Two problems at the intersection of statistics and machine learning: Calibration and Data Integration

Pragya Sur, Dept. of Statistics, Harvard University

Statistical Physics and Machine Learning: moving forward, Aug, 2025

[Joint works with Yufan Li (Harvard → Voleon), Yanke Song (Harvard → Apple), Sohom Bhattacharya (UFL), Subhabrata Sen (Harvard), Sagnik Nandy (Ohio State), Samriddha Lahiry (NUS), Kenny Gu (Harvard → Stanford)]

Outline

Part I: Calibration

- Angular Calibration
- Bregman Optimality
- Platt Scaling and its Connection to Angular Calibration

Outline

Part I: Calibration

- Angular Calibration
- Bregman Optimality
- Platt Scaling and its Connection to Angular Calibration

Part II: Data Integration (Distribution Shift)

- Max-margin/min-norm interpolation behavior
- Anisotropic Local Laws
- Applications beyond Interpolation Learning

Outline

Part I: Calibration

- Angular Calibration
- Bregman Optimality
- Platt Scaling and its Connection to Angular Calibration

Part II: Data Integration (Distribution Shift)

- Max-margin/min-norm interpolation behavior
- Anisotropic Local Laws
- Applications beyond Interpolation Learning

Part III: Beyond Distribution Shift: Multimodal Learning

Part I: Calibration

Calibration: The Quest for Confidence

Calibration of a classifier:

Probabilistic predictions should match true empirical probabilities.

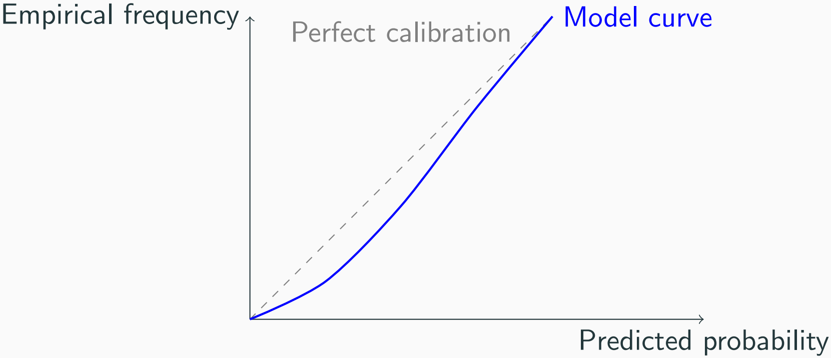
E.g., if a model says $P(class = 1) = 0.8$ to 100 samples, about 80 should be positive.

Calibration: The Quest for Confidence

Calibration of a classifier:

Probabilistic predictions should match true empirical probabilities.

E.g., if a model says $P(class = 1) = 0.8$ to 100 samples, about 80 should be positive.

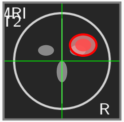


Why is calibration important?

Reliable decision-making in high-stakes problems → trustworthy algorithms

Why is calibration important?

Reliable decision-making in high-stakes problems → trustworthy algorithms



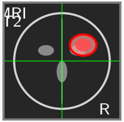
Brain MRI Scan

Alg: **80%** confidence

“*malignant*”

Why is calibration important?

Reliable decision-making in high-stakes problems → trustworthy algorithms



Brain MRI Scan

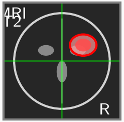
Alg: **80%** confidence

"*malignant*"

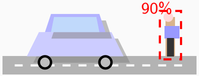
Does this mean **8 out of 10**
such cases are truly
malignant?

Why is calibration important?

Reliable decision-making in high-stakes problems → trustworthy algorithms



Brain MRI Scan



Autonomous Vehicle

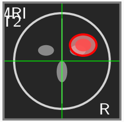
Alg: **80%** confidence
“*malignant*”

Alg: **90%** confidence
pedestrian detected

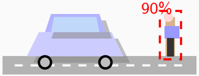
Does this mean **8 out of 10** such cases are truly malignant?

Why is calibration important?

Reliable decision-making in high-stakes problems → trustworthy algorithms



Brain MRI Scan



Autonomous Vehicle

Alg: **80%** confidence
“*malignant*”

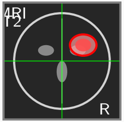
Alg: **90%** confidence
pedestrian detected

Does this mean **8 out of 10** such cases are truly malignant?

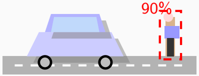
Does this imply **9 out of 10** such detections are accurate?

Why is calibration important?

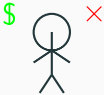
Reliable decision-making in high-stakes problems → trustworthy algorithms



Brain MRI Scan



Autonomous Vehicle



Credit Applicant

Alg: **80%** confidence "*malignant*"

Does this mean **8 out of 10** such cases are truly malignant?

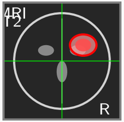
Alg: **90%** confidence pedestrian detected

Does this imply **9 out of 10** such detections are accurate?

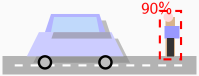
Alg: **95%** probability of default

Why is calibration important?

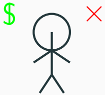
Reliable decision-making in high-stakes problems → trustworthy algorithms



Brain MRI Scan



Autonomous Vehicle



Credit Applicant

Alg: **80%** confidence "*malignant*"

Does this mean **8 out of 10** such cases are truly malignant?

Alg: **90%** confidence pedestrian detected

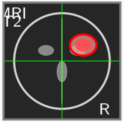
Does this imply **9 out of 10** such detections are accurate?

Alg: **95%** probability of default

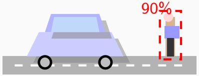
Would **19 out of 20** such applicants default in real life?

Why is calibration important?

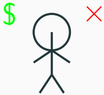
Reliable decision-making in high-stakes problems → trustworthy algorithms



Brain MRI Scan



Autonomous Vehicle



Credit Applicant

Alg: **80%** confidence
“*malignant*”

Does this mean **8 out of 10** such cases are truly malignant?

Alg: **90%** confidence
pedestrian detected

Does this imply **9 out of 10** such detections are accurate?

Alg: **95%** probability
of default

Would **19 out of 20** such applicants default in real life?

Calibrated machine learning algorithms ensure this is true!

Defining Calibration

A model is well-calibrated if:

Among all instances where the model predicts a probability of p, the true fraction of positives is exactly p.

