Towards Modern Datasets: laying mathematical foundations to streamline machine learning

Chen Cheng

Department of Statistics, Stanford University

Stanford University

Evolution of datasets

- **Datasets are central** to the success of **statistical machine learning**

Evolution of datasets

- **Datasets are central** to the success of **statistical machine learning**

David Donoho, "50 years of Data Science"

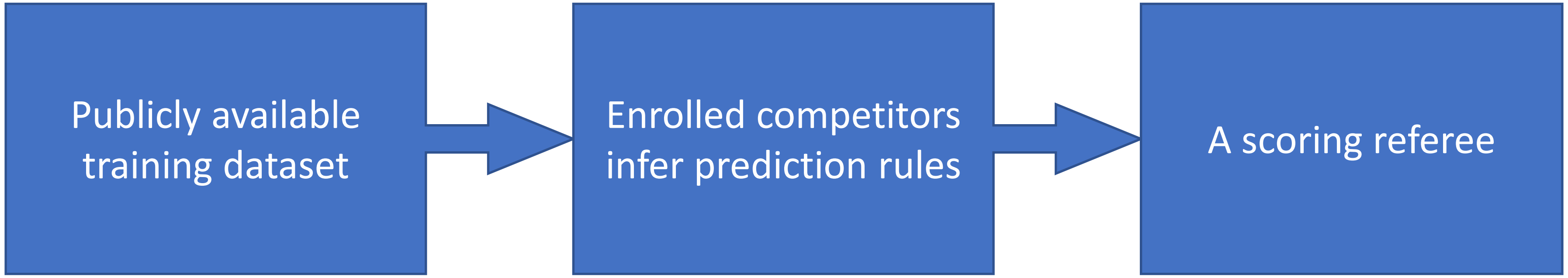
*"…those fields where machine learning has scored successes are essentially those fields where CTF (**common task framework**) has been applied systematically."*

Evolution of datasets

- **Datasets are central** to the success of **statistical machine learning**

David Donoho, “50 years of Data Science”

*“…those fields where machine learning has scored successes are essentially those fields where CTF (**common task framework**) has been applied systematically.”*

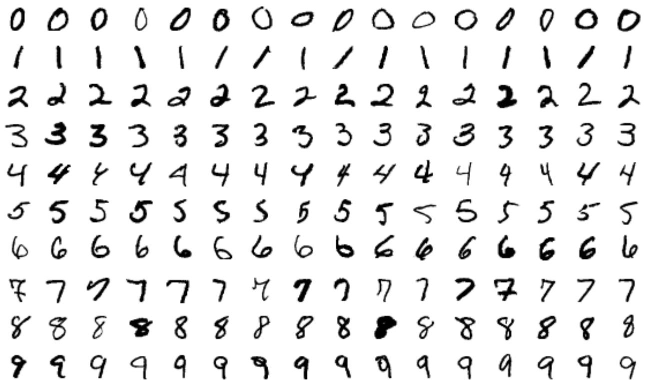


Evolution of datasets

- **Datasets are central** to the success of **statistical machine learning**
- Vast growth of modern datasets leads to the success of modern ML

Evolution of datasets

- **Datasets are central** to the success of **statistical machine learning**
- Vast growth of modern datasets leads to the success of modern ML



MNIST [LeCun et al. 94]

N~10^5, d~10^3

ImageNet [Deng et al. 09]

N~10^8, d~10^7

Q: An armchair that looks like an apple

C: Green Apple Chair

LAION-5B [Schuhmann et al. 22]

N~10^10, d~10^7*(text)

Neural network → Deep learning → Generative AI

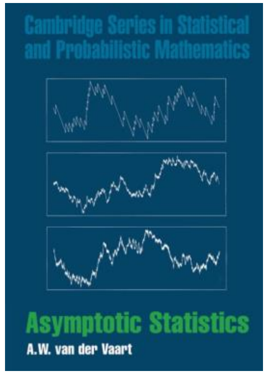
Evolution of datasets

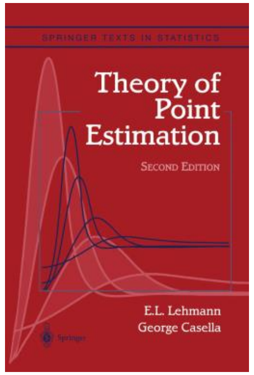
- **Datasets are central** to the success of **statistical machine learning**
- Vast growth of modern datasets leads to the success of modern ML
- **Breakdown** of standard statistical assumptions

Evolution of datasets

- **Datasets are central** to the success of **statistical machine learning**
- Vast growth of modern datasets leads to the success of modern ML
- **Breakdown** of standard statistical assumptions

$$\{(X_i, y_i)\}_{i=1}^n, X_i \in \mathbb{R}^d, y_i \in \mathbb{R}$$



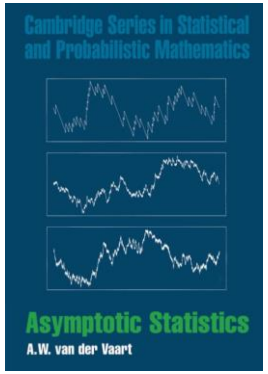


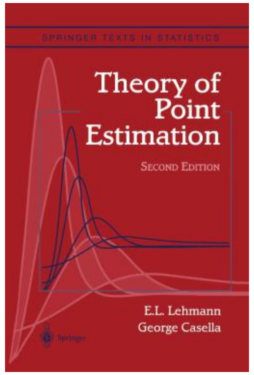
(Classical) textbook fairyland

Evolution of datasets

- **Datasets are central** to the success of **statistical machine learning**
- Vast growth of modern datasets leads to the success of modern ML
- **Breakdown** of standard statistical assumptions

$$\{(X_i, y_i)\}_{i=1}^n, X_i \in \mathbb{R}^d, y_i \in \mathbb{R}$$





Husky

High dimensionality?

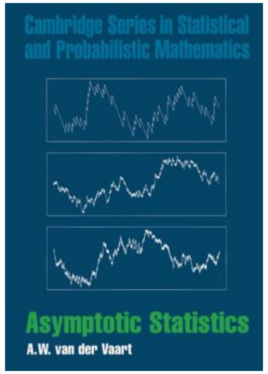
(Classical) textbook fairyland

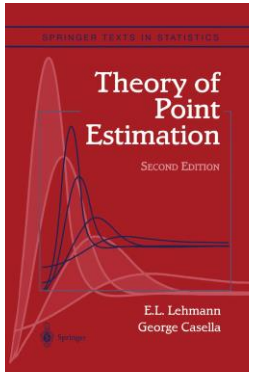
Complexity of modern datasets

Evolution of datasets

- **Datasets are central** to the success of **statistical machine learning**
- Vast growth of modern datasets leads to the success of modern ML
- **Breakdown** of standard statistical assumptions

$$\{(X_i, y_i)\}_{i=1}^n, X_i \in \mathbb{R}^d, y_i \in \mathbb{R}$$





(Classical) textbook fairyland

Husky

High dimensionality?

Husky

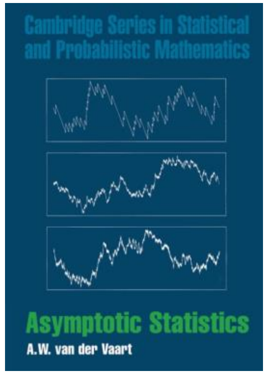
Missing data?

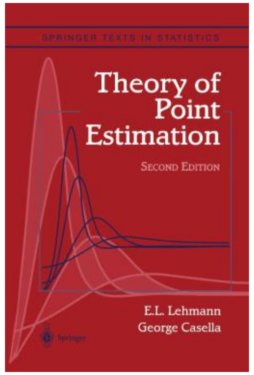
Complexity of modern datasets

Evolution of datasets

- **Datasets are central** to the success of **statistical machine learning**
- Vast growth of modern datasets leads to the success of modern ML
- **Breakdown** of standard statistical assumptions

$$\{(X_i, y_i)\}_{i=1}^n, X_i \in \mathbb{R}^d, y_i \in \mathbb{R}$$





(Classical) textbook fairyland

Husky

High dimensionality?

Husky

Missing data?

Husky
Samoyed
Snow

Multilabeling?

