# STAT 946 - Topics in Probability and Statistics: Mathematical Foundations of Deep Learning
# Lecture 7

Lucas Noritomi-Hartwig
University of Waterloo

September 24, 2025 from 16h00 to 17h20 in M3 3103

October 6th: Majid (Deep RL), Marty (MF Transformers) presenting articles.

October 20th: ? ?

October 8th: Last day to decide on a topic and direction for a project.

<u>Recall</u>

$$
\begin{aligned}
f\left(x^{\alpha};\ \theta\right) &= \frac{1}{\sqrt{n}} W_d \varphi\left(h_d\right) \\
h_{l+1}^{\alpha} &= \frac{1}{\sqrt{n}} W_l \varphi\left(h_l\right) \\
h_1^{\alpha} &= \frac{1}{\sqrt{n_0}} W_0 x^{\alpha}
\end{aligned}
$$

Backward neuron:

$$
\begin{aligned}
g_l^{\alpha} &= \sqrt{n} \frac{\partial f^{\alpha}}{\partial h_l^{\alpha}} \\
z_l^{\alpha} &= \frac{1}{\sqrt{n}} W_l^{\top} g_{l+1}^{\alpha} \\
&= \frac{1}{\sqrt{n}} W_l^{\top} \underbrace{\operatorname{diag}\left(\varphi'\left(h_{l+1}^{\alpha}\right)\right)}_{=: D_{l+1}^{\alpha}} z_{l+1}^{\alpha} \\
\mathcal{F}_l^z &= \sigma\left(\{h_k^{\alpha}\}_{\alpha,\, k},\ \{z_k\}_{k \geq l}\right)
\end{aligned}
$$

$$
\begin{aligned}
W | \sigma\left(W \varphi\right) &\overset{\mathrm{d}}{=} W P_{\varphi} + \underbrace{\tilde{W}}_{\text{indep. copy}} P_{\varphi}^{\perp} \\
z_l^{\alpha} | \mathcal{F}_l^z &= \frac{1}{\sqrt{n}} \left(P_{\varphi_l} W_l^{\top} + P_{\varphi_l}^{\perp} \tilde{W}_l^{\top}\right) D_{l+1}^{\alpha} z_{l+1}^{\alpha} | \mathcal{F}_l^z && \left(h_{l+1} = W_l \varphi\left(h_l\right)\right) \\
&= P_{\varphi_l} z_l^{\alpha} + P_{\varphi_l}^{\perp} \underbrace{\frac{1}{\sqrt{n}} \tilde{W}_l^{\top} g_{l+1}^{\alpha}}_{\tilde{z}_l^{\alpha} \sim \mathcal{N}(0,\, G_{l+1} \otimes I)} | \mathcal{F}_l^z
\end{aligned}
$$

$$
\begin{aligned}
G_l^{\alpha\beta} &= \frac{1}{n}\left\langle g_l^{\alpha},\, g_l^{\beta}\right\rangle \\
&= \frac{1}{n}\left\langle D_l^{\alpha} z_l^{\alpha},\, D_l^{\beta} z_l^{\beta}\right\rangle \\
&= \frac{1}{n}\left\langle D_l^{\alpha}\left(P_{\varphi_l} z_l^{\alpha} + P_{\varphi_l}^{\perp}\tilde{z}_l^{\alpha}\right),\, D_l^{\beta}\left(P_{\varphi_l} z_l^{\beta} + P_{\varphi_l}^{\perp}\tilde{z}_l^{\beta}\right)\right\rangle \\
&= \frac{1}{n}\left\langle D_l^{\alpha} P_{\varphi_l} z_l^{\alpha},\, D_l^{\beta} P_{\varphi_l} z_l^{\beta}\right\rangle && (1)\\
&\quad + \frac{1}{n}\left\langle D_l^{\alpha} P_{\varphi_l}^{\perp}\tilde{z}_l^{\alpha},\, D_l^{\beta} P_{\varphi_l} z_l^{\beta}\right\rangle && (2)\\
&\quad + \frac{1}{n}\left\langle D_l^{\alpha} P_{\varphi_l} z_l^{\alpha},\, D_l^{\beta} P_{\varphi_l}^{\perp}\tilde{z}_l^{\beta}\right\rangle && (3)\\
&\quad + \frac{1}{n}\left\langle D_l^{\alpha} P_{\varphi_l}^{\perp}\tilde{z}_l^{\alpha},\, D_l^{\beta} P_{\varphi_l}^{\perp}\tilde{z}_l^{\beta}\right\rangle && (4)
\end{aligned}
$$

$$
(1) = \frac{1}{n}\sum_{j=1}^{n} \underbrace{\varphi'\left(h_{l,j}^{\alpha}\right)}_{\in\Theta(1)} \left(P_{\varphi_l} z_l^{\alpha}\right)_j
$$

