Lecture 1/2 What is Machine Learning?

IEMS 402 Statistical Learning

Northwestern

Logistics

Logistics

- Course Website: https://2prime.github.io/teaching/2025-Statistical-Learning
- Grading: Problem Sets (15%) + Exams (80%) + Scribe Note (5%)

max(HW1,HW8)+max(HW2,HW3)+max(HW4,HW5)+max(HW6,HW7).

- [Homework 1] Review of Probability and Optimization — Review of technical basic
- [Homework 2] Bias and Variance Trade-off 1
- [Homework 3] Bias and Variance Trade-off 2
- [Homework 4] Asymptotic Theory 1
- [Homework 5] Asymptotic Theory 2
- [Homework 6] Non-Asymptotic Theory 1
- [Homework 7] Non-Asymptotic Theory 2
- [Homework 8] Advanced Topics — Advanced research in OR

- Latex and overleaf (not required)

Logistics

- Course Website: https://2prime.github.io/teaching/2025-Statistical-Learning
- Grading: Problem Sets (15%) + Exams (80%) + Scribe Note (5%)

max(HW1,HW8)+max(HW2,HW3)+max(HW4,HW5)+max(HW6,HW7).

- [Homework 1] Review of Probability and Optimization
- [Homework 2] Bias and Variance Trade-off 1 — Easy
- [Homework 3] Bias and Variance Trade-off 2
- [Homework 4] Asymptotic Theory 1 — Easy
- [Homework 5] Asymptotic Theory 2
- [Homework 6] Non-Asymptotic Theory 1 — Easy
- [Homework 7] Non-Asymptotic Theory 2
- [Homework 8] Advanced Topics

- Latex and overleaf (not required)

Logistics

- Course Website: https://2prime.github.io/teaching/2025-Statistical-Learning
- Grading: Problem Sets (15%) + $\boxed{\text{Exams (80\%)}}$ + Scribe Note (5%)

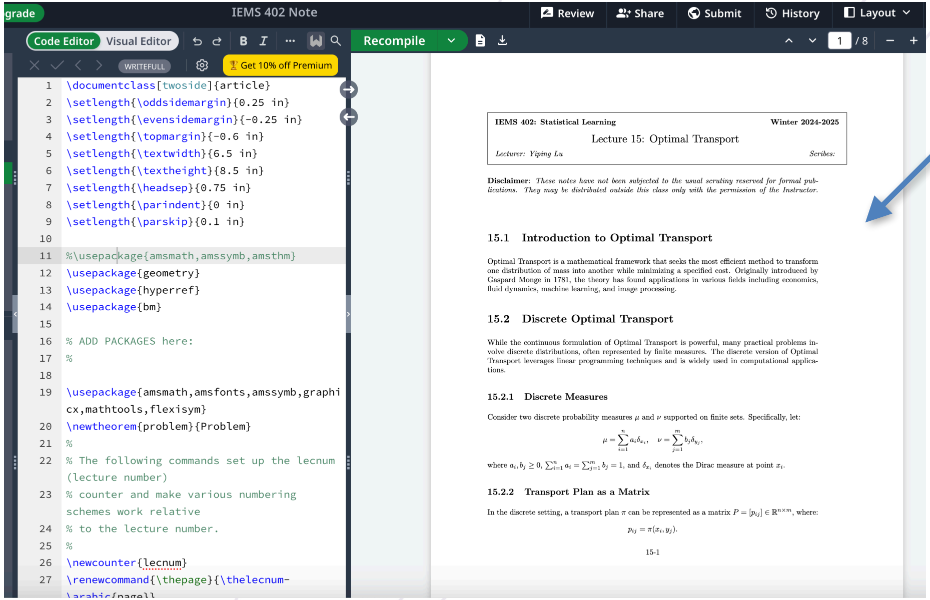
Exams

The same technique as the exam

- [Practice Mid-Term Exam]
  - Modern Machine Learning Concepts, Bias and Variance Trade-off
  - Kernel Smoothing, Asymptotic Theory, Influnce Function Concentration Inequality, Uniform Bound
- [Practice Final Exam]
  - Rademacher complexity, Covering Number, Dudley's theorem
  - RKHS, Optimal Transport, Robust Learning

Logistics

- Course Website: https://2prime.github.io/teaching/2025-Statistical-Learning
- Grading: Problem Sets (15%) + Exams (80%) + Scribe Note (5%)



Refine my note

Logistics

- Course Website: https://2prime.github.io/teaching/2025-Statistical-Learning
- Grading: Problem Sets (15%) + Exams (80%) + Scribe Note (5%)
- Textbook: Bach, Francis. Learning theory from first principles. MIT press, 2024.
  - https://www.di.ens.fr/~fbach/ltfp_book.pdf

Gradescope
Campuswire
ChatGPT Tutor!

Late Work Policy

- For your first late assignment within 12 hours after the deadline (as indicated on Gradescope), no point deductions.
- All subsequent assignments submitted within 12 hours after the deadline will convert to a zero at the end of semester.
- In all cases, work submitted 12 hours or more after the deadline will not be accepted.

Prelinminary

Review Document:

https://2prime.github.io/files/IEMS402/IEMS402ProbOptReview.pdf

Calculus, Linear Algebra

IEMS 302 Probability Probability and Statistics: Strong Law of Large Numbers, Central Limit Theorem, Big-O, little-o notation,

Optimization Theory: **Lagrangian Duality Theory IEMS 450-2: Mathematical Optimization II**