Complexity of modern datasets

Therefore the task

Unusual properties of modern datasets

call for

Novel statistical and mathematical foundations

that we can leverage to

Streamline machine learning

Therefore the task

Unusual properties of modern datasets

call for

Novel statistical and mathematical foundations

that we can leverage to

Streamline machine learning

- **Two stories today:**
  - Multilabeled dataset
  - Infinite dimensional regression

Story one: multilabeled dataset

- **ImageNet construction** [Deng et al. 09]

🥹 🥹 🥹 🥹 🥹

Mechanical Turk workers

Data cleaning and label aggregation

Story one: multilabeled dataset

- **ImageNet construction** [Deng et al. 09]

Mechanical Turk workers

$$\{(X_i, (y_{i1}, y_{i2}, \cdots, y_{im}))\}_{i=1}^n$$

$$\{(X_i, y_i)\}_{i=1}^n$$

Data cleaning and label aggregation

Story one: multilabeled dataset

- **ImageNet construction** [Deng et al. 09]

🥺 🥺 🥺 🥺 🥺

Mechanical Turk workers

$$\{(X_i, (y_{i1}, y_{i2}, \cdots, y_{im}))\}_{i=1}^n$$

$$\{(X_i, y_i)\}_{i=1}^n$$

Data cleaning and label aggregation

🤓 🤖

Statisticians & Engineers

Story one: multilabeled dataset

- **ImageNet construction** [Deng et al. 09]
- **Question:** Is this the right thing to do? (Calibration? Efficiency? etc.)

$$\{(X_i, (y_{i1}, y_{i2}, \cdots, y_{im}))\}_{i=1}^n$$

$$\{(X_i, y_i)\}_{i=1}^n$$

Data cleaning and label aggregation

Statisticians & Engineers

Story one: multilabeled dataset

- **ImageNet construction** [Deng et al. 09]
- **Question:** Is this the right thing to do? (Calibration? Efficiency? etc.)

$$\{(X_i, (y_{i1}, y_{i2}, \cdots, y_{im}))\}_{i=1}^n$$

OUTSIDE THE BOX!

$$\{(X_i, y_i)\}_{i=1}^n$$

Data cleaning and label aggregation

Statisticians & Engineers

Story one: multilabeled dataset

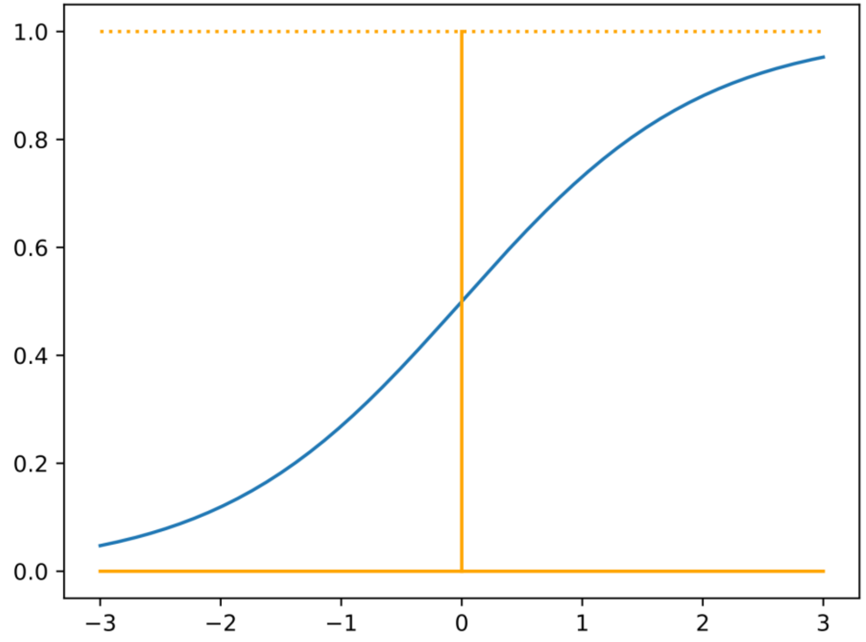
- **The model for** $\{(X_i, (y_{i1}, y_{i2}, \cdots, y_{im}))\}_{i=1}^{n}$

Story one: multilabeled dataset

- **The model for** $\{(X_i, (y_{i1}, y_{i2}, \cdots, y_{im}))\}_{i=1}^{n}$
  - Binary classification $y_{ij} \in \{\pm 1\}$

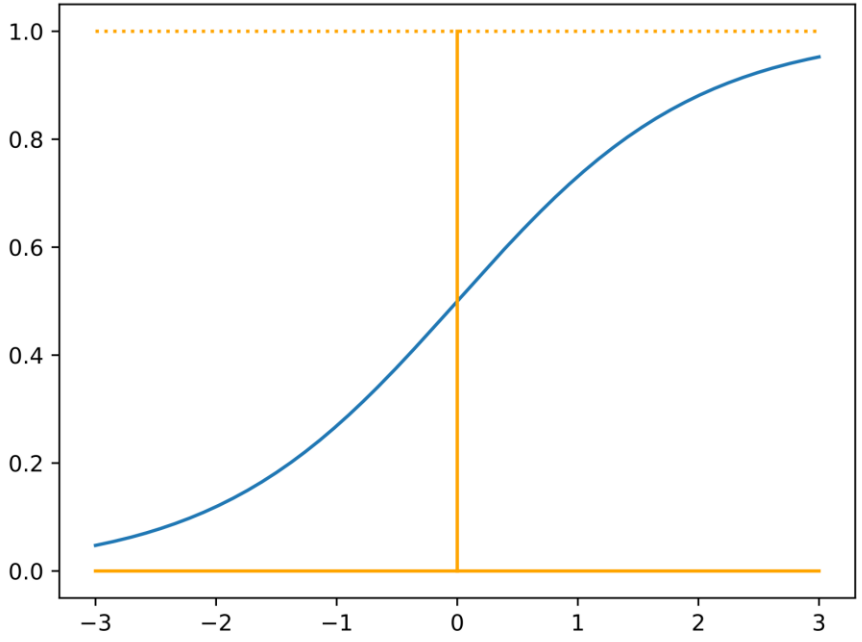
Story one: multilabeled dataset

- **The model for** $\{(X_i, (y_{i1}, y_{i2}, \cdots, y_{im}))\}_{i=1}^n$
  - Binary classification $y_{ij} \in \{\pm 1\}$
  - Single index model with symmetric link $\mathbb{P}(y_{ij} = y \mid X_i = x) = \sigma(y\langle x, \theta^\star \rangle)$



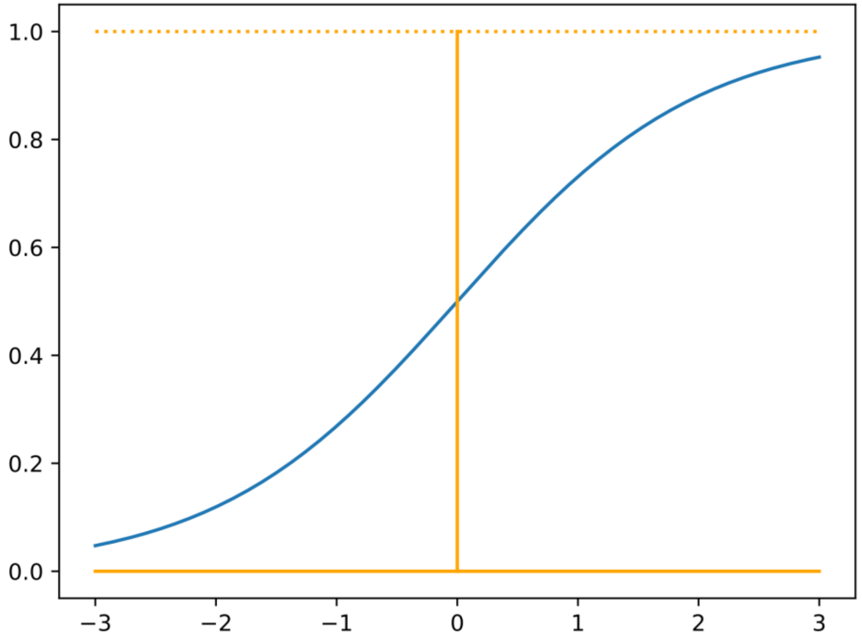
Story one: multilabeled dataset

- **The model for** $\{(X_i, (y_{i1}, y_{i2}, \cdots, y_{im}))\}_{i=1}^n$
    - Binary classification $y_{ij} \in \{\pm 1\}$
    - Single index model with symmetric link $\mathbb{P}(y_{ij} = y \mid X_i = x) = \sigma(y\langle x, \theta^\star \rangle)$
    - Isotropic covariate $X_i \overset{\text{iid}}{\sim} \mathsf{N}(0, I_d)$