$\|P_{\varphi_l} z_l^{\alpha}\|^2$ is a projection on the column space of $\varphi_l$, namely $\operatorname{col}(\varphi_l)$, where $\varphi_l$ is $n \times m$.

$$
\varphi_l = \begin{bmatrix} \varphi\left(\underbrace{h_l^{(1)}}_{n\times 1}\right) & \varphi\left(\underbrace{h_l^{(2)}}_{n\times 1}\right) & \cdots & \varphi\left(\underbrace{h_l^{(m)}}_{n\times 1}\right) \end{bmatrix}
$$

In the trivial case, i.e., $m = 1$,

$$
\varphi_l = e_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^n
$$

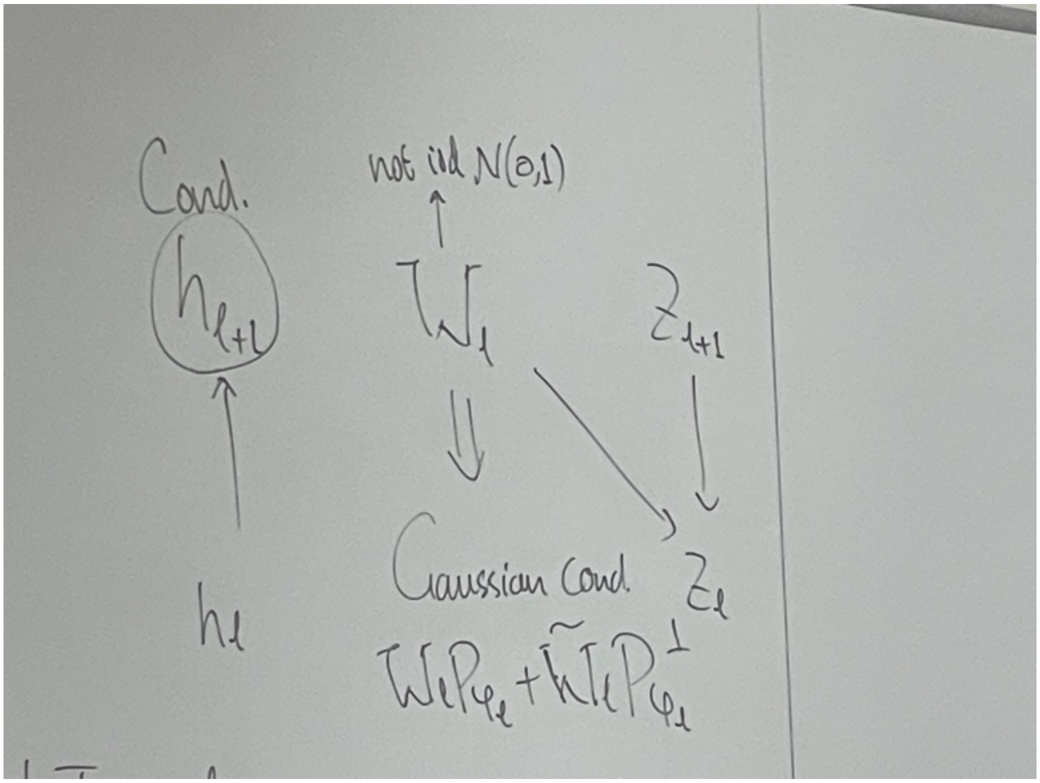
$$
P_{\varphi_l} z_l^{\alpha} = \begin{bmatrix} z_{l,1}^{\alpha} \\ 0 \\ \vdots \\ 0 \end{bmatrix}
$$