Defining Calibration

A model is well-calibrated if:

Among all instances where the model predicts a probability of p, the true fraction of positives is exactly p.

Formally: Let $\hat{f}$ denote a trained model in a supervised learning binary classification problem with i.i.d. training samples $\{y_i, \mathbf{x}_i\}_{i=1,\ldots,n}$; $\hat{f}$ is well-calibrated if

$$\mathbb{P}[y_{\text{new}} = 1 \mid \hat{f}(\mathbf{x}_{\text{new}}) = p] = p, \quad \forall p \in [0, 1]$$

Reformulating in single-index models: this talk

Data Generating Process: Assume

$$y_i \sim \mathrm{Bern}\left(\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_i\right)\right), \ \mathbf{x}_i \sim F$$

where $\mathbf{w}_\star$ is an unknown deterministic vector

Reformulating in single-index models: this talk

Data Generating Process: Assume

$$y_i \sim \mathrm{Bern}\left(\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_i\right)\right), \; \mathbf{x}_i \sim F$$

where $\mathbf{w}_\star$ is an unknown deterministic vector

Calibration Error: For any predictor $\hat{f}$, define for $\mathbf{x}_{\mathrm{new}} \sim F$, independent of $\mathbf{x}_i' s$

$$\Delta_p^{\mathrm{cal}}(\hat{f}) = p - \mathbb{E}\left[\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}\right) \,\middle|\, \hat{f}(\mathbf{x}_{\mathrm{new}}) = p\right]$$

Reformulating in single-index models: this talk

Data Generating Process: Assume

$$y_i \sim \mathrm{Bern}\left(\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_i\right)\right), \ \mathbf{x}_i \sim F$$

where $\mathbf{w}_\star$ is an unknown deterministic vector

Calibration Error: For any predictor $\hat{f}$, define for $\mathbf{x}_{\mathrm{new}} \sim F$, independent of $\mathbf{x}_i' s$

$$\Delta_p^{\mathrm{cal}}(\hat{f}) = p - \mathbb{E}\left[\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}\right) \,\middle|\, \hat{f}(\mathbf{x}_{\mathrm{new}}) = p\right]$$

Reformulating in single-index models: this talk

Data Generating Process: Assume

$$y_i \sim \mathrm{Bern}\left(\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_i\right)\right), \; \mathbf{x}_i \sim F$$

where $\mathbf{w}_\star$ is an unknown deterministic vector

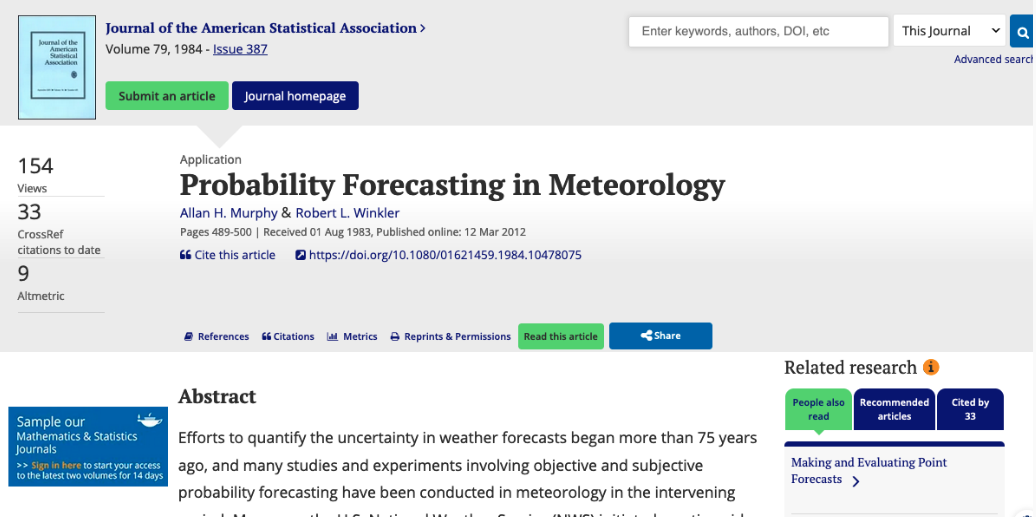
Calibration Error: For any predictor $\hat{f}$, define for $\mathbf{x}_{\mathrm{new}} \sim F$, independent of $\mathbf{x}_i's$

$$\Delta_p^{\mathrm{cal}}(\hat{f}) = p - \mathbb{E}\left[\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}\right) \,\middle|\, \hat{f}(\mathbf{x}_{\mathrm{new}}) = p\right]$$

A predictor is well–calibrated if $\Delta_p^{\mathrm{cal}}(\hat{f}) = 0$ for all p.

Old Roots in Statistics and Meteorology

- Original roots in weather forecasting and statistics



Journal of the American Statistical Association ›

Volume 79, 1984 - Issue 387

154 Views

33 CrossRef citations to date

9 Altmetric

Application

Probability Forecasting in Meteorology

Allan H. Murphy & Robert L. Winkler

Pages 489-500 | Received 01 Aug 1983, Published online: 12 Mar 2012

Cite this article https://doi.org/10.1080/01621459.1984.10478075

Related research

Making and Evaluating Point Forecasts ›

Abstract

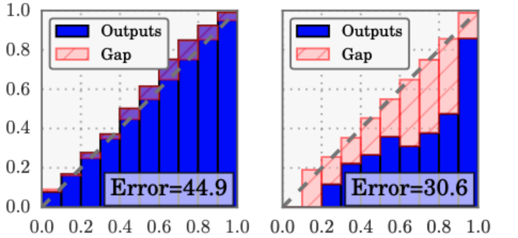
Efforts to quantify the uncertainty in weather forecasts began more than 75 years ago, and many studies and experiments involving objective and subjective probability forecasting have been conducted in meteorology in the intervening

The Problem: Modern Models are Often Miscalibrated

Modern neural networks are notoriously overconfident. [Guo et al., 2017]

The Problem: Modern Models are Often Miscalibrated

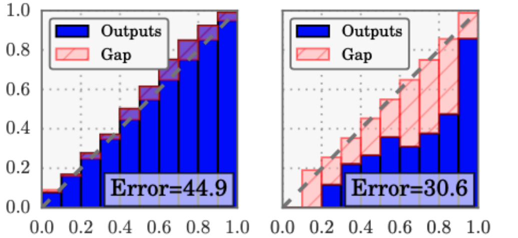
Modern neural networks are notoriously overconfident. [Guo et al., 2017]



Empirical frequency vs predicted probability: Old neural networks (LeNet-left, 5-layers) are well-calibrated, whereas modern deep networks (ResNet, right-110 layers) are overconfident.

