A Geometrical Phenomenon: Support Vector Machines and Linear Regression Coincide With Very High Dimensional Features

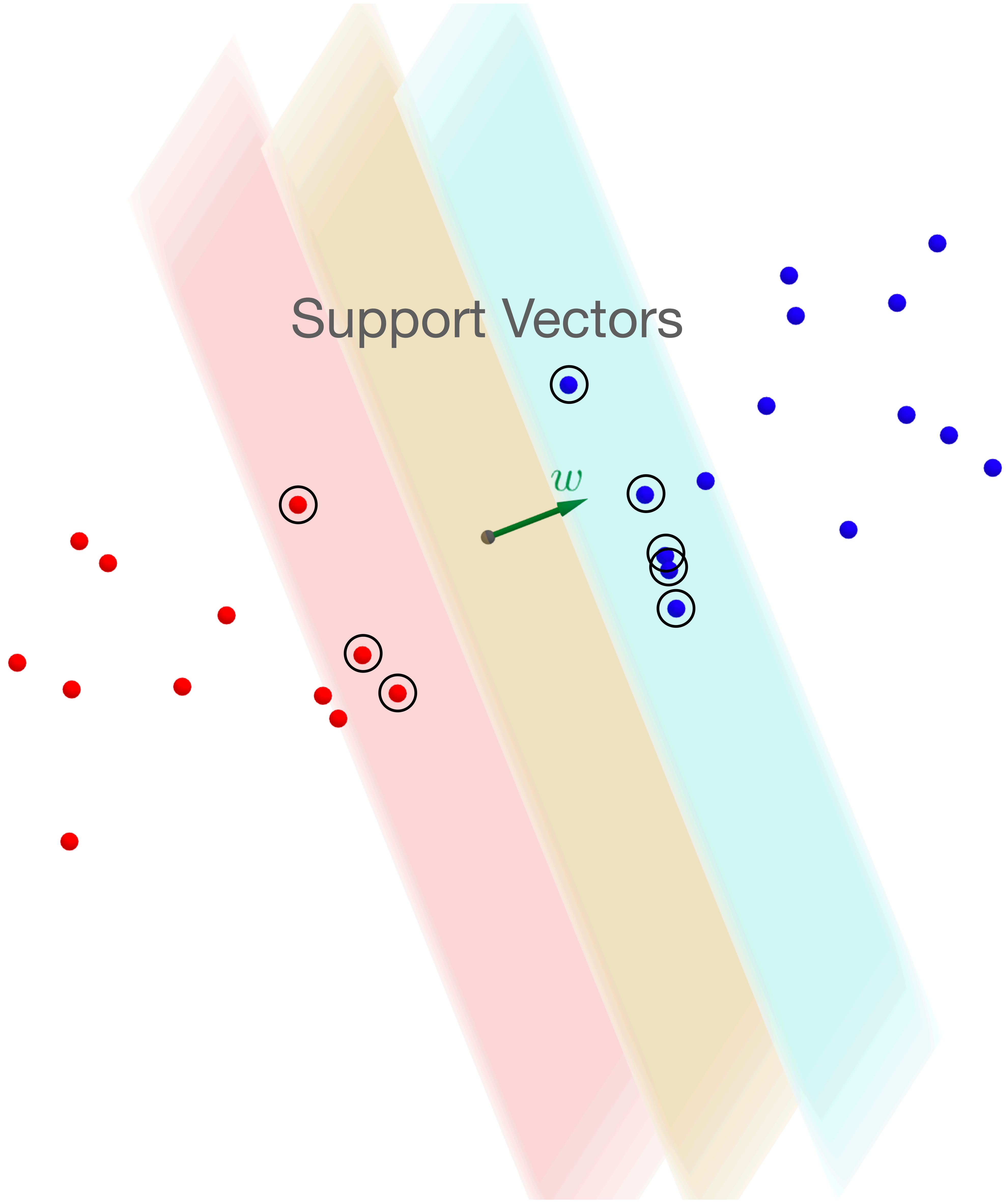
Navid Ardeshir

Columbia University, Department of Statistics

Based on joint work with Clayton Sanford and Daniel Hsu

Introduction

- **High Dimensional Regression and Classification**
  - **Regression:**
    Min Norm Linear Regression (OLS)
  - **Classification:**
    Max Margin Linear Classifier (SVM)



Surprise in High Dimensional Regression and Classification:

OLS = SVM

Support Vector Proliferation (SVP)

Outline

- Inductive bias of learning
  - Linear and logistic regression
- Classification vs. regression
  - **OLS** = **SVM** and its implications
  - Our results
  - Key lemma and geometrical intuition
  - Proof ideas
  - Empirical universality

Inductive bias

- The inductive bias is simply the set of assumptions that learner makes about inherent properties of the data.
- Deep learning practice:
  - Choice of architecture, e.g. CNN, Resnet18, etc.
  - Choice of **loss function**, e.g. square loss, logistic loss, etc.
  - Choice of optimization procedure, e.g. GD, SGD, Mirror Descent, etc.
- All these choices constitutes as inductive bias!

Inductive bias - regression

- **Question:** Given samples $(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n) \in \mathbb{R}^d \times \mathbb{R}$ what is the inductive bias of certain optimization procedures for **linear regression**?
- Linear Regression: $\mathscr{H} = \left\{x \mapsto w^\intercal x\right\}$.
- The goal of ERM learner is to find an estimator/classifier $h_w(x) = w^\intercal x$ such that it minimizes the empirical risk, $\hat{R}(h_w) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - h_w(\boldsymbol{x}_i))^2$.
- When $d > n$, there could be infinitely-many minimizers in $\arg\min_{h \in \mathscr{H}} \hat{R}(h)$.