So as $n \to \infty$, $\|P_{\varphi_l} z_l^{\alpha}\|^2 \lesssim \frac{m}{n}\|z_l^{\alpha}\|^2$

$$
(2) + (3) = \frac{1}{n}\sum_i \cdots \left(\underbrace{P_{\varphi_l}^{\perp}\tilde{z}_l^{\alpha}}_{\text{zero mean, indep.}}\right)_i \cdots \qquad (\dagger)
$$

$$
\tilde{z}_l^{\alpha} = \frac{1}{\sqrt{n}} \underbrace{\tilde{W}_l^{\top}}_{\text{indep. copy of } W_l^{\top}} g_{l+1}^{\alpha}
$$

$$
\begin{aligned}
(\dagger) &= \frac{1}{\sqrt{n}}\Theta(1) \to \text{CLT scaling} \\
&\to 0
\end{aligned}
$$




$$
\begin{aligned}
(4) &= \frac{1}{n}\sum_{i=1}^{n} \varphi'\left(h_{l,i}^{\alpha}\right) \tilde{Z}_{l,i}^{\alpha} \varphi'\left(h_{l,i}^{\beta}\right) \tilde{z}_{l,i}^{\beta} - \cdots \underbrace{\left(P_{\varphi_l} \tilde{z}_l^{\alpha}\right)}_{\Theta(m\ \ n) \to 0} \cdots \left(P_{\varphi_l} \tilde{z}_{l,i}^{\beta}\right) \\
&\to \mathbb{E}\left[\varphi'\left(h_{l,i}^{\alpha}\right) \varphi'\left(h_{l,i}^{\beta}\right) \mathbb{E}\left[\underbrace{\tilde{z}_{l,i}^{\alpha} \tilde{z}_{l,i}^{\beta}}_{\to G_{l+1}^{\alpha\beta}} | \mathcal{F}_{l+1}^{z}\right]\right]
\end{aligned}
$$

$$
P_{\varphi_l}^{\perp} \tilde{z}_l^{\alpha} = P_{\varphi_l}^{\perp} \frac{1}{\sqrt{n}} \tilde{W}_l^{\top} g_{l+1}^{\alpha} \sim \mathcal{N}\left(0,\, G_{l+1}^{\alpha\beta} \otimes \underbrace{P_{\varphi_l}^{\perp}}_{= I - P_{\varphi_l}}\right)
$$

Aside:

$$
\sum_{i=1}^{n} X_i \to X \implies \sum_{i=1, i \neq 2}^{n} X_i \to X,
$$

i.e., finitely many random varibales contribute infinitesimally small amounts to the limit.

$$
\begin{aligned}
\dot{\Phi}_l^{\alpha\beta} &:= \mathbb{E}\left[\varphi'\left(h_{l,i}^{\alpha}\right) \varphi'\left(h_{l,i}^{\beta}\right)\right] \\
G_l^{\alpha\beta} &= \dot{\Phi}_l^{\alpha\beta} \odot G_{l+1}^{\alpha\beta}
\end{aligned}
$$

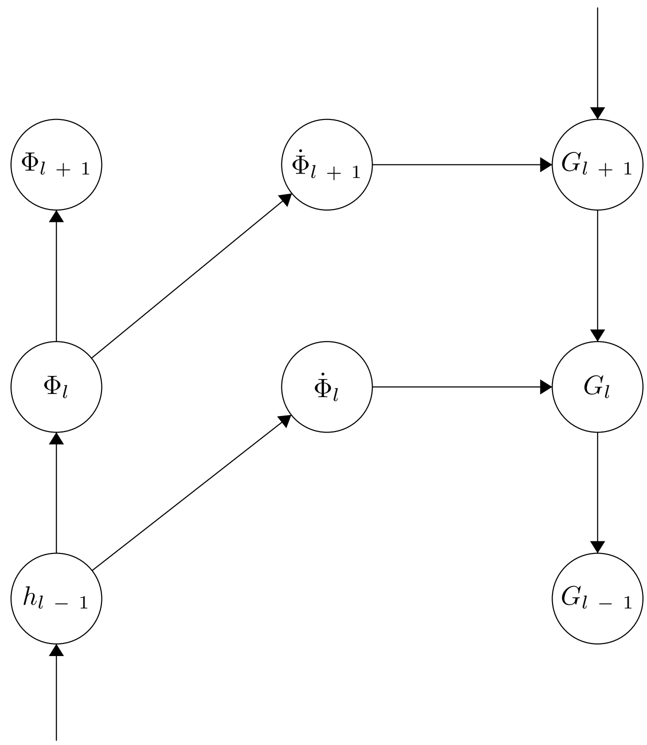
where $A \odot B = \left[a_{ij} b_{ij}\right]_{ij}$ is the Hadamard product of $A$ and $B$.

