Task Shift: From Classification to Regression in Overparameterized Linear Models

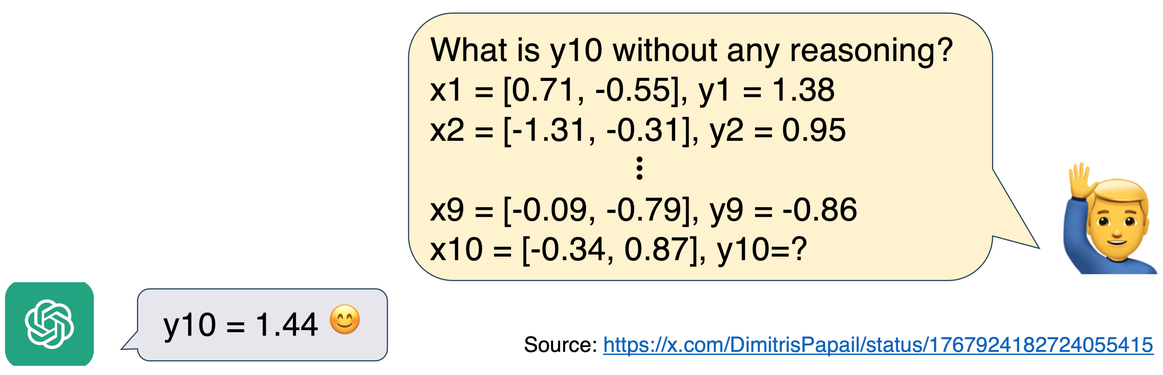
Tyler LaBonte*, Kuo-Wei Lai*, Vidya Muthukumar

Georgia Institute of Technology
*Equal contribution

arxiv link

Motivation

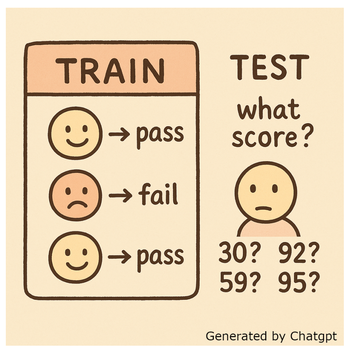
Modern classifiers adapt in-context to new tasks!



Source: https://x.com/DimitrisPapail/status/1767924182724055415

Q: Can we characterize this *easy-to-hard* task generalization in a tractable linear setting?

⇒ We focus on a classification to regression task shift problem.



Problem Setting

- Data $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $d \gg n$.
- Ground-truth regressor $\boldsymbol{\theta}^\star \in \mathbb{R}^d$.
- Classification labels $\hat{\mathbf{y}} \coloneqq (\operatorname{sign}(\mathbf{x}_i^\top \boldsymbol{\theta}^\star))_{i=1}^n$.
- Minimum ℓ_2-norm interpolator (MNI):

$$\hat{\boldsymbol{\theta}} = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \hat{\mathbf{y}}.$$

- Regression risk of classification MNI:

$$L(\hat{\boldsymbol{\theta}}) \coloneqq \mathbb{E}_{\mathbf{x}} (\mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \boldsymbol{\theta}^\star)^2.$$

- **Q:** Is the classification MNI consistent for the regression problem: $\lim_{n,d \to \infty} L(\hat{\boldsymbol{\theta}}) = 0$.