Inductive bias - regression

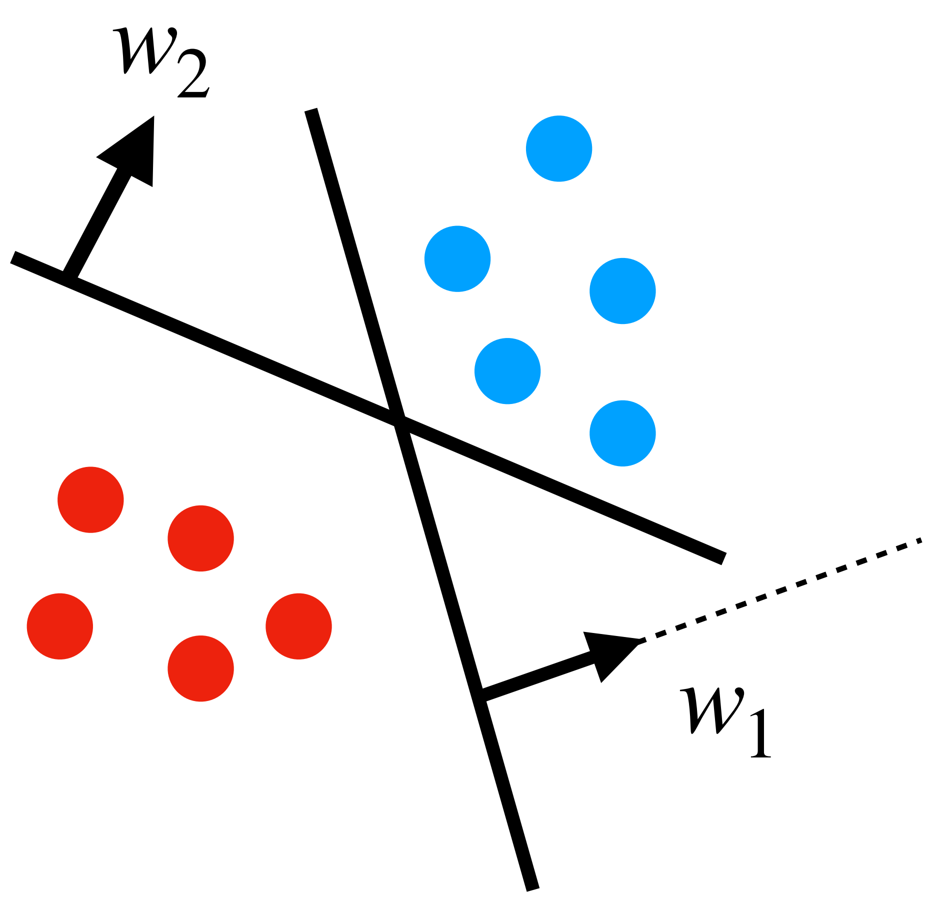
Theorem: [Werner Engl, et al. '96]
For a feasible set of linear equations, the evolution of GD with initialization at zero converges to the **minimum Euclidean norm linear interpolator** (**OLS**),

$$\lim_{t\to\infty} w_t = \arg\min_{w\in\mathbb{R}^d} \|w\|_2 \quad \text{s.t.} \quad w^\top \boldsymbol{x}_i = \boldsymbol{y}_i \, .$$

- Minimum ℓ_p-norm interpolators (ℓ_p-**OLS**) can be obtained by Steepest Descent on the dual norm [Gunasekar, et al. '18].

Inductive bias - classification

- **Question:** Given $(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n) \in \mathbb{R}^d \times \{\pm 1\}^n$ samples, what is the inductive bias of certain optimization procedures for **logistic regression**?
- Logistic regression: $\mathscr{H} = \left\{ x \mapsto w^\intercal x \mid w \in \mathbb{R}^d \right\}$.
- The goal is to minimize $\hat{R}(h_w) = \dfrac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-\boldsymbol{y}_i h_w(\boldsymbol{x}_i)})$.



- Linear separable: there exists a linear classifier with zero training classification error.
- When data is separable there may be infinitely-many empirical minimizers at infinity.

Inductive bias - classification

Theorem: [Soudry, et al. ’18]
For linearly separable data, the evolution of GD with any initialization converges to the **hard margin support vector machine** (**SVM**),

$$\lim_{t\to\infty} \frac{w_t}{\|w_t\|_2} = \frac{w^*}{\|w^*\|_2}, \quad w^* = \arg\min_{w\in\mathbb{R}^d} \|w\|_2 \quad \text{s.t.} \quad \boldsymbol{y}_i w^\intercal \boldsymbol{x}_i \geq 1.$$

- Similar result hold for ℓ_p-norm hard margin support vector machines (ℓ_p-**SVM**) with Steepest Descent dynamics [Gunasekar, et al. ’18].

- In particular ℓ_1-**SVM** is closely related to **infinitely wide 2-layer networks** [Neyshabur, et al. ’14] [Chizat, et al. ’18] and **Adaboost** [Rosset, et al. ’04].

Inductive bias

- Generalization properties of **OLS** in high dimensions is widely studied and characterized.
  - **Benign overfitting in ℓ_2-OLS**
    [Bartlett, et al. '19]
    [Hastie, et al. '19]
  - Benign overfitting in ℓ_1-**OLS**
    [Wang, et al. '22][Li, et al. '21]
  - Benign overfitting in ℓ_p-**OLS**
    [Wang, et al. '22]

- Less is known regarding generalization properties of hard margin **SVM** in high dimensions.
  - **Generalization behavior for ℓ_2-SVM**
    [Muthukumar, et al. '21]
    [Chatterji, et al. '20]
  - Generalization behavior for ℓ_1-**SVM**
    [Donhauser, et al. '22][Chinot, et al. '21]
  - Generalization behavior for ℓ_p-**SVM**
    [Donhauser, et al. '22]

- Inductive bias of learning
  - Linear and logistic regression
- **Classification vs. regression**
  - **OLS** = **SVM** and its implications
  - Our results
  - Key lemma and geometrical intuition
  - Proof ideas
  - Empirical universality

Classification vs. regression

Classification task

$(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i \le n} \in \mathbb{R}^d \times \{\pm 1\}$