NTK:

$$K^{\alpha\beta} = \sum_{l=0}^{d} \left\langle \nabla_{W_l} f^{\alpha},\, \nabla_{W_l} f^{\beta} \right\rangle$$

$$\sum_{l=0}^{d} \frac{1}{n} \left\langle g_{l+1}^{\alpha},\, g_{l+1}^{\beta} \right\rangle \frac{1}{n} \left\langle \varphi_l^{\alpha},\, \varphi_l^{\beta} \right\rangle$$

$$= \sum_{l=0}^{d} G_{l+1}^{\alpha\beta} \Phi_l^{\alpha\beta}$$




<u>Theorem</u>

$$\begin{cases} \Phi_{l+1} = f_1\left(\Phi_l\right) \\ G_l = \dot{\Phi}_l \odot G_{l+1} \\ \Phi_0 = \left[\frac{1}{n_0}\left\langle x^{\alpha},\, x^{\beta}\right\rangle\right]_{\alpha,\,\beta} \\ G_{d+1} = 11^{\top} \end{cases}$$

$$K = \sum_{l=0}^{d} \Phi_l \odot G_{l+1}$$

$$\dot{\Phi}_{l+1} = f_2\left(\Phi_l\right)$$

$$\dot{\Phi}_{l+1} = \mathbb{E}\left[\varphi'\left(h^{\alpha}\right)\varphi'\left(h^{\beta}\right)\right]$$

$$\begin{bmatrix} h^{\alpha} \\ h^{\beta} \end{bmatrix} \sim \mathcal{N}\left(0,\, \begin{bmatrix} \Phi_l^{\alpha\alpha} & \Phi_l^{\alpha\beta} \\ \Phi_l^{\beta\alpha} & \Phi_l^{\beta\beta} \end{bmatrix}\right)$$

$$z_l^{\alpha} = P_{\varphi_l} z_l^{\alpha} + P_{\varphi_l}^{\perp} \tilde{z}_l^{\alpha}$$

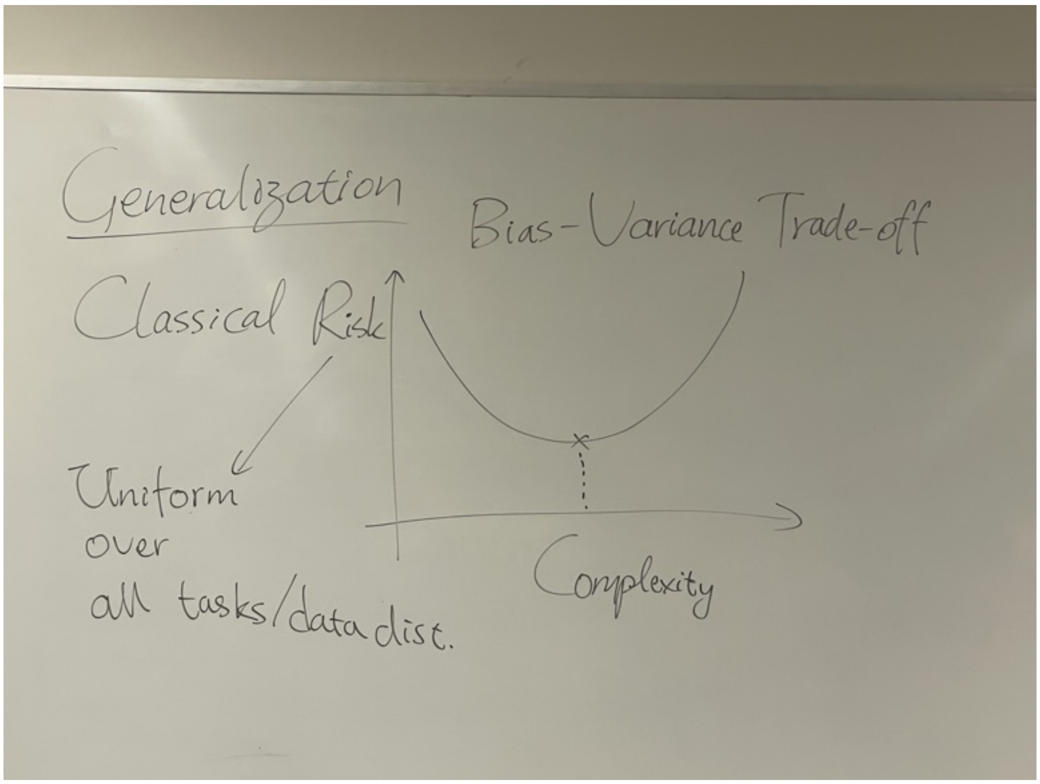
## Generalization

Something of which we have a completely new understanding since 2018-2019.