Story one: multilabeled dataset

- **The model for** $\{(X_i, (y_{i1}, y_{i2}, \cdots, y_{im}))\}_{i=1}^n$
  - Binary classification $y_{ij} \in \{\pm 1\}$
  - Single index model with symmetric link $\mathbb{P}(y_{ij} = y \mid X_i = x) = \sigma(y\langle x, \theta^\star \rangle)$
  - Isotropic covariate $X_i \overset{\text{iid}}{\sim} \mathsf{N}(0, I_d)$
- Low dimension (fixed d)



Story one: multilabeled dataset

- **The learning algorithms: all labels vs. majority vote**

Story one: multilabeled dataset

- **The learning algorithms: all labels vs. majority vote**
  - Margin-based loss $\ell'_\theta(y \mid x) = -\sigma(-y\langle x, \theta\rangle)$

Story one: multilabeled dataset

- **The learning algorithms: all labels vs. majority vote**
  - Margin-based loss $\ell'_\theta(y \mid x) = -\sigma(-y\langle x, \theta \rangle)$
  - e.g. Logistic regression

$$\ell_\theta(y \mid x) = \log(1 + e^{y\langle x, \theta \rangle})$$

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

Story one: multilabeled dataset

- **The learning algorithms: all labels vs. majority vote**
  - Margin-based loss $\ell'_\theta(y \mid x) = -\sigma(-y\langle x, \theta \rangle)$
  - e.g. Logistic regression

    $$\ell_\theta(y \mid x) = \log(1 + e^{y\langle x, \theta \rangle})$$

    $$\sigma(t) = \frac{1}{1 + e^{-t}}$$

  - Two estimators

Story one: multilabeled dataset

- **The learning algorithms: all labels vs. majority vote**
  - Margin-based loss $\ell'_\theta(y \mid x) = -\sigma(-y\langle x, \theta\rangle)$
  - e.g. Logistic regression

    $$\ell_\theta(y \mid x) = \log(1 + e^{y\langle x, \theta\rangle})$$

    $$\sigma(t) = \frac{1}{1 + e^{-t}}$$
  - Two estimators

Full label information

$$\hat{\theta}_n := \arg\min \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \ell_\theta(y_{ij} \mid X_i)$$

Story one: multilabeled dataset

- **The learning algorithms: all labels vs. majority vote**
  - Margin-based loss $\ell'_\theta(y \mid x) = -\sigma(-y\langle x, \theta \rangle)$
  - e.g. Logistic regression

    $$\ell_\theta(y \mid x) = \log(1 + e^{y\langle x, \theta \rangle})$$

    $$\sigma(t) = \frac{1}{1 + e^{-t}}$$

  - Two estimators

Full label information

$$\hat{\theta}_n := \arg\min \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \ell_\theta(y_{ij} \mid X_i)$$

Majority vote aggregation

$$y_i := \mathsf{maj}(y_{i1}, \cdots, y_{im})$$

$$\hat{\theta}_n^{\text{mv}} := \arg\min \frac{1}{n} \sum_{i=1}^{n} \ell_\theta(y_i \mid X_i)$$

Story one: multilabeled dataset

- **Quantifiers of interest: calibration and classification**

$$\hat{\theta} - \theta \qquad\qquad \hat{u} - u := \hat{\theta}/\|\hat{\theta}\| - \theta/\|\theta\|$$

Story one: multilabeled dataset

- **Quantifiers of interest: calibration and classification**

$$\hat{\theta} - \theta \qquad\qquad \hat{u} - u := \hat{\theta}/\|\hat{\theta}\| - \theta/\|\theta\|$$

Theorem (Cheng, Asi, Duchi, 22) Under regularity assumptions, label aggregation yields

Story one: multilabeled dataset

- **Quantifiers of interest: calibration and classification**

$$\hat{\theta} - \theta \qquad\qquad \hat{u} - u := \hat{\theta}/\|\hat{\theta}\| - \theta/\|\theta\|$$

Theorem (Cheng, Asi, Duchi, 22) Under regularity assumptions, label aggregation yields

- (**Mis-calibration and slower rate**) For well-specified models (loss link is the true link)

$$\sqrt{n}(\hat{\theta}_n - \theta) \approx \mathsf{N}(0, m^{-1}\Sigma_{\sigma,\theta}), \qquad \sqrt{n}(\hat{u}_n - u) \approx \mathsf{N}(0, m^{-1}C_{\sigma,\theta})$$

$$\hat{\theta}_n^{\mathrm{mv}} \approx \sqrt{m}t_\sigma\theta, \qquad \sqrt{n}(\hat{u}_n^{\mathrm{mv}} - u) \approx \mathsf{N}(0, m^{-1/2}C_{\sigma,\theta}^{\mathrm{mv}})$$

Story one: multilabeled dataset

- **Quantifiers of interest: calibration and classification**

$$\hat{\theta} - \theta \qquad\qquad \hat{u} - u := \hat{\theta}/\|\hat{\theta}\| - \theta/\|\theta\|$$

Theorem (Cheng, Asi, Duchi, 22) Under regularity assumptions, label aggregation yields

- (**Mis-calibration and slower rate**) For well-specified models (loss link is the true link)

$$\sqrt{n}(\hat{\theta}_n - \theta) \approx \mathsf{N}(0, m^{-1}\Sigma_{\sigma,\theta}), \qquad \sqrt{n}(\hat{u}_n - u) \approx \mathsf{N}(0, m^{-1}C_{\sigma,\theta})$$

$$\hat{\theta}_n^{\mathrm{mv}} \approx \sqrt{m}t_\sigma\theta, \qquad \sqrt{n}(\hat{u}_n^{\mathrm{mv}} - u) \approx \mathsf{N}(0, m^{-1/2}C_{\sigma,\theta}^{\mathrm{mv}})$$

- (**Robustness**) For mis-specified models (the loss link is not the true link)

$$\hat{\theta}_n \approx t_{\sigma,\sigma^{\mathrm{loss}}}\theta, \qquad \sqrt{n}(\hat{u}_n - u) \approx \mathsf{N}(0, C_{\sigma,\sigma^{\mathrm{loss}},\theta})$$

$$\hat{\theta}_n^{\mathrm{mv}} \approx \sqrt{m}t_{\sigma,\sigma^{\mathrm{loss}}}^{\mathrm{mv}}\theta, \qquad \sqrt{n}(\hat{u}_n^{\mathrm{mv}} - u) \approx \mathsf{N}(0, m^{-1/2}C_{\sigma,\sigma^{\mathrm{loss}},\theta}^{\mathrm{mv}})$$

Story one: multilabeled dataset

- **Quantifiers of interest: calibration and classification**

$$\hat{\theta} - \theta \qquad \hat{u} - u := \hat{\theta}/\|\hat{\theta}\| - \theta/\|\theta\|$$

Theorem (Cheng, Asi, Duchi, 22) Under regularity assumptions, label aggregation yields

- (**Mis-calibration and slower rate**) For well-specified models (loss link is the true link)

$$\sqrt{n}(\hat{\theta}_n - \theta) \approx \mathsf{N}(0, m^{-1}\Sigma_{\sigma,\theta}), \qquad \sqrt{n}(\hat{u}_n - u) \approx \mathsf{N}(0, m^{-1}C_{\sigma,\theta})$$

$$\hat{\theta}_n^{\mathrm{mv}} \approx \sqrt{m} t_\sigma \theta, \qquad \sqrt{n}(\hat{u}_n^{\mathrm{mv}} - u) \approx \mathsf{N}(0, m^{-1/2}C_{\sigma,\theta}^{\mathrm{mv}})$$

- (**Robustness**) For mis-specified models (the loss link is not the true link)

$$\hat{\theta}_n \approx t_{\sigma,\sigma^{\mathrm{loss}}}\theta, \qquad \sqrt{n}(\hat{u}_n - u) \approx \mathsf{N}(0, C_{\sigma,\sigma^{\mathrm{loss}},\theta})$$

$$\hat{\theta}_n^{\mathrm{mv}} \approx \sqrt{m} t_{\sigma,\sigma^{\mathrm{loss}}}^{\mathrm{mv}} \theta, \qquad \sqrt{n}(\hat{u}_n^{\mathrm{mv}} - u) \approx \mathsf{N}(0, m^{-1/2}C_{\sigma,\sigma^{\mathrm{loss}},\theta}^{\mathrm{mv}})$$