Square loss

Linear regression

OLS

$$\min \ \|w\|_2 \quad \text{s.t.} \ \boldsymbol{x}_i^{\intercal} w = y_i$$

Logistic loss

Logistic regression

SVM

$$\min \ \|w\|_2 \quad \text{s.t.} \ y_i \boldsymbol{x}_i^{\intercal} w \ge 1$$

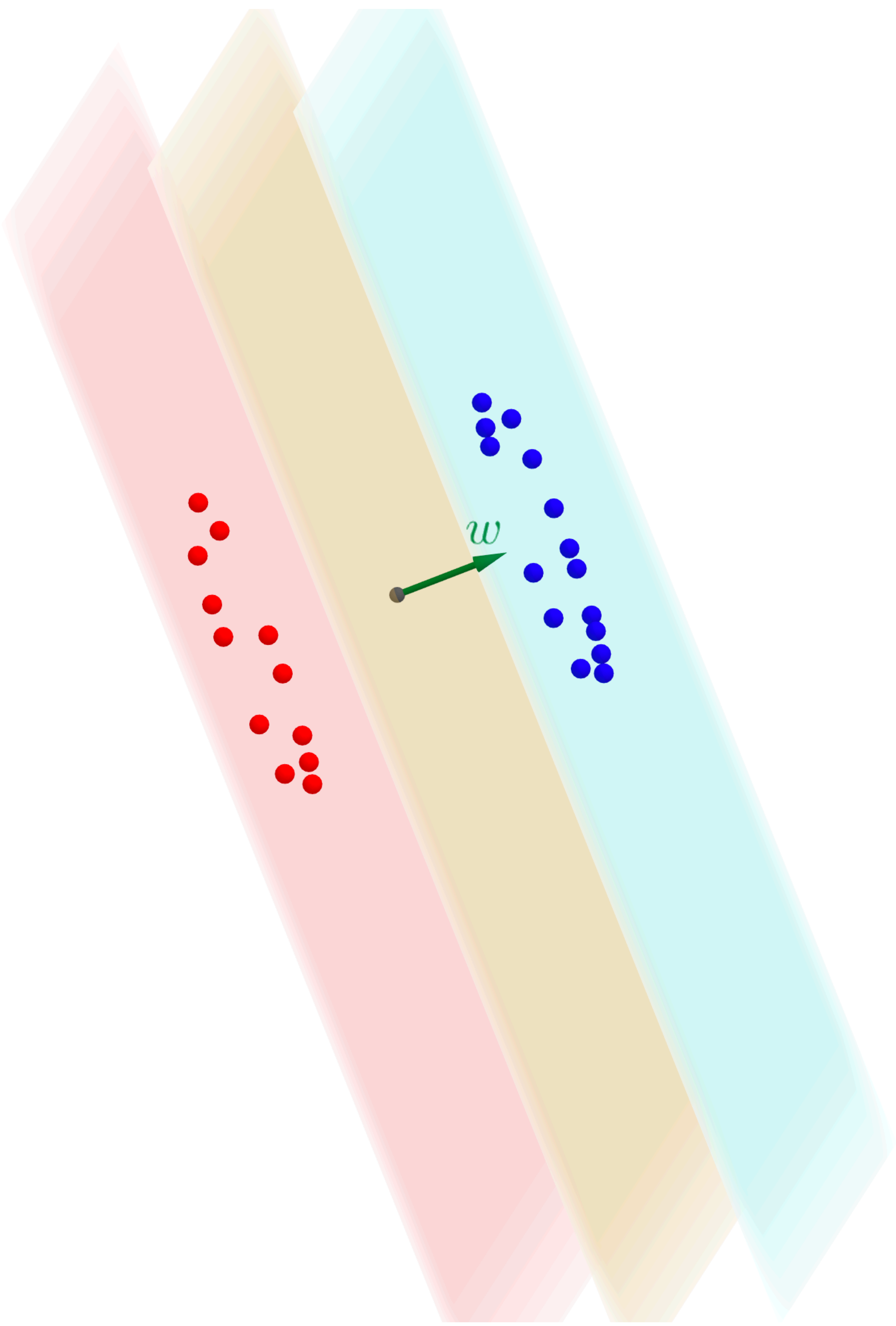
[Muthukumar, et al. ’21]

Classification vs. regression

Support vector proliferation

- What does **OLS**=**SVM** mean?
    - **SVM** classifier **interpolates** the data.
    - All samples must become **support vectors.**
- This situation was classically considered to generalize poorly,

  SVM Complexity ↔ # Support Vectors
- However, "Good" generalization properties of **OLS** carries over to **SVM** in these regimes.

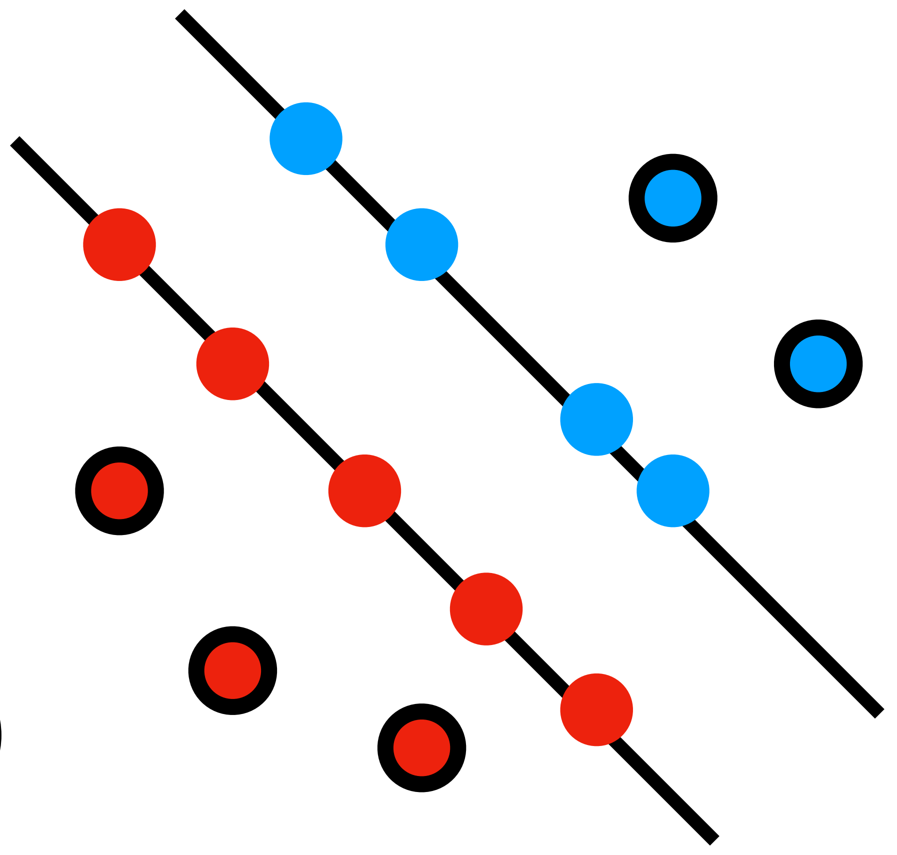


$$\min \|w\|_2 \quad \text{s.t.} \quad \boldsymbol{x}_i^\top w = y_i$$

$$\min \|w\|_2 \quad \text{s.t.} \quad y_i \boldsymbol{x}_i^\top w \geq 1$$

Classical SVM generalization bounds

- **SVM** Complexity $\leftrightarrow$ # Support Vectors
- When fraction of support vectors is $o(1)$, then **SVM** generalizes. [Graepel, et al. '05]
  - Sample compression based bounds.
  - Dropping non support vector samples still yields the **SVM** same classifier
  - Distribution free, thus widely applicable.
- This sparsity in #SV can happen in **underparameterized** asymptotic regimes.
- Different story in overparameterized regimes (e.g. when **OLS**=**SVM**)



OLS = SVM and its implications

- "Good" generalization properties of **OLS** carries over to **SVM** in these regimes.

$$\mathbb{P}\left[y h_w(\boldsymbol{x}) < 0\right] \leq \mathbb{E}\left[(1 - y h_w(x))^2\right]$$

- Classification is (thought of to be) "easier" than regression.

$$\text{Regression consistency} \implies \text{Classification consistency}$$

- Using this coincidence [Muthukumar, et al. '21] shows a regime where classification is consistent but not regression, under a spiked covariance model on features.

