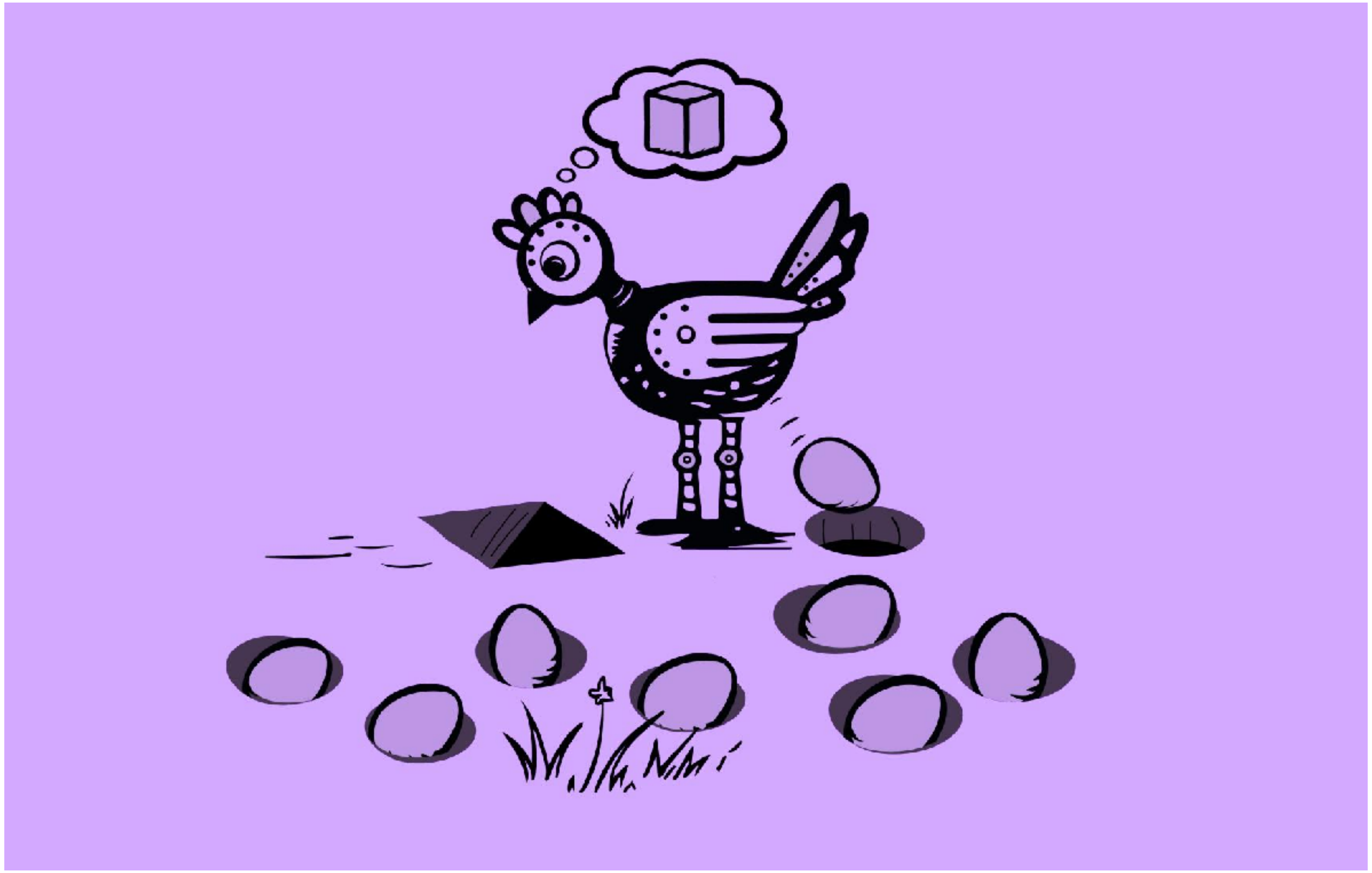
(Unexpected Test) Losses from Generalization Theory?

Frederic Koehler (University of Chicago)
Simons Institute, September 2024

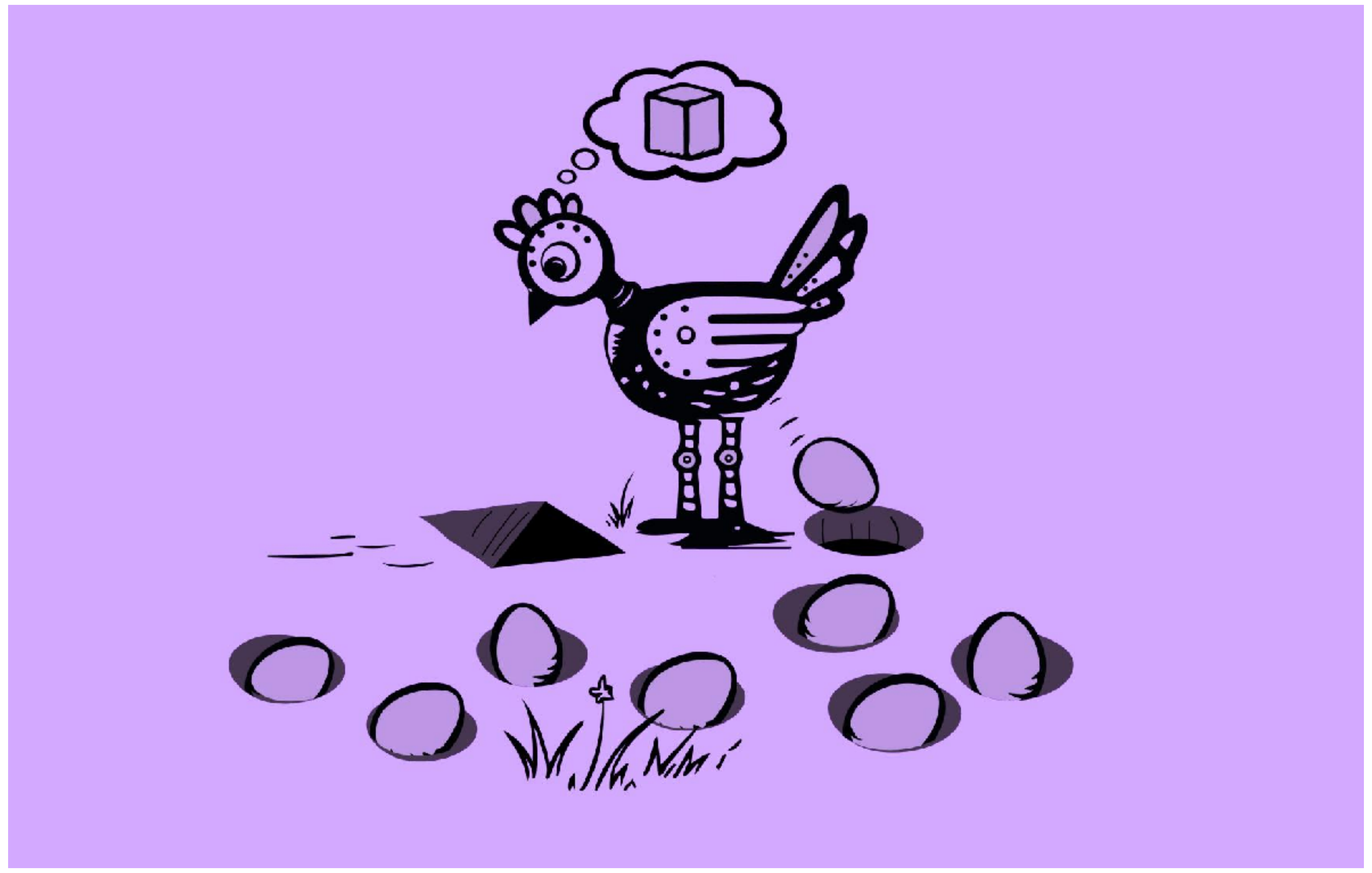
“The workshop will bring together applied researchers and theorists, with the goal of understanding how each **understands notions of generalization**.”



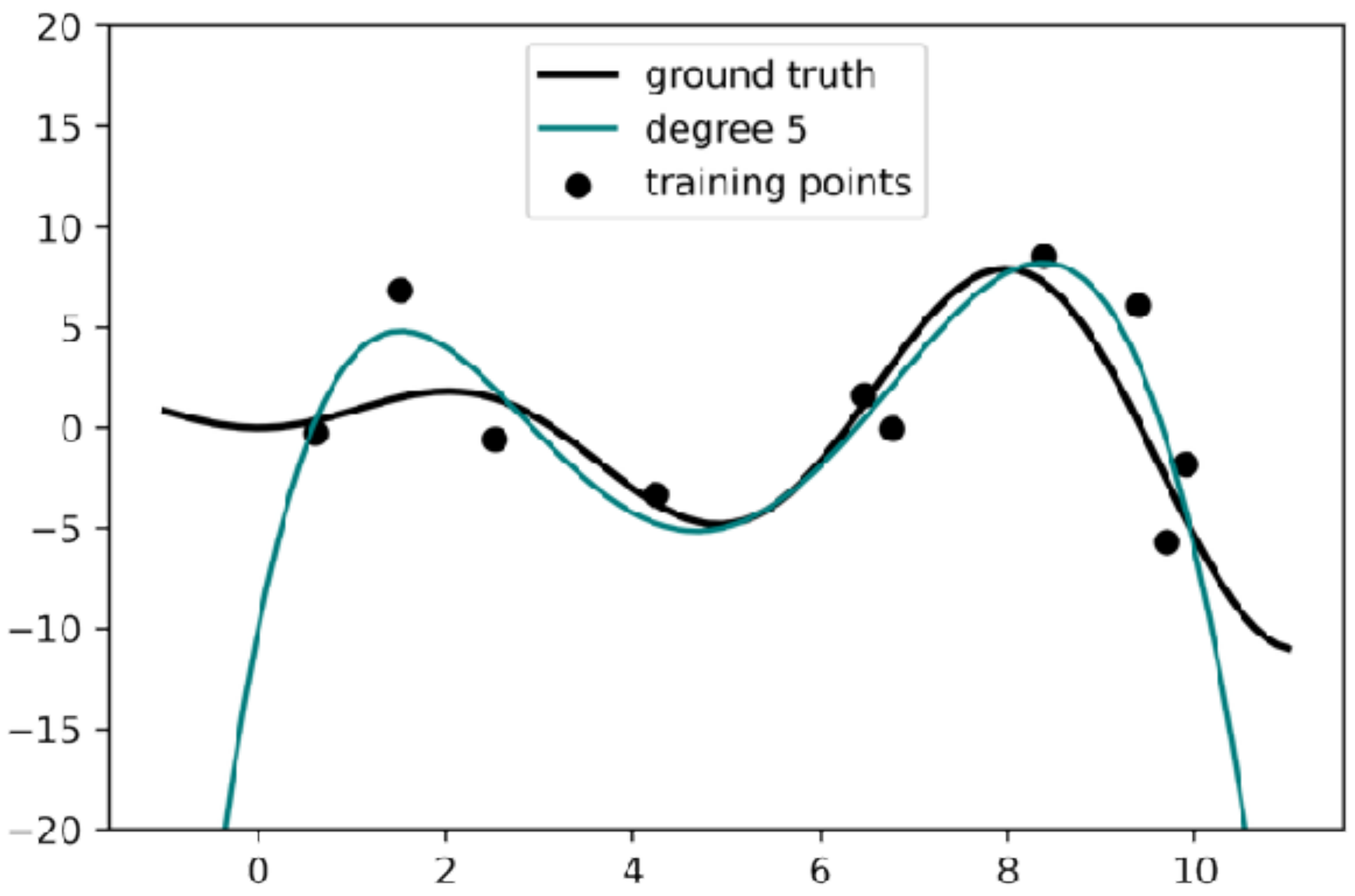
What is **generalization?**

Different notions of generalization

- Generalization: prediction on future examples which you haven't seen before (?)
- **Classical setting**: additionally assume training data is iid from test distribution $\mathscr{D}$.
  - "Generalization" <-> understanding overfitting
  - <-> distortion between training loss and test loss
- OOD Generalization, interactive settings, dependent samples, …. not strictly within scope of classical generalization theory.
- Nevertheless, very natural setting to understand *pattern-formation/recognition ability!* Even most basic case of a broader theory is worth understanding (and has caused people concern).

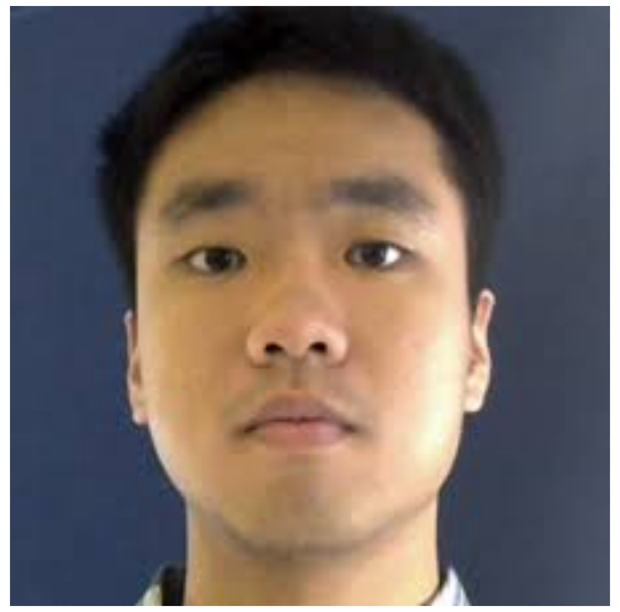


?



Collaborators, etc.

- This talk is largely based on two recent joint works.
  - “A Non-Asymptotic Moreau Envelope Theory for High-Dimensional Generalized Linear Models” with Lijia Zhou, Danica Sutherland, Pragya Sur, and Nati Srebro
  - “Uniform Convergence with Square-Root Lipschitz Loss” with Lijia Zhou, Zhen Dai, and Nati Srebro

Danica J. Sutherland

- Also, it is related to lots of other interesting works and efforts.
  - For example, it is related to words like “proportional asymptotics”, “CGMT” statistical physics methods in high-dimensional statistics (e.g. AMP), “benign overfitting”, “implicit bias”, …

Presentation vs papers

- The papers have the **rigorous results**.
  - Nonasymptotic, appropriate for different noise models, etc.
- But it could look a bit unfriendly…
- This talk won't be that formal (relatively)
  - e.g. drop error terms/low sample size & dimension effects (even though they are not complicated)
  - simple examples
  - deemphasize minor assumptions
  - hopefully an enticement for papers…

Definition 2. Under the model assumptions (2), define a (possibly oblique) projection matrix Q onto the space orthogonal to $w_1^*, ..., w_k^*$ and a mapping ϕ from $\mathbb{R}^d$ to $\mathbb{R}^{k+1}$ by

$$Q = I_d - \sum_{i=1}^{k} w_i^*(w_i^*)^T\Sigma, \quad \phi(w) = (\langle w, \Sigma w_1^* \rangle, ..., \langle w, \Sigma w_k^* \rangle, \|\Sigma^{1/2} Q w\|_2)^T. \tag{4}$$

We let $\Sigma^{\perp} = Q^T \Sigma Q$ denote the covariance matrix of $Q^T x$. We also define a low-dimensional *surrogate distribution* $\tilde{\mathcal{D}}$ over $\mathbb{R}^{k+1} \times \mathbb{R}$ by

$$\tilde{x} \sim \mathcal{N}(0, I_{k+1}), \quad \tilde{\xi} \sim \mathcal{D}_{\xi}, \quad \text{and} \quad \tilde{y} = g(\tilde{x}_1, ..., \tilde{x}_k, \tilde{\xi}). \tag{5}$$

This surrogate distribution compresses the "meaningful part" of x while maintaining the test loss, as shown by our main result Theorem 1 (proved in Appendix D). Note that as a non-asymptotic statement, the functions $\epsilon_{\lambda,\delta}$ and C_δ only need hold for a specific choice of n and $\mathcal{D}$.

