τ.

$$\mu_\tau(A) = \begin{cases} \left(1 - \frac{1}{\tau}\right) 1_{0\in A} + \nu_\tau(A), & \text{if } \tau > 1 \\ \nu_\tau(A), & \text{if } 0 \le \tau \le 1, \end{cases}$$



where

$$d\nu_\tau(x) = \frac{1}{2\pi} \frac{\sqrt{(\tau_+ - x)(x - \tau_-)}}{\tau x} 1_{x\in[\tau_-,\tau_+]} dx,$$

and

$$\tau_\pm = (1 \pm \sqrt{\tau})^2.$$

Deterministic equivalents

Definition (Serdobolskii, Girko, etc)

We say that the (deterministic or random) not necessarily symmetric matrix sequences A_n, B_n of growing dimensions are equivalent, and write

$$A_n \asymp B_n$$

if

$$\lim_{n\to\infty} \left|\mathrm{tr}\left[C_n\left(A_n - B_n\right)\right]\right| = 0 \tag{1}$$

almost surely, for any sequence C_n of not necessarily symmetric matrices with bounded trace norm, i.e., such that

$$\limsup \left\|C_n\right\|_{tr} < \infty.$$

Moreover, if (1) only holds almost surely for any sequence $C_n \in \mathbb{R}^{d_n \times d_n}$ of positive semidefinite matrices with $O\left(1/d_n\right)$ spectral norm, A_n and B_n are said to be *weak deterministic equivalents* and denoted by $A_n \overset{w}{\asymp} B_n$.

Some techniques in the proof

Example 1. (Mestre et al., 2011)

Let $\hat{\Sigma} = X^\top X/n$, where $X = Z\Sigma^{1/2}$ and Z is an $n \times p$ random matrix with iid entries of zero mean, unit variance and finite $8+\eta$ moment. Also, $\Sigma^{1/2}$ is any sequence of $p \times p$ positive semi-definite matrices satisfying $\sup \|\Sigma\|_2 < \infty$. As $n, p \to \infty$ proportionally, for any $\lambda > 0$

$$(\widehat{\Sigma} + \lambda I_p)^{-1} \asymp (q_p \Sigma + \lambda I_p)^{-1},$$

where q_p is the solution of a fixed point equation.

Some techniques in the proof

Example 2. (Couillet et al., 2012)

Let $W \in \mathbb{R}^{p\times d}$ be the first p rows of a unitary Haar distributed random matrix. Suppose $R^{d\times d}$ is a sequence of positive semi-definite random matrices such that $\sup \|R\|_2 < \infty$, almost surely. As $p, d \to \infty$ proportionally, for any $\lambda > 0$

$$(R^{1/2}W^\top WR^{1/2} + \lambda I_d)^{-1} \overset{w}{\asymp} (\bar{e}_d R + \lambda I_d)^{-1},$$

where $\bar{e}_d$ is the solution of a fixed point equation.

Summary

- ANOVA decomposition of test error
  1. Surprising finding: interaction effect is large — need to take into account interaction effects between initialization, input randomness, label noise
  2. Monotonicity, Unimodality
- What causes the test error? Going beyond bias-variance via ANOVA: `arxiv.org/abs/2010.05170`, *to appear in JMLR*
- code to reproduce numerical results: `github.com/licong-lin/VarianceDecomposition`
- Thanks!

References I

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. *arXiv preprint arXiv:2011.03321, NeurIPS 2020*, 2020.

David Barber, David Saad, and Peter Sollich. Finite-size effects and optimal test set size in linear perceptrons. *Journal of Physics A: Mathematical and General*, 28(5):1325, 1995.

Romain Couillet, Jakob Hoydis, and Mérouane Debbah. Random beamforming over quasi-static and fading channels: A deterministic equivalent approach. *IEEE Transactions on Information Theory*, 58(10):6392–6425, 2012.

Robert PW Duin. Small sample size generalization. In *Proceedings of the Scandinavian Conference on Image Analysis*, volume 2, pages 957–964. PROCEEDINGS PUBLISHED BY VARIOUS PUBLISHERS, 1995.

Lars Kai Hansen. Stochastic linear learning: Exact test and training error averages. *Neural Networks*, 6(3):393–396, 1993.

References II

JA Hertz, A Krogh, and GI Thorbergsson. Phase transitions in simple learning. *Journal of Physics A: Mathematical and General*, 22(12):2133, 1989.

M Opper, W Kinzel, J Kleinz, and R Nehl. On the ability of the optimal perceptron to generalise. *Journal of Physics A: Mathematical and General*, 23(11):L581, 1990.

Manfred Opper. Statistical mechanics of learning: Generalization. *The Handbook of Brain Theory and Neural Networks,*, pages 922–925, 1995.

Manfred Opper and Wolfgang Kinzel. Statistical mechanics of generalization. In *Models of neural networks III*, pages 151–209. Springer, 1996.

Sarunas Raudys and Robert PW Duin. Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. *Pattern recognition letters*, 19(5-6):385–392, 1998.