Classical view: bias-variance tradeoff, i.e., in order to perform better (have low risk), we must regularize our model's complexity.

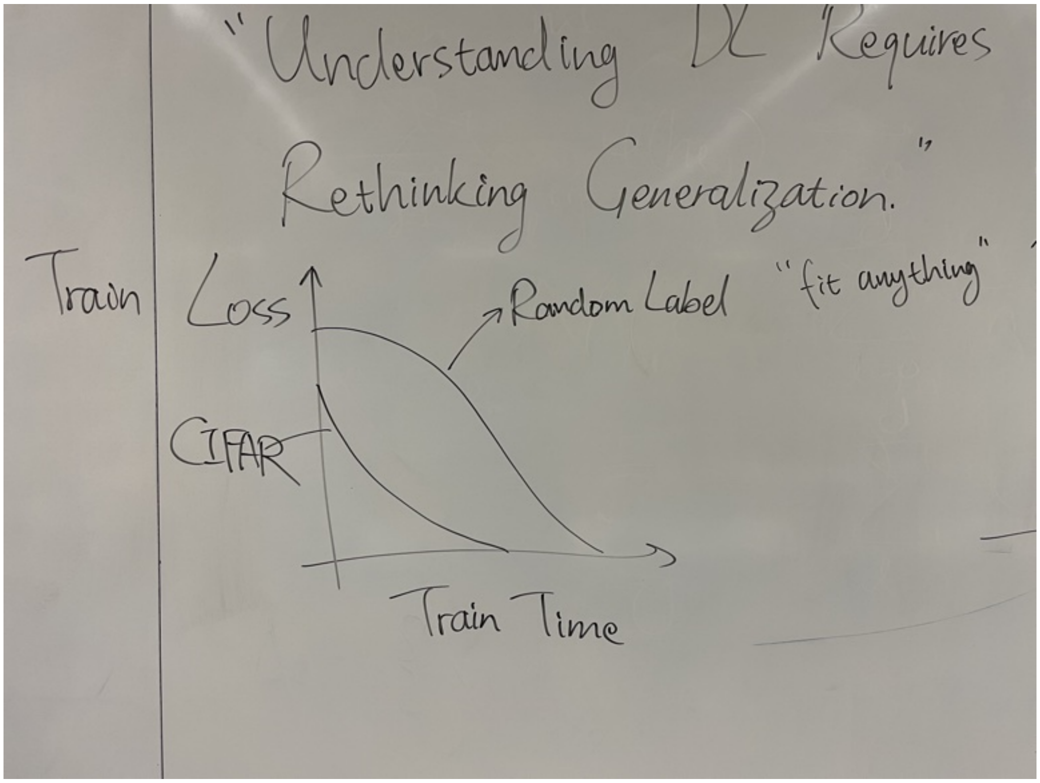
What is risk? In classical theories, "risk" is defined in a very restrictive way.

No Free Lunch Theorem
Informally: If algorithm $A$ can fit any data, then $\exists$ a task (data distribution) on which $A$ fails, i.e., there is some fixed amount of risk under which $A$ cannot output a model that performs better than the risk.