Fig. From [Muthukumar, et al. '21]

- Inductive bias of learning
  - Linear and logistic regression
- Classification vs. regression
  - OLS = SVM and its implications
  - **Our results**
  - Key lemma and geometrical intuition
  - Proof ideas
  - Empirical universality

Our results

- **Data model**: Labels are fixed and features are anisotropic Gaussian

$$\boldsymbol{x}_i \sim \mathcal{N}(0,\Sigma) \in \mathbb{R}^d,\ y_i \in \{\pm 1\},\ 1 \le i \le n$$

- **Effective ranks:** let $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_d$ be the eigenvalues of Σ

$$d_{eff} = \left(\frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|_F}\right)^2 = \left(\frac{\|\lambda\|_1}{\|\lambda\|_2}\right)^2, \quad d_\infty = \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|_{op}} = \frac{\|\lambda\|_1}{\|\lambda\|_\infty}$$

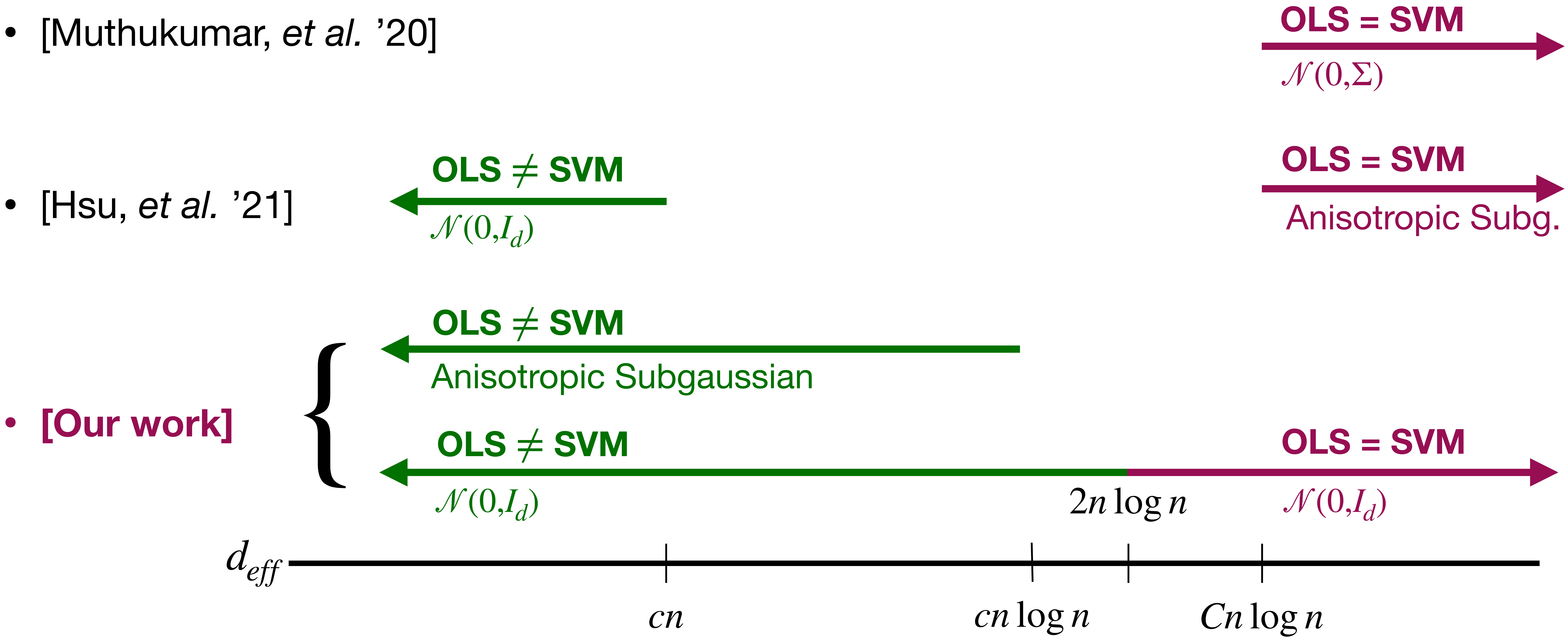
Theorem: [Our work]

Given n samples (as above) assume $d_{eff} = O(n \log n)$, $d_\infty = \Omega(n)$, and the absence of a single strong feature, then w.h.p. **OLS** $\neq$ **SVM**.

- For isotropic Gaussian features $d_{eff} = d_\infty = d$

Comparison with previous works

Question: For what $d_{eff} = d_{eff}(n)$ do we have **OLS**=**SVM** with high probability?



Asymptotic comparisons

Theorem: [Buhot, et al. '01]

For isotropic Gaussian features in the proportional regime $d(n) = \alpha n$, then the fraction of support vectors in the **SVM** converges w.h.p to,

$$\lim_{n\to\infty} \frac{\#\text{SV}}{n} = \begin{cases} 0.952\alpha & \alpha \ll 1 \quad \text{(Underparameterized)} \\ 1 - \sqrt{\frac{2}{\pi\alpha}} e^{-\frac{\alpha}{2}} & \alpha \gg 1 \quad \text{(Overparameterized)} \end{cases}$$

Theorem: [Our work]

For isotropic Gaussian data in the regime where $d(n) = \tau n \log n$,

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{\#\text{SV}}{n} = 1\right] = \begin{cases} 0 & \tau < 2 \\ 1 & \tau > 2 \end{cases}$$

- Inductive bias of learning
    - Linear and logistic regression
- Classification vs. regression
    - **OLS** = **SVM** and its implications
    - Our results
    - **Key lemma and geometrical intuition**
    - Proof ideas
    - Empirical universality