- (**Lower bound**) For logistic models, the Fisher information matrix for the majority vote $\{(X_i, y_i)\}_{i=1}^n$

$$\mathsf{I}^{\mathrm{mv}}(\theta) \approx \frac{a}{\|\theta\|^3\sqrt{m}} uu^\top + \frac{b\sqrt{m}}{\|\theta\|}(I - uu^\top)$$

Story one: multilabeled dataset

- **Quantifiers of interest: calibration and classification**

$$\hat{\theta} - \theta \qquad\qquad \hat{u} - u := \hat{\theta}/\|\hat{\theta}\| - \theta/\|\theta\|$$

Theorem (Cheng, Asi, Duchi, 22) Under regularity assumptions, label aggregation yields

- (**Mis-calibration and slower rate**) For well-specified models (loss link is the true link)

$$\sqrt{n}(\hat{\theta}_n - \theta) \approx \mathsf{N}(0, m^{-1}\Sigma_{\sigma,\theta}), \qquad \sqrt{n}(\hat{u}_n - u) \approx \mathsf{N}(0, m^{-1}C_{\sigma,\theta})$$

$$\hat{\theta}_n^{\mathrm{mv}} \approx \sqrt{m}t_\sigma\theta, \qquad \sqrt{n}(\hat{u}_n^{\mathrm{mv}} - u) \approx \mathsf{N}(0, m^{-1/2}C_{\sigma,\theta}^{\mathrm{mv}})$$

- (**Robustness**) For mis-specified models (the loss link is not the true link)

$$\hat{\theta}_n \approx t_{\sigma,\sigma^{\mathrm{loss}}}\theta, \qquad \sqrt{n}(\hat{u}_n - u) \approx \mathsf{N}(0, C_{\sigma,\sigma^{\mathrm{loss}},\theta})$$

$$\hat{\theta}_n^{\mathrm{mv}} \approx \sqrt{m}t_{\sigma,\sigma^{\mathrm{loss}}}^{\mathrm{mv}}\theta, \qquad \sqrt{n}(\hat{u}_n^{\mathrm{mv}} - u) \approx \mathsf{N}(0, m^{-1/2}C_{\sigma,\sigma^{\mathrm{loss}},\theta}^{\mathrm{mv}})$$

- (**Lower bound**) For logistic models, the Fisher information matrix for the majority vote $\{(X_i, y_i)\}_{i=1}^n$

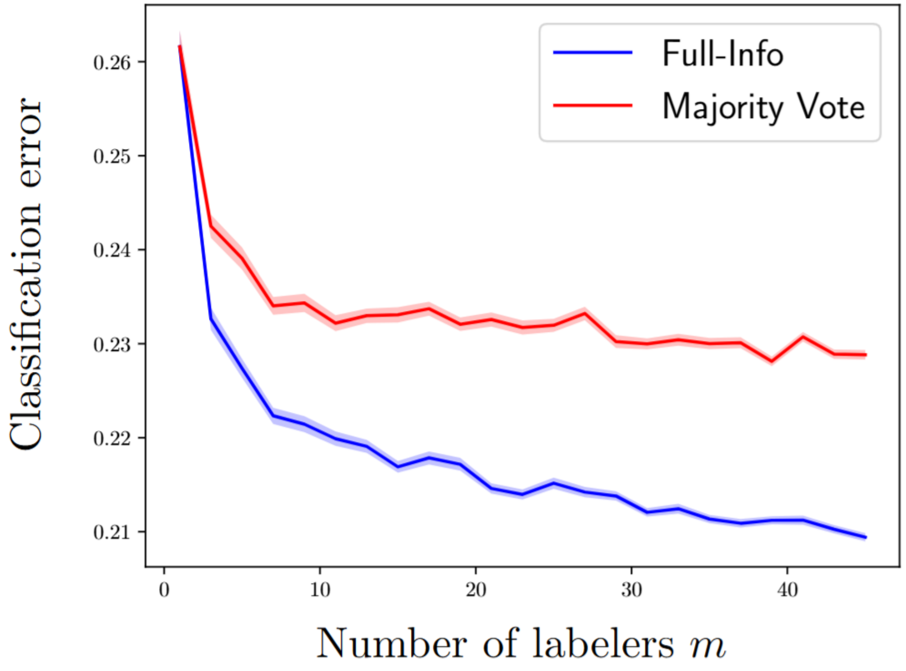
$$\mathsf{I}^{\mathrm{mv}}(\theta) \approx \frac{a}{\|\theta\|^3\sqrt{m}}uu^\top + \frac{b\sqrt{m}}{\|\theta\|}(I - uu^\top) \longleftrightarrow \text{Mis-calibration term + Slower rate term}$$

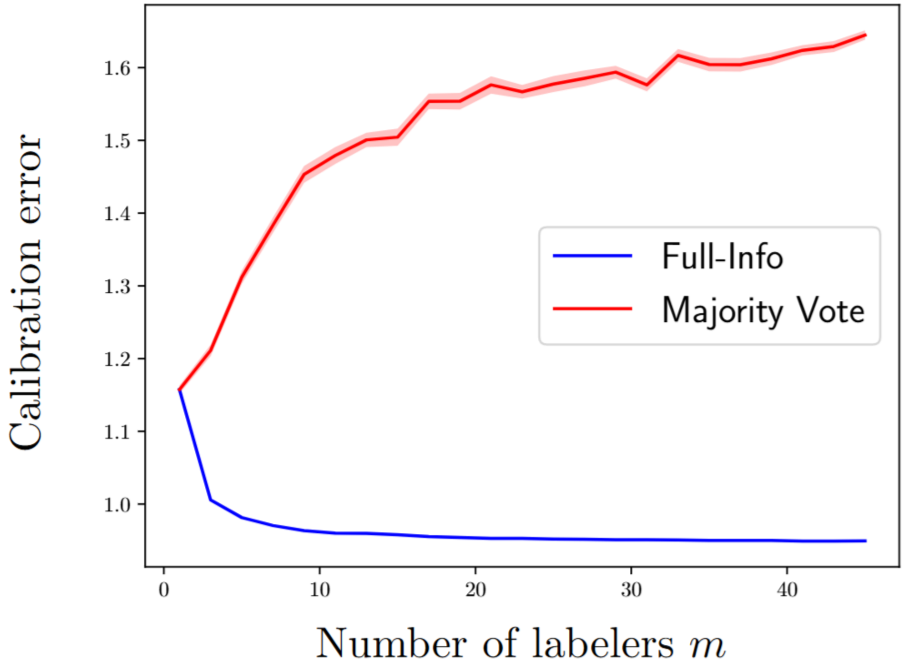
Story one: multilabeled dataset

- **Conclusion**

Story one: multilabeled dataset

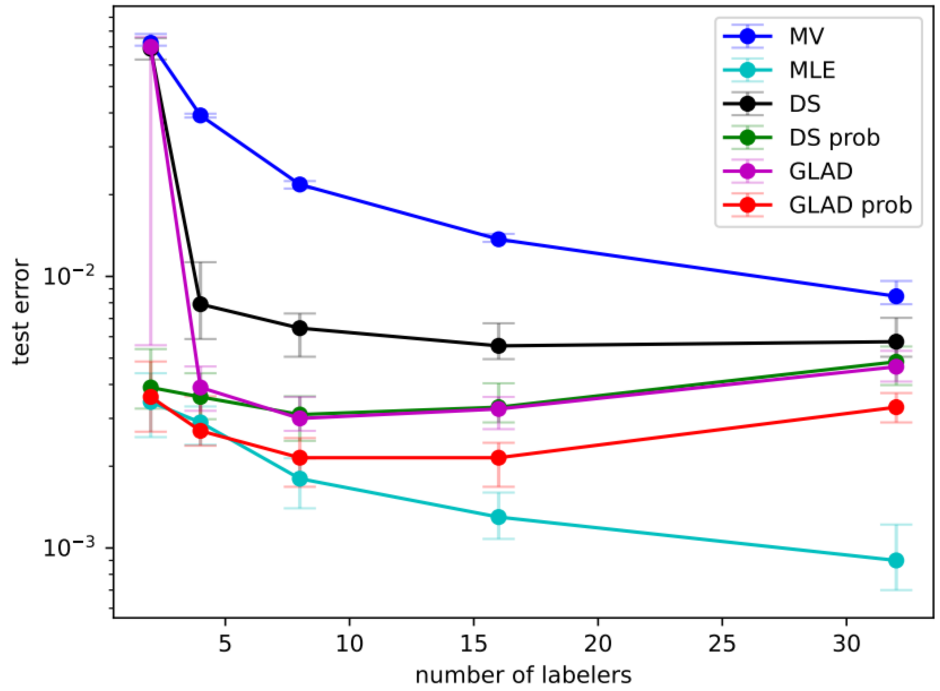
- **Conclusion**
  - Majority vote is mis-calibrated and overconfident, less efficient but more robust when mis-specified.