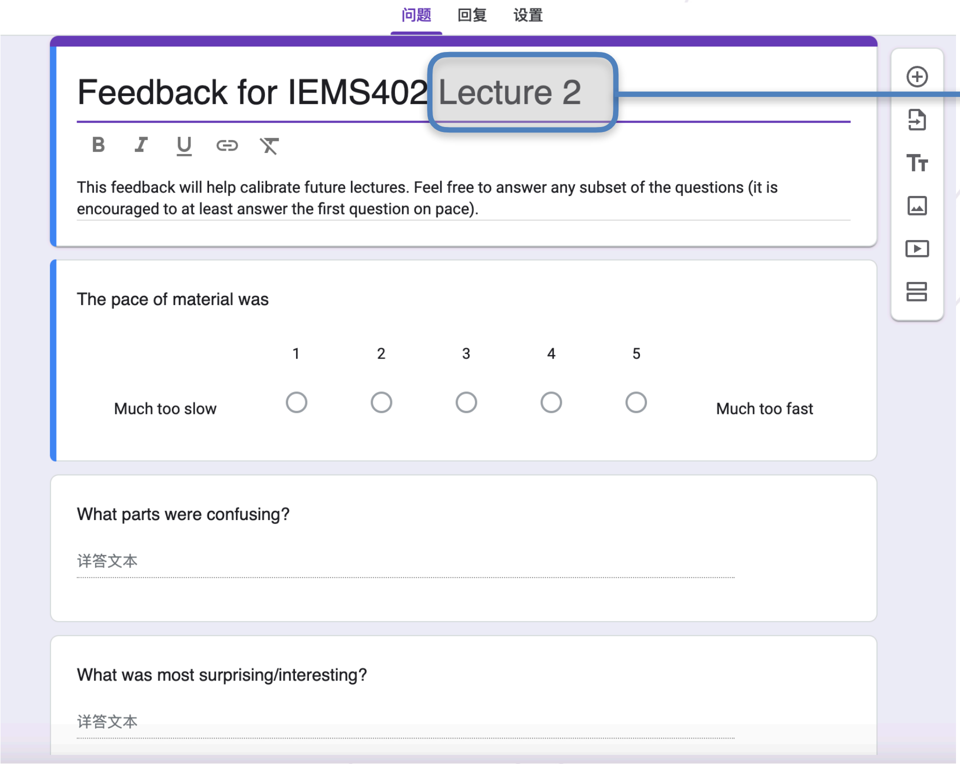
(Interestingly, IEMS 450-1 is not required)

You need to know

- Law of strong numbers, Central Limit Theorem, Continuous Map Theorem, Slutsky Theorem, Markov's Inequality

You don't need to distinguish Convergence in Probability/Covergence in distribution, you just need to write $\rightarrow$

Online Calibration with Human Feedback


问题 回复 设置

Feedback for IEMS402 Lecture 2

B I U

This feedback will help calibrate future lectures. Feel free to answer any subset of the questions (it is encouraged to at least answer the first question on pace).

The pace of material was

	1	2	3	4	5	
Much too slow	○	○	○	○	○	Much too fast

What parts were confusing?

详答文本

What was most surprising/interesting?

详答文本

Feedback for each lecture

Other Course

Stats 300b - Stanford

1. Introduction
2. Convergence of random variables (January 14)
3. Delta method (January 14)
4. Basics of asymptotic normality (January 18 and 20)
5. Moment method (January 20)
6. Uniform laws of large numbers (January 26)
7. Basics of concentration (January 28 and February 2)
8. Sub Gaussian processes and chaining (February 2 and February 4)
9. VC Dimension (February 4)
10. Uniform central limit theorems and convergence in distribution (February 9 and February 11)
11. Applications of Uniform Central Limit Theorems (February 16 and February 18)
12. Relative efficiency and basic tests (February 18 and February 23)
13. Asymptotic level and relative efficiency in testing (February 23 and 25)
14. Contiguity and Asymptotics (February 25)
15. Local Asymptotic Normality (March 2 and 4)
16. Regular estimators and consequences (March 8 and 10)
17. U statistics (March 11 and 16)
18. Parting thoughts (March 18)

Stats 705 - CMU

Date	Lecture Topic
August 31	Review
September 2	Concentration Inequalities
September 4	Concentration Inequalities
September 7	**No Class (Labor Day)**
September 9	Convergence
September 11	Convergence
September 14	Central Limit Theorem
September 18	Uniform Laws and Empirical Process Theory
September 18	Uniform Laws and Empirical Process Theory
September 21	Uniform Laws and Empirical Process Theory
September 23	Review
September 25	**TEST 1**
September 28	Likelihood and Sufficiency
September 30	Point Estimation (MLE)
October 2	Point Estimation (Method of Moments, Bayes)
October 5	Decision Theory
October 7	Decision Theory
October 9	Asymptotic Theory
October 12	Asymptotic Theory
October 14	Hypothesis Testing
October 16	NO CLASS (Community Engagement)
October 19	Goodness-of-fit, two-sample, independence
October 21	Multiple testing
October 23	**NO CLASS (Mid-Semester Break)**
October 26	Multiple testing
October 28	Confidence Intervals
October 30	Confidence Intervals
November 2	Confidence Intervals
November 4	Review
November 6	**TEST 2**
November 9	Bootstrap
November 11	Bootstrap
November 13	Bayesian Inference
November 16	Bayesian Inference
November 18	Linear Regression
November 20	Non-parametric Regression
November 23	NO CLASS
November 25	NO CLASS (Thanksgiving)
November 27	NO CLASS
November 30	Minimax Lower Bounds
December 2	Minimax Lower Bounds
December 4	High-dimensional Statistics
December 7	High-dimensional Statistics
December 9	Model Selection
December 11	Model Selection

Other Course

Stanford: Stats 300b/ CS229T

Berkeley: Stats 241/Stats 241B

MIT IDS.160/9.521/18.656/6.S988

CMU Stat705, 10-072

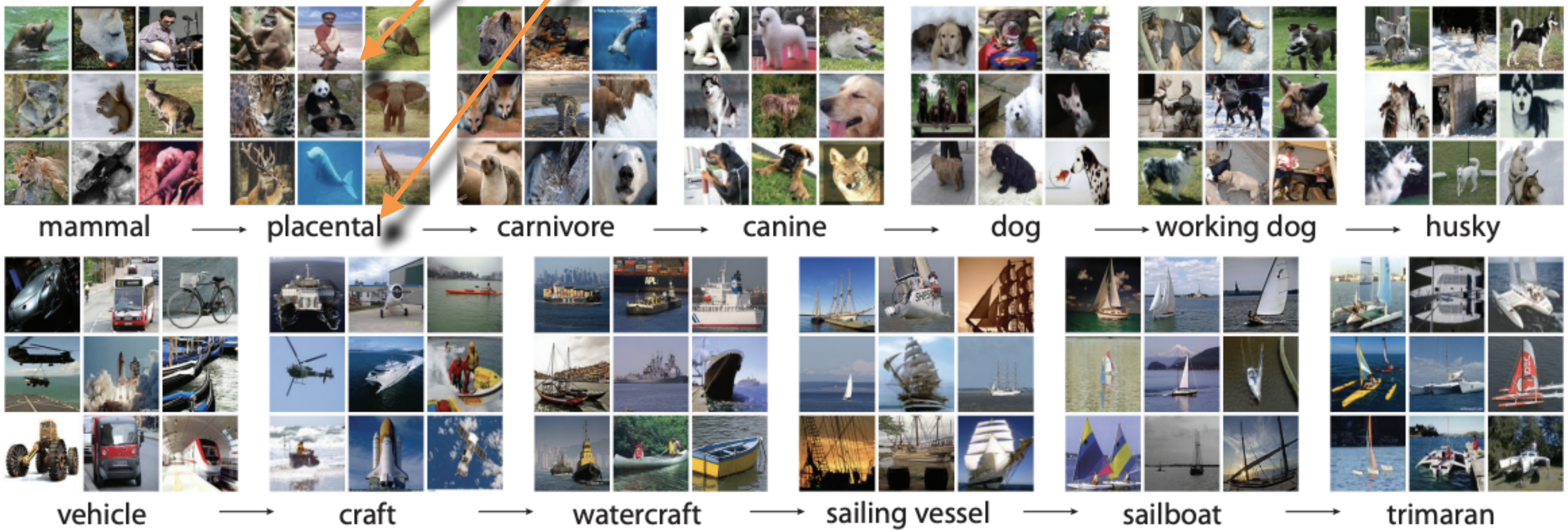
Umich EECS598, UW Madison CS 839, UofT STA3000F

Good machine learning courses are open source!