The Problem: Modern Models are Often Miscalibrated

Modern neural networks are notoriously overconfident. [Guo et al., 2017]



Empirical frequency vs predicted probability: Old neural networks (LeNet-left, 5-layers) are well-calibrated, whereas modern deep networks (ResNet, right-110 layers) are overconfident. Their predicted probabilities grossly overestimate empirical frequencies although the prediction error is lower by 15%.

Overparametrization hurts calibration: why?

Insights from a simple model: logistic regression

For high-dimensional logistic regression, where dim./sample size $\to \gamma > 0$, the MLE (the ERM solution with logistic loss), in suitable high-dimensional sense, satisfies (S. and Candès, PNAS '19)

$$\hat{\mathbf{w}} \sim \alpha_\star \mathbf{w}_\star + \sigma_\star \mathbf{Z},$$

See also: Barbier et al. PNAS '19

Insights from a simple model: logistic regression

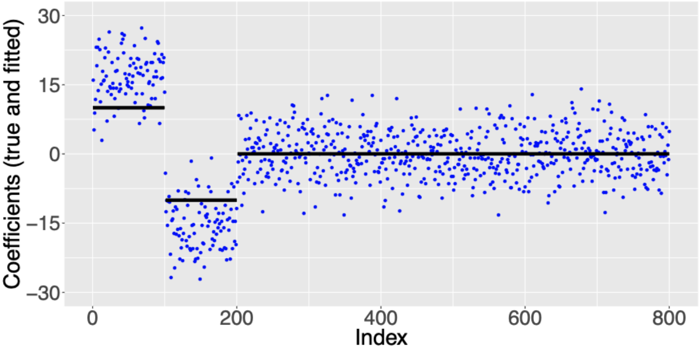
For high-dimensional logistic regression, where dim./sample size $\to \gamma > 0$, the MLE (the ERM solution with logistic loss), in suitable high-dimensional sense, satisfies (S. and Candès, PNAS '19)

$$\hat{\mathbf{w}} \sim \alpha_\star \mathbf{w}_\star + \sigma_\star \mathbf{Z},$$

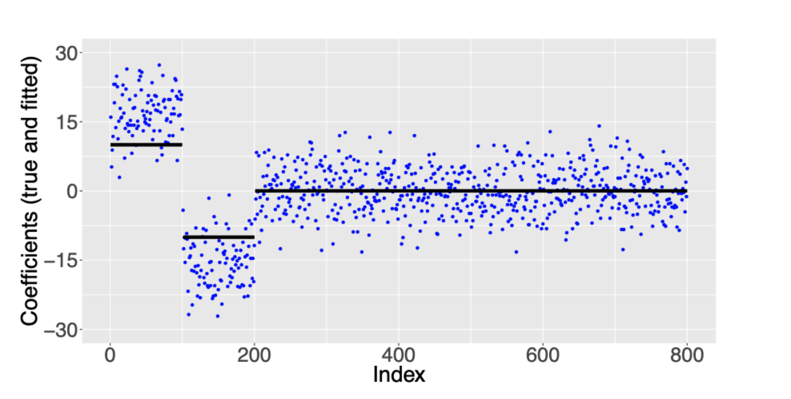
Crucial: $\alpha_\star > 1, \sigma_\star > \sigma_{\text{classical}}$ as soon as $\gamma > 0$, grows larger as γ increases.

See also: Barbier et al. PNAS '19

Over-confident estimates...

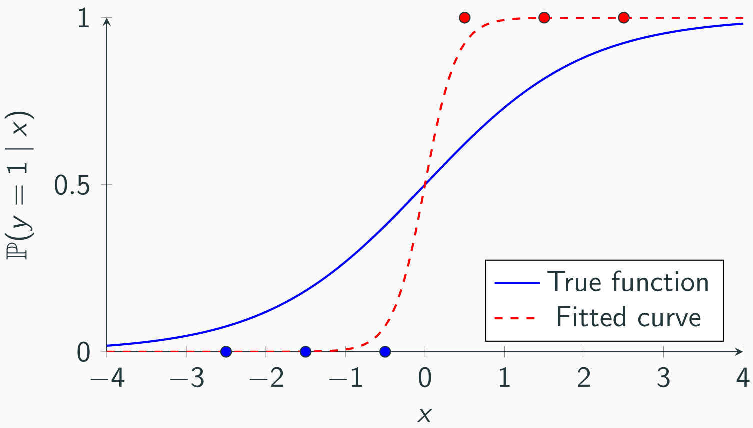


Over-confident estimates...

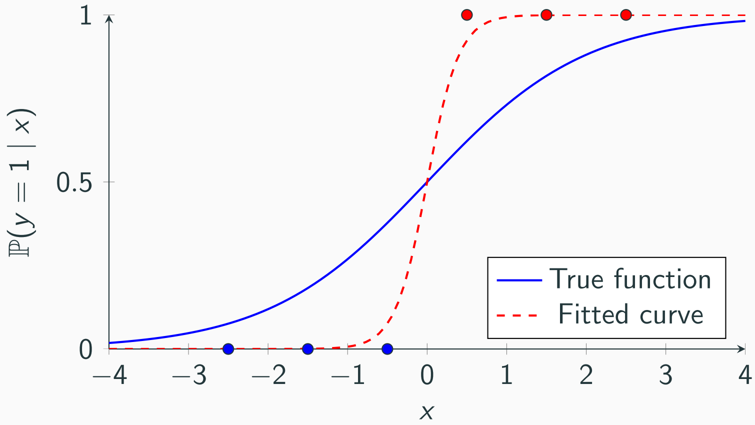


The multiplicative bias $\alpha_\star > 1 \rightarrow$ ERM is over-confident: coefficient estimates seriously biased upward!