Story one: multilabeled dataset

- **Conclusion**
  - Majority vote is mis-calibrated and overconfident, less efficient but more robust when mis-specified.
  - Soft labels can be more beneficial (experiments for generative model-based crowdsourcing approaches).



Story one: multilabeled dataset

- **Conclusion**
  - Majority vote is mis-calibrated and overconfident, less efficient but more robust when mis-specified.
  - Soft labels can be more beneficial (experiments for generative model-based crowdsourcing approaches).
  - Semiparametric approaches.

Theorem (Cheng, Asi, Duchi, 22) Within an appropriate link function class $\sigma \in \mathcal{F}^{\text{link}}$, the two stage semiparametric estimator achieves optimal rate for classification

$$\hat{\sigma}_n := \arg\min \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} (\sigma(\langle \hat{u}_n^{\text{mv}}, X_i \rangle) - y_{ij})^2, \qquad \ell_\theta^{\text{sp}\prime}(t) := -\hat{\sigma}_n(-t)$$

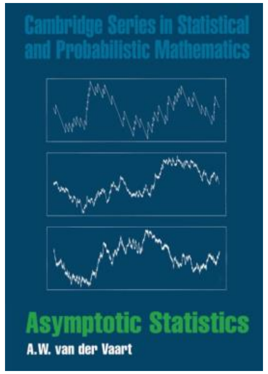
$$\hat{u}_n^{\text{sp}} := \arg\min \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \ell_\theta^{\text{sp}}(y_{ij} \mid X_i)$$

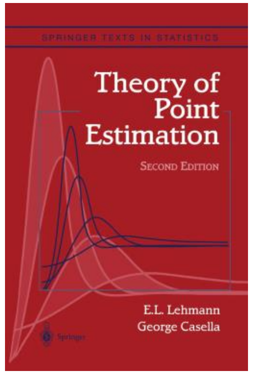
Story two: infinite dimensional regression

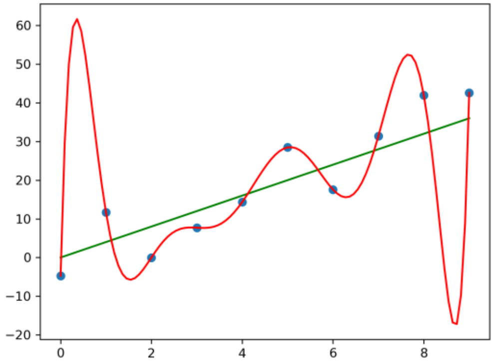
- **Paradoxical behavior of “overfitting” vs. “benign overfitting”**

Story two: infinite dimensional regression

- **Paradoxical behavior of “overfitting” vs. “benign overfitting”**





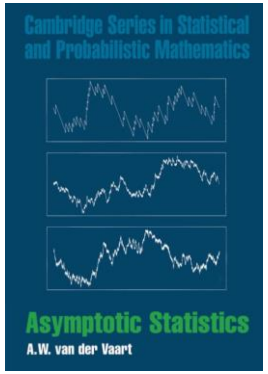


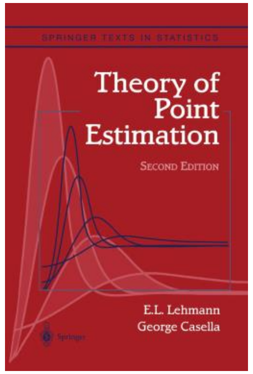
The classical “overfitting” phenomenon

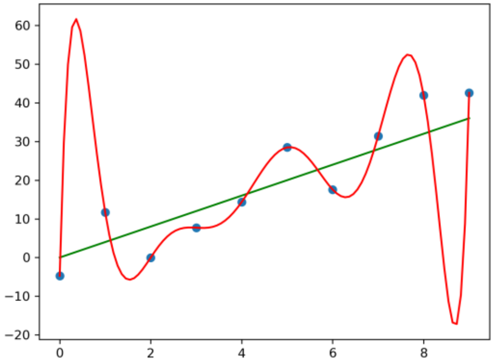
Variance-bias trade-off

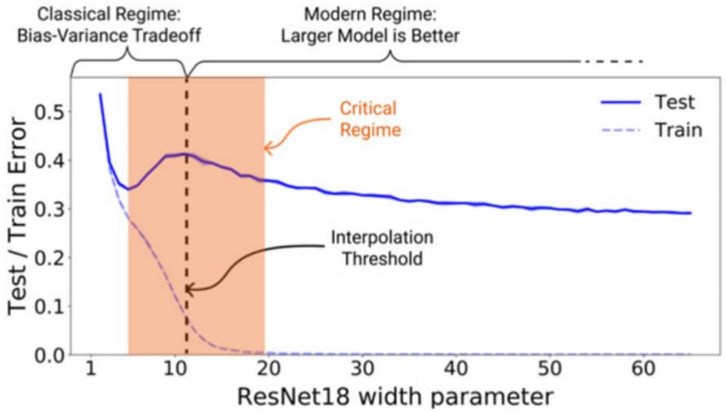
Story two: infinite dimensional regression

- **Paradoxical behavior of "overfitting" vs. "benign overfitting"**









The classical "overfitting" phenomenon

Variance-bias trade-off

The modern "benign overfitting" phenomenon

[Nakkiran et al. 21]

Double descent, implicit regularization etc.

Story two: infinite dimensional regression

Story two: infinite dimensional regression

- **“Weak” benign overfitting: overparameterization gives “good interporlators” that don’t overfit. (implicit bias)**
  - Ridge regression [Hastie et al. 18, Tsigler and Bartlett 20]
  - Max-margin classifiers [Montanari et al. 19]
  - Kernel ridge regression [Liang and Rakhlin 20]

Story two: infinite dimensional regression

- **“Weak” benign overfitting: overparameterization gives “good interporlators” that don’t overfit. (implicit bias)**
  - Ridge regression [Hastie et al. 18, Tsigler and Bartlett 20]
  - Max-margin classifiers [Montanari et al. 19]
  - Kernel ridge regression [Liang and Rakhlin 20]
- **Necessity of “weak” benign overfitting: algorithms that don’t overfit will perform worse.**
  - Linear model [Cheng, Duchi and Kuditipudi 22]

Story two: infinite dimensional regression

- **“Weak” benign overfitting: overparameterization gives “good interporlators” that don’t overfit. (implicit bias)**
    - Ridge regression [Hastie et al. 18, Tsigler and Bartlett 20]
    - Max-margin classifiers [Montanari et al. 19]
    - Kernel ridge regression [Liang and Rakhlin 20]
- **Necessity of “weak” benign overfitting: algorithms that don’t overfit will perform worse.**
    - Linear model [Cheng, Duchi and Kuditipudi 22]
- **“Sharp” benign overfitting: overparameterization gives “sharp interpolators” with vanishing generalization error.**
    - Ridge regression [Tsigler and Bartlett 20]

Story two: infinite dimensional regression

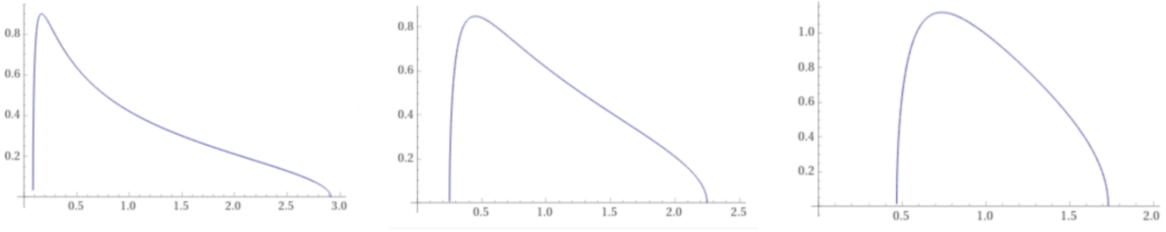
- **However…**

Story two: infinite dimensional regression

- **However…**
  - Previous results rely crucially on (by far) classical proportional asymptotics (such as classical random matrix theory)