Theorem 1. *Suppose $\lambda \in \mathbb{R}^+$ satisfies that for any $\delta \in (0,1)$, there exists a continuous function $\epsilon_{\lambda,\delta} : \mathbb{R}^{k+1} \to \mathbb{R}$ such that with probability at least $1 - \delta/4$ over independent draws $(\tilde{x}_i, \tilde{y}_i)$ from the surrogate distribution $\tilde{\mathcal{D}}$ defined in (5), we have uniformly over all $(\tilde{w}, \tilde{b}) \in \mathbb{R}^{k+2}$ that*

$$\frac{1}{n}\sum_{i=1}^{n} f_\lambda(\langle \tilde{w}, \tilde{x}_i \rangle + \tilde{b}, \tilde{y}_i) \geq \mathop{\mathbb{E}}_{(\tilde{x},\tilde{y}) \sim \tilde{\mathcal{D}}} [f_\lambda(\langle \tilde{w}, \tilde{x} \rangle + \tilde{b}, \tilde{y})] - \epsilon_{\lambda,\delta}(\tilde{w}, \tilde{b}). \tag{6}$$

Further, assume that for any $\delta \in (0,1)$, there exists a continuous function $C_\delta : \mathbb{R}^d \to [0, \infty]$ such that with probability at least $1 - \delta/4$ over $x \sim \mathcal{N}(0, \Sigma)$, uniformly over all $w \in \mathbb{R}^d$,

$$\langle Qw, x \rangle \leq C_\delta(w). \tag{7}$$

4

Then it holds with probability at least $1 - \delta$ that uniformly over all $(w, b) \in \mathbb{R}^{d+1}$, we have

$$L_{f_\lambda}(w, b) \leq \hat{L}_f(w, b) + \epsilon_{\lambda,\delta}(\phi(w), b) + \frac{\lambda C_\delta(w)^2}{n}. \tag{8}$$

If we additionally assume that (6) holds uniformly for all $\lambda \in \mathbb{R}^+$, then (8) does as well.

Overview: two high-level phenomena

- Observe some interesting (to me) phenomena in very simple models
- Not intuitively obvious (IMO) but natural output of mathematical analysis
- Food for future thought: these high-level phenomena may occur in other settings?

Phenomena 1: mismatch of training and test losses

- Classical generalization story (e.g. Vapnik-Chervonenkis ’71):

$$\text{test}(f) \leq \text{train}(f) + \text{complexity}(f)$$

  - where $\text{test}(f) = \mathbb{E}\ell(f(x), y)$ and $\text{train}(f) = \dfrac{1}{n}\sum_{i=1}^{n} \ell(f(x^{(i)}), y^{(i)})$
  - Seems natural because empirical mean concentrates about true mean?
- Today’s story:
  - Explain why we sometimes would have **test loss differing from training loss!**
    - I.e. model seeks to emulate an “oracle” minimizing a *different* loss than ℓ.
  - Naturally arises when we look at **optimal generalization bounds.**

Phenomena 2: “implicit bias” of overfitting, emergent losses

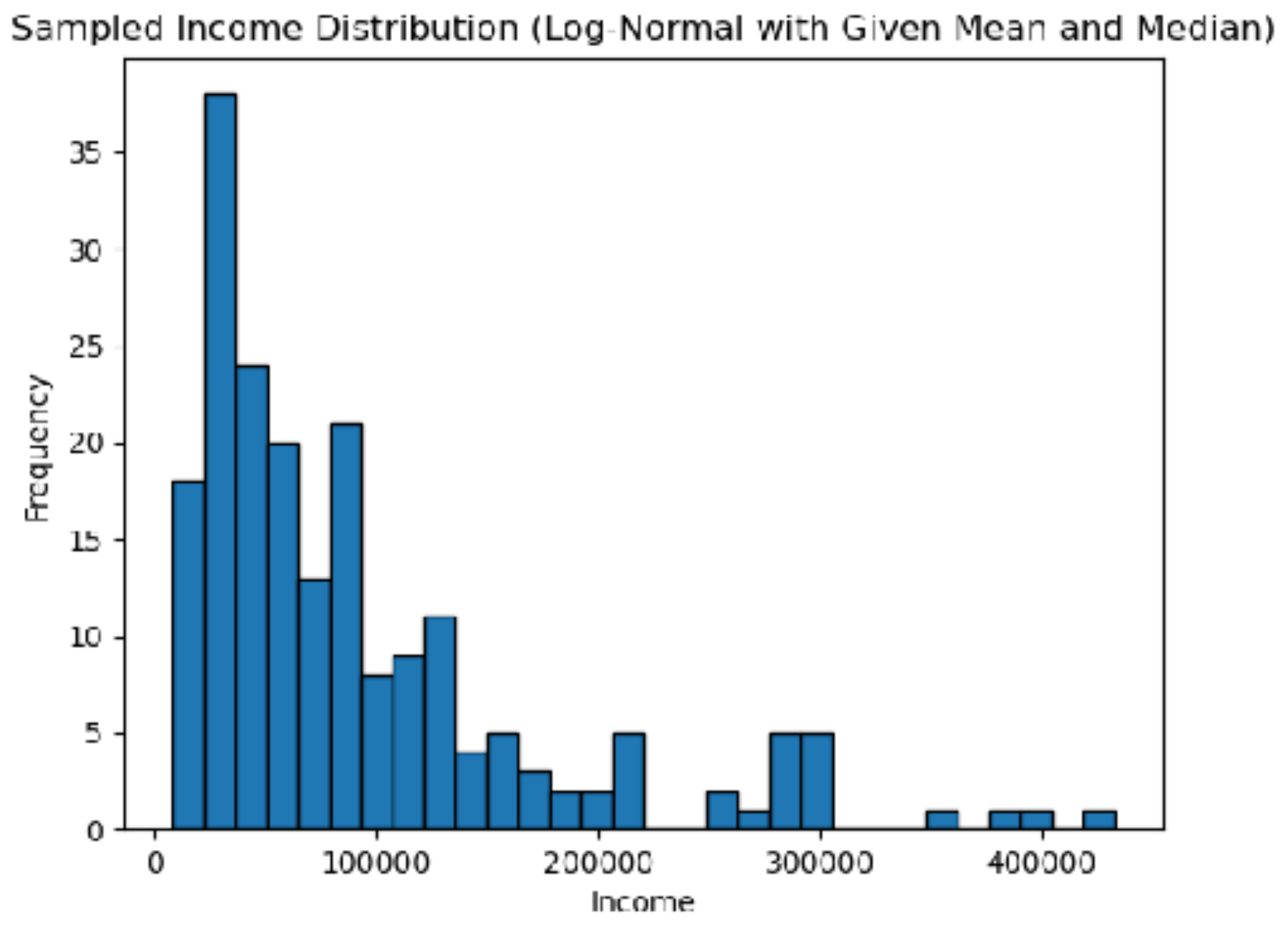
- Name borrowed from a parallel work of Ohad Shamir
- A new kind of “implicit bias” in learning:
    - Previous work focuses on **implicit regularization**
    - Choice of optimization method & model parameterization lead to a preference in **regularization**, e.g. low ℓ_2-norm or low rank solutions.
- **New phenomena:** model capacity <-> implicit change in **loss function**
    - Important because typically, different losses to lead to different minimizers.
    - I will give a very simple example to explain this point.
    - Especially discover some new unexpected/emergent “sqrt-lipschitz” losses.

An example to illustrate the idea

- A made-up story:
  - You: have some user data.
  - Surprisingly, you want to make money off the users.
  - You want to know the income of each user (e.g. to target ads), but you don't know what it is.
  - Train a model f to predict income Y based of feature vector X.
- RMK: the story is just a pedagogical tool. Don't take it too seriously…

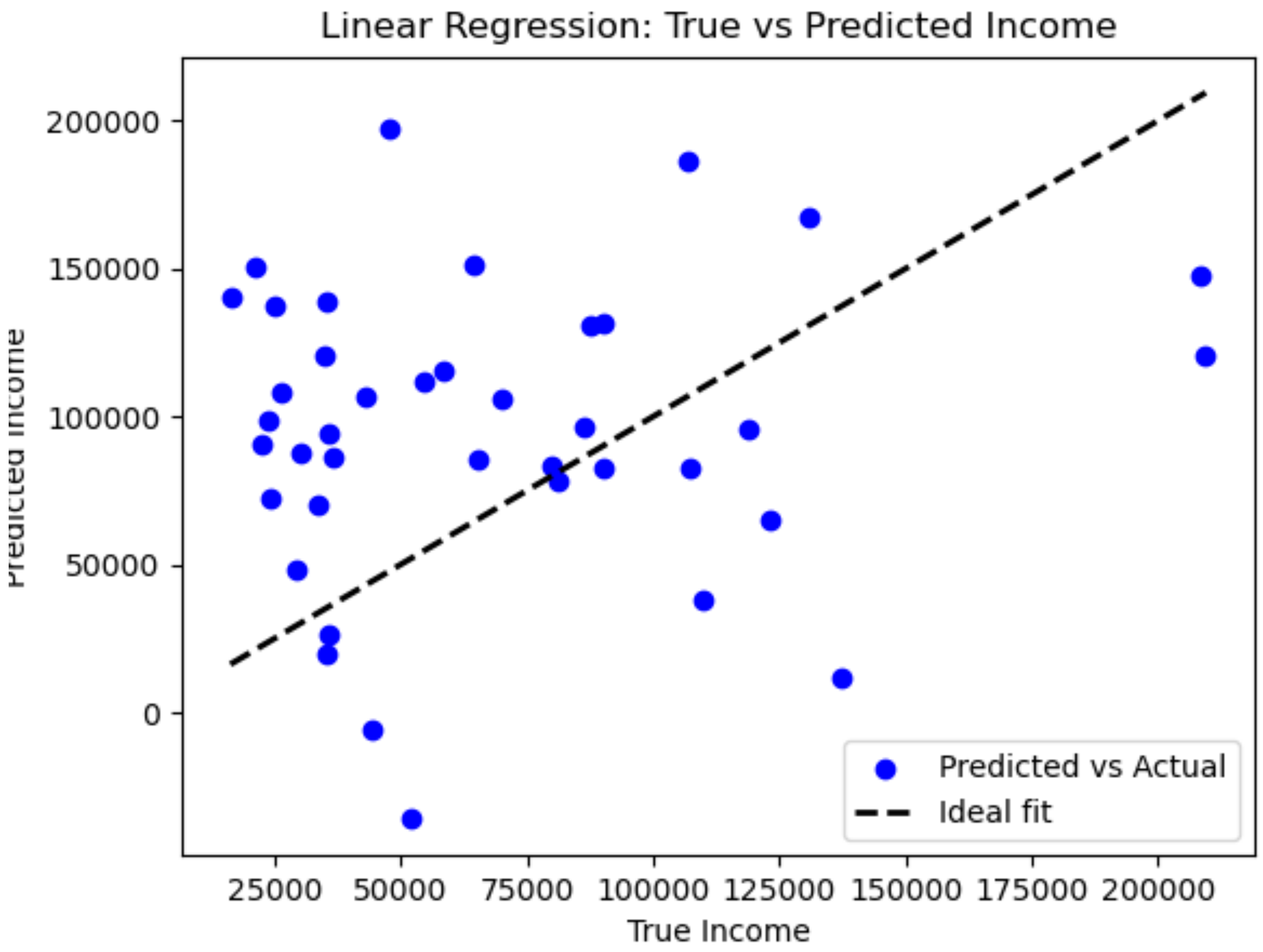
The example, continued

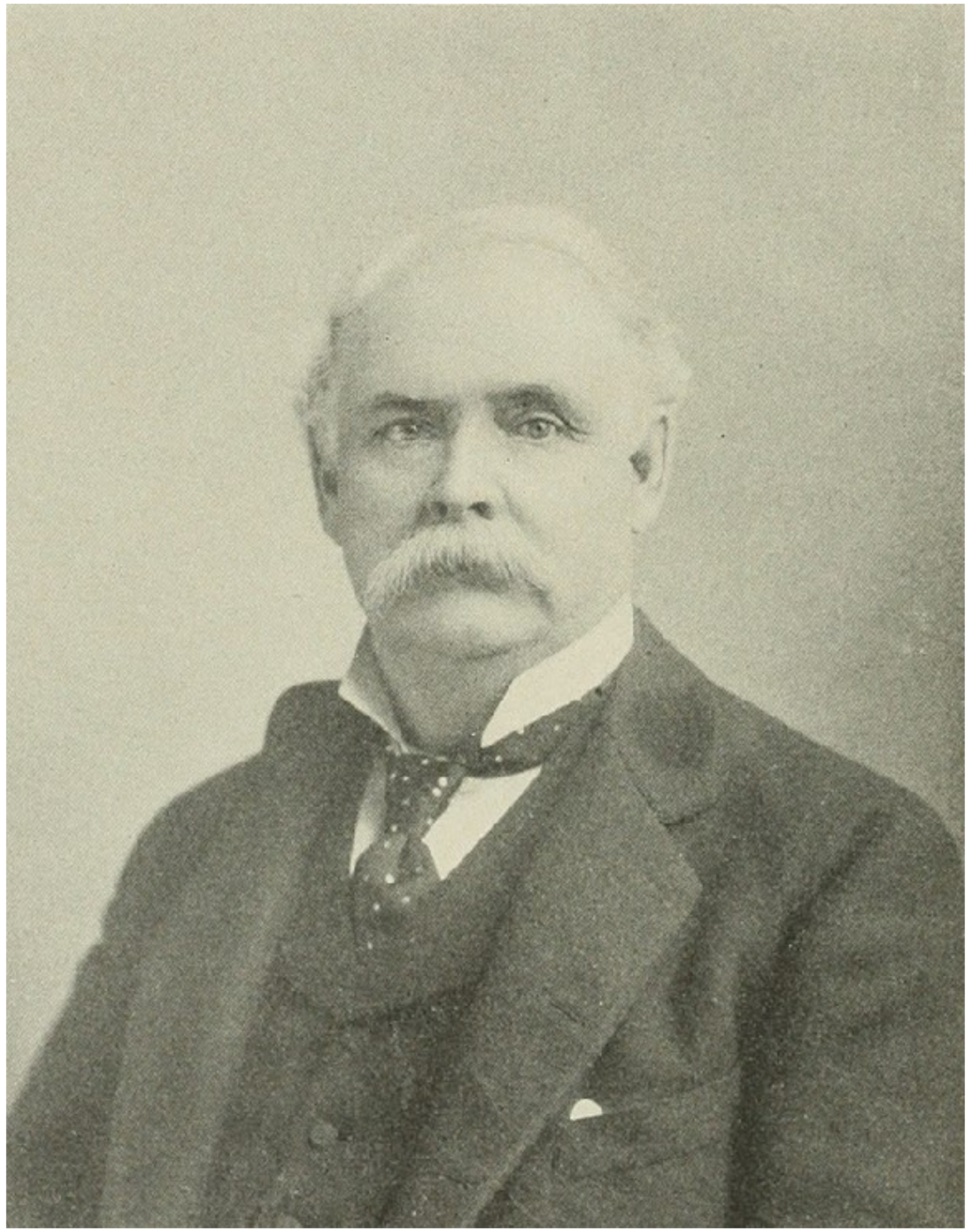
- You collect a small subset of labeled examples (e.g. by asking the users). For simplicity: assume they are iid.
- Now you want to train a model f to map X -> Y = income.
- Natural to pick f by “Empirical Risk Minimization”:
- $$\min \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(X_i)|$$
- **To maximize simplicity, I consider an unlucky setting where** $X \sim N(0,I)$ **independent of Y.** i.e. features are actually **useless**. and I fit a linear model.