Ridge is asy Bayes optimal

Theorem 2.8 (Ridge is optimal). *Suppose that the samples are drawn from the standard normal distribution, i.e., x and X both have i.i.d. $\mathcal{N}(0,1)$ entries. Given the projection W, projected matrix $XW^\top$ and response Y, we define the optimal regression parameter β_{opt} as the one minimizing the MSE over the posterior distribution $p(\theta|XW^\top, W, Y)$ of the parameter θ,*

$$\beta_{opt} := \operatorname{argmin}_\beta \mathbb{E}_{p(\theta|XW^\top,W,Y)} \mathbb{E}_{x,\varepsilon}[(Wx)^\top \beta - (x^\top \theta + \varepsilon)]^2, \tag{10}$$

where $x \sim \mathcal{N}(0, I_d)$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and x, ε are independent. We will check that this can be expressed in terms of the posterior of θ as

$$\beta_{opt} = W \cdot \mathbb{E}_{p(\theta|XW^\top,W,Y)} \theta. \tag{11}$$

The optimal ridge estimator $\hat{\beta} = (n^{-1}WX^\top XW^\top + \lambda^* I_p)^{-1} WX^\top Y/n$ (Theorem 2.3) satisfies the almost sure convergence in the mean squared error

$$\lim_{d\to\infty} \mathbb{E}_{XW^\top,W,Y} \|\hat{\beta} - \beta_{opt}\|_2^2 = 0, \tag{12}$$

and is thus asymptotically optimal. Here $d \to \infty$ means $p, d, n \to \infty$ proportionally as in Theorem 2.3.

Remarks:

Ridge is asy Bayes optimal

Theorem 2.8 (Ridge is optimal). *Suppose that the samples are drawn from the standard normal distribution, i.e., x and X both have i.i.d. $\mathcal{N}(0,1)$ entries. Given the projection W, projected matrix $XW^\top$ and response Y, we define the optimal regression parameter β_{opt} as the one minimizing the MSE over the posterior distribution $p(\theta|XW^\top, W, Y)$ of the parameter θ,*

$$\beta_{opt} := \operatorname{argmin}_\beta \mathbb{E}_{p(\theta|XW^\top,W,Y)} \mathbb{E}_{x,\varepsilon}[(Wx)^\top \beta - (x^\top \theta + \varepsilon)]^2, \tag{10}$$

where $x \sim \mathcal{N}(0, I_d)$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and x, ε are independent. We will check that this can be expressed in terms of the posterior of θ as

$$\beta_{opt} = W \cdot \mathbb{E}_{p(\theta|XW^\top,W,Y)}\theta. \tag{11}$$

The optimal ridge estimator $\hat{\beta} = (n^{-1}WX^\top XW^\top + \lambda^* I_p)^{-1}WX^\top Y/n$ (Theorem 2.3) satisfies the almost sure convergence in the mean squared error

$$\lim_{d\to\infty} \mathbb{E}_{XW^\top,W,Y}\|\hat{\beta} - \beta_{opt}\|_2^2 = 0, \tag{12}$$

and is thus asymptotically optimal. Here $d \to \infty$ means $p, d, n \to \infty$ proportionally as in Theorem 2.3.

Remarks:

- $\mathbb{E}\|\hat{\beta}\|^2 \to c_0 > 0$. Thus, the asymptotic result is non-trivial.

Ridge is asy Bayes optimal

Theorem 2.8 (Ridge is optimal). *Suppose that the samples are drawn from the standard normal distribution, i.e., x and X both have i.i.d. $\mathcal{N}(0,1)$ entries. Given the projection W, projected matrix $XW^\top$ and response Y, we define the optimal regression parameter β_{opt} as the one minimizing the MSE over the posterior distribution $p(\theta|XW^\top, W, Y)$ of the parameter θ,*

$$\beta_{opt} := \operatorname{argmin}_\beta \mathbb{E}_{p(\theta|XW^\top,W,Y)}\mathbb{E}_{x,\varepsilon}[(Wx)^\top\beta - (x^\top\theta + \varepsilon)]^2, \tag{10}$$

where $x \sim \mathcal{N}(0, I_d)$, $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ and x, ε are independent. We will check that this can be expressed in terms of the posterior of θ as

$$\beta_{opt} = W \cdot \mathbb{E}_{p(\theta|XW^\top,W,Y)}\theta. \tag{11}$$

The optimal ridge estimator $\hat\beta = (n^{-1}WX^\top XW^\top + \lambda^* I_p)^{-1}WX^\top Y/n$ (Theorem 2.3) satisfies the almost sure convergence in the mean squared error

$$\lim_{d\to\infty} \mathbb{E}_{XW^\top,W,Y}\|\hat\beta - \beta_{opt}\|_2^2 = 0, \tag{12}$$

and is thus asymptotically optimal. Here $d \to \infty$ means $p, d, n \to \infty$ proportionally as in Theorem 2.3.

Remarks:

- $\mathbb{E}\|\hat\beta\|^2 \to c_0 > 0$. Thus, the asymptotic result is non-trivial.
- Given X instead of $XW^\top$, ridge is not Bayes optimal.