Translate to over-confident predictions

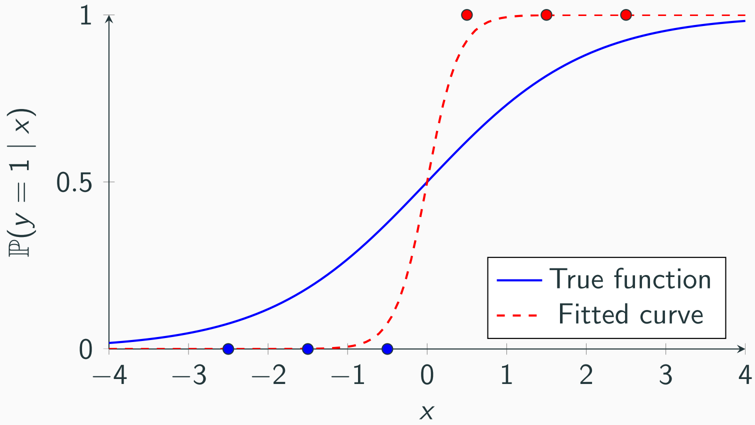


Translate to over-confident predictions



The true sigmoid transitions smoothly from class 0 to 1, but

Translate to over-confident predictions



The true sigmoid transitions smoothly from class 0 to 1, but the fitted curve is overly steep: predicts hard 0's and 1's with unnecessarily high confidence

Consequences for calibration

- The calibration error $\Delta_p^{\mathrm{cal}}(\hat{f})$ grows as feature to sample size ratio increases.
- The error is positive even for small values of feature dim./sample size ratio.
- Calibration error analyses in Bai et al. '21, Clarté et al. '22.

The fix?

Post-hoc calibration

Typical approach: train a model, then calibrate its outputs on a held-out set.

Post-hoc calibration

Typical approach: train a model, then calibrate its outputs on a held-out set.

Platt Scaling [Platt, 1999]

- Fits a logistic regression model on the original model's scores.
- Learns two parameters: a scaling factor and a shift.
- Simple and often effective, but assumes a sigmoid shape for the miscalibration.

Post-hoc calibration

Typical approach: train a model, then calibrate its outputs on a held-out set.

Platt Scaling [Platt, 1999]

- Fits a logistic regression model on the original model's scores.
- Learns two parameters: a scaling factor and a shift.
- Simple and often effective, but assumes a sigmoid shape for the miscalibration.

Many Other Approaches

- Histogram Binning [Zadronsky & Elkan, 2001]
- Isotonic Regression [Zadronsky & Elkan, 2002]

⋮

- Expectation Consistency Calibration [Clarte et al. 2023]

The Gap: Where Theory Meets Practice

Despite empirical success, substantial gaps persist.

- **Do these methods provably work in overparametrized problems?** We often rely on empirical validation. Prior theory (Kumar et al. '2019, Gupta et al. '2020, Jung et al. '2021, Sun et al' '2024) does not capture the precise impact of dimensionality on the performance of calibration methods; An exception: Clarte et al. '2023

The Gap: Where Theory Meets Practice

Despite empirical success, substantial gaps persist.

- **Do these methods provably work in overparametrized problems?** We often rely on empirical validation. Prior theory (Kumar et al. '2019, Gupta et al. '2020, Jung et al. '2021, Sun et al' '2024) does not capture the precise impact of dimensionality on the performance of calibration methods; An exception: Clarte et al. '2023
- **Do calibration and prediction face inherent trade-offs?** Can a method simultaneously achieve low calibration and prediction errors, especially under overparametrization?
- **Is there a "best" calibrated predictor?** Are there calibration methods that perform provably better than others in specific contexts?

The Gap: Where Theory Meets Practice

Despite empirical success, substantial gaps persist.

- **Do these methods provably work in overparametrized problems?** We often rely on empirical validation. Prior theory (Kumar et al. '2019, Gupta et al. '2020, Jung et al. '2021, Sun et al' '2024) does not capture the precise impact of dimensionality on the performance of calibration methods; An exception: Clarte et al. '2023
- **Do calibration and prediction face inherent trade-offs?** Can a method simultaneously achieve low calibration and prediction errors, especially under overparametrization?
- **Is there a "best" calibrated predictor?** Are there calibration methods that perform provably better than others in specific contexts?

Our work bridges this gap for an important class of models.

Our Contribution: Angular Calibration

Setting: High-Dimensional Linear Classification

- **Data Model:** Feature vectors $\mathbf{x}_i \in \mathbb{R}^d$ are Gaussian, $\mathbf{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$.
- **True Labels:** Follow a single-index model with true signal $\mathbf{w}_\star \in \mathbb{R}^d$:

$$y_i \sim \text{Bernoulli}(\sigma(\langle \mathbf{w}_\star, \mathbf{x} \rangle))$$

where σ is a link function (e.g., logistic). W.l.o.g. $\mathbf{w}_\star^\top \Sigma \mathbf{w}_\star = 1$

Setting: High-Dimensional Linear Classification

- **Data Model:** Feature vectors $\mathbf{x}_i \in \mathbb{R}^d$ are Gaussian, $\mathbf{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$.
- **True Labels:** Follow a single-index model with true signal $\mathbf{w}_\star \in \mathbb{R}^d$:

  $$y_i \sim \text{Bernoulli}(\sigma(\langle \mathbf{w}_\star, \mathbf{x} \rangle))$$

  where σ is a link function (e.g., logistic). W.l.o.g. $\mathbf{w}_\star^\top \Sigma \mathbf{w}_\star = 1$
- **Estimator:** Obtain $\hat{\mathbf{w}}$ using regularized logistic regression (convex loss and penalty) :

  $$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} \ell_{y_i}(\mathbf{w}^\top \mathbf{x}) + \sum_{i=1}^{n} g(w_i)$$
- **Raw Predictor:** The initial, uncalibrated probability of success for a new $\mathbf{x}_{\text{test}}$ is $S(\mathbf{x}_{\text{test}}) = \sigma(\langle \hat{\mathbf{w}}, \mathbf{x}_{\text{test}} \rangle)$

Setting: High-Dimensional Linear Classification

- **Data Model:** Feature vectors $\mathbf{x}_i \in \mathbb{R}^d$ are Gaussian, $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{\Sigma})$.
- **True Labels:** Follow a single-index model with true signal $\mathbf{w}_\star \in \mathbb{R}^d$:

$$y_i \sim \text{Bernoulli}(\sigma(\langle \mathbf{w}_\star, \mathbf{x} \rangle))$$

  where σ is a link function (e.g., logistic). W.l.o.g. $\mathbf{w}_\star^\top \Sigma \mathbf{w}_\star = 1$
- **Estimator:** Obtain $\hat{\mathbf{w}}$ using regularized logistic regression (convex loss and penalty) :

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} \ell_{y_i}(\mathbf{w}^\top \mathbf{x}) + \sum_{i=1}^{n} g(w_i)$$

- **Raw Predictor:** The initial, uncalibrated probability of success for a new $\mathbf{x}_{\text{test}}$ is $S(\mathbf{x}_{\text{test}}) = \sigma(\langle \hat{\mathbf{w}}, \mathbf{x}_{\text{test}} \rangle)$
- **High-dimensional asymptotic regime:** Assume $n, d \to \infty$, $d/n \to \gamma \in (0, \infty)$

The Core Idea: What's Wrong with the Raw Predictor?