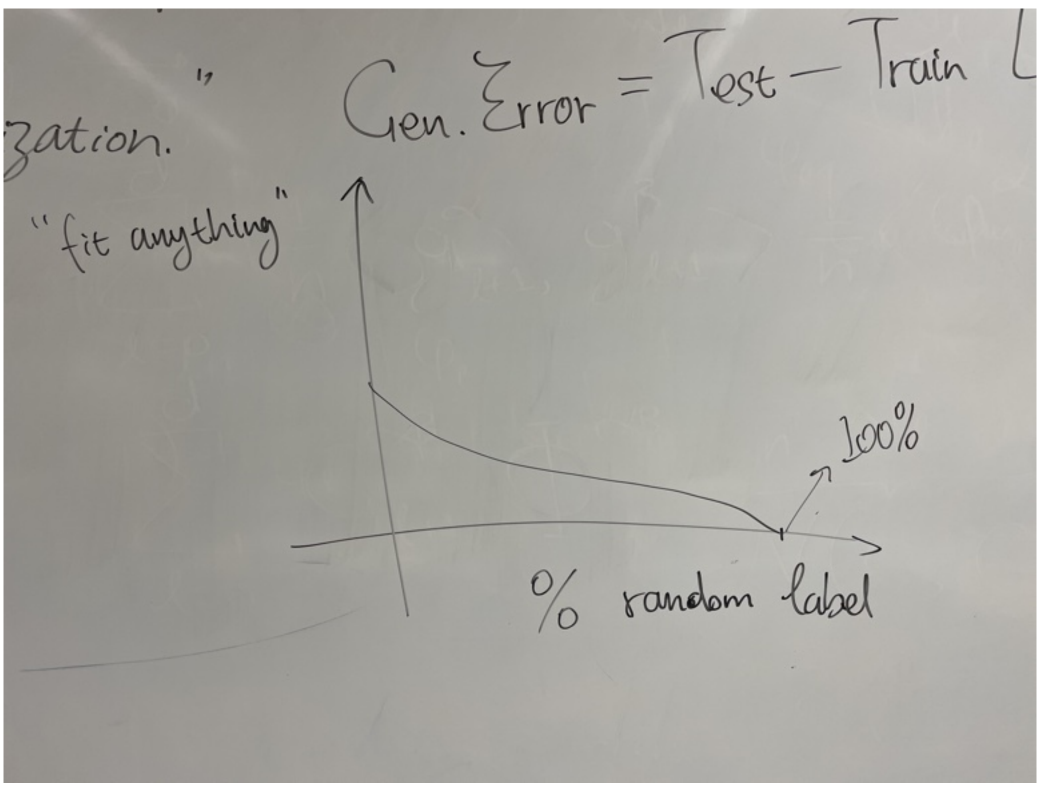



"Understanding Deep Learning Requires Rethinking Generalization" - Zhang et al (2016)

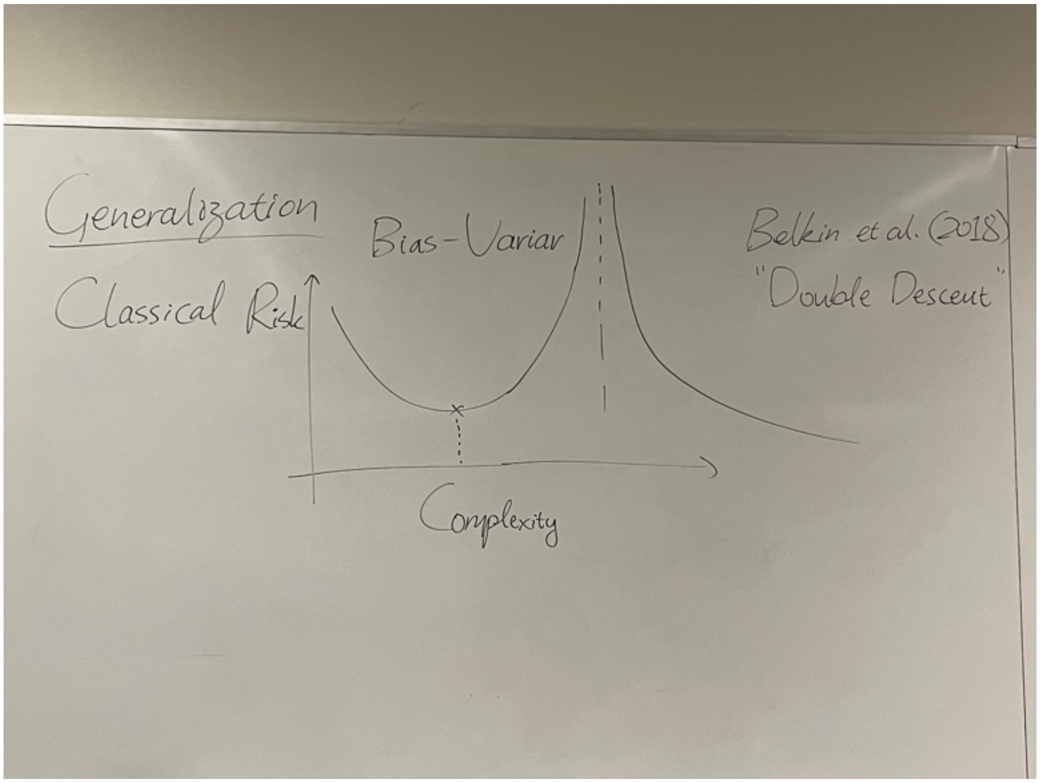
First plot: Training loss




Second plot: Generalization error (Test - Training loss)




Belkin et al. (2018) "Double Descent"


Generalization
Bias-Variar
Belkin et al. (2018)
"Double Descent"
Classical Risk
Complexity