Useless features lead to bad predictions…

(test set performance)





Trivia: Who is this?

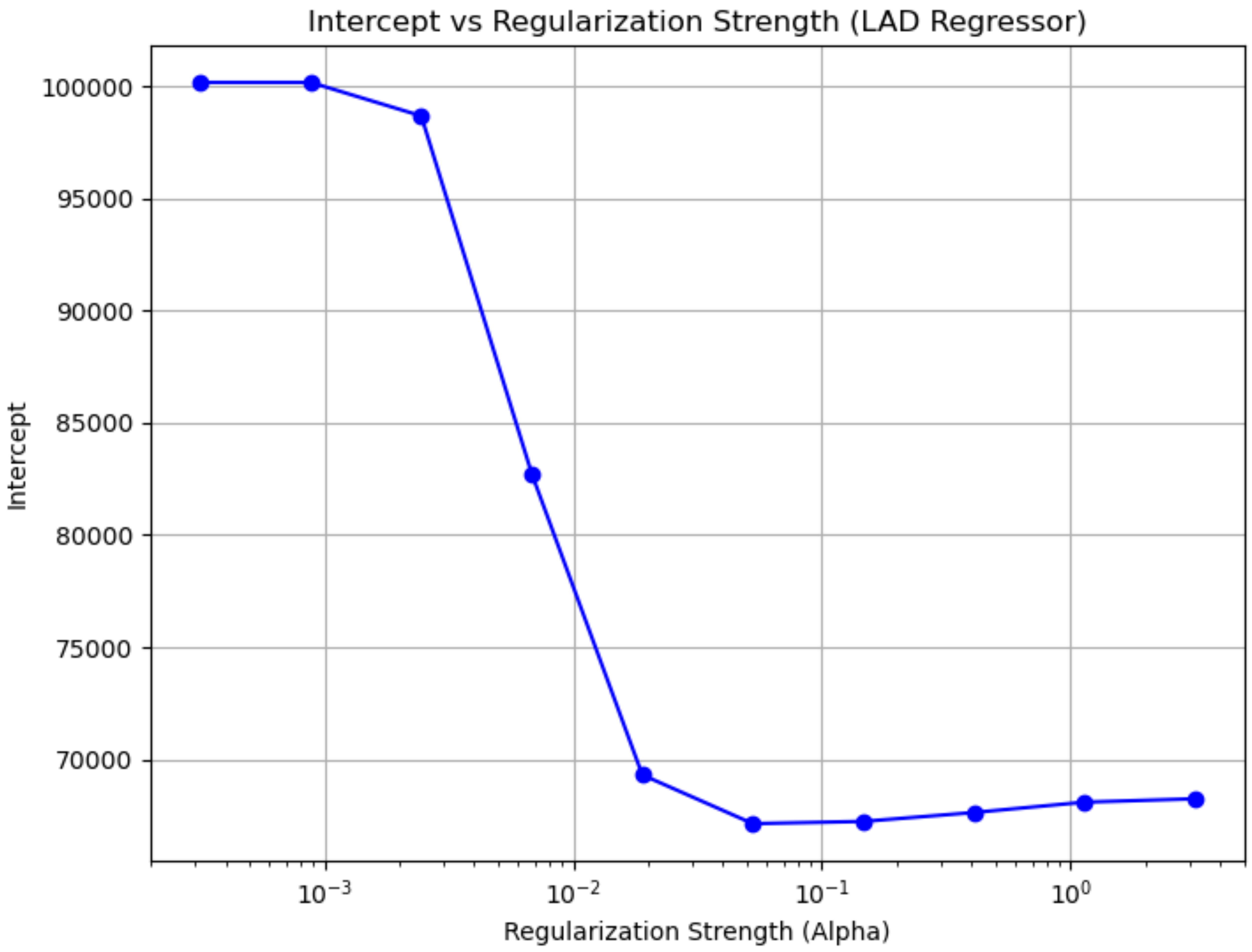
- (by itself, not surprising or interesting)

Any interesting phenomena?

(interesting to me)

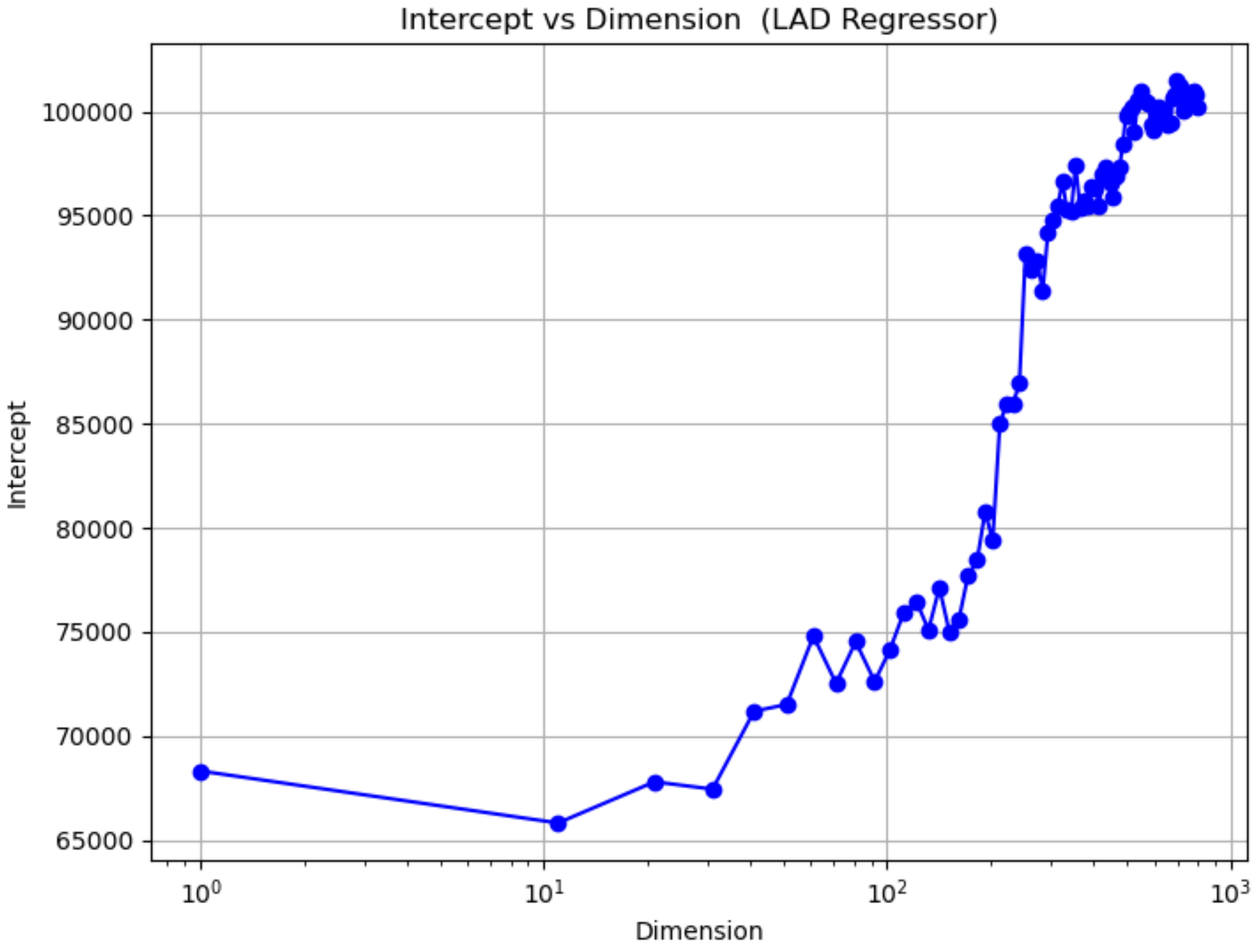
- I trained linear model $f(x) = \langle w, x \rangle + b$ with different levels of ridge regularization and n = 200 examples.
- Since (for simplicity) $X \sim N(0, I_{800})$ **independently of label** Y, the only predictive part of the model is the intercept b.
- Next slide: plot of b for different regularization levels.
  - What will it look like?

Intercept vs regularization level



Intercept vs # of parameters (unregularizd)

a diferent experiment



Why interesting? (to me)

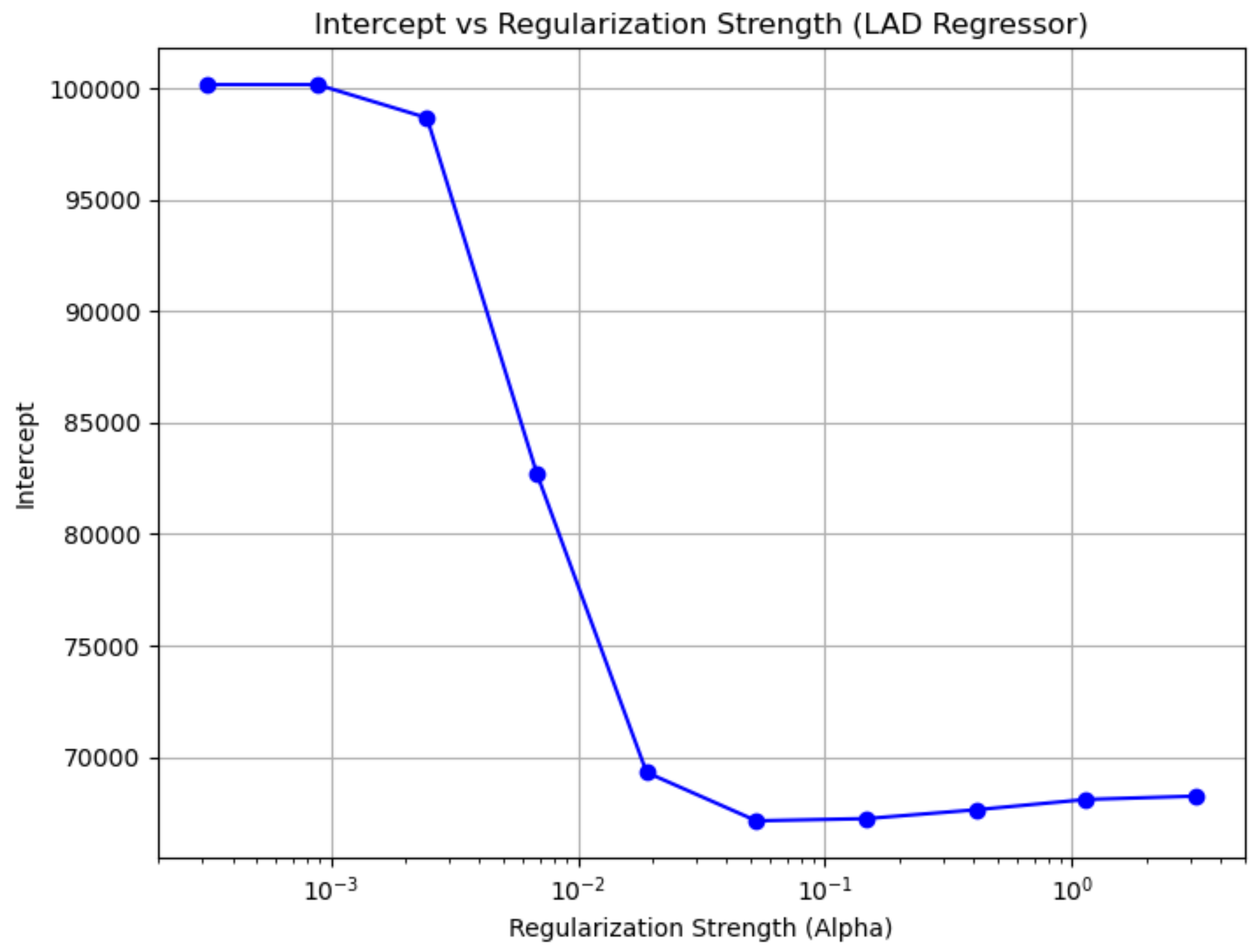
- Each plot has two extreme sides.
- In "low capacity" case (e.g. high ridge penalty):
  - Classical statistical theory tells us

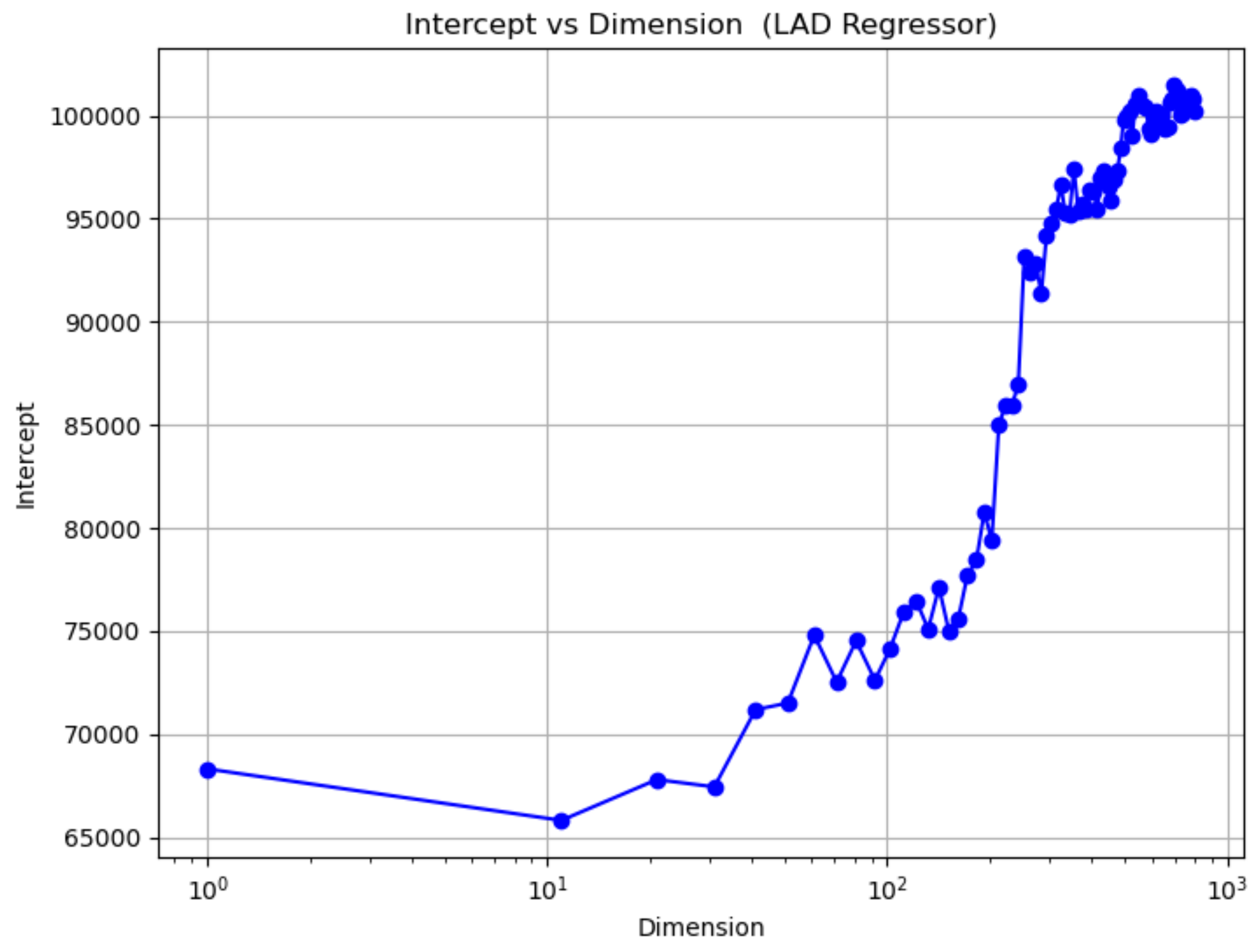
$$b \approx \text{median} = \arg\min_m \frac{1}{n} \sum_{i=1}^{n} |y_i - b|$$

- But it turns out in **high capacity case:**

$$b \approx \text{mean} = \arg\min \frac{1}{n} \sum_{i=1}^{n} (y_i - b)^2$$

- I.e. model seeks to emulate a *different "oracle" !*
- **Where did the square loss come from?** 🤔





Why is it happening?