Supervised Learning

Supervised Learning

- Aim: learn a predictor $f : \mathscr{X} \rightarrow \mathscr{Y}$



Supervised Learning

- Aim: learn a predictor $f : \mathscr{X} \to \mathscr{Y}$
- What is a good predictor? -> evaluation criteria

$$\mathscr{R}(f) = \mathbb{E}[\ell(y, f(x))] = \int_{\mathscr{X}\times\mathscr{Y}} \ell(\boxed{y}, \boxed{f(x)})dp(x, y)\,.$$

Evaluate the error of label and predicition

Supervised Learning

- Aim: learn a predictor $f : \mathcal{X} \to \mathcal{Y}$
- What is a good predictor? -> evaluation criteria

$$\mathcal{R}(f) = \mathbb{E}[\ell(y, f(x))] = \boxed{\int_{\mathcal{X}\times\mathcal{Y}}} \ell(\boxed{y}, \boxed{f(x)})dp(x, y)\,.$$

Evaluate the error of label and predicition

If I want to know the risk, I need to have all the data in the univers?

Empirical Risk: $\hat{\mathcal{R}}(f) = \boxed{\frac{1}{n}\sum_{i=1}^{n}} \ell(y_i, f(x_i))$, where $\{(x_i, y_i)\}_{i=1}^{n}$ is a collected dataset

Conditional Risk

$$\mathscr{R}(f) = \mathbb{E}_{x' \sim p}\left[\mathbb{E}\left[\ell(y, f(x'))\,\middle|\, x = x'\right]\right] = \int_{\mathscr{X}} \underbrace{\mathbb{E}\left[\ell(y, f(x'))\,\middle|\, x = x'\right]}_{\text{Conditional Risk: } r(z \mid x') = \mathbb{E}\left[\ell(y,z)\,\middle|\, x = x'\right]} dp(x')\,.$$

* means the best

- Bayes Predictor: $f^{\boxed{*}}(x') \in \arg\min_{z \in \mathscr{Y}} \mathbb{E}\left[\ell(y, z)\,\middle|\, x = x'\right] = \arg\min_{z \in \mathscr{Y}} r(z \mid x')\,.$

Conditional Risk

$$\mathscr{R}(f) = \mathbb{E}_{x' \sim p}\left[\mathbb{E}\left[\ell(y, f(x'))\,\middle|\, x = x'\right]\right] = \int_{\mathscr{X}} \underbrace{\mathbb{E}\left[\ell(y, f(x'))\,\middle|\, x = x'\right]}_{\text{Conditional Risk: } r(z \mid x') = \mathbb{E}\left[\ell(y,z)\,\middle|\, x = x'\right]} dp(x')\,.$$

- Bayes Predictor: $f^*(x') \in \arg\min_{z \in \mathscr{Y}} \mathbb{E}\left[\ell(y, z)\,\middle|\, x = x'\right] = \arg\min_{z \in \mathscr{Y}} r(z \mid x')\,.$

What is the Bayes Predictor of ℓ_2 loss or ℓ_1 loss?

How to design a loss function

- Method 1: Know what is your Bayes Predictor! Homework 1 Question 1.

How to design a loss function

- Method 1: Know what is your Bayes Predictor! Homework 1 Question 1.
- Method 2: Use Max likelihood
  - Step 1: understand what is your $p(y \mid x)$, e.g. Gaussian, heavy tail distribution
  - Step 2: What is the log-likelihood of dataset $\{(x_i, y_i)\}_{i=1}^n$?

How to design a loss function

- Method 1: Know what is your Bayes Predictor! [Homework 1 Question 1.]
- Method 2: Use Max likelihood
  - Step 1: understand what is your $p(y \mid x)$, e.g. Gaussian, heavy tail distribution
  - Step 2: What is the log-likelihood of dataset $\{(x_i, y_i)\}_{i=1}^n$?
    - $\log \Pi_{i=1}^n p(y_i \mid x_i) = \sum_{i=1}^{n} \log p(y_i \mid x_i)$
  - Step 3: use $\log p(\cdot \mid x_i)$ as your loss function!

How can I get the ℓ_2 loss using this methods?

Example: Logistic Regression

Consider a binary classification with $p(y_i = 1 \mid \mathbf{x}_i, \theta) = \sigma(\mathbf{x}_i^\top \theta) = \dfrac{1}{1 + e^{-\mathbf{x}_i^\top \theta}}$

Example: Gaussian with Learned Variance

Not Required

Example (*Gaussian with Learned Variance Leads to Sparsity*)

$$\begin{aligned}
\ell(\mu, \sigma^2) &= \sum_{i=1}^{n} \log P(y_i | \mu(x_i), \sigma(x_i)^2) \\
&= \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma(x_i)^2) - \frac{(y_i - \mu(x_i))^2}{2\sigma(x_i)^2} \right) \\
&= -\frac{n}{2}\ln(2\pi(x_i)) - \underbrace{\frac{n}{2}\ln(\sigma(x_i)^2)}_{\text{sparse regularization}} - \underbrace{\sum_{i=1}^{n} \frac{(y_i - \mu(x_i))^2}{2\sigma(x_i)^2}}_{\text{weighted } \ell_2 \text{ loss}}
\end{aligned}$$

Empirical Risk Minimization

I want an estimator to minimize the risk, but I can only get the empirial risk? What's the best thing I can do?