Zero-shot I: sparse signal

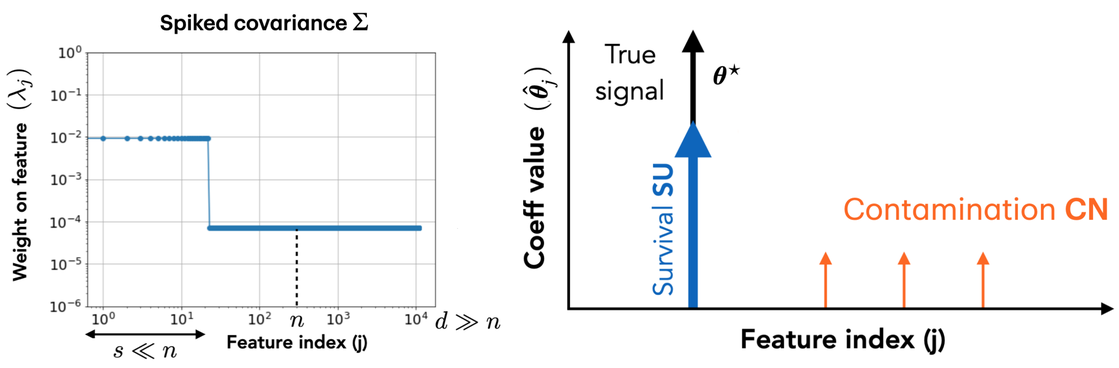


Image from Muthukumar et al., JMLR 2021

- Let $\boldsymbol{\theta}^\star$ be t-sparse and write $\boldsymbol{\theta}_j^\star \coloneqq a_j \lambda_j^{-1/2}$ where $\{\lambda_j\}_{j=1}^d$ is the spectrum of $\boldsymbol{\Sigma}$.
- *Survival* and *contamination*:

$$\mathsf{SU}_j \coloneqq \frac{\hat{\theta}_j}{\theta_j^\star} \qquad \mathsf{CN} \coloneqq \sqrt{\sum_{j \in \mathcal{S}^\mathsf{c}} \lambda_j \hat{\theta}_j^2}.$$

- Error decomposition:

$$L(\hat{\boldsymbol{\theta}}) = \sum_{j \in \mathcal{S}} a_j^2 (\mathsf{SU}_j - 1)^2 + \mathsf{CN}^2.$$

- Plug in our general SU_j and CN bounds for spiked covariance:

$$\lim_{n,d \to \infty} L(\hat{\boldsymbol{\theta}}) = \sum_{j \le s, j \in \mathcal{S}} a_j^2 \left(\sqrt{\frac{2}{\pi \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^\star\|_2^2}} - 1 \right)^2 + \sum_{j > s, j \in \mathcal{S}} a_j^2.$$

- Regression consistency *iff* $\|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^\star\|_2^2 = \frac{2}{\pi}$ and $a_j = 0$ for $j > s, j \in \mathcal{S}$.

Findings

- Zero-shot task shift is *impossible* for both sparse and random (dense) signals.
- Few-shot task shift is possible via *postprocessing* algorithm: no finetuning necessary.
- Structured attenuation of classification MNI: support recovery is guaranteed even when *generalization* is impossible!

Zero-shot II: random signal

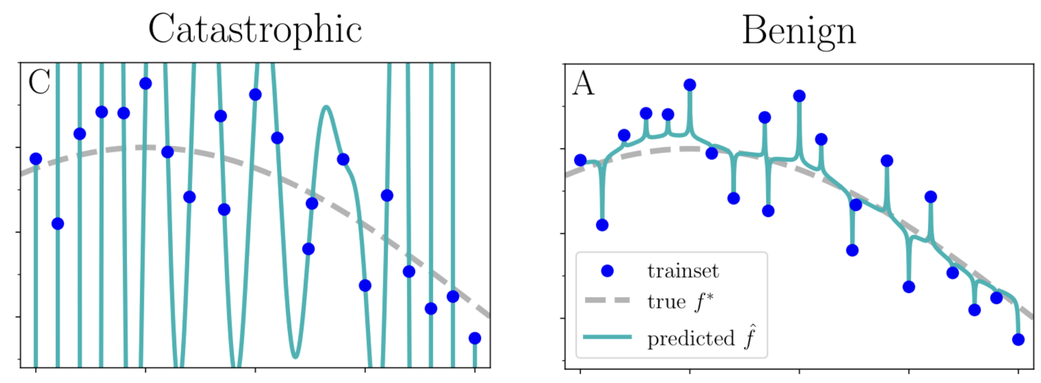


Image from Mallinar et al., NeurIPS 2022.