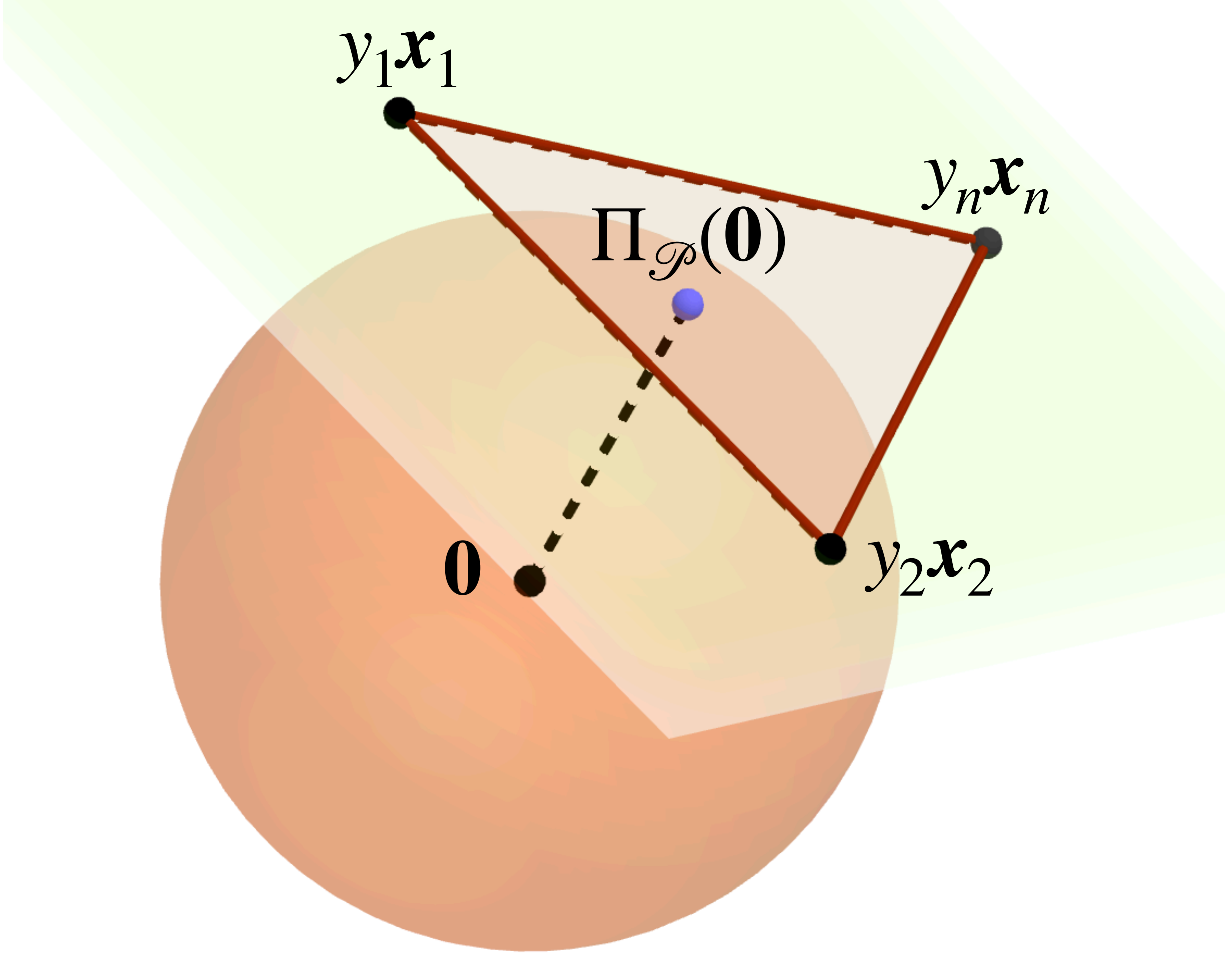
Geometrical Intuition

- The **OLS**=**SVM** occurrence is equivalent to,

$$\Pi_{\mathscr{P}}(\mathbf{0}) \in \text{ConvHull}(y_1\boldsymbol{x}_1, \ldots, y_n\boldsymbol{x}_n)\,.$$

- For isotropic gaussian features with $d \gg n$, all samples ($y_i\boldsymbol{x}_i$'s) are on the convex hull.
- $\|\boldsymbol{x}_i\|_2$ is roughly the same for all the samples.
- The convex hull is almost a regular polygon.
- Intuitively, larger d increases the probability of this occurrence.

$\mathscr{P} = \text{AffineHull}(y_1\boldsymbol{x}_1, y_2\boldsymbol{x}_2, \ldots, y_n\boldsymbol{x}_n) \subset \mathbb{R}^d$



- Inductive bias of learning
  - Linear and logistic regression
- Classification vs. regression
  - **OLS** = **SVM** and its implications
  - Our results
  - Key lemma and geometrical intuition
  - **Proof ideas**
  - Empirical universality

Features for sample i	$\boldsymbol{x}_i$
Label for sample i	y_i
Affine space when sample i is excluded.	$\mathscr{P}_{\setminus i}$

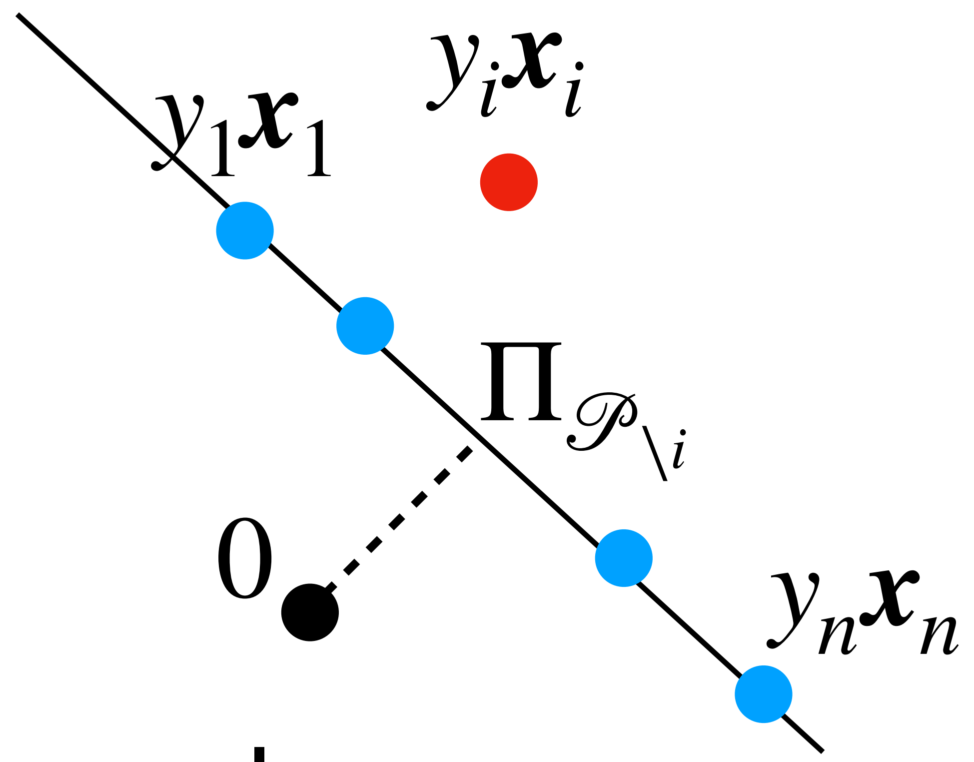
Proof ideas

Key lemma [Hsu, et al. '20][Our work]

Let $\Pi_{\mathscr{P}}$ be the projection onto $\mathscr{P} = \text{AffineHull}(y_1\boldsymbol{x}_1, \ldots, y_n\boldsymbol{x}_n)$ using ℓ_2-norm.

$$\max_{i \le n} \left\{ \left\langle y_i\boldsymbol{x}_i , \frac{\Pi_{\mathscr{P}_{\setminus i}}(\mathbf{0})}{\|\Pi_{\mathscr{P}_{\setminus i}}(\mathbf{0})\|_2^2} \right\rangle \right\} < 1 \iff \textbf{OLS} = \textbf{SVM} \iff \Pi_{\mathscr{P}}(\mathbf{0}) \in \text{ConvHull}(y_1\boldsymbol{x}_1, \ldots, y_n\boldsymbol{x}_n)$$