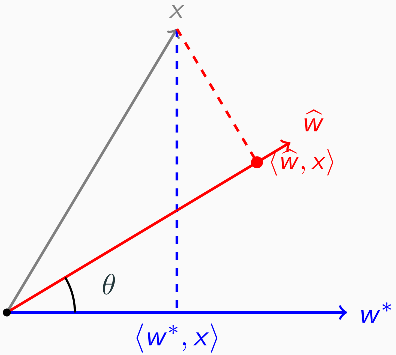
The raw predictor $\sigma(\langle \hat{\mathbf{w}}, \mathbf{x} \rangle)$ uses $\hat{\mathbf{w}}$ as a proxy for $\mathbf{w}_\star$. But $\hat{\mathbf{w}}$ is just an *estimate*! In fact, an inconsistent one, in the sense that $\|\hat{\mathbf{w}} - \mathbf{w}_\star\| = O(1)$.

The Core Idea: What's Wrong with the Raw Predictor?

The raw predictor $\sigma(\langle \hat{\mathbf{w}}, \mathbf{x} \rangle)$ uses $\hat{\mathbf{w}}$ as a proxy for $\mathbf{w}_\star$. But $\hat{\mathbf{w}}$ is just an *estimate*! In fact, an inconsistent one, in the sense that $\|\hat{\mathbf{w}} - \mathbf{w}_\star\| = O(1)$.

The Core Idea: What's Wrong with the Raw Predictor?

The raw predictor $\sigma(\langle \hat{\mathbf{w}}, \mathbf{x} \rangle)$ uses $\hat{\mathbf{w}}$ as a proxy for $\mathbf{w}_\star$. But $\hat{\mathbf{w}}$ is just an *estimate*! In fact, an inconsistent one, in the sense that $\|\hat{\mathbf{w}} - \mathbf{w}_\star\| = O(1)$.



The core of miscalibration lies in the discrepancy between $\hat{\mathbf{w}}$ and $\mathbf{w}_\star$, captured by the angle $\angle(\hat{\mathbf{w}}, \mathbf{w}_\star)$

Angular calibration

We identify a well-calibrated predictor, called **angular calibration**, that adjusts prediction logits using the angle between the estimator and the true signal $\theta_* = \arccos\left(\frac{\langle \mathbf{w}_\star, \hat{\mathbf{w}} \rangle_\Sigma}{\|\hat{\mathbf{w}}\|_\Sigma \|\mathbf{w}_\star\|_\Sigma}\right)$

Define

$$\hat{f}_{\mathrm{ang}}\left(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}; \theta_\star\right) = \mathbb{E}_{Z \sim N(0,1)}\left[\sigma\left(\cos(\theta_\star)\,\frac{\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}}{\|\hat{\mathbf{w}}\|_\Sigma} + \sin(\theta_\star)\, Z\right)\right],$$

The logit is an interpolation between the informative component $\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}$ and the noninformative Gaussian noise Z.

Angular predictor

With $\hat{\theta}$ a consistent estimator of $\theta_\star$, define

$$\hat{f}_{\mathrm{ang}}\left(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}; \hat{\theta}\right) = \mathbb{E}_{Z\sim N(0,1)}\left[\sigma\left(\cos(\hat{\theta})\,\frac{\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}}{\|\hat{\mathbf{w}}\|_\Sigma} + \sin(\hat{\theta})\,Z\right)\right],$$

Angular predictor

With $\hat{\theta}$ a consistent estimator of $\theta_\star$, define

$$\hat{f}_{\mathrm{ang}}\left(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}; \hat{\theta}\right) = \mathbb{E}_{Z\sim N(0,1)}\left[\sigma\left(\cos(\hat{\theta})\,\frac{\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}}{\|\hat{\mathbf{w}}\|_\Sigma} + \sin(\hat{\theta})\,Z\right)\right],$$

- The Gaussian noise balances the bias in $\hat{\mathbf{w}}$: the poorer the alignment between $\mathbf{w}_\star$ and $\hat{\mathbf{w}}$, the greater the magnitude of noise required to maintain calibration.
- If $\hat{\mathbf{w}}$ is perfectly aligned with $\mathbf{w}_\star$ ($\cos^2 = 1$), then the Gaussian component vanishes: we trust our original predictions completely and vice versa.

Main Results: Provable Calibration & Optimality

Result 1: Well–Calibration Theorem

Theorem (Li and S. ’25+)

Assume the link function is continuous. Then, the predictor $\hat{f}_{\mathrm{ang}}(\cdot;\hat{\theta})$ is well–calibrated as $d, n \to \infty$ with $n/d \to (0,\infty)$; that is, for any p in its range,

$$\Delta_p^{\mathrm{cal}}\left(\hat{f}_{\mathrm{ang}}(\cdot;\hat{\theta})\right) = p - \mathbb{E}\left[\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}\right) \,\middle|\, \hat{f}_{\mathrm{ang}}(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}};\hat{\theta}) = p\right] \to 0,$$

when $\hat{\theta}$ is a consistent estimator for $\theta_\star$.

Result 1: Well–Calibration Theorem

Theorem (Li and S. '25+)

Assume the link function is continuous. Then, the predictor $\hat{f}_{\mathrm{ang}}(\cdot;\hat{\theta})$ is well–calibrated as $d, n \to \infty$ with $n/d \to (0,\infty)$; that is, for any p in its range,

$$\Delta_p^{\mathrm{cal}}\left(\hat{f}_{\mathrm{ang}}(\cdot;\hat{\theta})\right) = p - \mathbb{E}\left[\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}\right) \,\middle|\, \hat{f}_{\mathrm{ang}}(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}};\hat{\theta}) = p\right] \to 0,$$

when $\hat{\theta}$ is a consistent estimator for $\theta_\star$.