NATURAL GUESS: squared loss due to **ridge penalty?** $\|w\|_2 = \sqrt{\sum_i w_i^2}$

- This wouldn't really explain the same phenomena happening as you vary # of parameters.
- It turns out, you could also get the same phenomena to happen with ℓ_1 (LASSO) regularization instead of ridge.

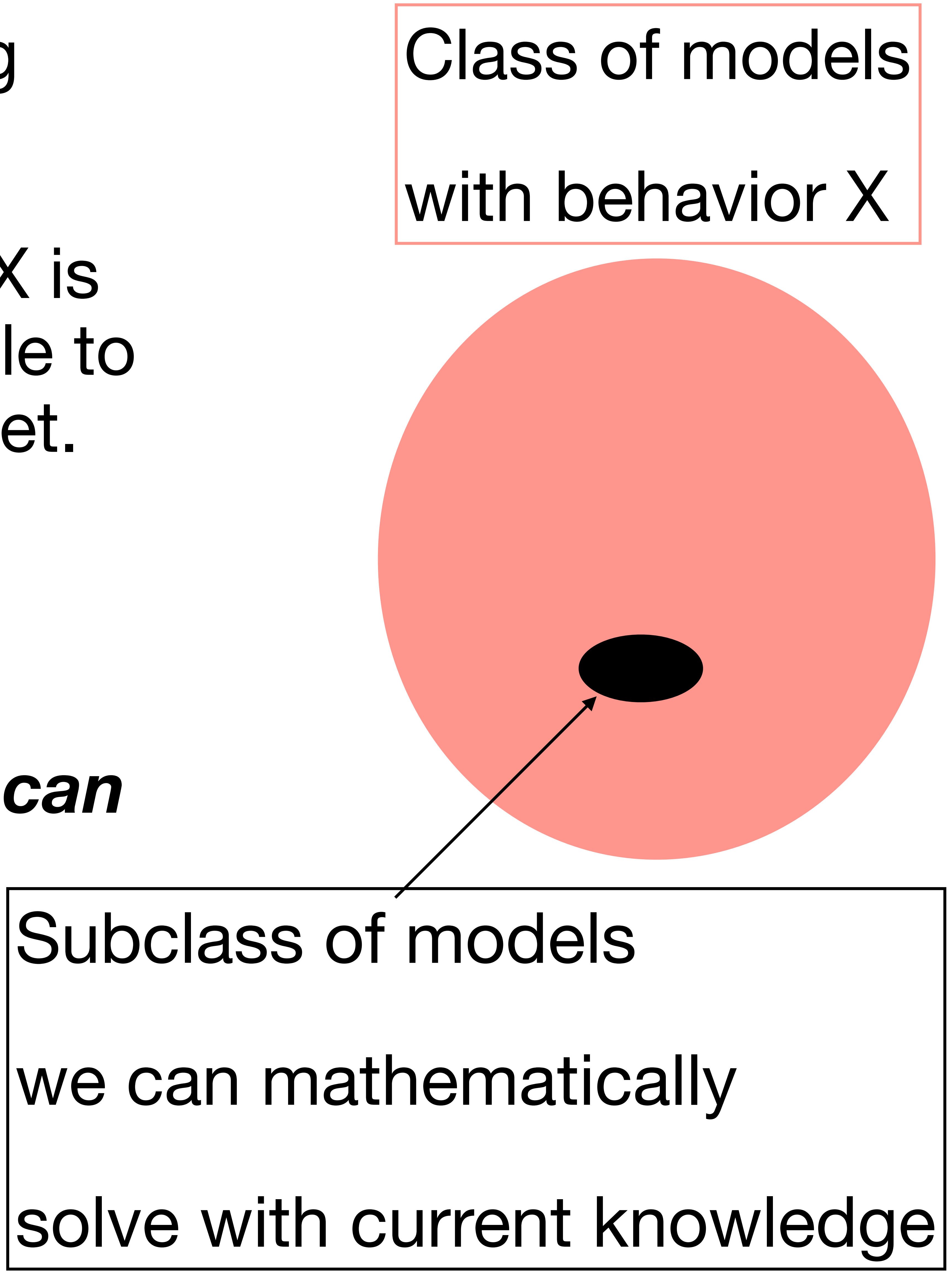
Actually, we will see: ℓ_2 arises purely from geometry of **overfitting.**

- More precise story: overfitting transforms the LAD loss into the Huber and ultimately the squared loss.

We will see this through more general results…

How general will we go?

- The focus of this talk/line of work is understanding **optimal generalization bounds**.
- Common reality of mathematics: even if we think X is true for a **big class** of models, we may only be able to **rigorously prove** X for a small well-behaved subset.
    - e.g. special solvable/integrable models in “universality classes” in physics
- ***I will make somewhat strong assumptions so I can solve for what happens precisely.***
    - In particular, Gaussianity of data…



Rmk: some universality observable

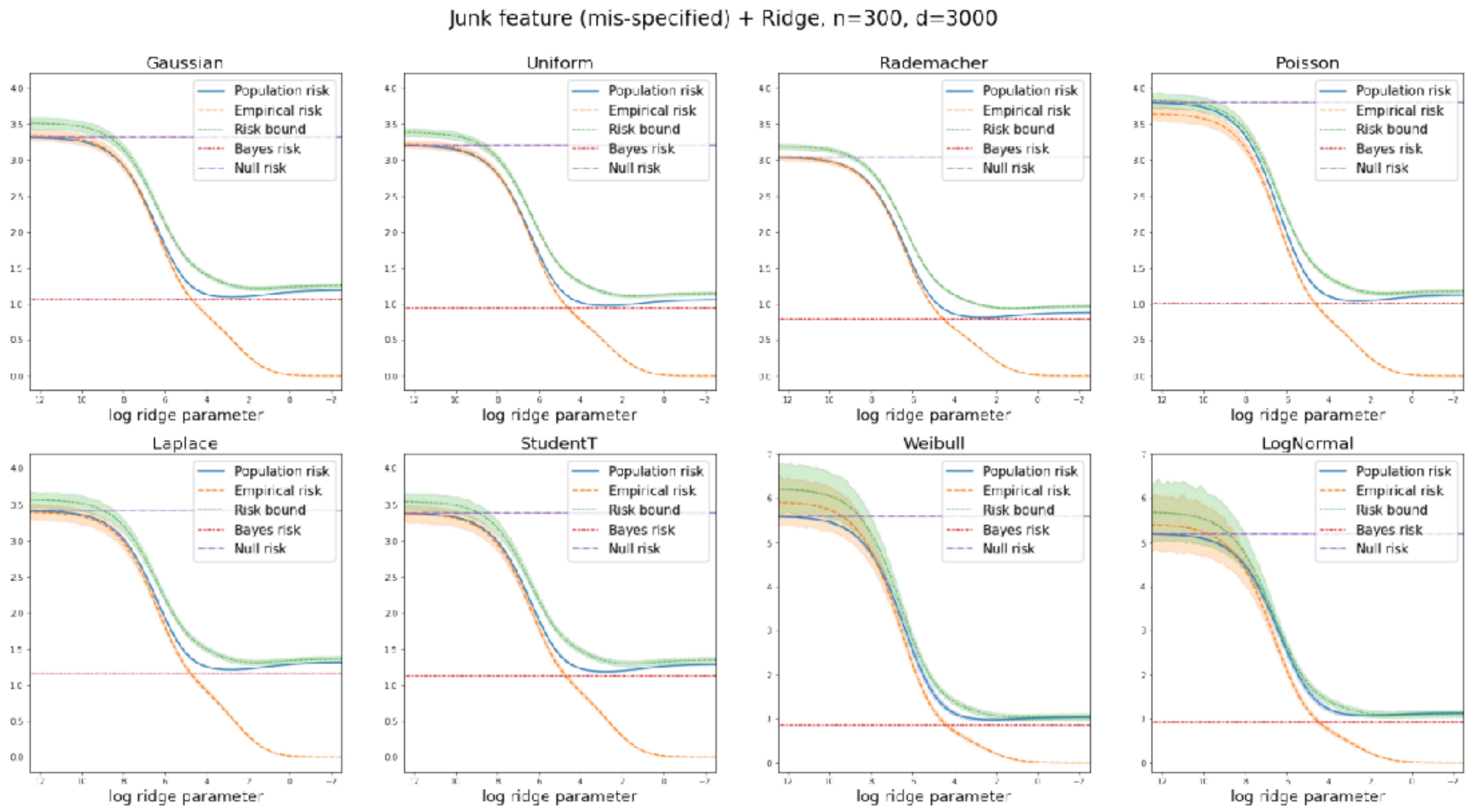
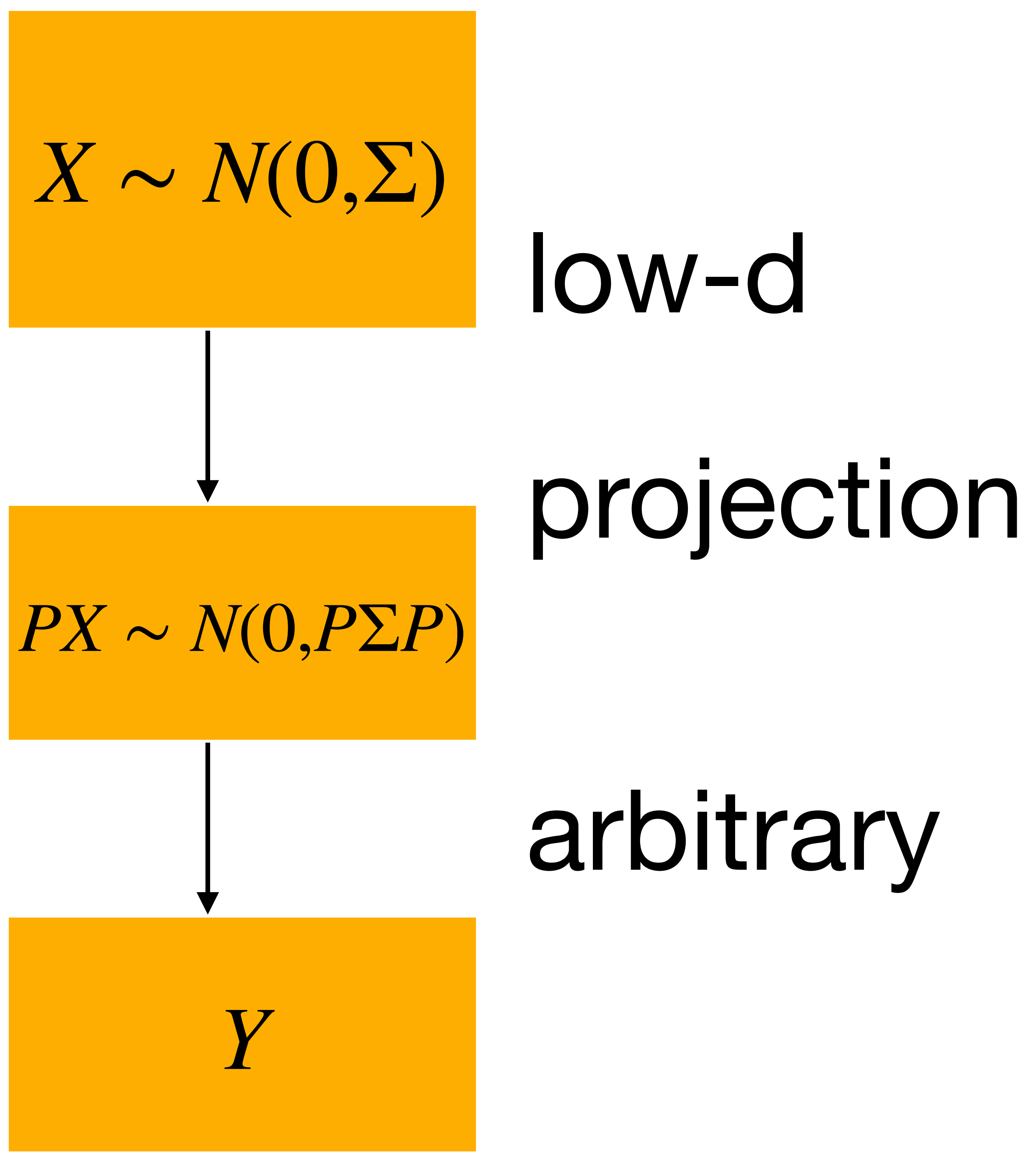


Figure 4: Ridge regression with junk features ($n = 300, d = 3000$). In the junk features setting, as predicted in section 6, the test error curve is essentially flat once the regularization is small enough to fit the signal, and we get nearly optimal population risk as long as we do not over-regularize the predictor. The test error curve can be expected to be more flat with increasing d. This phenomenon is also consistent across different feature distributions and label generating processes. Our bound (19) closely tracks the performance of ridge regression along the entire regularization path.

Formal generative setting

- Data points are $X \sim N(0, \Sigma)$
- Label Y is generated in an arbitrary way based on a low-dimensional projection of X.
  - I.e. $Y = f(\xi, \langle v_1^*, X \rangle, \ldots, \langle v_k^*, X \rangle)$ for a noise variable ξ and for k vectors $v_1^*, \ldots, v_k^*$.
  - This is called a **multi-index model** in statistics.
  - TODO: some trick to get rid of this assumption?
- Our fit models are generally *misspecified*



Moreau envelope + generalization bound

(will be explained!!)