Story two: infinite dimensional regression

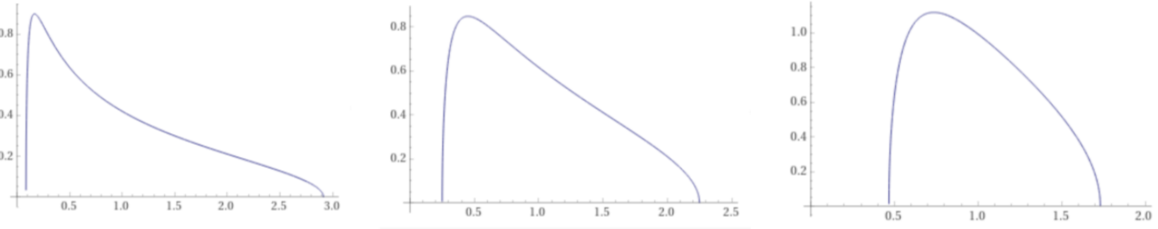
- **However…**
  - Previous results rely crucially on (by far) classical proportional asymptotics (such as classical random matrix theory)



The celebrated Marchenko-Pastur law

Story two: infinite dimensional regression

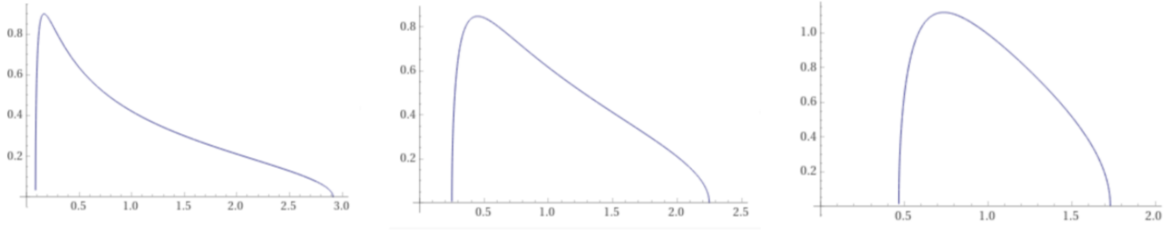
- **However…**
  - Previous results rely crucially on (by far) classical proportional asymptotics (such as classical random matrix theory)
  - In reality, the statisticians don't decide $d \ll n, d \asymp n$ or $d \gg n$



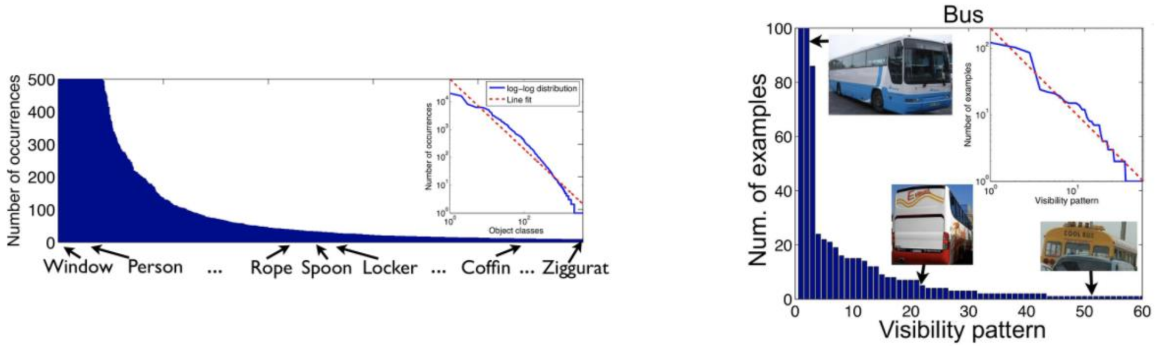
The celebrated Marchenko-Pastur law

Story two: infinite dimensional regression

- **However…**
  - Previous results rely crucially on (by far) classical proportional asymptotics (such as classical random matrix theory)
  - In reality, the statisticians don't decide $d \ll n, d \asymp n$ or $d \gg n$
  - Nature doesn't have a finite inherent dimension nor a well-conditioned covariance



The celebrated Marchenko-Pastur law



Real world data showing Zipf's law ($\lambda_i \asymp i^{-\alpha}$) decay [Feldman 19]

Story two: infinite dimensional regression

- **However…**
  - Previous results rely crucially on (by far) classical proportional asymptotics (such as classical random matrix theory)
  - In reality, the statisticians don't decide $d \ll n, d \asymp n$ or $d \gg n$
  - Nature doesn't have a finite inherent dimension nor a well-conditioned covariance

Story two: infinite dimensional regression

- **However…**
  - Previous results rely crucially on (by far) classical proportional asymptotics (such as classical random matrix theory)
  - In reality, the statisticians don't decide $d \ll n, d \asymp n$ or $d \gg n$
  - Nature doesn't have a finite inherent dimension nor a well-conditioned covariance

- **The ambient manifold doesn't change but we recover better with more data!**

Story two: infinite dimensional regression

- **However…**
  - Previous results rely crucially on (by far) classical proportional asymptotics (such as classical random matrix theory)
  - In reality, the statisticians don't decide $d \ll n, d \asymp n$ or $d \gg n$
  - Nature doesn't have a finite inherent dimension nor a well-conditioned covariance

- **The ambient manifold doesn't change but we recover better with more data!**
- **Question:** How do we develop "dimensional free" tools to understand this learning procedure?

Story two: infinite dimensional regression

- **However…**
  - Previous results rely crucially on (by far) classical proportional asymptotics (such as classical random matrix theory)
  - In reality, the statisticians don't decide $d \ll n, d \asymp n$ or $d \gg n$
  - Nature doesn't have a finite inherent dimension nor a well-conditioned covariance

- **The ambient manifold doesn't change but we recover better with more data!**
- **Question:** How do we develop "dimensional free" tools to understand this learning procedure?
- $d = \infty$! (my cat is not finite dimensional)

Story two: infinite dimensional regression

Story two: infinite dimensional regression

- **A unified theoretical framework: ridge(less) regression**

Story two: infinite dimensional regression

- **A unified theoretical framework: ridge(less) regression**
- **Data**

$$\{(X_i, y_i)\}_{i=1}^n, X_i \in \mathbb{R}^d, y_i \in \mathbb{R}$$

$$y_i = \langle X_i, \theta \rangle + \epsilon_i$$

Story two: infinite dimensional regression

- **A unified theoretical framework: ridge(less) regression**
- **Data**

$$\{(X_i, y_i)\}_{i=1}^n, X_i \in \mathbb{R}^d, y_i \in \mathbb{R}$$

$$y_i = \langle X_i, \theta \rangle + \epsilon_i$$

- **We allow** $d = \infty$ **, consider i.i.d. noise and features from a trace class** $\mathrm{Tr}(\Sigma) := \mathrm{Tr}(\mathbb{E}[X_i X_i^\top]) = \mathbb{E}[\|X_i\|^2] < \infty$

Story two: infinite dimensional regression

- **A unified theoretical framework: ridge(less) regression**
- **Data**

$$\{(X_i, y_i)\}_{i=1}^n, X_i \in \mathbb{R}^d, y_i \in \mathbb{R}$$

$$y_i = \langle X_i, \theta \rangle + \epsilon_i$$

- **We allow** $d = \infty$ **, consider i.i.d. noise and features from a trace class** $\mathrm{Tr}(\Sigma) := \mathrm{Tr}(\mathbb{E}[X_i X_i^\top]) = \mathbb{E}[\|X_i\|^2] < \infty$
- **Featurization of RKHS** $f(z_i) = \langle X_i, \theta \rangle, \qquad f \in \mathcal{H}$

Story two: infinite dimensional regression

- **The ridge estimator**

$$X = \begin{bmatrix} — & X_1^\top & — \\ — & X_2^\top & — \\ & \vdots & \\ — & X_n^\top & — \end{bmatrix} \in \mathbb{R}^{n \times d} \qquad \hat{\theta}_\lambda = (X^\top X + \lambda I)^{-1} X^\top y$$