- Consider a parameterized family of prediction functions (often referred to as models) $f_\theta : \mathscr{X} \to \mathscr{Y}$, e.g.
  - Linear prediction
  - Neural Network

^ means empirical

- Empirical Risk Minimization: $\hat{\theta} \in \hat{\mathscr{R}}(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))\,.$

The only theorem: Risk Decompoisition

$$\mathscr{R}(f_{\hat{\theta}}) - \mathscr{R}^* = \underbrace{\left\{ \mathscr{R}(f_{\hat{\theta}}) - \inf_{\theta' \in \Theta} \mathscr{R}(f_{\theta'}) \right\}}_{\text{Estimation error}} + \underbrace{\left\{ \inf_{\theta' \in \Theta} \mathscr{R}(f_{\theta'}) - \mathscr{R}^* \right\}}_{\text{Approximation error}}$$

The only theorem: Risk Decompoisition

$$\mathscr{R}(f_{\hat{\theta}}) - \mathscr{R}^* = \underbrace{\left\{ \mathscr{R}(f_{\hat{\theta}}) - \inf_{\theta' \in \Theta} \mathscr{R}(f_{\theta'}) \right\}}_{\text{Estimation error}} + \underbrace{\left\{ \inf_{\theta' \in \Theta} \mathscr{R}(f_{\theta'}) - \mathscr{R}^* \right\}}_{\text{Approximation error}}$$

For an ERM Estimator:

$$\underbrace{\mathscr{R}(f_{\hat{\theta}}) - \hat{R}(f_{\hat{\theta}})}_{\text{Generalization error}} + \underbrace{\hat{R}(f_{\hat{\theta}}) - \inf_{\theta' \in \Theta} \hat{\mathscr{R}}(f_{\theta'})}_{\text{Optimization error}} + \underbrace{\inf_{\theta' \in \Theta} \hat{\mathscr{R}}(f_{\theta'}) - \inf_{\theta' \in \Theta} \mathscr{R}(f_{\theta'})}_{\text{Generalization error}}$$

The only theorem: Risk Decompoisition

$$\mathscr{R}(f_{\hat{\theta}}) - \mathscr{R}^* = \underbrace{\left\{ \mathscr{R}(f_{\hat{\theta}}) - \inf_{\theta' \in \Theta} \mathscr{R}(f_{\theta'}) \right\}}_{\text{Estimation error}} + \underbrace{\left\{ \inf_{\theta' \in \Theta} \mathscr{R}(f_{\theta'}) - \mathscr{R}^* \right\}}_{\text{Approximation error}}$$

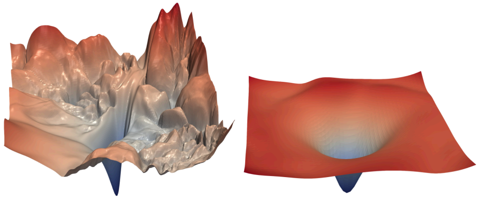
For an ERM Estimator: $\parallel$

$$\boxed{\underbrace{\mathscr{R}(f_{\hat{\theta}}) - \hat{R}(f_{\hat{\theta}})}_{\text{Generalization error}}} + \underbrace{\hat{R}(f_{\hat{\theta}}) - \inf_{\theta' \in \Theta} \hat{\mathscr{R}}(f_{\theta'})}_{\text{Optimization error}} + \boxed{\inf_{\theta' \in \Theta} \hat{\mathscr{R}}(f_{\theta'}) - \underbrace{\inf_{\theta' \in \Theta} \mathscr{R}(f_{\theta'})}_{\text{Generalization error}}}$$

$$\leq 2 \sup_{\theta \in \Theta} | R(f_\theta) - \hat{R}(f_{\hat{\theta}}) |$$

Uniform Bound!

Pro and Con of ERM



(a) without skip connections (b) with skip connections

Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.

- Pro:
  - Flexible
  - Algorithms are available (e.g. SGD)
- Con:
  - can be relatively hard to optimize when the optimization formulation is not convex (e.g.,neural networks);
  - the dependence on parameters can be complex (e.g., neural networks);
  - need some capacity control to avoid overfitting

Our course is about overfitting!

Difference between 401 and 402

Statistics Learning

- Difference 1: Parameter Convergence and Risk Convergence
- Difference 2: Parametric and Non-parametric

You use a parameterized family in Empirical risk minimization, why you call “non-parametric”?

Hardness of ERM

Error of ERM

IEMS 402 Focus

Assume to be 0

Approximation Error + Generalization Error + Optimization Error

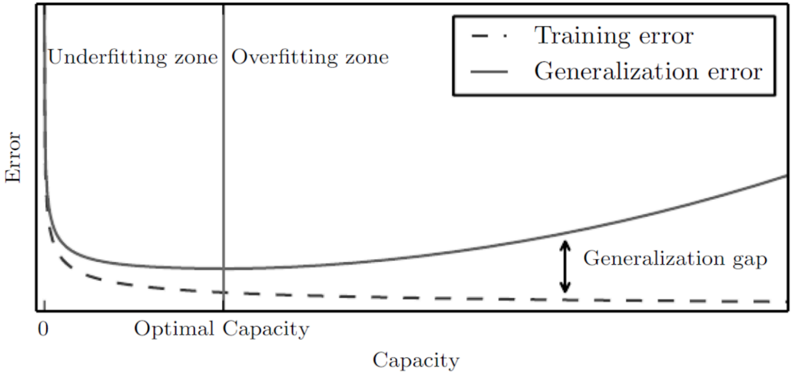
$$\inf_{\theta' \in \Theta} \mathscr{R}(f_{\theta'}) - R^*$$

$$\sup_{\theta \in \Theta} |R(f_\theta) - \hat{R}(f_{\hat{\theta}})|$$

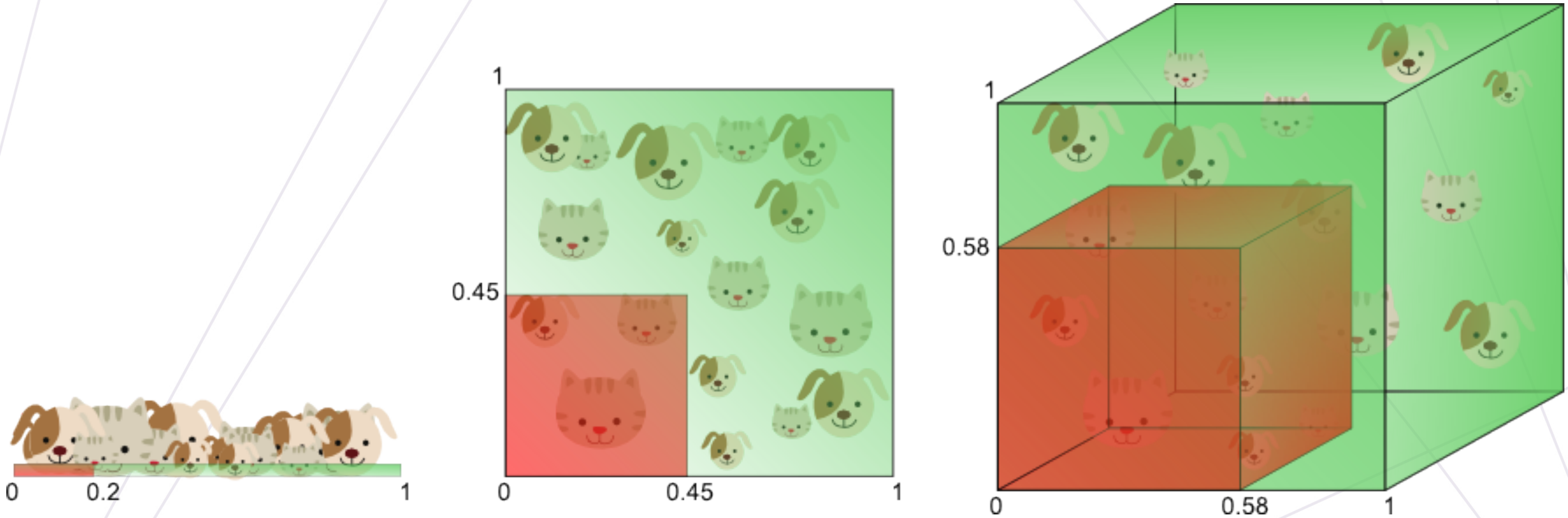
When we use more powerful parameterized family, e.g. Θ is larger:

- Approximation error is smaller!
- Generalization error is larger!

Bias-Variance Trade-off



Approximation: Curse of Dimensionality



Formulation: Approximate a smooth function

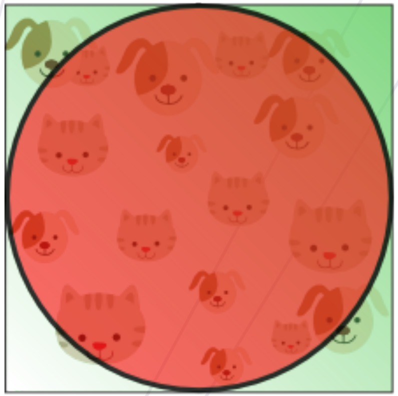
Fact. The number of parameters N required to achieve an approximation error of at most ϵ can be estimated by:

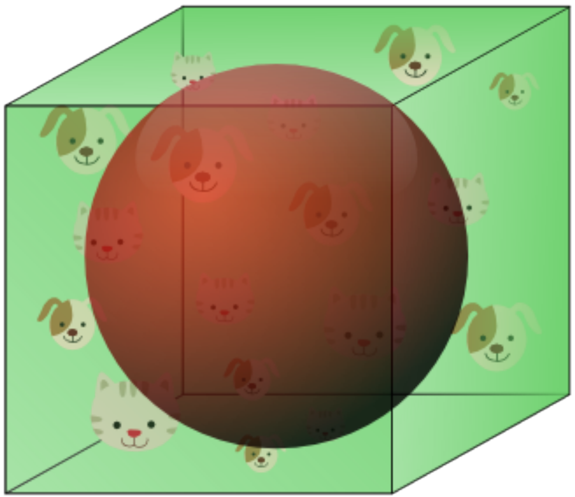
$$N \approx \left(\frac{1}{\epsilon}\right)^{\boxed{d} / \boxed{s}}$$

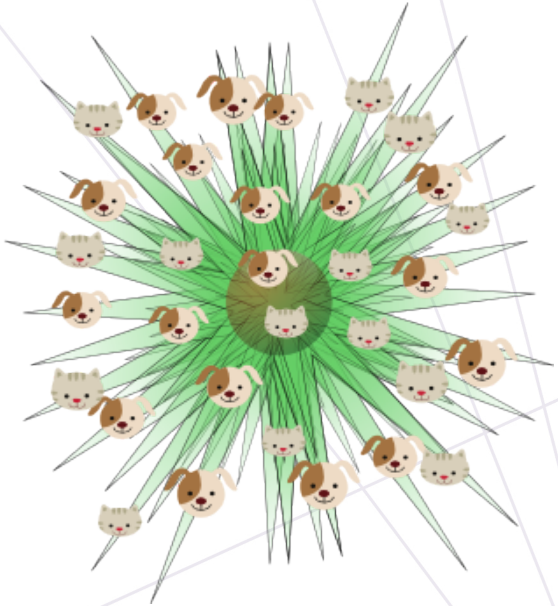
Dimension
smoothness

- Another Formulation see Homework 1 Question 3.

How to think about High Dimension







Generalization: Overfitting?

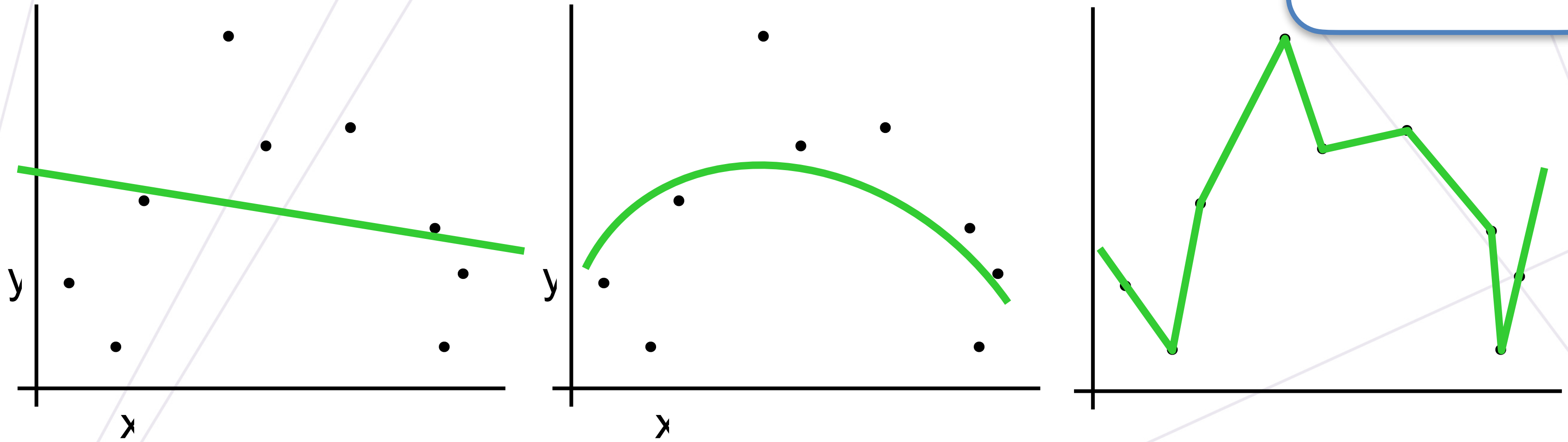
y = f(x) + noise

Can we learn f from this data?