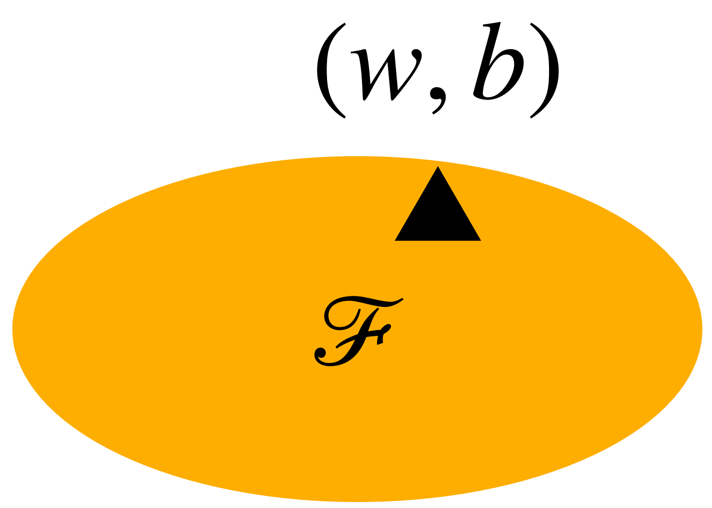
Definition 1. The *Moreau envelope* of $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with parameter $\lambda \in \mathbb{R}^+$ is defined as the function $f_\lambda : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ given by

$$f_\lambda(\hat{y}, y) = \inf_u \, f(u, y) + \lambda(u - \hat{y})^2. \tag{3}$$

- The minimizer is called the “proximal operator”. Standard objects in convex analysis.
- Then $\forall (w, b) \in \mathcal{F}$ one can prove the following (one-sided) generalization bound $\forall \lambda \geq 0$:

$$\mathbb{E} f_\lambda(\langle w, X\rangle + b, Y) \leq \hat{\mathbb{E}} f(\langle w, X\rangle + b, Y) + \lambda \mathcal{R}_n^2$$

Test error (envelope) Train error Rad. Complexity



where $(w, b) \in \mathcal{F}$ and $\mathcal{R}_n = \mathbb{E} \sup_{w,b} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i(\langle w, X_i\rangle + b)$ for $\epsilon \sim Uni\{\pm 1\}^n$

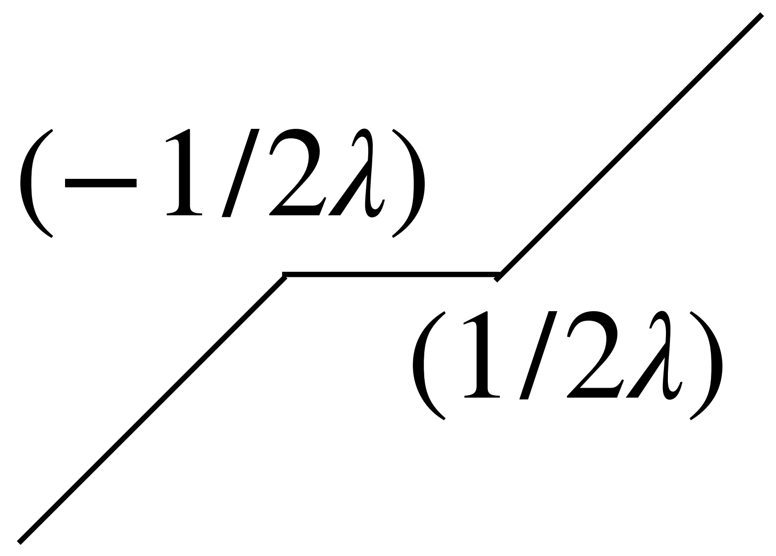
Example!

Moreau Envelope

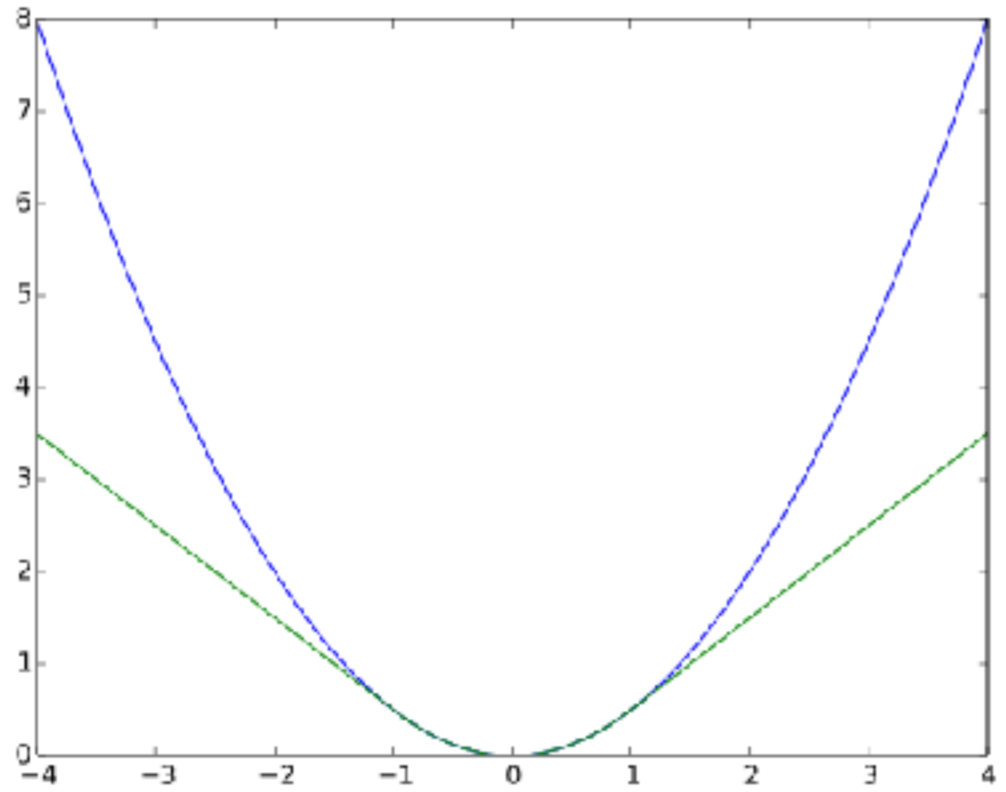
Definition 1. The *Moreau envelope* of $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with parameter $\lambda \in \mathbb{R}^+$ is defined as the function $f_\lambda : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ given by

$$f_\lambda(\hat{y}, y) = \inf_u f(u, y) + \lambda(u - \hat{y})^2. \tag{3}$$

- Consider $f(\hat{y}, y) = |\hat{y} - y|$ the LAD loss
- Solve for the proximal operator by setting derivative to zero
  - Minimizer is 0 for $|\hat{y}| \leq 1/2\lambda$ and otherwise $\hat{y} - sgn(\hat{y})(1/2\lambda)$



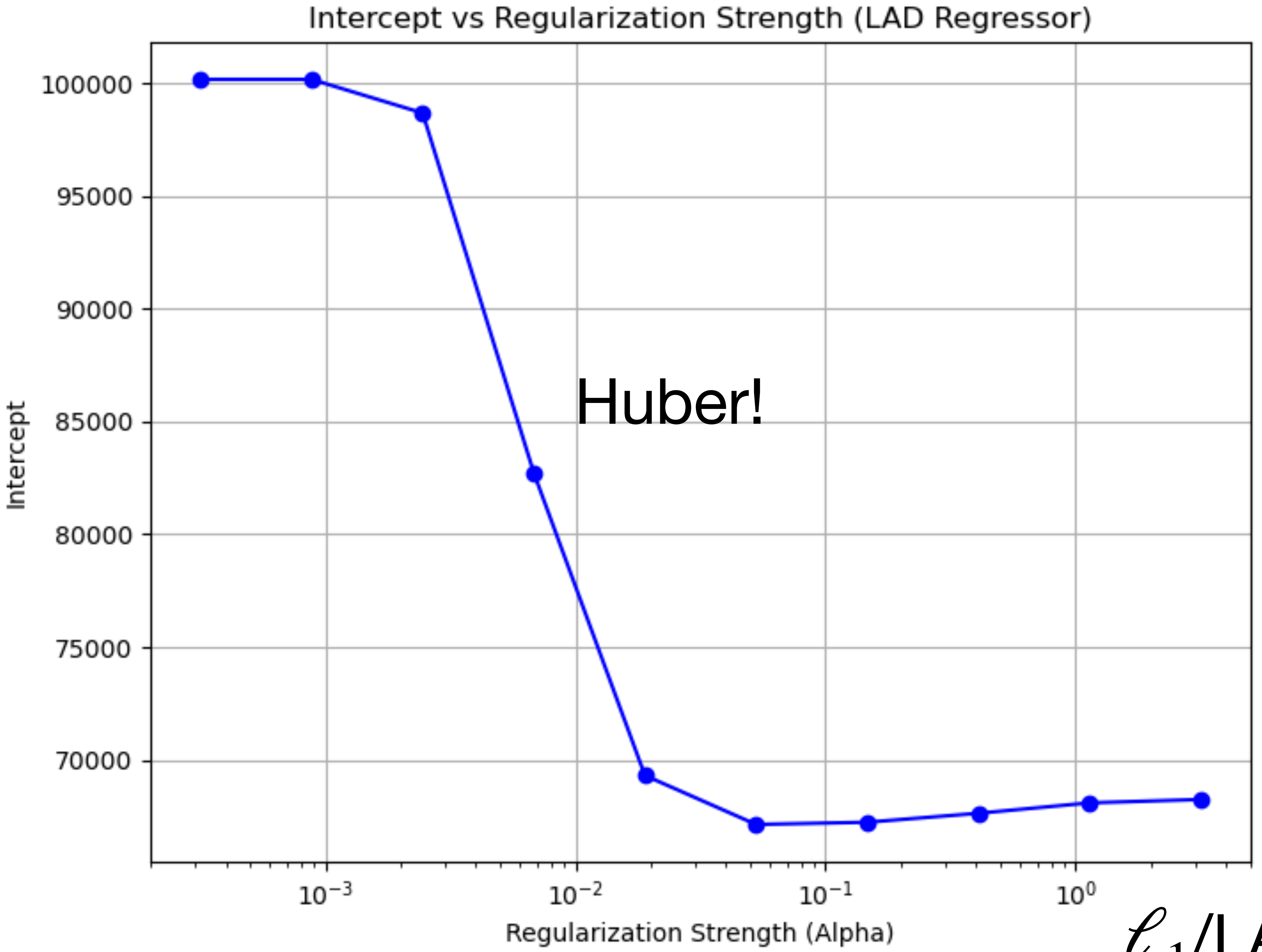
- Moreau envelope is 2λ times the $1/2\lambda$-**Huber loss**!
  - Quadratic for $|\hat{y}| \leq 1/2\lambda$, then linear
  - $\lambda \to 0$ then $f_\lambda \approx \lambda \times (\hat{y})^2$