- **Generalization error**

$$R_X(\lambda) = \mathbb{E}_{X_{\text{new}} \sim P}[\|\langle X_{\text{new}}, \hat{\theta}_\lambda \rangle - \langle X_{\text{new}}, \theta \rangle\|^2]$$

Story two: infinite dimensional regression

Story two: infinite dimensional regression

- **The equivalent sequence model**

$$y^s = \Sigma^{1/2}\theta + \frac{w}{\sqrt{n}} g \qquad w > 0, g \sim \mathsf{N}(0, I)$$

$$\hat{\theta}^s_{\lambda_\star} = \arg\min\{\|y^s - \Sigma^{1/2}\beta\|^2 + \lambda_\star\|\beta\|^2\}$$

Story two: infinite dimensional regression

- **The equivalent sequence model**

$$y^s = \Sigma^{1/2}\theta + \frac{w}{\sqrt{n}} g \qquad w > 0, g \sim \mathsf{N}(0, I)$$

$$\hat{\theta}^s_{\lambda_\star} = \arg\min\{\|y^s - \Sigma^{1/2}\beta\|^2 + \lambda_\star\|\beta\|^2\}$$

- **The generalization error (deterministic)**

$$R(\lambda_\star) = \mathbb{E}_g\|\hat{\theta}^s_{\lambda_\star} - \theta\|^2_\Sigma$$

Story two: infinite dimensional regression

- **The equivalent sequence model**

$$y^s = \Sigma^{1/2}\theta + \frac{w}{\sqrt{n}} g \qquad w > 0, g \sim \mathsf{N}(0, I)$$

$$\hat{\theta}^s_{\lambda_\star} = \arg\min\{\|y^s - \Sigma^{1/2}\beta\|^2 + \lambda_\star \|\beta\|^2\}$$

- **The generalization error (deterministic)**

$$R(\lambda_\star) = \mathbb{E}_g \|\hat{\theta}^s_{\lambda_\star} - \theta\|^2_\Sigma$$

Theorem (Cheng and Montanari, 22) (Informal) Under appropriate assumptions, for $\lambda_\star = \lambda_\star(\lambda)$ (suppressing the dependence on n and covariance),

$$R_X(\lambda) = \mathbb{E}_{X_{\text{new}} \sim P}[\|\langle X_{\text{new}}, \hat{\theta}_\lambda\rangle - \langle X_{\text{new}}, \theta\rangle\|^2] = (1 + \mathsf{err}_n) \cdot R(\lambda_\star) = \mathbb{E}_g \|\hat{\theta}^s_{\lambda_\star} - \theta\|^2_\Sigma (1 + \mathsf{err}_n)$$

Story two: infinite dimensional regression

- **More precisely**

Theorem (Cheng and Montanari, 22) (Informal) Under appropriate assumptions, for $\lambda_\star = \lambda_\star(\lambda)$ (suppressing the dependence on n and covariance),

$$R_X(\lambda) = \mathbb{E}_{X_{\text{new}} \sim P}[\|\langle X_{\text{new}}, \hat{\theta}_\lambda \rangle - \langle X_{\text{new}}, \theta \rangle\|^2] = (1 + \text{err}_n) \cdot R(\lambda_\star) = \mathbb{E}_g \|\hat{\theta}^s_{\lambda_\star} - \theta\|_\Sigma^2 (1 + \text{err}_n)$$

Story two: infinite dimensional regression

- **More precisely**

$$y = X\theta + \epsilon \qquad \mathbb{E}[\epsilon_i^2] = \tau^2$$
$$\hat{\theta}_\lambda = \arg\min\{\|y - X\beta\|^2 + \lambda\|\beta\|^2\}$$

Random design

$$y^s = \Sigma^{1/2}\theta + \frac{w}{\sqrt{n}}g \qquad w > 0, g \sim \mathsf{N}(0, I)$$
$$\hat{\theta}^s_{\lambda_\star} = \arg\min\{\|y^s - \Sigma^{1/2}\beta\|^2 + \lambda_\star\|\beta\|^2\}$$

Fixed design

Theorem (Cheng and Montanari, 22) (Informal) Under appropriate assumptions, for $\lambda_\star = \lambda_\star(\lambda)$ (suppressing the dependence on n and covariance),

$$R_X(\lambda) = \mathbb{E}_{X_{\text{new}}\sim P}[\|\langle X_{\text{new}}, \hat{\theta}_\lambda\rangle - \langle X_{\text{new}}, \theta\rangle\|^2] = (1 + \mathsf{err}_n) \cdot R(\lambda_\star) = \mathbb{E}_g\|\hat{\theta}^s_{\lambda_\star} - \theta\|^2_\Sigma (1 + \mathsf{err}_n)$$

Story two: infinite dimensional regression

- **More precisely**

$$w^2 = \tau^2 + R(\lambda_\star)$$

$$n - \frac{\lambda}{\lambda_\star} = \mathrm{Tr}(\Sigma(\Sigma + \lambda_\star I)^{-1})$$

$$y = X\theta + \epsilon \qquad \mathbb{E}[\epsilon_i^2] = \tau^2$$

$$\hat{\theta}_\lambda = \arg\min\{\|y - X\beta\|^2 + \lambda\|\beta\|^2\}$$

Random design

Deterministic equivalence

$$y^s = \Sigma^{1/2}\theta + \frac{w}{\sqrt{n}} g \qquad w > 0, g \sim \mathsf{N}(0, I)$$

$$\hat{\theta}^s_{\lambda_\star} = \arg\min\{\|y^s - \Sigma^{1/2}\beta\|^2 + \lambda_\star\|\beta\|^2\}$$

Fixed design

Theorem (Cheng and Montanari, 22) (Informal) Under appropriate assumptions, for $\lambda_\star = \lambda_\star(\lambda)$ (suppressing the dependence on n and covariance),

$$R_X(\lambda) = \mathbb{E}_{X_{\mathrm{new}} \sim P}[\|\langle X_{\mathrm{new}}, \hat{\theta}_\lambda\rangle - \langle X_{\mathrm{new}}, \theta\rangle\|^2] = (1 + \mathsf{err}_n) \cdot R(\lambda_\star) = \mathbb{E}_g\|\hat{\theta}^s_{\lambda_\star} - \theta\|^2_\Sigma (1 + \mathsf{err}_n)$$

Story two: infinite dimensional regression

$$w^2 = \tau^2 + R(\lambda_\star)$$

- **More precisely** $$n - \frac{\lambda}{\lambda_\star} = \mathrm{Tr}(\Sigma(\Sigma + \lambda_\star I)^{-1})$$

$$y = X\theta + \epsilon \qquad \mathbb{E}[\epsilon_i^2] = \tau^2$$

$$\hat{\theta}_\lambda = \arg\min\{\|y - X\beta\|^2 + \lambda\|\beta\|^2\}$$

Random design

Deterministic equivalence

$$y^s = \Sigma^{1/2}\theta + \frac{w}{\sqrt{n}} g \qquad w > 0, g \sim \mathsf{N}(0, I)$$

$$\hat{\theta}^s_{\lambda_\star} = \arg\min\{\|y^s - \Sigma^{1/2}\beta\|^2 + \lambda_\star\|\beta\|^2\}$$

Fixed design

Typically nontrivial behavior and vanishing multiplicative error term when the effective rank parameter is comparable to λ/σ_n

$$\sum_{l=k}^{d} \sigma_l \le d_\Sigma(n)\sigma_k, k = 1, 2, \cdots, n \qquad d_\Sigma/n \asymp \lambda/\sigma_n$$

Theorem (Cheng and Montanari, 22) (Informal) Under appropriate assumptions, for $\lambda_\star = \lambda_\star(\lambda)$ (suppressing the dependence on n and covariance),