- **Proof intuition:** $w_{\mathbf{OLS}}^{(i)} = \dfrac{\Pi_{\mathscr{P}_{\setminus i}}(\mathbf{0})}{\|\Pi_{\mathscr{P}_{\setminus i}}(\mathbf{0})\|_2^2}$ using duality.
- $\langle u , w_{\mathbf{OLS}}^{(i)} \rangle - 1$ is a hyperplane passing through $\mathscr{P}_{\setminus i}$
- Origin and the i'th sample should be on the same side of this hyperplane.
  - Otherwise the i'th sample is "unnecessary" for **SVM**



Features for sample i	$\boldsymbol{x}_i$
Label for sample i	y_i
Affine space when sample i is excluded	$\mathscr{P}_{\setminus i}$
Collection of samples except I	$\mathbf{X}_{\setminus i}$

Proof ideas

Key lemma [Hsu, et al. '20][Our work]

Let $\Pi_{\mathscr{P}}$ be the projection onto $\mathscr{P} = \text{AffineHull}(y_1\boldsymbol{x}_1, \ldots, y_n\boldsymbol{x}_n)$ using ℓ_2-norm.

$$\max_{i \le n} \left\{ \left\langle y_i \boldsymbol{x}_i \,,\, \frac{\Pi_{\mathscr{P}_{\setminus i}}(\mathbf{0})}{\|\Pi_{\mathscr{P}_{\setminus i}}(\mathbf{0})\|_2^2} \right\rangle \right\} < 1 \iff \textbf{OLS} = \textbf{SVM} \iff \Pi_{\mathscr{P}}(\mathbf{0}) \in \text{ConvHull}(y_1\boldsymbol{x}_1, \ldots, y_n\boldsymbol{x}_n)$$

- For ℓ_2 explicit solutions for **OLS** is known:

$$w_{\mathbf{OLS}}^{(i)} = \frac{\Pi_{\mathscr{P}_{\setminus i}}(\mathbf{0})}{\|\Pi_{\mathscr{P}_{\setminus i}}(\mathbf{0})\|_2^2} = \boldsymbol{X}_{\setminus i}^{\intercal} \left(\boldsymbol{X}_{\setminus i} \boldsymbol{X}_{\setminus i}^{\intercal} \right)^{-1} y_{\setminus i}$$

- We use this lemma to prove lower bounds on the dimension.

Features for sample i	$\boldsymbol{x}_i$
Label for sample i	y_i
Collection of Features except	$\boldsymbol{X}_{\backslash i}$
Collection of labels except sample i	$y_{\backslash i}$

Proof ideas

- **Question:** For what values $d = d(n)$ do we have the following with high probability?

$$\max_{i \le n} \underbrace{\left\{ \left\langle y_i \boldsymbol{x}_i \, , \boldsymbol{X}_{\backslash i}^{\intercal} \left(\boldsymbol{X}_{\backslash i} \boldsymbol{X}_{\backslash i}^{\intercal} \right)^{-1} y_{\backslash i} \right\rangle \right\}}_{z_i} < 1$$

- z_i behaves roughly as a $\mathcal{N}\left(0, \dfrac{n}{d}\right)$.

- **if** z_i's were independent: $\max_{i \le n} z_i = \Theta_p\left(\sqrt{\dfrac{2n \log n}{d}}\right)$

Critical threshold

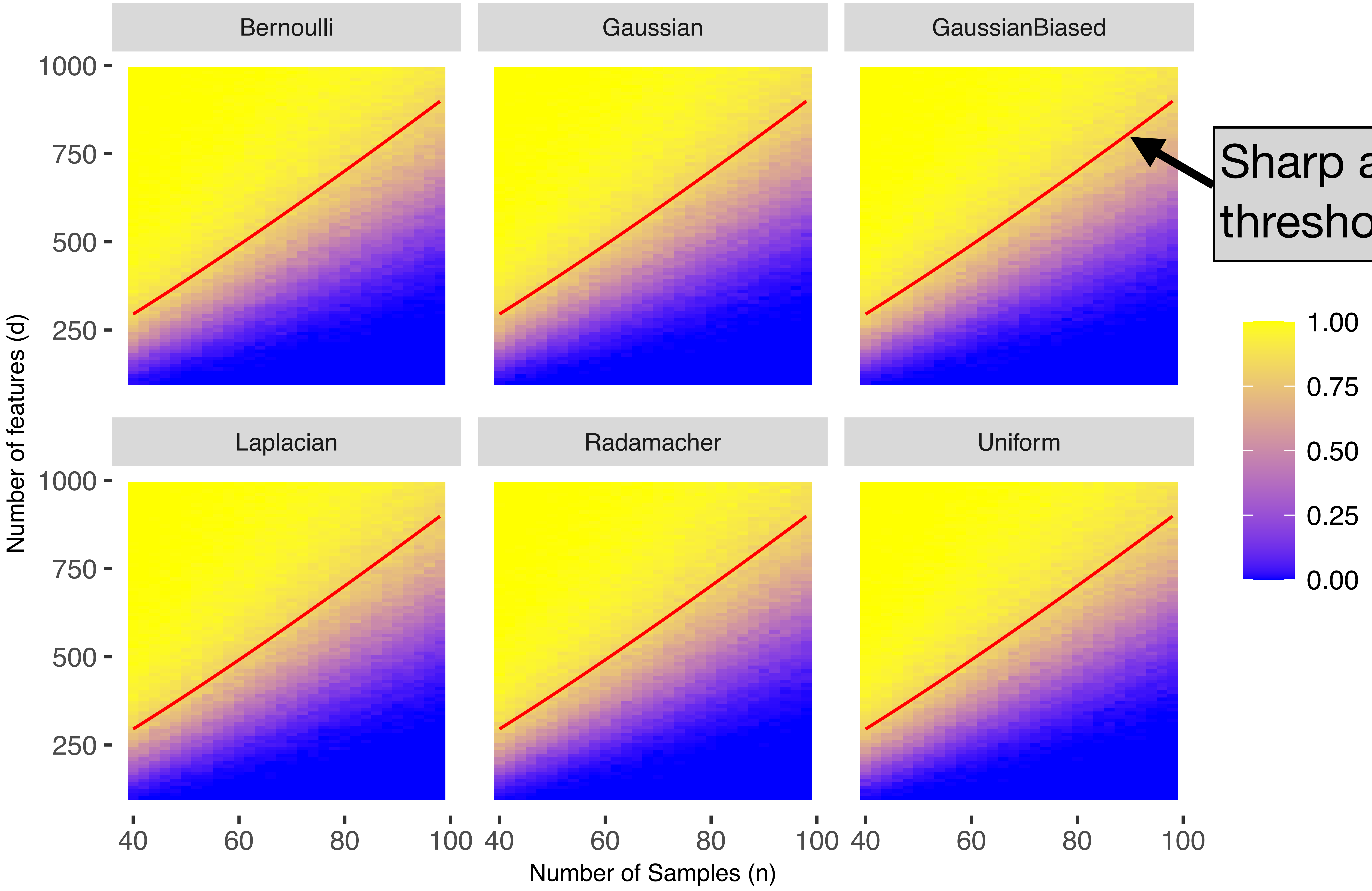
$$\boxed{\Longrightarrow d = \Theta(n \log(n))}$$

- The **correlation** among z_i's are weak $\Theta(\dfrac{1}{d})$, thus behavior stays the same.