- Provable calibration in high dimensions
- It holds for a wide class of estimators $\hat{\mathbf{w}}$ and link functions σ

Result 1: Well–Calibration Theorem

Theorem (Li and S. ’25+)

Assume the link function is continuous. Then, the predictor $\hat{f}_{\mathrm{ang}}(\cdot;\hat{\theta})$ *is well–calibrated as* $d, n \to \infty$ *with* $n/d \to (0,\infty)$*; that is, for any* p *in its range,*

$$\Delta_p^{\mathrm{cal}}\left(\hat{f}_{\mathrm{ang}}(\cdot;\hat{\theta})\right) = p - \mathbb{E}\left[\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}\right) \,\middle|\, \hat{f}_{\mathrm{ang}}(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}};\hat{\theta}) = p\right] \to 0,$$

when $\hat{\theta}$ *is a consistent estimator for* $\theta_\star$.

- Provable calibration in high dimensions
- It holds for a wide class of estimators $\hat{\mathbf{w}}$ and link functions σ
- Consistent angle estimator can be developed from Bellec (2022)

Proof idea using tower property

Let us define the following event

$$\mathcal{A} := \hat{f}_{\mathrm{ang}}\left(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}; \theta_\star\right) = p.$$

We have when $\hat{\theta} = \theta_*$

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}\right) \mid \mathcal{A}\right] \\
&\overset{(i)}{=} \mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}\right) \mid \mathbf{x}_{\mathrm{new}}^\top \hat{\mathbf{w}}\right] \mid \mathcal{A}\right]
\end{aligned}$$

Proof idea using tower property

Let us define the following event

$$\mathcal{A} := \hat{f}_{\mathrm{ang}}\left(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}; \theta_\star\right) = p.$$

We have when $\hat{\theta} = \theta_*$

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}\right) \mid \mathcal{A}\right] \\
&\overset{(i)}{=} \mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}\right) \mid \mathbf{x}_{\mathrm{new}}^\top \hat{\mathbf{w}}\right] \mid \mathcal{A}\right] \\
&\overset{(ii)}{=} \mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\mathbb{E}_Z\left[\sigma\left(\frac{1}{\|\hat{\mathbf{w}}\|_\Sigma} \cdot \cos\left(\theta_*\right) \cdot \mathbf{x}_{\mathrm{new}}^\top \hat{\mathbf{w}} + \sin\left(\theta_*\right) \cdot Z\right)\right] \mid \mathcal{A}\right]
\end{aligned}$$

Proof idea using tower property

Let us define the following event

$$\mathcal{A} := \hat{f}_{\mathrm{ang}}\left(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}; \theta_\star\right) = p.$$

We have when $\hat{\theta} = \theta_*$

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}\right) \mid \mathcal{A}\right] \\
&\overset{(i)}{=} \mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}\right) \mid \mathbf{x}_{\mathrm{new}}^\top \hat{\mathbf{w}}\right] \mid \mathcal{A}\right] \\
&\overset{(ii)}{=} \mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\mathbb{E}_{Z}\left[\sigma\left(\frac{1}{\|\hat{\mathbf{w}}\|_\Sigma} \cdot \cos\left(\theta_*\right) \cdot \mathbf{x}_{\mathrm{new}}^\top \hat{\mathbf{w}} + \sin\left(\theta_*\right) \cdot Z\right)\right] \mid \mathcal{A}\right] \\
&= \mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\hat{f}_{\mathrm{ang}}\left(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}; \theta_\star\right) \mid \mathcal{A}\right]
\end{aligned}$$

Proof idea using tower property

Let us define the following event

$$\mathcal{A} := \hat{f}_{\mathrm{ang}}\left(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}; \theta_\star\right) = p.$$

We have when $\hat{\theta} = \theta_*$

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}\right) \mid \mathcal{A}\right] \\
&\quad \overset{(i)}{=} \mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\sigma\left(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}\right) \mid \mathbf{x}_{\mathrm{new}}^\top \hat{\mathbf{w}}\right] \mid \mathcal{A}\right] \\
&\quad \overset{(ii)}{=} \mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\mathbb{E}_Z\left[\sigma\left(\frac{1}{\|\hat{\mathbf{w}}\|_\Sigma} \cdot \cos\left(\theta_*\right) \cdot \mathbf{x}_{\mathrm{new}}^\top \hat{\mathbf{w}} + \sin\left(\theta_*\right) \cdot Z\right)\right] \mid \mathcal{A}\right] \\
&\quad = \mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[\hat{f}_{\mathrm{ang}}\left(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}; \theta_\star\right) \mid \mathcal{A}\right] \\
&\quad = p.
\end{aligned}$$

where (i) follows from tower property, (ii) follows from conditional Gaussian distribution. The rest is checking convergence as $|\hat{\theta} - \theta_*| \to 0$.

Is calibration enough?

While calibration is important, does our predictor perform well otherwise?

- We want a predictor that is not only calibrated but has desirable properties in terms of other performance metrics as well

Is calibration enough?

While calibration is important, does our predictor perform well otherwise?

- We want a predictor that is not only calibrated but has desirable properties in terms of other performance metrics as well
- Once choice: ask for calibrated predictors as “close” as possible to the true conditional probability $\sigma(\langle \mathbf{w}_\star, \mathbf{x} \rangle)$.

Is calibration enough?

While calibration is important, does our predictor perform well otherwise?

- We want a predictor that is not only calibrated but has desirable properties in terms of other performance metrics as well
- Once choice: ask for calibrated predictors as “close” as possible to the true conditional probability $\sigma(\langle \mathbf{w}_\star, \mathbf{x} \rangle)$.

We establish this through the lens of Bregman divergence.

