Support vector machines and linear regression coincide with very high-dimensional features

Navid Ardeshir*, Clayton Sanford*, Daniel Hsu (*Columbia University*)

Support Vector Machines $\stackrel{?}{=}$ Ordinary Least Squares

Suppose we have a classification task with n independent observations

$$(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{\pm 1\},$$

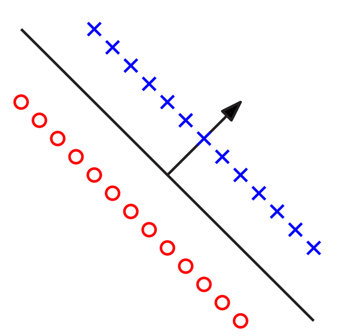
where the labels y_i are fixed. (Assume dataset is linearly separable.)

Hard ℓ_p-Norm Margin Suppor Vector Machine (SVM): Linear classifier $x \mapsto \text{sign}(x^\top w_{\text{SVM}})$ that maximizes ℓ_p-norm margin

$$w_{\text{SVM}} = \arg\min \|w\|_p \quad s.t.\ y_i \mathbf{x}_i^\top w \geq 1$$

Minimum ℓ_p-Norm Ordinary Least Squares (OLS): Linear function $x \mapsto x^\top w_{\text{OLS}}$ of minimum ℓ_p norm that interpolates data

$$w_{\text{OLS}} = \arg\min \|w\|_p \quad s.t.\ y_i \mathbf{x}_i^\top w = 1$$



Question: For what values $d = d(n)$ do we have **SVM** $=$ **OLS** with high probability?

Remark: Equiv. to all inequality constraints being tight; all samples are support vectors, a.k.a. "support vector proliferation (**SVP**)".

Implications on SVM generalization

Implicit Bias of optimization procedure: Gradient descent (coordinate descent) on logistic loss converges to the solution of ℓ_2-norm (ℓ_1-norm) hard margin **SVM** [1, 2].

Generalization: Classical bounds tied good generalization properties of **SVM** to paucity of support vectors (or large margins).

$$\#\textbf{Support Vectors} \downarrow \Rightarrow \textbf{Model Complexity} \downarrow$$

- Recent line of work, demonstrates high dimensional regimes for **SVM** with high complexity (and vanishing margins) but good generalization using this **SVM** $=$ **OLS** coincidence [3].
- "Benign overfitting" in over-parameterized linear regression provides generalization bounds for **OLS** [4, 5, 6, 7].
- **SVP** translates benign overfitting bounds to SVM for $d = \Omega(n \log n)$ under Gaussian data [3].