- Let $\boldsymbol{\theta}^\star$ be random such that $\mathbb{E}\theta_j^{\star 2} \ge \sigma^2$ for all j.
- Error decomposition:

$$L(\hat{\boldsymbol{\theta}}) = \underbrace{L(\tilde{\boldsymbol{\theta}})}_{\text{bias}} + \underbrace{\mathbb{E}_{\mathbf{x}} (\mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}})^2}_{\text{task shift error}}.$$

- Modified *benign overfitting* analysis:

$$\mathbb{E}_{\boldsymbol{\theta}^\star, \mathbf{x}} (\mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}})^2 \gtrsim \sigma^2 \left(\sum_{j=1}^{k^\star} \lambda_j + \frac{n}{R_{k^\star}(\boldsymbol{\Sigma})} \sum_{j=k^\star+1}^{d} \lambda_j \right).$$

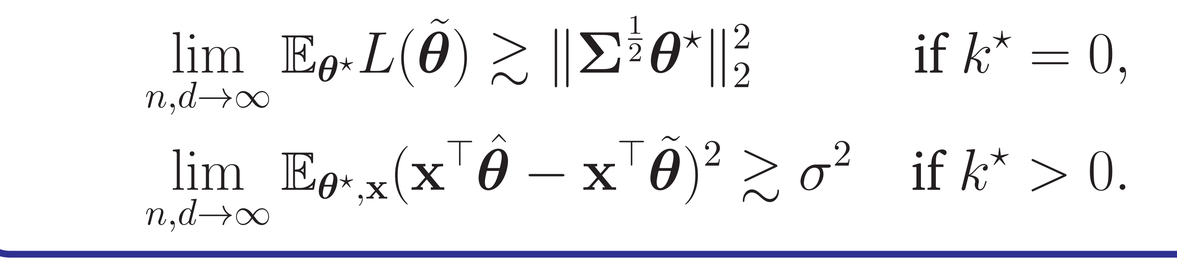
- Together with standard bounds on the bias:

$$\lim_{n,d \to \infty} \mathbb{E}_{\boldsymbol{\theta}^\star} L(\tilde{\boldsymbol{\theta}}) \gtrsim \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^\star\|_2^2 \qquad \text{if } k^\star = 0,$$

$$\lim_{n,d \to \infty} \mathbb{E}_{\boldsymbol{\theta}^\star, \mathbf{x}} (\mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}})^2 \gtrsim \sigma^2 \qquad \text{if } k^\star > 0.$$

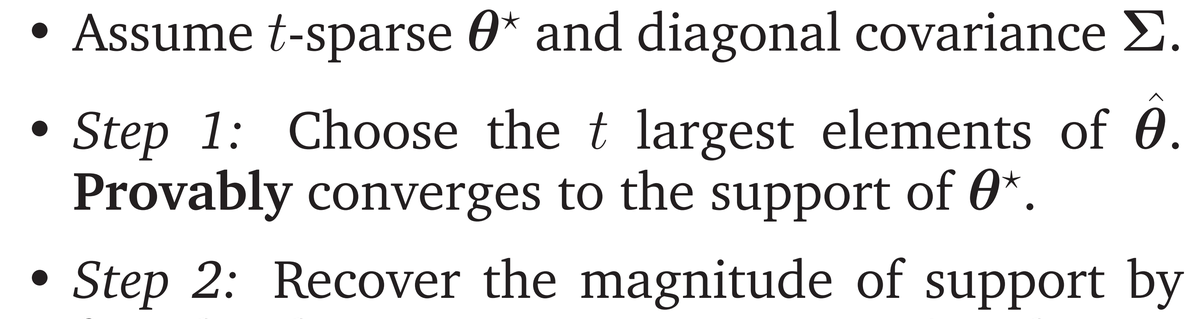
Few-shot postprocessing

- Assume t-sparse $\boldsymbol{\theta}^\star$ and diagonal covariance $\boldsymbol{\Sigma}$.
- *Step 1:* Choose the t largest elements of $\hat{\boldsymbol{\theta}}$. **Provably** converges to the support of $\boldsymbol{\theta}^\star$.
- *Step 2:* Recover the magnitude of support by few-shot least-squares on m regression data to obtain $\mathcal{O}\left(\frac{t}{m}\right)$ regression error.



Support recovery *m-shot least squares* *Regression risk*

Task shift for spiked $\boldsymbol{\Sigma}$ when regression generalizes.



Support recovery *m-shot least squares* *Regression risk*

Task shift for spiked $\boldsymbol{\Sigma}$ when regression does not generalize.