A Quick Detour: Bregman Divergence

For a strictly convex, differentiable function $\phi : \mathbb{R}^2 \to \mathbb{R}$, define the Bregman divergence between two vectors $P, Q \in \mathbb{R}^2$ to be

$$D_\phi(P, Q) = \phi(P) - \phi(Q) - \langle P - Q, \nabla\phi(Q)\rangle.$$

- Take $\phi(x) = \|x\|^2$: generates squared Euclidean distance
- Take $\phi(x) = \sum_j x_j \log x_j$: recovers KL divergence

Three random probability vectors of interest

With $F : \mathbb{R} \to [0, 1]$, define

$$\mathbf{q}_\star = \underbrace{\begin{pmatrix} \sigma(\mathbf{w}_\star^\top \mathbf{x}_{\text{new}}) \\ 1 - \sigma(\mathbf{w}_\star^\top \mathbf{x}_{\text{new}}) \end{pmatrix}}_{\text{True label prob.}}, \quad \hat{\mathbf{q}}_F = \underbrace{\begin{pmatrix} F(\hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}) \\ 1 - F(\hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}) \end{pmatrix}}_{\text{Prob. based on a general } F \text{ and estimator } \hat{\mathbf{w}}},$$

Three random probability vectors of interest

With $F : \mathbb{R} \to [0, 1]$, define

$$\mathbf{q}_\star = \underbrace{\begin{pmatrix} \sigma(\mathbf{w}_\star^\top \mathbf{x}_{\text{new}}) \\ 1 - \sigma(\mathbf{w}_\star^\top \mathbf{x}_{\text{new}}) \end{pmatrix}}_{\text{True label prob.}}, \quad \hat{\mathbf{q}}_F = \underbrace{\begin{pmatrix} F(\hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}) \\ 1 - F(\hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}) \end{pmatrix}}_{\text{Prob. based on a general } F \text{ and estimator } \hat{\mathbf{w}}},$$

$$\hat{\mathbf{q}}_{\text{ang}}(\hat{\theta}) := \underbrace{\begin{pmatrix} \hat{f}_{\text{ang}}(\hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}; \hat{\theta}) \\ 1 - \hat{f}_{\text{ang}}(\hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}; \hat{\theta}) \end{pmatrix}}_{\text{Prob. obtained from our angular predictor}}$$

Three random probability vectors of interest

With $F : \mathbb{R} \to [0,1]$, define

$$\mathbf{q}_\star = \underbrace{\begin{pmatrix} \sigma(\mathbf{w}_\star^\top \mathbf{x}_{\text{new}}) \\ 1 - \sigma(\mathbf{w}_\star^\top \mathbf{x}_{\text{new}}) \end{pmatrix}}_{\text{True label prob.}}, \quad \hat{\mathbf{q}}_F = \underbrace{\begin{pmatrix} F(\hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}) \\ 1 - F(\hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}) \end{pmatrix}}_{\text{Prob. based on a general } F \text{ and estimator } \hat{\mathbf{w}}},$$

$$\hat{\mathbf{q}}_{\text{ang}}(\hat{\theta}) := \underbrace{\begin{pmatrix} \hat{f}_{\text{ang}}(\hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}; \hat{\theta}) \\ 1 - \hat{f}_{\text{ang}}(\hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}; \hat{\theta}) \end{pmatrix}}_{\text{Prob. obtained from our angular predictor}}$$

All three are random probability vectors, randomness coming from $\mathbf{x}_{\text{new}}, \hat{\mathbf{w}}, \hat{\theta}$

Result 2: Bregman Optimality

Theorem (Li & S. '25+)

Let $\phi : \mathbb{R}^2 \to \mathbb{R}$ be any strictly convex differentiable function with finite $\mathbb{E}_{\mathbf{x}_{\mathrm{new}}}[\phi(\mathbf{q}_\star)]$.
(i) The expected Bregman loss $\mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[D_\phi\left(\mathbf{q}_\star, \hat{\mathbf{q}}_F\right)\right]$ admits a unique minimizer (upto a.s. equivalence) among all predictors of the form $F(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}})$.

Result 2: Bregman Optimality

Theorem (Li & S. ’25+)

Let $\phi : \mathbb{R}^2 \to \mathbb{R}$ be any strictly convex differentiable function with finite $\mathbb{E}_{\mathbf{x}_{\mathrm{new}}}[\phi(\mathbf{q}_\star)]$. *(i) The expected Bregman loss $\mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[D_\phi\left(\mathbf{q}_\star, \hat{\mathbf{q}}_F\right)\right]$ admits a unique minimizer (upto a.s. equivalence) among all predictors of the form $F(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}})$.*

(ii) Denote this minimizer by $F_\star = \arg\min_F \mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[D_\phi\left(\mathbf{q}_\star, \hat{\mathbf{q}}_F\right)\right]$. As $n, d \to \infty$, in probability,

$$\|\hat{\mathbf{q}}_{\mathrm{ang}}(\hat{\theta}) - F_\star(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}})\|_2^2 \to 0$$

Result 2: Bregman Optimality

Theorem (Li & S. '25+)

Let $\phi : \mathbb{R}^2 \to \mathbb{R}$ be any strictly convex differentiable function with finite $\mathbb{E}_{\mathbf{x}_{\mathrm{new}}}[\phi(\mathbf{q}_\star)]$.
(i) The expected Bregman loss $\mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[D_\phi\left(\mathbf{q}_\star, \hat{\mathbf{q}}_F\right)\right]$ admits a unique minimizer (upto a.s. equivalence) among all predictors of the form $F(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}})$.

(ii) Denote this minimizer by $F_\star = \arg\min_F \mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[D_\phi\left(\mathbf{q}_\star, \hat{\mathbf{q}}_F\right)\right]$. As $n, d \to \infty$, in probability,

$$\|\hat{\mathbf{q}}_{\mathrm{ang}}(\hat{\theta}) - F_\star(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}})\|_2^2 \to 0$$

Informally, among a natural class of predictors that are functions of $\langle \hat{\mathbf{w}}, \mathbf{x} \rangle$, our Angular Calibration predictor is ***uniquely optimal*** *in a Bregman divergence sense.*

Proof idea

Application of Banerjee et al. (2005) : for random vectors $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ with suitable expectations finite

$$\arg\min_{\mathbf{Y} \in f(\mathbf{Z})} \mathbb{E}\left[D_\phi(\mathbf{X}, \mathbf{Y})\right] = \mathbb{E}[\mathbf{X} \mid \mathbf{Z}].$$

Assign

$$\mathbf{X} \leftarrow \begin{pmatrix} (\sigma(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}) \\ 1 - \sigma(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}) \end{pmatrix}, \quad \mathbf{Y} \leftarrow \begin{pmatrix} (F(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}) \\ 1 - F(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}) \end{pmatrix}, \quad \mathbf{Z} \leftarrow \hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}.$$

Proof idea

Application of Banerjee et al. (2005) : for random vectors $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ with suitable expectations finite

$$\arg\min_{\mathbf{Y}\in f(\mathbf{Z})} \mathbb{E}\left[D_\phi(\mathbf{X},\mathbf{Y})\right] = \mathbb{E}[\mathbf{X} \mid \mathbf{Z}].$$