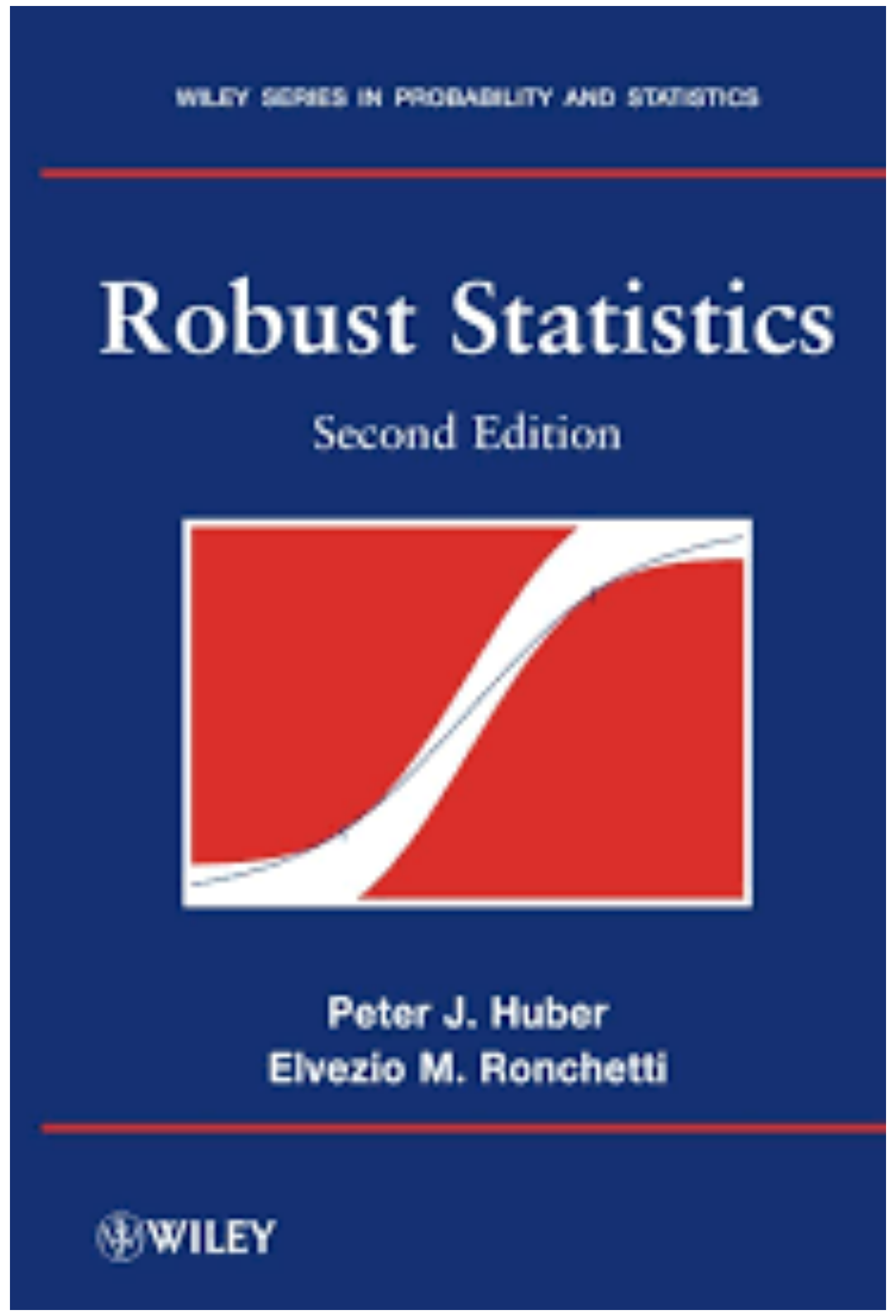
Matching the losses to example…

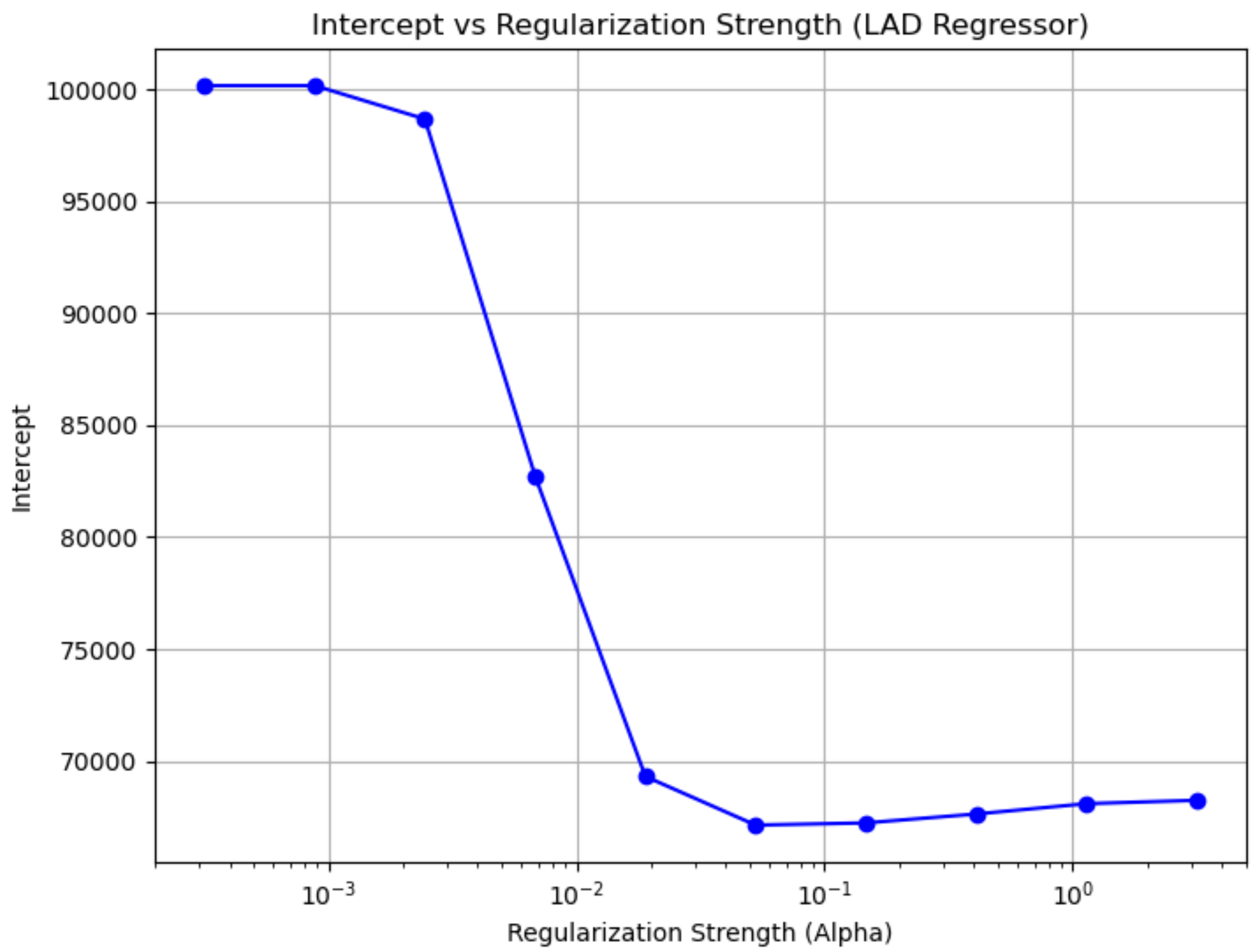
(not fully explained yet…)

ℓ_2/OLS



ℓ_1/LAD





For each point along the curve,

there is a corresponding value of λ

such that intercept is Huber median.

(moreover such that for “localized $\mathscr{F}$”,

below inequality is close

to equality…)

$$\mathbb{E} f_\lambda(\langle w, X\rangle + b, Y) \leq \hat{\mathbb{E}} f(\langle w, X\rangle + b, Y) + \lambda \mathscr{R}_n^2$$

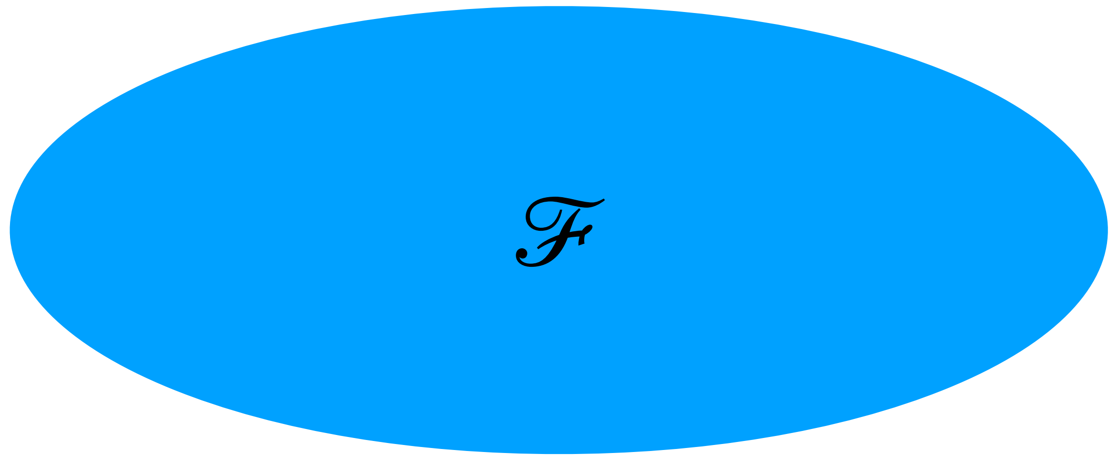
Test error (envelope) Train error Rad. Complexity

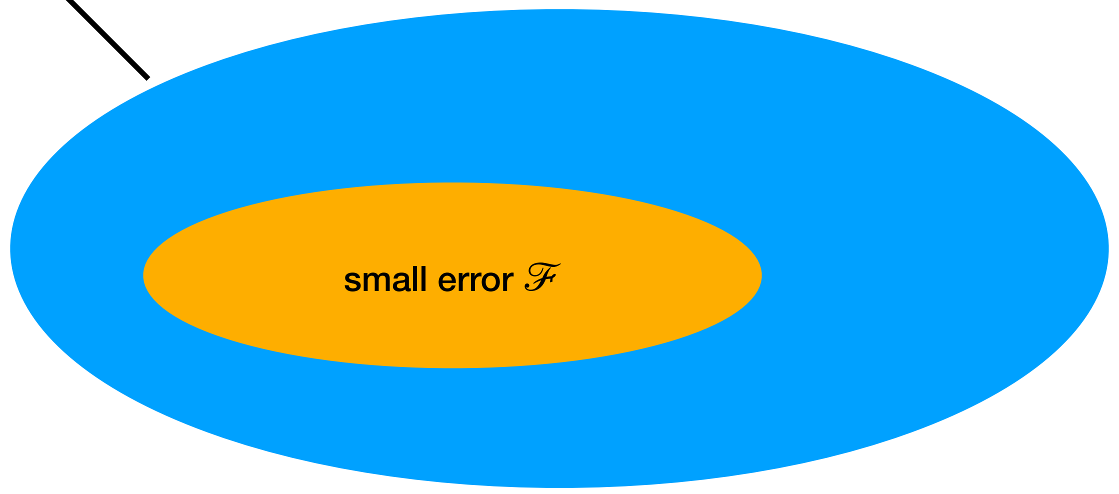
A bit more precisely…

- In this setting, “optimal generalization argument” is given by combining our bound with complexity of *localized* sets
  - I.e. we can show “localized Rademacher complexities” determine **sharp** bound in proportional asymptotics and other settings
- Localization: here means you apply the bound (for all λ) over and over…

$$\mathbb{E} f_\lambda(\langle w, X\rangle + b, Y) \leq \hat{\mathbb{E}} f(\langle w, X\rangle + b, Y) + \lambda \mathcal{R}_n(\mathcal{F}_r)^2$$

Test error (envelope) Train error Rad. Complexity





Saturation is an upper bound on train error

Generalization bound is a **lower bound on training error**.

- By "optimal generalization bound", we mean it is saturated by ERM.

$$\boxed{\max_{\lambda \geq 0} \left[\underbrace{\mathbb{E} f_\lambda(\langle w, X \rangle + b, Y)}_{\text{Test error (envelope)}} - \lambda \underbrace{\mathscr{R}_n(\mathscr{F}_{r(w)})^2}_{\text{Local Rad. Complexity}} \right] \leq \underbrace{\hat{\mathbb{E}} f(\langle w, X \rangle + b, Y)}_{\text{Train error}}}$$

For **convex ERM** $(\hat{w}, \hat{b}) = \arg\min \hat{\mathbb{E}} f(\langle w, X \rangle + b, Y)$ there is a **dual upper bound** on the training error for some localized ball $\mathscr{F}_r = \{f \in \mathscr{F} : \|f - f^*\|_{L_2} \leq r\}$ …

The statement and proof is closely related to "Convex Gaussian Minmax Theorem" (CGMT) and also predictions obtainable from Approximate Message Passing (AMP). (These existing frameworks tell us that Moreau envelopes/proximal operators are key.)

For reference. High-dimensional M-estimation <-> Moreau envelopes, proximal operators has a big literature… (e.g. [Stojnic '12])

Our goal is to extend this to the Moreau envelope generalization theory…

Convex Analysis: For a convex function $f : \mathbb{R}^n \to \mathbb{R}$, we let $\partial f(\mathbf{x})$ denote the subdifferential of f at $\mathbf{x}$ and $f^*(\mathbf{y}) = \sup_{\mathbf{x}} \mathbf{y}^T\mathbf{x} - f(\mathbf{x})$ its Fenchel conjugate. The Moreau envelope function of f at $\mathbf{x}$ with parameter τ is defined by

$$\mathrm{e}_f(\mathbf{x};\tau) := \min_{\mathbf{v}} \frac{1}{2\tau}\|\mathbf{x}-\mathbf{v}\|_2^2 + f(\mathbf{v}).$$

Sequence of problem instances: Formally, our result applies on a sequence of problem instances $\{\mathbf{x}_0, \mathbf{A}, \mathbf{z}, \mathcal{L}, f, m\}_{n\in\mathbb{N}}$ indexed by n such that the properties listed above hold for all members of the sequence and for all $n \in \mathbb{N}$. (We do not write out the subscripts n for arguments of the sequence to not overload notation). Every such sequence generates a sequence $\{\mathbf{y}, \hat{\mathbf{x}}\}_{n\in\mathbb{N}}$ where $\mathbf{y} := \mathbf{A}\mathbf{x}_0 + \mathbf{z}$, and,

$$\hat{\mathbf{x}} := \arg\min_{\mathbf{x}} \mathcal{L}(\mathbf{y}-\mathbf{A}\mathbf{x}) + \lambda f(\mathbf{x}). \tag{2}$$

Assumption 1 (Summary functionals L and F). *We say that Assumption 1 holds if:*

(a) *For all $c \in \mathbb{R}$ and $\tau > 0$, there exist continuous functions $L : \mathbb{R}\times\mathbb{R}_{>0} \to \mathbb{R}$ and $F : \mathbb{R}\times\mathbb{R}_{>0} \to \mathbb{R}$ such that*[5]

$$\frac{1}{m}\{\mathrm{e}_{\mathcal{L}}(c\mathbf{g}+\mathbf{z};\tau) - \mathcal{L}(\mathbf{z})\} \xrightarrow{P} L(c,\tau) \quad \text{and} \quad \frac{1}{n}\{\mathrm{e}_f(c\mathbf{h}+\mathbf{x}_0;\tau) - f(\mathbf{x}_0)\} \xrightarrow{P} F(c,\tau),$$

(b) *At least one of the following holds. There exists constant $C > 0$ such that $\frac{\|\mathbf{z}\|_2}{\sqrt{m}} \le C$ with probability approaching 1 (w.p.a.1), or, $\sup_{\mathbf{v}\in\mathbb{R}^m} \sup_{\mathbf{s}\in\partial\mathcal{L}(\mathbf{v})} \|\mathbf{s}\|_2 < \infty$ for all $m \in \mathbb{N}$.*

Theorem 3.1 (Master Theorem). *Let $\hat{\mathbf{x}}$ be a minimizer of the Generalized M-estimator in (2) for fixed $\lambda > 0$. Further let Assumptions 1 and 2 hold. If the following convex-concave minimax scalar optimization*

$$\inf_{\substack{\alpha\ge0\\ \tau_g>0}} \sup_{\substack{\beta\ge0\\ \tau_h>0}} \mathcal{D}(\alpha,\tau_g,\beta,\tau_h) := \frac{\beta\tau_g}{2} + \delta\cdot L\left(\alpha,\frac{\tau_g}{\beta}\right) - \frac{\alpha\tau_h}{2} - \frac{\alpha\beta^2}{2\tau_h} + \lambda\cdot F\left(\frac{\alpha\beta}{\tau_h},\frac{\alpha\lambda}{\tau_h}\right). \tag{3}$$

has a unique minimizer α_*, then, it holds in probability that

$$\lim_{n\to\infty}\frac{1}{n}\|\hat{\mathbf{x}}-\mathbf{x}_0\|_2^2 = \alpha_*^2.$$

[…,Thrampoulidis-Abbasi-Hassibi '16, …]

CGMT

see e.g. […,Berthier-Montanari-Nyugen '17] for analogous AMP theory

A key special case: the theory for sqrt-Lipschitz losses

Sqrt-Lipschitz Generalization Theory

- Bound simplifies dramatically if the loss is **sqrt-Lipschitz** (relaxation of smooth)
- Assume $\sqrt{f}$ is H-Lipschitz. EXERCISE: prove the below fact

Proposition 1. *A function $f : \mathbb{R} \to \mathbb{R}$ is non-negative and $\sqrt{H}$-square-root-Lipschitz if and only if for any $x \in \mathbb{R}$ and $\lambda \geq 0$, it holds that*

$$f_\lambda(x) \geq \frac{\lambda}{\lambda + H} f(x). \tag{35}$$

- So

$$\frac{\lambda}{\lambda + H}\mathbb{E}f(\langle w, X\rangle + b, Y) \leq \hat{\mathbb{E}}f(\langle w, X\rangle + b, Y) + \lambda \mathcal{R}_n^2$$

$$\mathbb{E}f(\langle w, X\rangle + b, Y) \leq (1 + H/\lambda)\hat{\mathbb{E}}f(\langle w, X\rangle + b, Y) + (1 + \lambda/H)H\mathcal{R}_n^2$$

- Choosing λ to balance terms yields

$$\sqrt{\mathbb{E}f(\langle w, X\rangle + b, Y)} \leq \sqrt{\hat{\mathbb{E}}f(\langle w, X\rangle + b, Y)} + \sqrt{H}\mathcal{R}_n$$