Repeated Parrot

vs

understanding



Degree of Freedom

Suppose that we observe $y_i = r(x_i) + \epsilon_i (i = 1, \ldots, n)$, where the errors ϵ_i are uncorrelated with common variance $\sigma^2 > 0$

Now consider the fitted values $\hat{y}_i = \hat{r}(x_i)$ from a regression estimator $\hat{r}$.

Degree of freedom is defined as $\boxed{\mathrm{df}(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(\hat{y}_i, y_i)\,.}$

"How much I remember the label"

Degree of freedom

Fact. $$\underbrace{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i' - \hat{y}_i)^2\right] - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]}_{\text{Generalization error}} = \frac{2\sigma^2}{n}\,\mathrm{df}(\hat{y})\,.$$

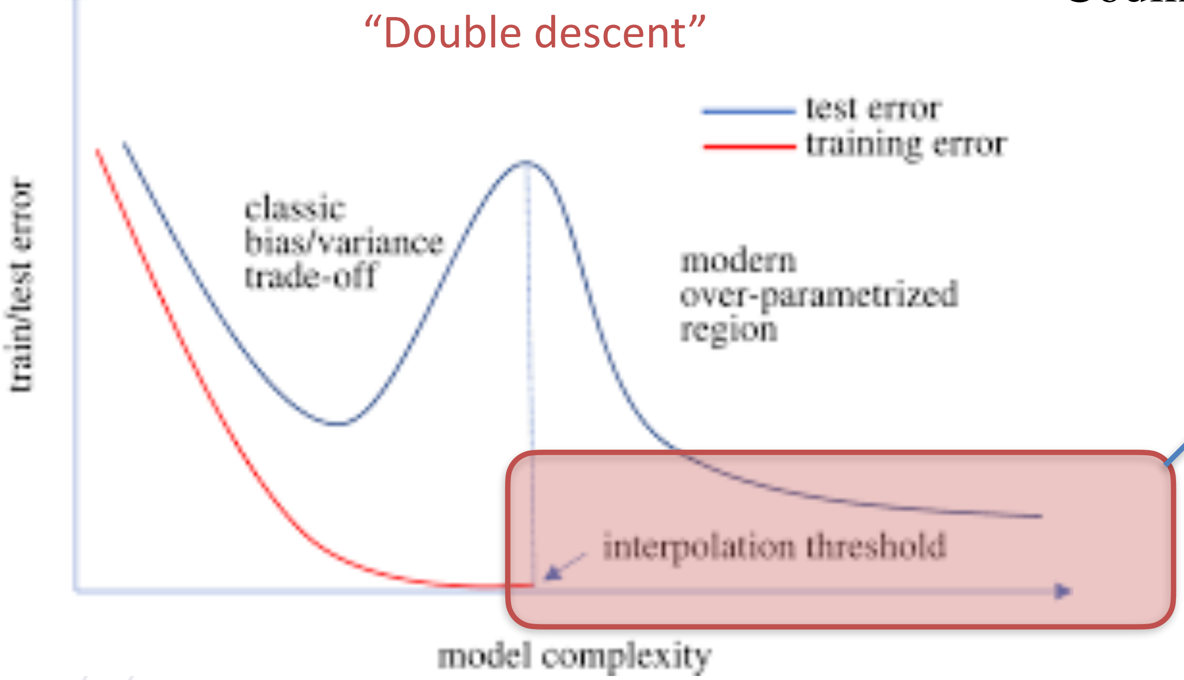
Example of DOF 1

Example of DOF 2

Not Required

However…

- Coding: [Homework 2 Question 3.]

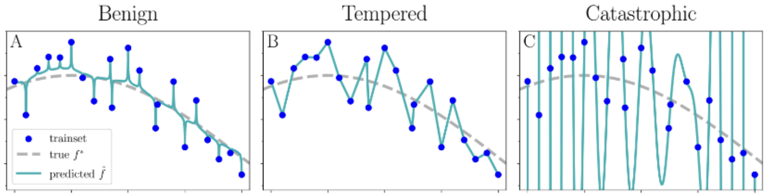


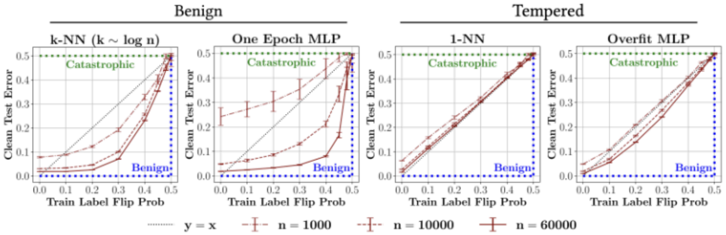
All the data can be remembered
#parameter > #data

Taxonomy of (over)fitting

	Regression	Classification
Benign	$\lim_{n\to\infty} \mathcal{R}_n = R^*$	$\lim_{n\to\infty} \mathcal{R}_n = R^*$
Tempered	$\lim_{n\to\infty} \mathcal{R}_n \in (R^*, \infty)$	$\lim_{n\to\infty} \mathcal{R}_n \in (R^*, 1 - \frac{1}{K})$
Catastrophic	$\lim_{n\to\infty} \mathcal{R}_n = \infty$	$\lim_{n\to\infty} \mathcal{R}_n = 1 - \frac{1}{K}$

Table -1.1: Taxonomy of (over)fitting.

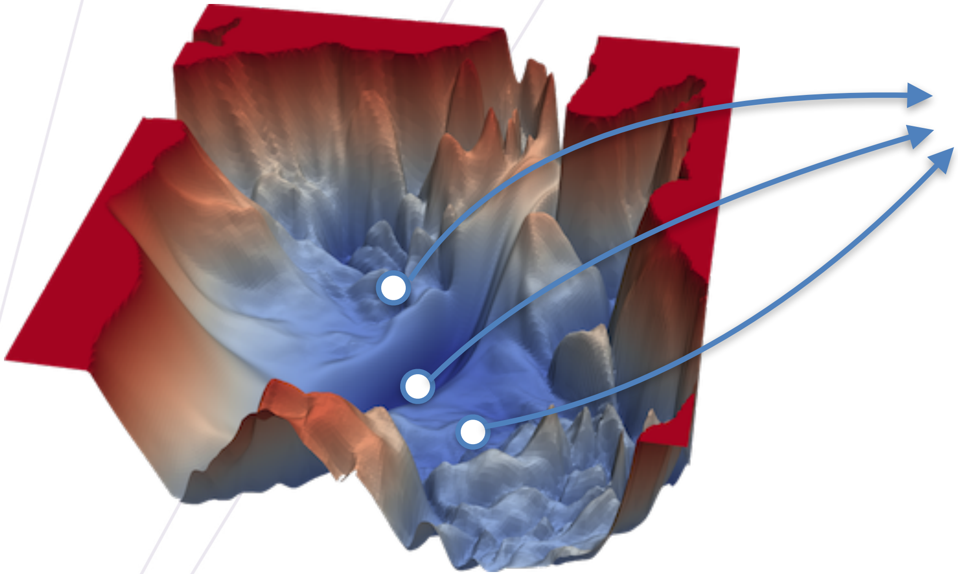




Mallinar, Neil, et al. "Benign, tempered, or catastrophic: A taxonomy of overfitting (2022)." arXiv preprint arXiv:2207.06569.