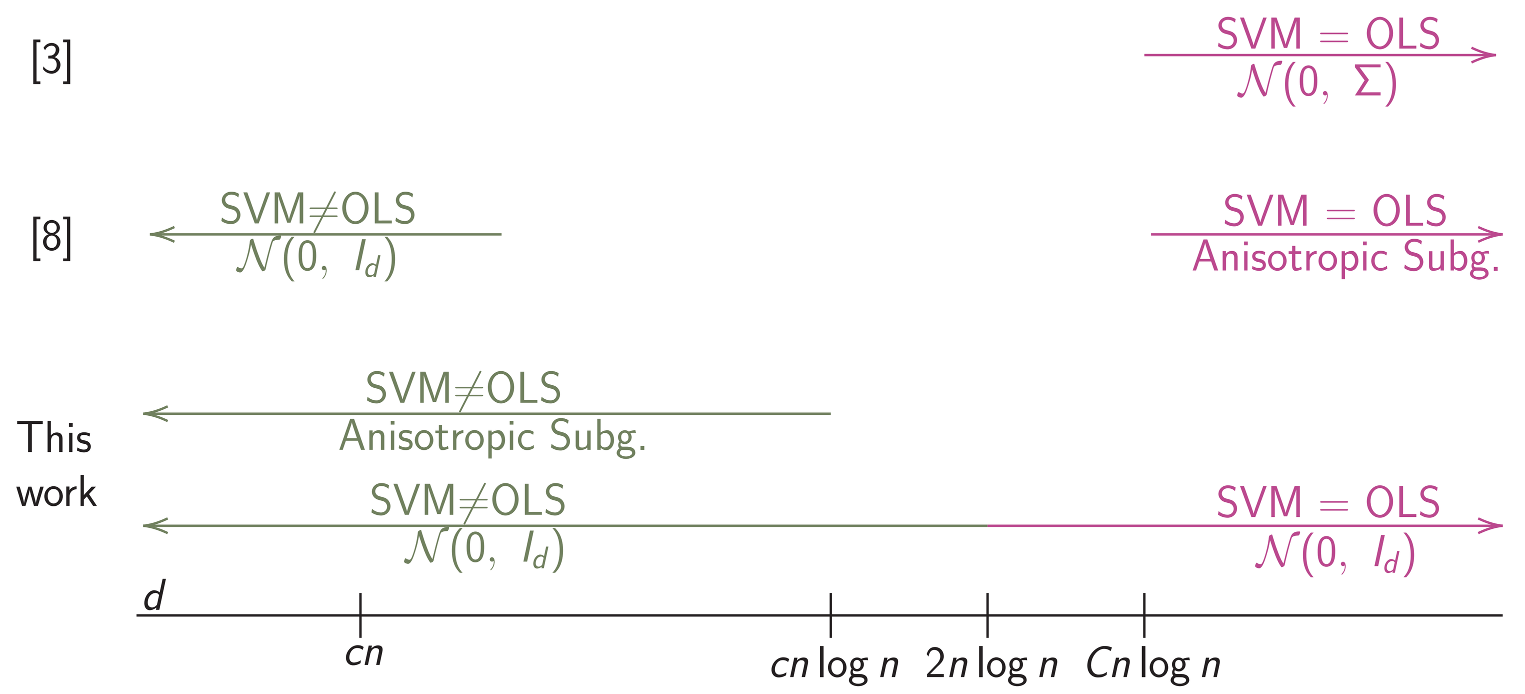
Previous work

- Further work found that **SVP** occurs when $d = \Omega(n \log n)$ for anisotropic Subgaussian data and *does not* occur when $d = O(n)$ for isotropic Gaussian data [8].
- Limitations:
  (1) There is a n vs $n \log n$ gap for SVP threshold
  (2) Unclear generality of lower bounds beyond isotropic Gaussian data.

Our contributions

We characterize the number of features d needed for **SVM=OLS** to occur when $p = 2$. Let $\Sigma \in \mathbb{R}^{d \times d}$ be an arbitrary covariance matrix.

- We provide **non-asymptotic** bounds for Subgaussian features,
  $$\mathbf{x}_i = \Sigma^{1/2}\mathbf{z}_i, \quad \mathbf{z}_i \stackrel{\text{i.i.d}}{\sim} \textbf{Subg}(1). \qquad \text{(anisotropic)}$$
- We show a **phase transition** occurs for standard Gaussian features,
  $$\mathbf{x}_i \stackrel{\text{i.i.d}}{\sim} \mathcal{N}(0, I_{d\times d}). \qquad \text{(isotropic)}$$
- We demonstrate an empirical **universality** of this phase transition.
- Conjecture about phase transition when $p = 1$.



Main results

Define effective dimensions via eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ of Σ:

$$d_2 := \left(\frac{\text{tr}(\Sigma)}{\|\Sigma\|_{\text{Fro}}}\right)^2 = \left(\frac{\|\lambda\|_1}{\|\lambda\|_2}\right)^2, \qquad d_\infty := \frac{\text{tr}(\Sigma)}{\|\Sigma\|_{\text{op}}} = \frac{\|\lambda\|_1}{\|\lambda\|_\infty}.$$

We show that this coincidence can be characterised in terms of these effective dimensions rather than actual ambient dimension d.

Theorem 1. Assume $d > n$ and $p = 2$.

(i) **Upper Bound ([8]):** Under anisotropic data model, there exist constants $c > 0$ such that
$$d_\infty \geq cn \log n \ \Rightarrow\ \mathbf{P}[\textbf{SVM} = \textbf{OLS}] \geq 0.9,$$

(ii) **Lower Bound:** Under anisotropic data model, there exist constants $c, c' > 0$ such that
$$d_2 \leq cn \log n,\ d_\infty \geq c'\sqrt{nd_2} \ \Rightarrow\ \mathbf{P}[\textbf{SVM} = \textbf{OLS}] \leq 0.1.$$

(iii) Under isotropic data model with identity covariance matrix $\Sigma = I_{d\times d}$, there exists constant $c, c' > 0$ such that
$$d \geq cn \log n \ \Rightarrow\ \mathbf{P}[\textbf{SVM} = \textbf{OLS}] \geq 0.9$$
$$d \leq c' n \log n \ \Rightarrow\ \mathbf{P}[\textbf{SVM} = \textbf{OLS}] \leq 0.1$$

(iv) **Phase Transition:** Under isotropic Gaussian model, there exists constant $c > 0$ such that
$$\lim_{n\to\infty} \mathbf{P}[\textbf{SVM} = \textbf{OLS}] = \begin{cases} 1 & \text{if } d \geq \left(2 + \frac{c}{\sqrt{\log n}}\right) n \log n \\ 0 & \text{if } d \leq \left(2 - \frac{c}{\sqrt{\log n}}\right) n \log n. \end{cases}$$

Proof ideas

Our results rely on the following algebraic characterization of **SVP**:

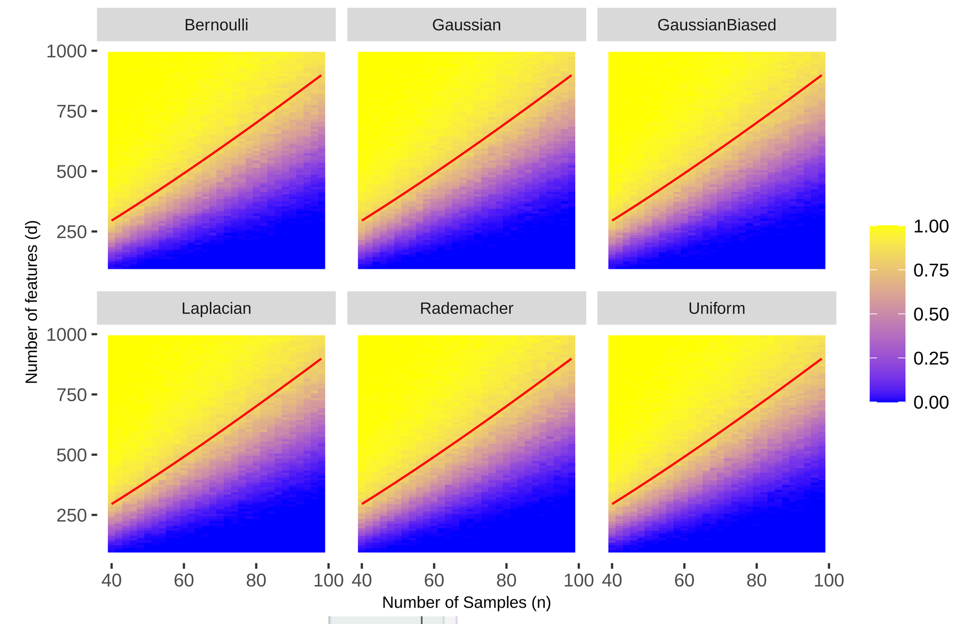
Lemma ([8]): All samples are support vectors if and only if

$$\max_{i \leq n} \underbrace{\left\langle y_i \mathbf{x}_i\ ,\ \mathbf{X}_{\setminus i}^\top \left(\mathbf{X}_{\setminus i}\mathbf{X}_{\setminus i}^\top\right)^{-1} y_{\setminus i} \right\rangle}_{\mathbf{u}_i} < 1.$$

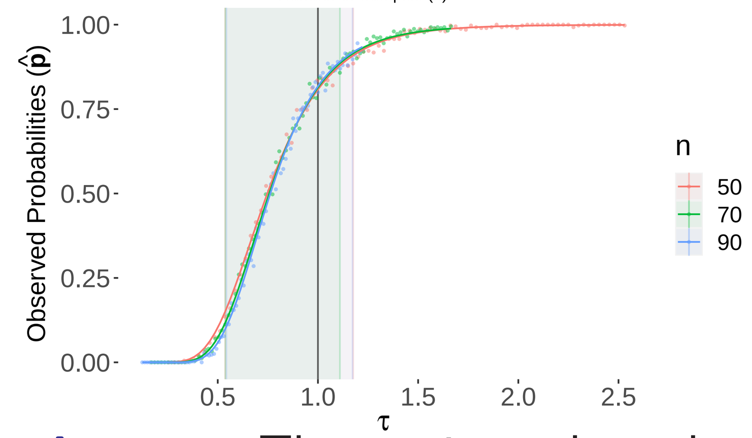
(We provide a new geometric proof that can be straightforwardly extended to infinite dimensional spaces.)

- For isotropic Gaussian case, $\mathbf{u}_i$ are marginally $\mathcal{N}(0, \frac{n}{d})$.
- If $\mathbf{u}_i$ were independent, $\max_i \mathbf{u}_i = \Theta_p(\sqrt{n \log n / d})$, so threshold should occur when $d = \Theta(n \log n)$.
- Despite lack of independence, same result follows by considering a subsample and showing that $(\mathbf{X}_{\setminus i}\mathbf{X}_{\setminus i}^\top)^{-1} \approx \frac{1}{d} I_{n-1}$.
- Result for anisotropic setting follows from subgaussian concentration and Berry-Esseen type bounds.

Empirical results



We empirically show that phase transition occurs at $d = 2n \log n$ (shown in red curve) rate for a wide range of distributions including ones with heavier tails than Subgaussians.



We model the behavior of the phase transition by $\tau = d/2n \log n$ and perform a parametric test to validate the **universality**.

ℓ_1 Conjecture: There exists a boundary $f(n) = \omega(n \log n)$ under isotropic data model such that,

$$\lim_{n\to\infty} \mathbf{P}[\textbf{SVM} = \textbf{OLS} \text{ for } p = 1] = \begin{cases} 1 & d > cf(n) \\ 0 & d < c'f(n) \end{cases}$$

where $c \geq c'$ are constants.

Bibliography

[1] Z. Ji and M. Telgarsky. The implicit bias of gradient descent on nonseparable data. In *COLT*, 2019.

[2] D. Soudry, E. Hoffer, M.S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *JMLR*, 19(1):2822–2878, 2018.

[3] V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. Hsu, and A. Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *JMLR*, 22(222):1–69, 2021.

[4] P. Bartlett, P. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *PNAS*, 117(48):30063–30070, 2020.

[5] M. Belkin, D. Hsu, and J. Xu. Two models of double descent for weak features. *SIMODS*, 2(4):1167–1180, 2020.

[6] T. Hastie, A. Montanari, S. Rosset, and R. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[7] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

[8] D. Hsu, V. Muthukumar, and J. Xu. On the proliferation of support vectors in high dimensions. In *AISTATS*, 2021.