“Easy” example: linear regression w/ squared loss

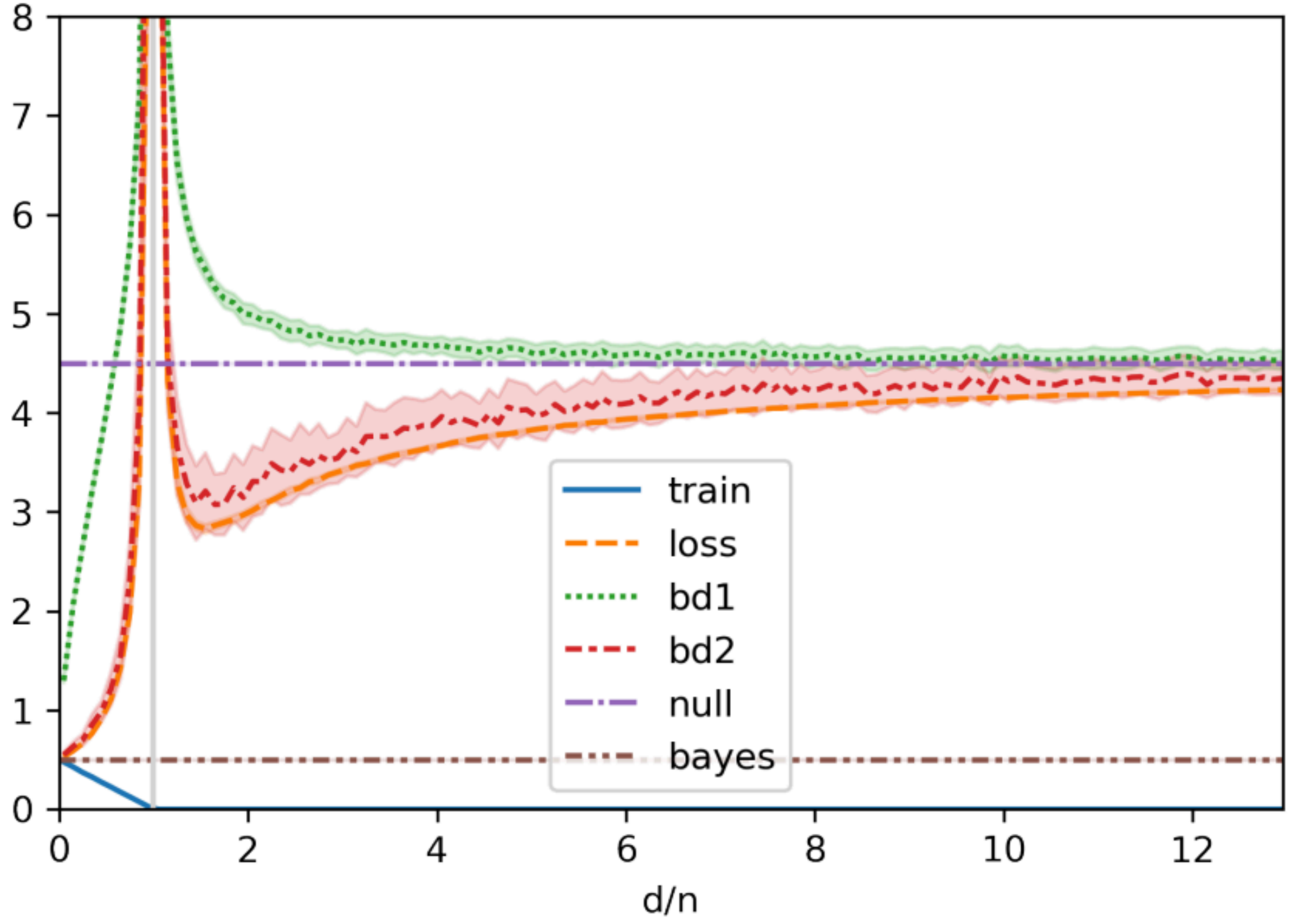
- Square loss = 1-sqrt Lipschitz (exercise: any H-smooth nonnegative loss is H-sqrt Lipschitz), so

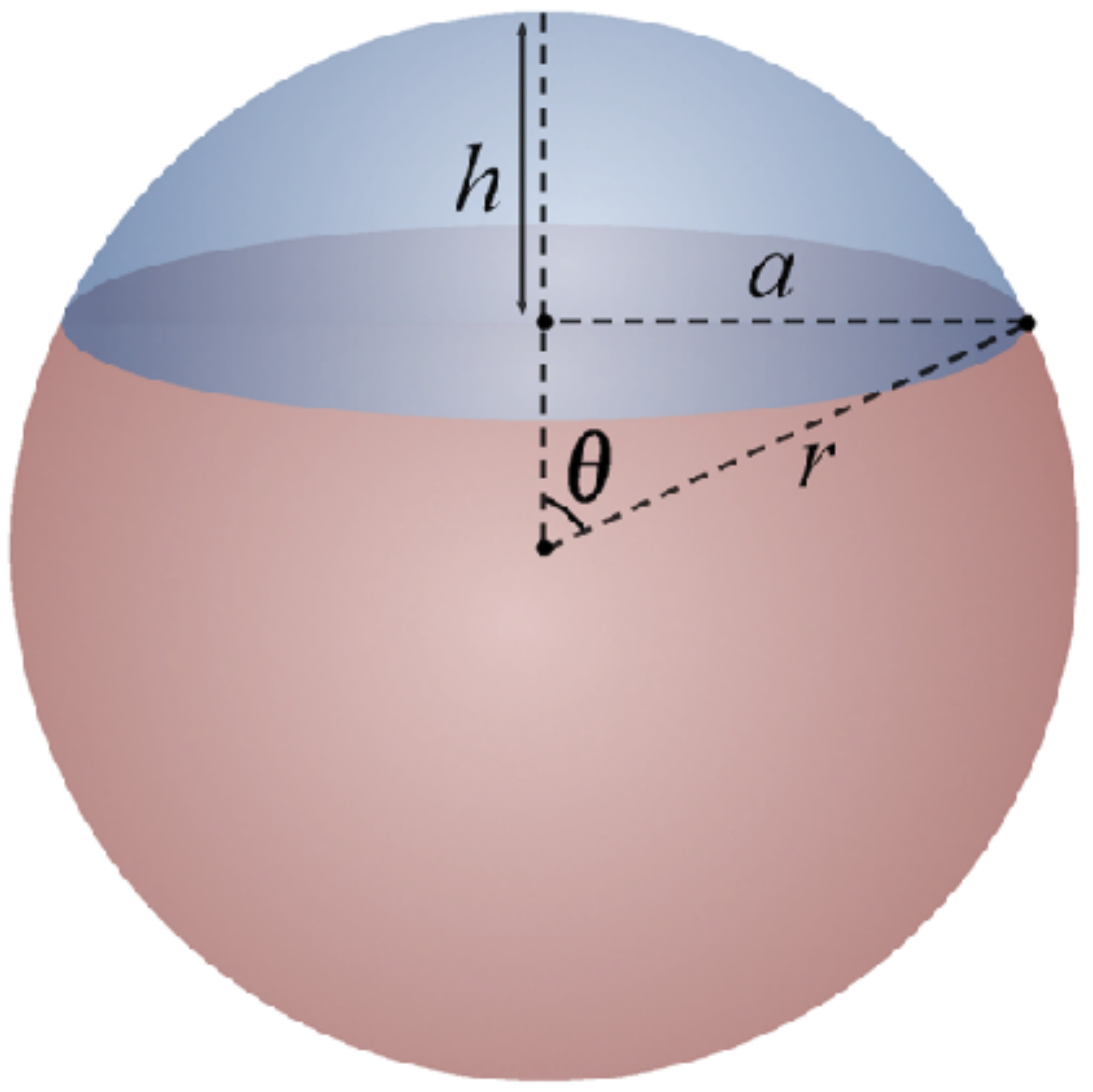
$$\sqrt{\mathbb{E}(\langle w, X\rangle + b - Y)^2} \leq \sqrt{\hat{\mathbb{E}}(\langle w, X\rangle + b - Y)^2} + \mathscr{R}_n$$

- Optimal “optimistic rate” for squared loss regression [Z**K**SS ’24]
- Recovers benign overfitting a la [BLTT ’21], etc. (next slide)

Example: OLS on $N(0, I_d)$ data

Localized bound (red) is close to saturated





Localization <->

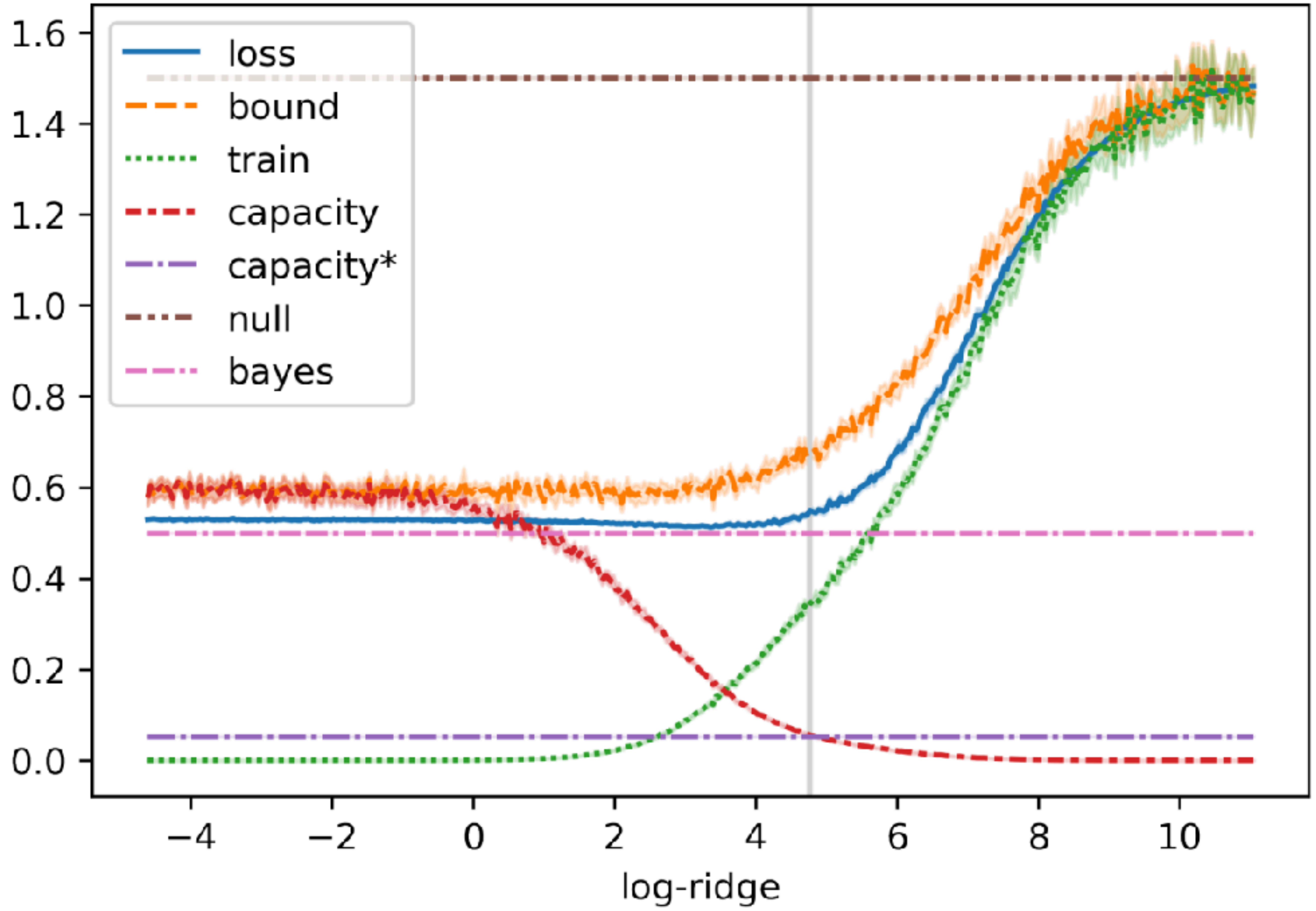
intersection of ℓ_2 ball

with L_2(P) ball

Example (benign overfitting)

(in general, can handle benign covariances as in [BLTT '19])

- Completely overfit, but test error close to optimal (and generalization bound gets it)



$n = 600 \quad d/n = 20$

Benign Overfitting Conditions

[Bartlett et al '19, Tsigler-Bartlett '20]: If $\|w^*\| = 1$, $\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}$ and:

$$\mathrm{rank}(\Sigma_1) = o(n) \qquad \mathrm{Tr}(\Sigma_2) = o(n) \qquad \frac{\mathrm{Tr}(\Sigma_2^2)^2}{\mathrm{Tr}(\Sigma_2^2)} = \omega(n)$$

Then $\hat{w} = \arg\min_{\mathbb{Y}=\mathbb{X}w} \|w\|$ is consistent: $\mathbb{E}(Y - \hat{w} \cdot X)^2 \to \min_w \mathbb{E}(Y - w \cdot X)^2$.

(Performance of squared loss oracle)

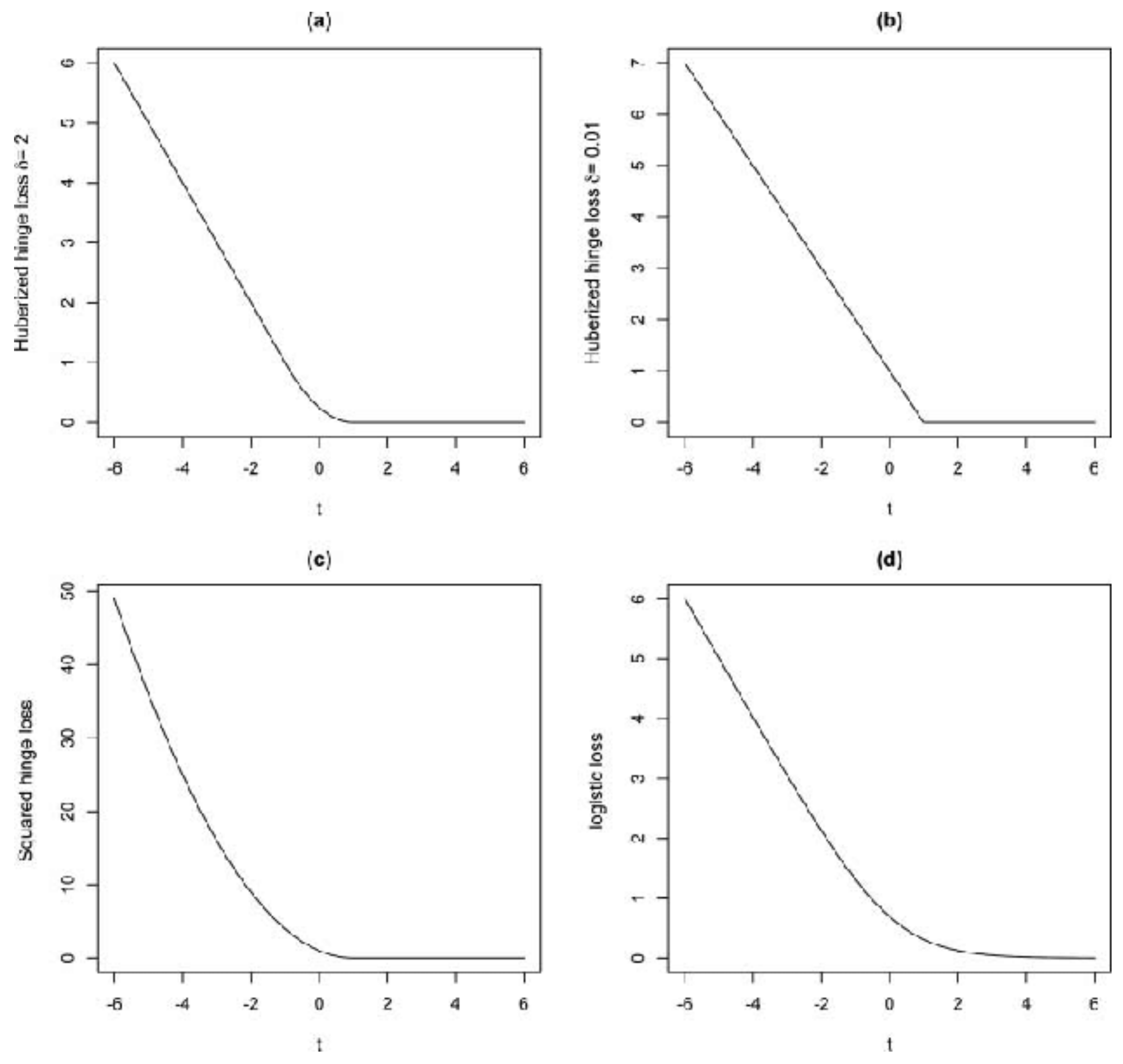
Bounding Rademacher w/ norm recovers this:

$$\mathbb{E}(Y - \hat{w} \cdot X)^2 \leq \mathscr{R}_n^2 \leq \frac{\|\hat{w}\|^2 \mathbb{E}_{x \sim N(0,\Sigma_2)} \|x\|^2}{n} \approx \sigma^2$$

because we prove: $\|\hat{w}\|^2 \leq (1 + o(1)) \dfrac{\sigma^2 n}{\mathbb{E}_{x \sim N(0,\Sigma_2)} \|x\|^2}$

Classification

- Analogue of squared loss for binary classification?
- **Squared hinge loss** $\ell(\hat{y}, y) = \max(0, 1 - \hat{y}y)^2$ **!**
    - "Huberization" if you optimize standard hinge loss in training
        - Similar behavior if you start with logistic loss
    - Key in our proof of benign overfitting results for classification (in fact, the same proof handles regression and classification….)