- Inductive bias of learning
  - Linear and logistic regression
- Classification vs. regression
  - **OLS** = **SVM** and its implications
  - Our results
  - Key lemma and geometrical intuition
  - Proof ideas
  - **Empirical universality**

Empirical evidence for universality

- Universality of **SVP** phenomenon (**SVM** = **OLS**) under different feature distributions



Sharp asymptotic threshold at $d = 2n \log n$

Number of samples.	n
Number of features	d
Distribution under which features are generated from.	$\mathscr{D}$
Probit link function	Φ

Empirical evidence for universality

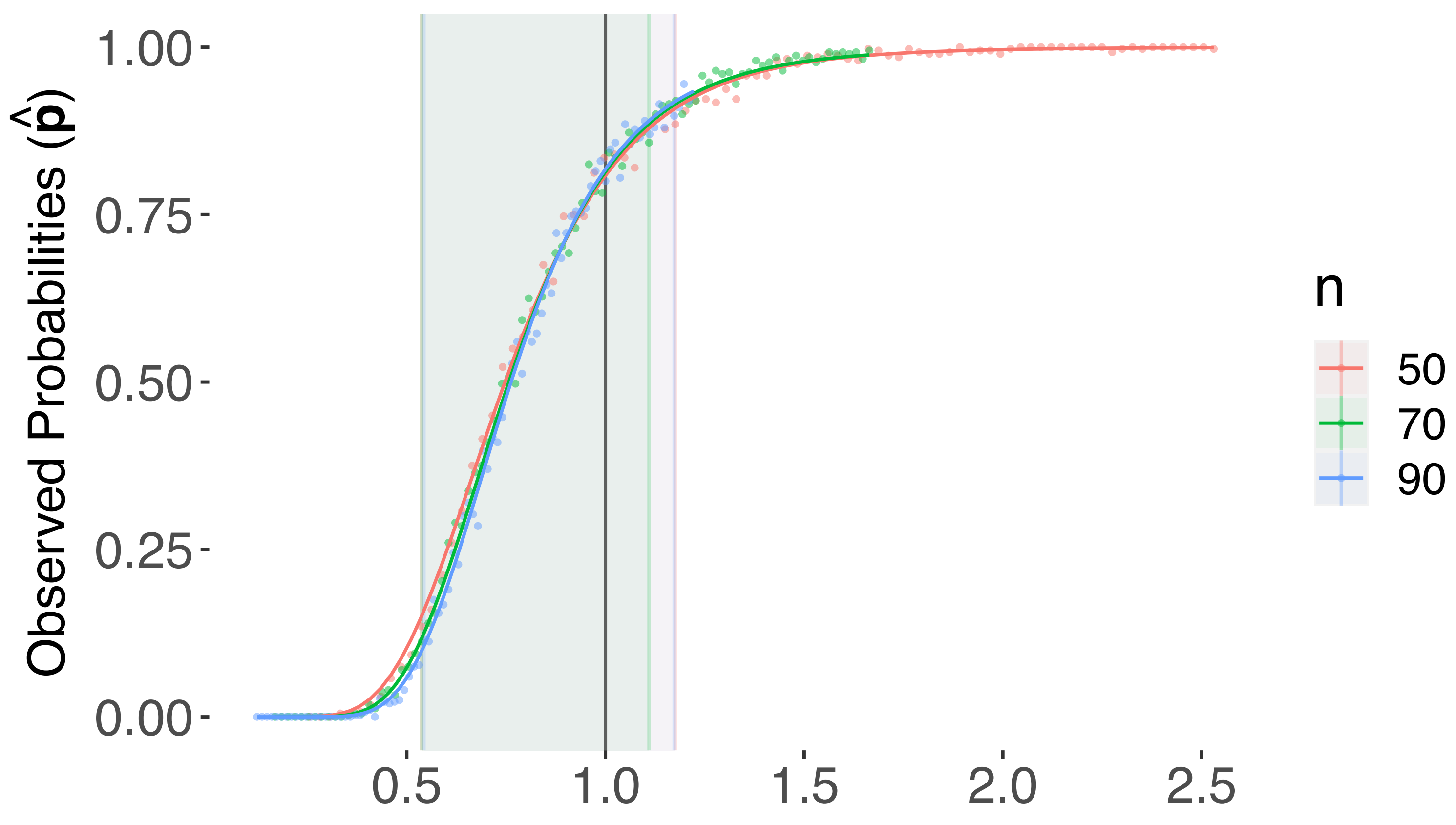
- Statistical methodology inspired by [Donoho, et al. 09]
- We use Probit regression to model the observed probability of **OLS**=**SVM**.

$$p(n, d; \mathscr{D}) = \Phi\left(\mu^{(0)}(n, \mathscr{D}) + \mu^{(1)}(n, \mathscr{D})\tau + \mu^{(2)}(n, \mathscr{D})\log\tau\right)$$