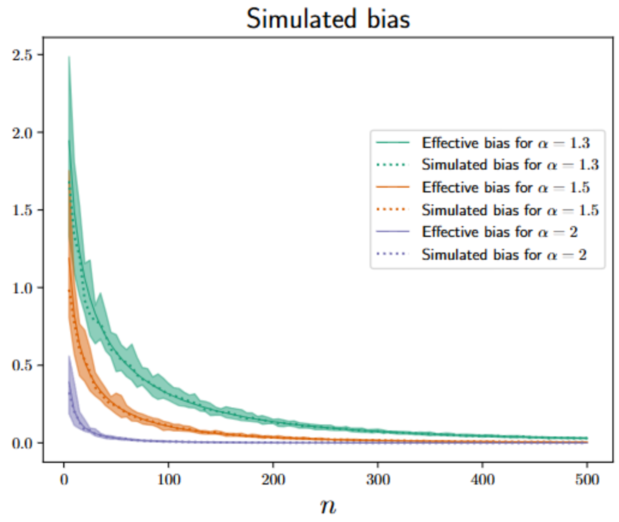
$$R_X(\lambda) = \mathbb{E}_{X_{\mathrm{new}}\sim P}[\|\langle X_{\mathrm{new}}, \hat{\theta}_\lambda\rangle - \langle X_{\mathrm{new}}, \theta\rangle\|^2] = (1 + \mathsf{err}_n)\cdot R(\lambda_\star) = \mathbb{E}_g\|\hat{\theta}^s_{\lambda_\star} - \theta\|^2_\Sigma (1 + \mathsf{err}_n)$$

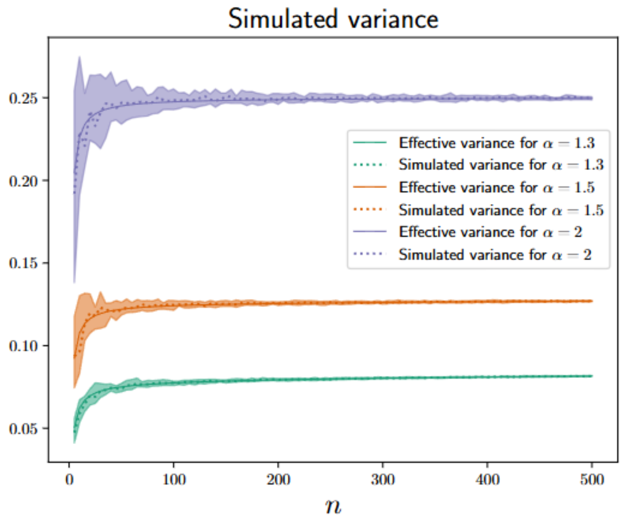
Story two: infinite dimensional regression

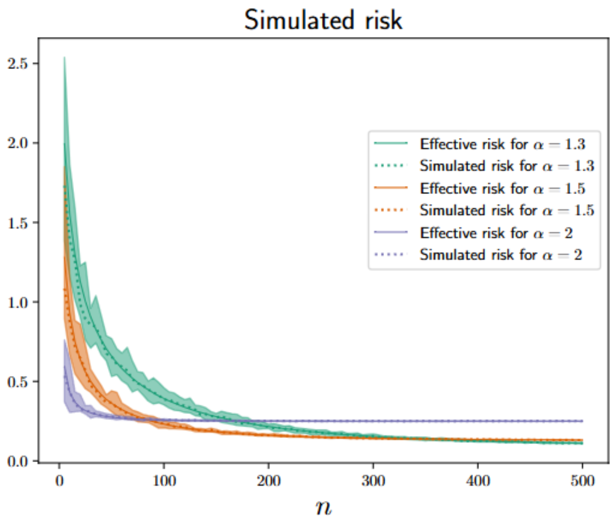
- **Experiments**

Story two: infinite dimensional regression

- **Experiments**
  - Zipf's law ("weak" benign overfitting)
    $\sigma_i = i^{-\alpha}$

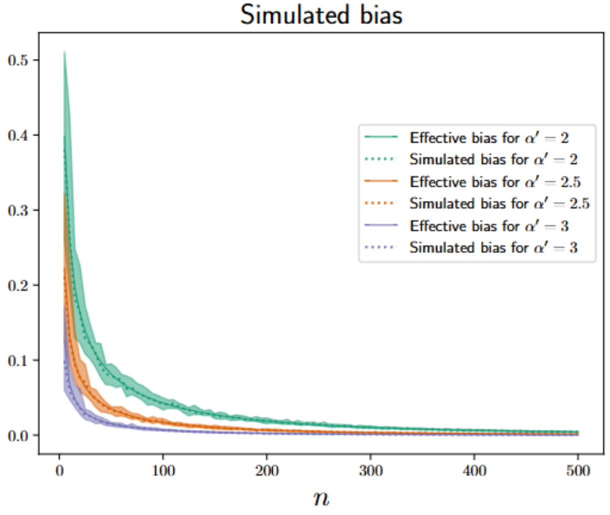


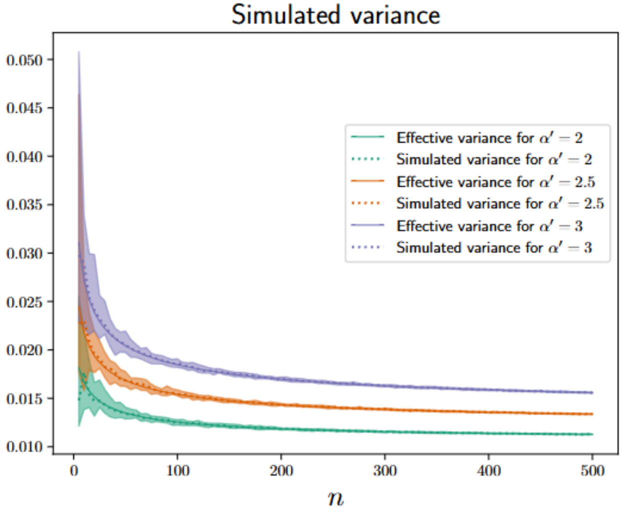


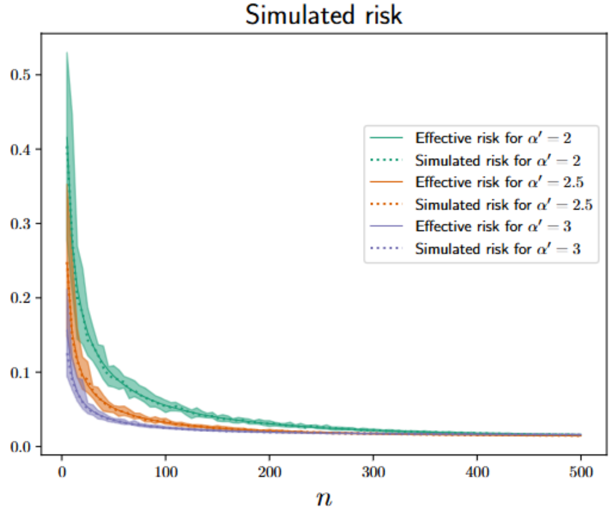


Story two: infinite dimensional regression

- **Experiments**
  - Zipf's law ("weak" benign overfitting)
    $\sigma_i = i^{-\alpha}$
  - Critical law ("sharp benign" overfitting)
    $\sigma_i = i^{-1}(1 + \log i)^{-\alpha'}$







Story two: infinite dimensional regression

- **A very high level proof**

Story two: infinite dimensional regression

- **A very high level proof**

$$\mathrm{R}_i(\boldsymbol{Q}) := \mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\zeta\boldsymbol{I} + \mu_i\boldsymbol{\Sigma} + \boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{X}_i\right)^{-1}\right)$$

Leave one out and appropriately interpolate from $\mu\Sigma$ to $\hat{\Sigma}$ through an martingale argument

Story two: infinite dimensional regression

- **A very high level proof**

$$\mathrm{R}_i(\boldsymbol{Q}) := \mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\zeta\boldsymbol{I} + \mu_i\boldsymbol{\Sigma} + \boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{X}_i\right)^{-1}\right)$$

  Leave one out and appropriately interpolate from $\mu\Sigma$ to $\hat{\Sigma}$
  through an martingale argument

- **Conclusion**
  - Dimension free deterministic equivalent risk for ridge(less) regression
  - Shed light to understanding real world data

Other work

- **High dimensional data**
  - Memorization [Cheng, Duchi, Kuditipudi, 22]
  - High dimensional gradient flow [Celentano, Cheng, Montanari, 21]
  - Low-rank matrix recovery [Cheng, Wei, Chen, 21]
- **Robustness quantification and fundamental limits**
  - Geometry and computational optimality [Cheng, Duchi, Levy, 24]
  - Weighted conformal inference [Areces, Cheng, Kuditipudi, 24]
  - Collaborative learning [Cheng, Cheng, Duchi, 23]
- **Reinforcement learning**
  - Entropy regularization [Cen, Cheng, Chen, Wei, Chi, 20]

Thank You!