Huberized hinge loss

Figure 2, [Yang-Zou '12]

Moreau theory key for sharp LAD analysis

- Old school generalization theory for LAD: by contraction principle, $\mathbb{E}\,|\,Y - \langle w, X\rangle\,| \leq \hat{\mathbb{E}}\,|\,Y - \langle w, X\rangle\,| + 2\mathscr{R}_n$. Can replace 2 by 1, but cannot shrink further.
- If we run unregularized LAD in above setting with dimension $d_J \to \infty$, it is consistent so $\mathbb{E}\,|\,Y - \langle \hat{w}, X\rangle\,| \to \mathbb{E}_{Z\sim N(0,\sigma^2)}|Z| = \sigma\sqrt{2/\pi} \approx 0.798\sigma$
- Moreau envelope theory can prove this, but **old school bound only gives generalization gap of $\sigma \gg 0.798\sigma$** on smallest norm ball containing $\hat{w}$. Why?
  - **Taking $\lambda \to 0$ focuses Moreau envelope bound on low-training error predictors.** Old-school bound is hurt by large-training error predictor.
  - broader theme: *models with different training errors have different generalization gaps. "optimism"*

E.4 Sharpness of Improved Lipschitz Contraction

In this section, we show that the Lipschitz contraction bound (11) for 1-Lipschitz loss functions f,

$$(1 - o(1))L_f(w) \leq \hat{L}_f(w) + \sqrt{\frac{C_\delta(w)^2}{n}}$$

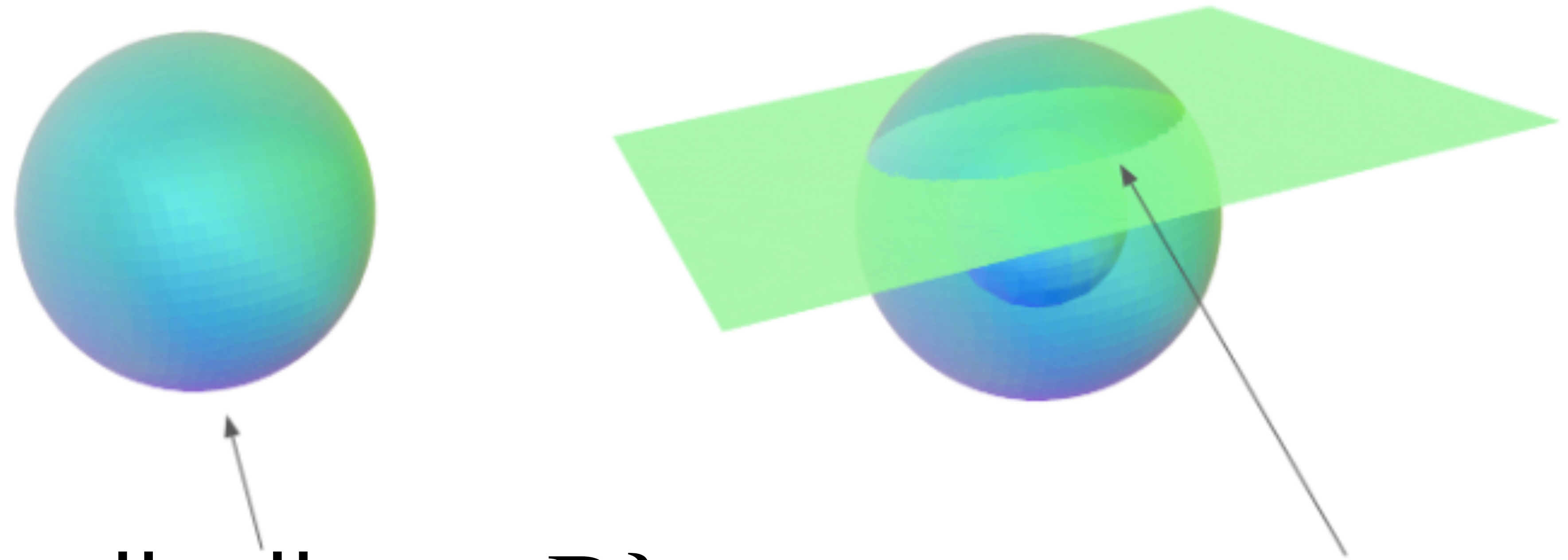
has sharp constants in the case of the L_1 loss $f(\hat{y}, y) := |y - \hat{y}|$. This shows that the only way to tighten the bound further is to consider one with a different functional form (e.g. the Moreau envelope bound with the Huber test loss). In particular, the Moreau envelope version of the bound is significantly more useful when looking at interpolators.

Data Distribution. We will show tightness in the setting of the junk features model. Let's consider

$$x \sim \mathcal{N}(0, \Sigma), \quad y \sim \mathcal{N}(0, \sigma^2)$$

where the response y is independent of the covariate x and the covariance Σ is given by

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & \frac{\lambda_n}{d_J} I_{d_J} \end{bmatrix}.$$



$\{w : \|w\|_2 \leq B\}$ $\{w : \|w\|_2 \leq B, \hat{L}(w) = 0\}$

“Hard” example: ReLU regression

- We saw that if you interpolate with a linear model, the ‘correct’ test loss is squared loss. (e.g. consistent under ‘benign overfitting’ covariance)
- What if you interpolate with a single ReLU neuron? $\sigma(\langle w, x \rangle + b)$
- “Obvious generalization” of linear case: loss is $(\sigma(\hat{y}) - y)^2$… it’s wrong!
- Correct answer is **discontinuous !** 😱 $y = \epsilon$ very different from $y = 0$

Our analysis will show that the consistent loss for benign overfitting with ReLU regression is

$$f(\hat{y}, y) = \begin{cases} (\hat{y} - y)^2 & \text{if} \quad y > 0 \\ \sigma(\hat{y})^2 & \text{if} \quad y = 0. \end{cases} \qquad (13)$$

Why discontinuous?

Our analysis will show that the consistent loss for benign overfitting with ReLU regression is

$$f(\hat{y}, y) = \begin{cases} (\hat{y} - y)^2 & \text{if} \quad y > 0 \\ \sigma(\hat{y})^2 & \text{if} \quad y = 0. \end{cases} \tag{13}$$

- Suppose $\hat{y} = -100$ and $y = 0$. Then $\sigma(\hat{y}) = y$ so we already interpolate.
- But if $\hat{y} = -100$ and $y = 0.01$, then $\sigma(\hat{y}) \neq y$ so if we want to change our prediction to interpolate, then we have to put in a lot of effort… (100.01^2)
- Note: loss is sqrt-Lipschitz but not twice-diferentiable (“smooth”)

The ReLU case: food for thought

Our analysis will show that the consistent loss for benign overfitting with ReLU regression is

$$f(\hat{y}, y) = \begin{cases} (\hat{y} - y)^2 & \text{if} \quad y > 0 \\ \sigma(\hat{y})^2 & \text{if} \quad y = 0. \end{cases} \qquad (13)$$

- **This loss is a function** of the *preactivation* $\hat{y}$
    - Not determined by model output $\sigma(\hat{y})$!
- Shows us that **model architecture** plays a key role in the loss
    - But agnostic to regularization (e.g. ℓ_1 vs ℓ_2 penalization)…
- Other settings? Only other solved case is phase retrieval model.

(Phase retrieval)

- fit a model $x \mapsto |\langle w, x\rangle + b|$ to nonnegative labels Y
- consistent loss is squared loss $(|\hat{y}| - y)^2$
- ERM is nonconvex, but we can analyze its generalization performance anyway
  - (even though e.g. Convex Gaussian Minmax Theorem requires convexity…)
  - can prove natural benign overfitting results in phase retrieval

Proof idea of main result

$$\boxed{\mathbb{E}f_\lambda(\langle w, X\rangle + b, Y) \leq \hat{\mathbb{E}}f(\langle w, X\rangle + b, Y) + \lambda\mathcal{R}_n(\mathcal{F})^2}$$

Test error (envelope) Train error Rad. Complexity

- Proved based on *Gordon's Theorem (gaussian minmax theorem)*
  - cf *"Gaussian Processes and Almost Spherical Sections of Convex Bodies"*
- View generalization as **lower bound on stochastic process (training error).**

$$\mathbb{E}f_\lambda(\langle w, X\rangle + b, Y) - \lambda\mathcal{R}_n(\mathcal{F})^2 \leq \hat{\mathbb{E}}f(\langle w, X\rangle + b, Y)$$

(assuming, b = 0, Y = pure noise model for simplicity)

Let F(w) = $\mathbb{E}f_\lambda(\langle w, X\rangle + b, Y) - \lambda \mathscr{R}_n(\mathscr{K})^2$**. Then**

$$\max_{w\in\mathscr{K},u} [F(w) - \hat{\mathbb{E}}f(\langle X, w\rangle, Y)]$$

$$= \max_{w\in\mathscr{K},u} \inf_{\gamma\in\mathbb{R}} [F(w) - \hat{\mathbb{E}}f(u, Y) + \langle \gamma, u - Xw\rangle]$$

$$\leq \max_{w\in\mathscr{K},u} \inf_{\gamma\in\mathbb{R}} [F(w) - \hat{\mathbb{E}}f(u, Y) + \langle \gamma, u\rangle - \|\gamma\|_2\langle G, \Sigma^{1/2}w\rangle - \|\Sigma^{1/2}w\|_2\langle H, \gamma\rangle] \qquad \text{(GMT!)}$$

$$= \max_{w\in\mathscr{K},u} \inf_{\gamma\in\mathbb{R}} [F(w) - \hat{\mathbb{E}}f(u, Y) + \langle \gamma, u - \|\Sigma^{1/2}w\|_2 H\rangle - \|\gamma\|_2\langle G, \Sigma^{1/2}w\rangle]$$

$$= \max_{w\in\mathscr{K},u,\|u-\|\Sigma^{1/2}w\|_2H\|_2\leq\langle G,\Sigma^{1/2}w\rangle} [F(w) - \hat{\mathbb{E}}f(u, Y)]$$

$$= \max_{w\in\mathscr{K}}[F(w) - \min_{r:\|r\|_2\leq\langle G,\Sigma^{1/2}w\rangle} \hat{\mathbb{E}}f(\|\Sigma^{1/2}w\|_2 H + r, Y)]$$

$$= \max_{w\in\mathscr{K}}[F(w) - \min_{r:\|r\|_2\leq\langle G,\Sigma^{1/2}w\rangle} [\hat{\mathbb{E}}f(\|\Sigma^{1/2}w\|_2 H + r, Y) + \lambda r^2 - \lambda r^2]]$$

$$= \max_{w \in \mathscr{K}} [F(w) - \min_{r: \|r\|_2 \le \langle G, \Sigma^{1/2} w \rangle} [\hat{\mathbb{E}} f(\|\Sigma^{1/2} w\|_2 H + r, Y) + \lambda \|r\|^2 - \lambda \|r\|^2]]$$

$$\le \max_{w \in \mathscr{K}} [F(w) - \min_{r} [\hat{\mathbb{E}} f(\|\Sigma^{1/2} w\|_2 H + r, Y) + \lambda \|r\|^2] + \lambda \max_{w} \langle G, \Sigma^{1/2} w \rangle^2]]$$

$$\le \max_{w \in \mathscr{K}} [F(w) - \hat{\mathbb{E}} f_{\lambda}(\|\Sigma^{1/2} w\|_2 H + r, Y) + \lambda \max_{w} \langle G, \Sigma^{1/2} w \rangle^2]]$$

Next slide: show this “Gaussian width”

equals Rademacher complexity $\mathscr{R}_n$

≈ 0 (by LLN)

$$\max_{w \in \mathscr{K}, u} [F(w) - \hat{\mathbb{E}} f(\langle X, w \rangle, Y)] \le 0$$

so F is a valid lower bound

on training error.

Gaussian Width is Rademacher Complexity of the Function Class

Rademacher complexity: how well can functions correlate with pure noise (**random signs**)? see e.g. [Bartlett-Mendelson '02], [SSS-SBD '14]

Expected Rademacher Complexity of $\mathscr{F} := \{x \mapsto \langle w, x\rangle : w \in \mathscr{K}\}$

$$\mathscr{R}_n(\mathscr{F}) := \mathbb{E}_{\substack{X_1,\ldots,X_n \sim \mathcal{N}(0,\Sigma) \\ \sigma \sim Uni(\{\pm 1\}^n)}} \left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \right| \right].$$

$$= \mathbb{E}_{\substack{X_1,\ldots,X_n \sim \mathcal{N}(0,\Sigma) \\ \sigma \sim Uni(\{\pm 1\}^n)}} \left[\sup_{w \in \mathscr{K}} \left| \frac{1}{\sqrt{n}} \left\langle w, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sigma_i X_i \right\rangle \right| \right].$$

$$= \mathbb{E}_{x \sim \mathcal{N}(0,\Sigma)} \left[\sup_{w \in \mathscr{K}} \left| \frac{1}{\sqrt{n}} \langle w, x \rangle \right| \right] = \frac{W(\Sigma^{1/2}\mathscr{K})}{\sqrt{n}}$$

Closing thoughts

- When we learn there is often some overfitting/memorization. When there is overfitting, may want to consider that the model's implicit/correct "test loss" may not be train…
- Interesting that overfitting can make model "care more" about outliers.
  - Moreau -> ℓ_2/squared hinge loss more sensitive than ℓ_1/logistic/hinge
  - Could be a good thing sometimes.
- Interesting that the "real objective" of model could be different from the training objective
- There is probably more to say…
- Thanks!