Implicit bias

"Multiple Minima"

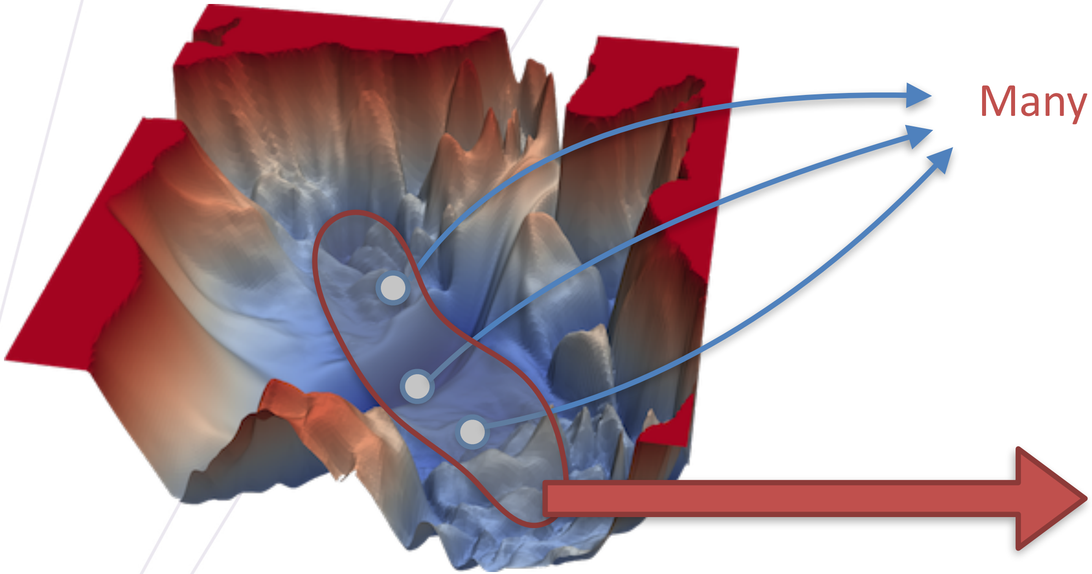


Many models can achieve low training loss

Loss landscape of VGG on CIFAR

Implicit bias

"Multiple Minima"



Many models can achieve low training loss

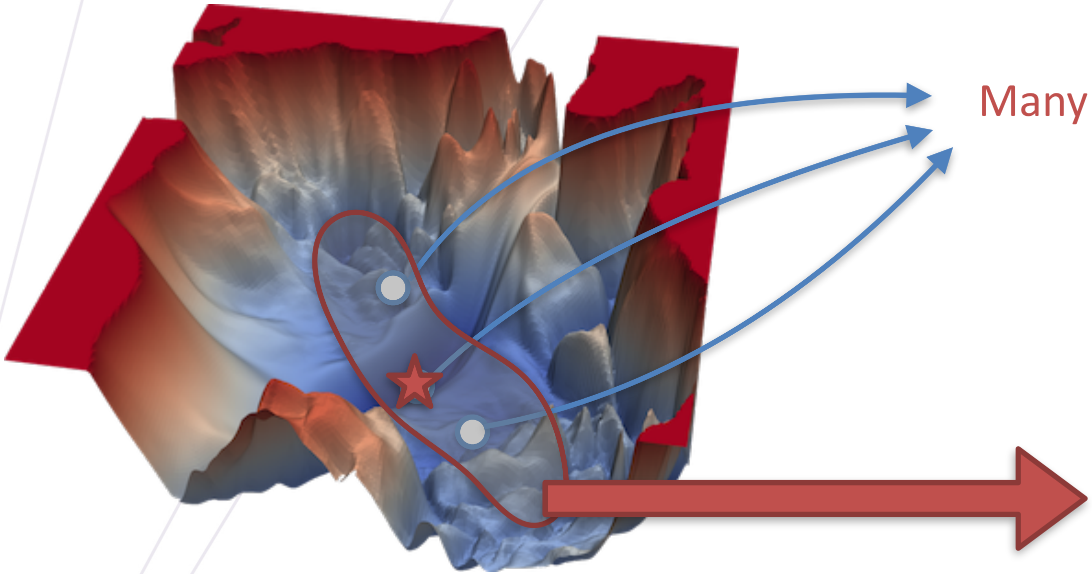
Traditional bounds:

$$\sup_{\theta \in \Theta} |R(f_\theta) - \hat{R}(f_{\hat{\theta}})|$$

Loss landscape of VGG on CIFAR

Implicit bias

"Multiple Minima"



Many models can achieve low training loss

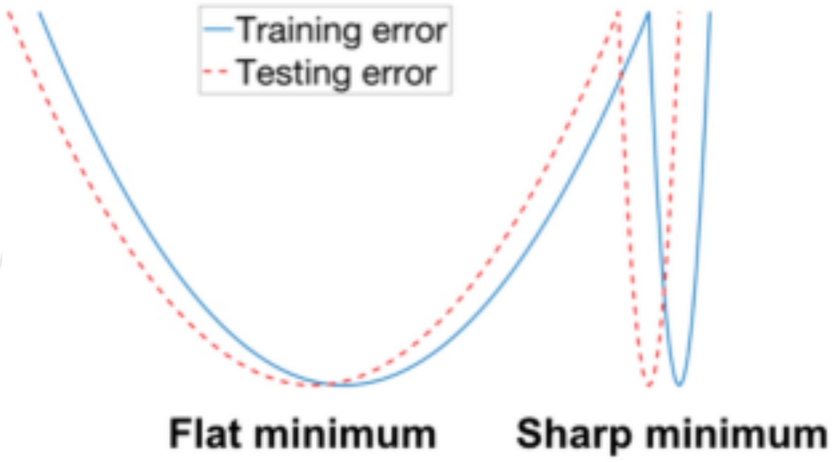
Traditional bounds:

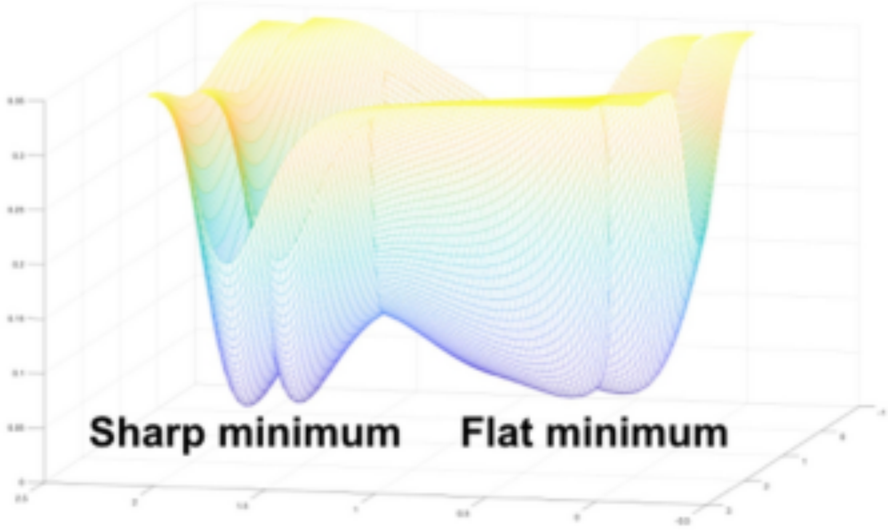
$$\sup_{\theta \in \Theta} | R(f_\theta) - \hat{R}(f_{\hat{\theta}}) |$$

Loss landscape of VGG on CIFAR

What’s special about over-para

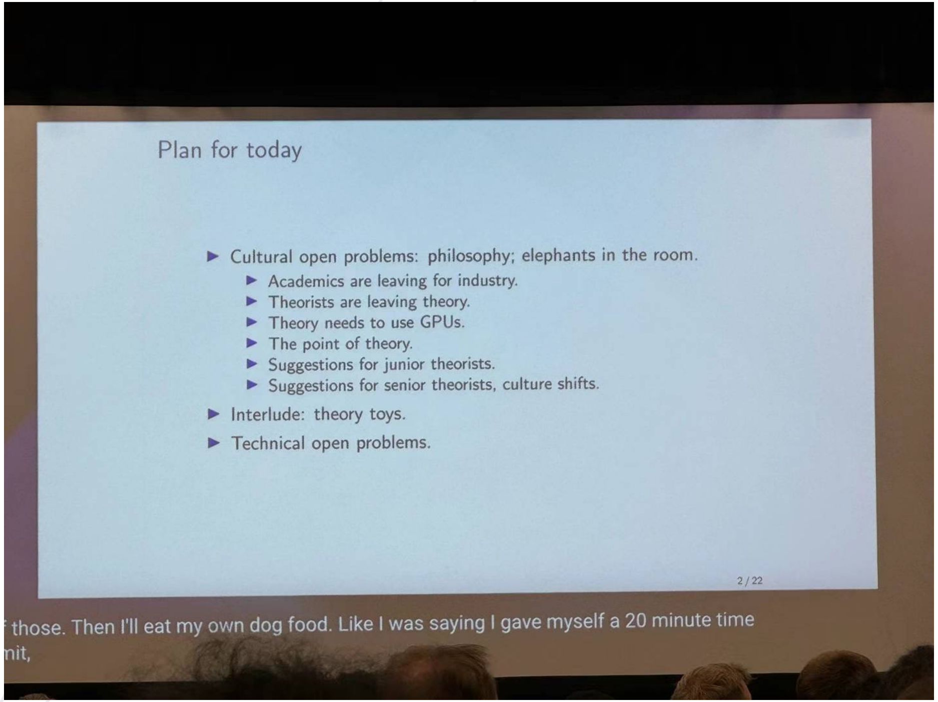
<u>“Multiple Minima”</u>





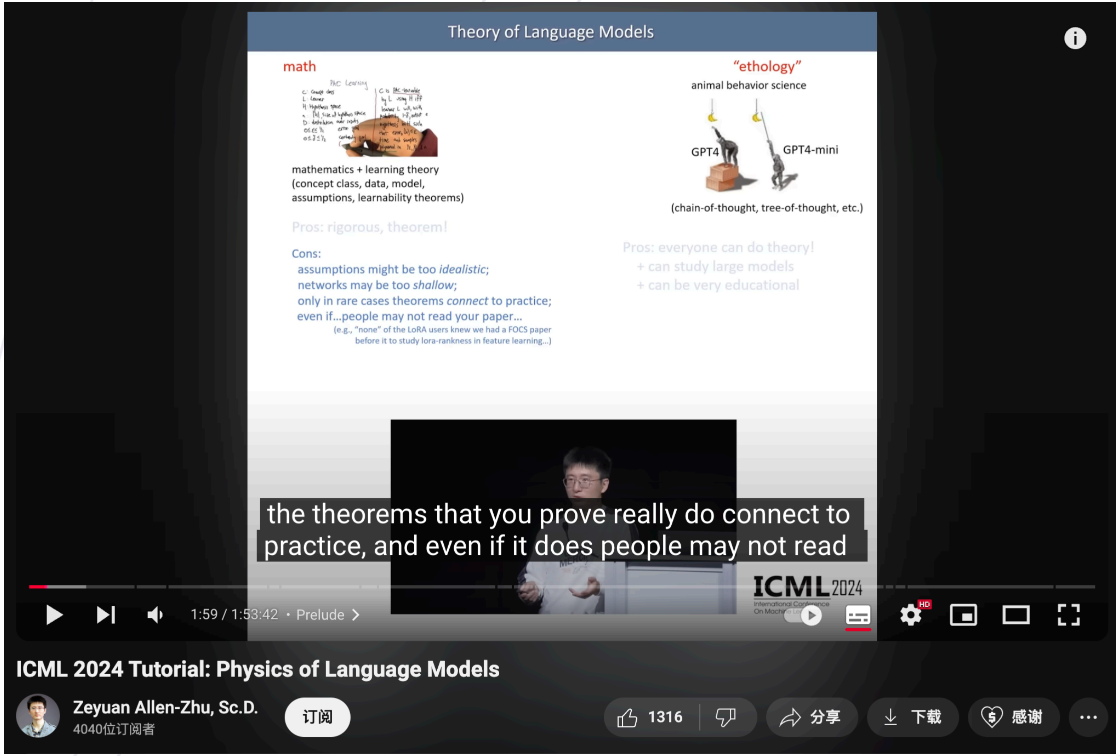
Last Note on Learning Theory

ML Theory workshop @Neurips24



https://cims.nyu.edu/~matus/neurips.2024.workshop/talk.pdf

Math-physics-ethology



Physics of language model
ICML 2024
https://shorturl.at/ZDwQE