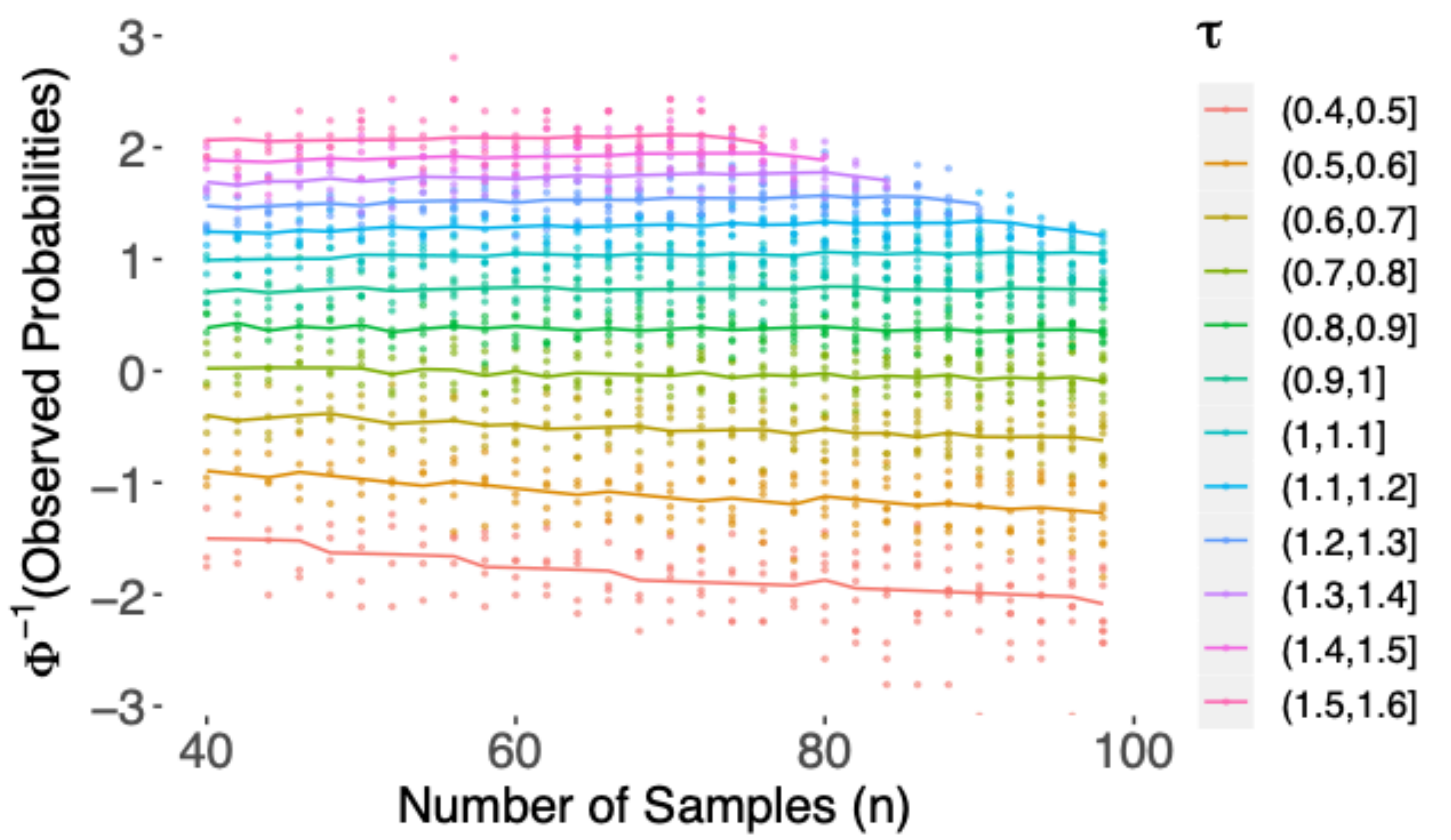
$$\mu^{(i)}(n, \mathscr{D}) = \mu_0^{(i)}(\mathscr{D}) + \frac{\mu_1^{(i)}(\mathscr{D})}{\sqrt{n}}$$

- Perform sequential hypothesis test using ANOVA.

$$\begin{cases} M_0 : \mu_j^{(i)}(\mathscr{D}) = \mu_j^{(i)} & \Longrightarrow \text{Reject} \\ M_1 : \mu_0^{(i)}(\mathscr{D}) = \mu_0^{(i)} & \Longrightarrow \text{Fail to reject} \\ M_2 : \text{O.W.} & \end{cases}$$



$$\tau = \frac{d}{2n\log n}$$



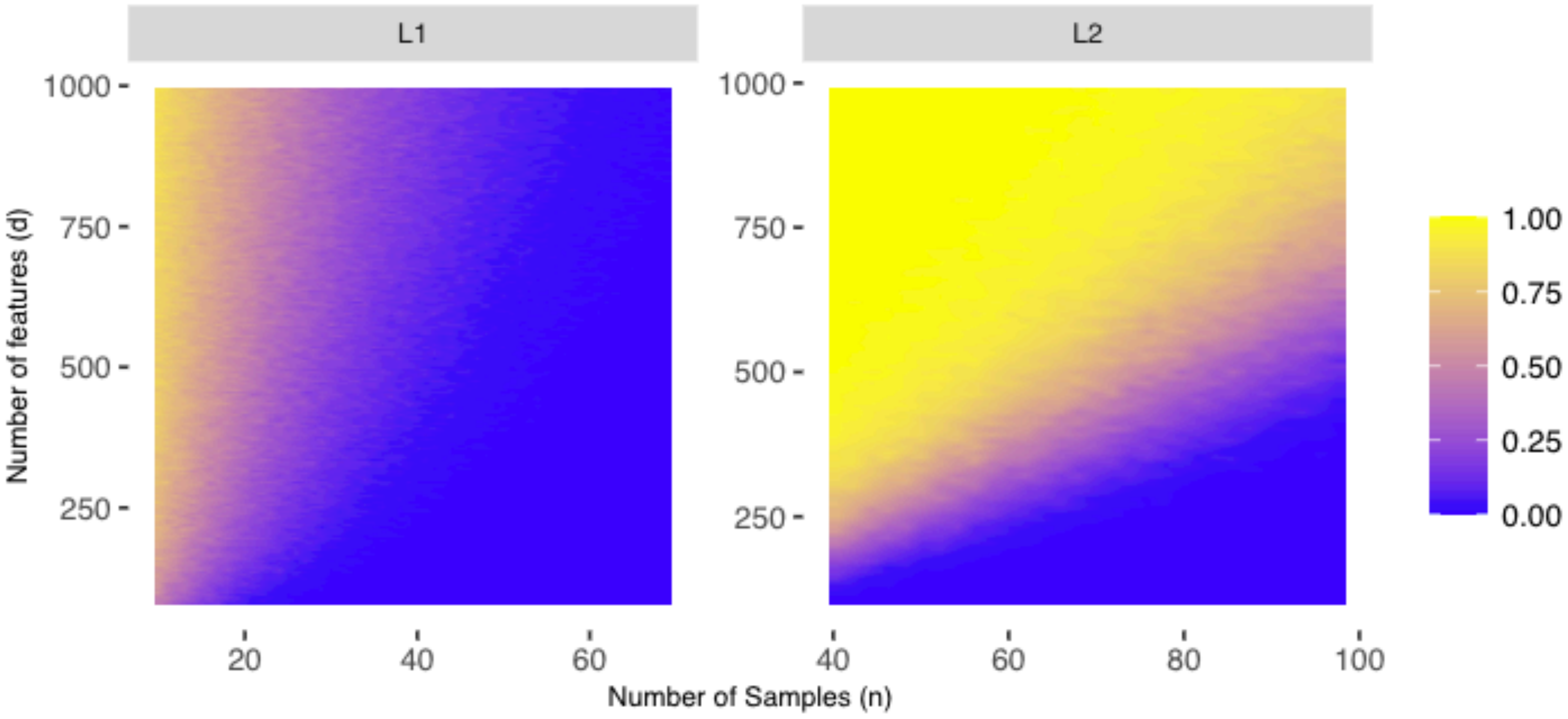
Open questions

- When does **SVM** = **OLS** for other norms?

$$\min \ \|w\|_p \qquad\qquad \min \ \|w\|_p$$
$$\text{s.t.} \ \ y_i \boldsymbol{x}_i^\top w \geq 1 \qquad \text{s.t.} \ \ \boldsymbol{x}_i^\top w = y_i$$

- Conjecture: For $p = 1$, **SVM** = **OLS** still occurs but the threshold is much larger function of number of samples n.

- Theoretical understanding of universality.



Thank you!