Assign

$$\mathbf{X} \leftarrow \begin{pmatrix} (\sigma(\mathbf{w}_\star^\top \mathbf{x}_{\text{new}}) \\ 1-\sigma(\mathbf{w}_\star^\top \mathbf{x}_{\text{new}}) \end{pmatrix}, \quad \mathbf{Y} \leftarrow \begin{pmatrix} (F(\hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}) \\ 1-F(\hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}) \end{pmatrix}, \quad \mathbf{Z} \leftarrow \hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}.$$

Then

$$\mathbb{E}[\mathbf{X} \mid \mathbf{Z}] = \mathbb{E}\left[\begin{pmatrix} (\sigma(\mathbf{w}_\star^\top \mathbf{x}_{\text{new}}) \\ 1-\sigma(\mathbf{w}_\star^\top \mathbf{x}_{\text{new}}) \end{pmatrix} \mid \mathbf{x}_{\text{new}}^\top \hat{\mathbf{w}} \right] = \begin{pmatrix} (\hat{f}_{\text{ang}}\left(\hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}; \theta_\star\right) \\ 1-\hat{f}_{\text{ang}}\left(\hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}; \theta_\star\right) \end{pmatrix}$$

by Gaussianity. The rest is checking convergence as $|\hat{\theta} - \theta_*| \to 0$.

Takeaway

Our angular predictor faces no trade-off between between calibration and optimality, if the latter is measured as minimizing the Bregman loss from the true label probabilities

Do popular calibration algorithms inherit such properties?

Platt Scaling

Platt scaling fits a mapping parameterized by A, B

$$F_{A,B}(u) = \sigma(Au + B), \quad A, B \in \mathbb{R}$$

by minimizing a logistic negative log–likelihood on a holdout set:

$$\hat{\ell}_{n_{\mathrm{ho}}}(F_{A,B}) = \sum_{i=1}^{n_{\mathrm{ho}}} \left[-\, y_{\mathrm{ho},i} \log\left(F_{A,B}(\hat{\mathbf{w}}^{\top}\mathbf{x}_{\mathrm{ho},i})\right) - (1 - y_{\mathrm{ho},i}) \log\left(1 - F_{A,B}(\hat{\mathbf{w}}^{\top}\mathbf{x}_{\mathrm{ho},i})\right)\right],$$

Platt Scaling

Platt scaling fits a mapping parameterized by A, B

$$F_{A,B}(u) = \sigma(Au + B), \quad A, B \in \mathbb{R}$$

by minimizing a logistic negative log–likelihood on a holdout set:

$$\hat{\ell}_{n_{\mathrm{ho}}}(F_{A,B}) = \sum_{i=1}^{n_{\mathrm{ho}}} \left[- y_{\mathrm{ho},i} \log\left(F_{A,B}(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{ho},i})\right) - (1 - y_{\mathrm{ho},i}) \log\Big(1 - F_{A,B}(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{ho},i})\Big)\right],$$

i.e., compute

$$\hat{A}^{n_{\mathrm{ho}}}, \hat{B}^{n_{\mathrm{ho}}} = \arg\min_{(A,B)\in\mathcal{H}} \hat{\ell}_{n_{\mathrm{ho}}}\Big(F_{A,B}\Big).$$

Platt Scaling

Platt scaling fits a mapping parameterized by A, B

$$F_{A,B}(u) = \sigma(Au + B), \quad A, B \in \mathbb{R}$$

by minimizing a logistic negative log–likelihood on a holdout set:

$$\hat{\ell}_{n_{\mathrm{ho}}}(F_{A,B}) = \sum_{i=1}^{n_{\mathrm{ho}}} \Big[- y_{\mathrm{ho},i} \log \left(F_{A,B}(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{ho},i}) \right) - (1 - y_{\mathrm{ho},i}) \log \Big(1 - F_{A,B}(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{ho},i}) \Big) \Big],$$

i.e., compute

$$\hat{A}^{n_{\mathrm{ho}}}, \hat{B}^{n_{\mathrm{ho}}} = \arg\min_{(A,B) \in \mathcal{H}} \hat{\ell}_{n_{\mathrm{ho}}} \Big(F_{A,B} \Big).$$

Platt Scaling

Platt scaling fits a mapping parameterized by A, B

$$F_{A,B}(u) = \sigma(Au + B), \quad A, B \in \mathbb{R}$$

by minimizing a logistic negative log–likelihood on a holdout set:

$$\hat{\ell}_{n_{\mathrm{ho}}}(F_{A,B}) = \sum_{i=1}^{n_{\mathrm{ho}}} \Big[-y_{\mathrm{ho},i} \log\left(F_{A,B}(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{ho},i})\right) - (1 - y_{\mathrm{ho},i}) \log\Big(1 - F_{A,B}(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{ho},i})\Big)\Big],$$

i.e., compute

$\hat{A}^{n_{\mathrm{ho}}}, \hat{B}^{n_{\mathrm{ho}}} = \arg\min_{(A,B)\in\mathcal{H}} \hat{\ell}_{n_{\mathrm{ho}}}\Big(F_{A,B}\Big)$. The Platt scaling predictor is given by

$$\hat{f}^{n_{\mathrm{ho}}}_{\mathrm{platt}}(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}) = \sigma\Big(\hat{A}^{n_{\mathrm{ho}}} \cdot \hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}} + \hat{B}^{n_{\mathrm{ho}}}\Big),$$

with $\hat{\mathbf{w}}$: regularized ERM as before computed on non-holdout data

Why Platt scaling works

- The logistic negative log-likelihood is the KL divergence between empirical and predicted probabilities.
- Intuitively, if the probability estimates are not calibrated, the negative log-likelihood will be higher.

Why Platt scaling works

- The logistic negative log-likelihood is the KL divergence between empirical and predicted probabilities.
- Intuitively, if the probability estimates are not calibrated, the negative log-likelihood will be higher.
- Platt scaling shifts and rescales the argument of the sigmoid so the predicted probabilities mimic the empirical frequencies.

Result 3: The Surprising Power of Platt Scaling

Theorem (Li and S. '25+)

If link function is approximately probit $(x) = \Phi(ax + b)$ *for some* $a \neq 0$ *and* $b \in \mathbb{R}$, *then we have under suitable technical conditions, as* $n_{\mathrm{ho}} \to \infty$,

$$\sup_{u \in \mathbb{R}} \left| \hat{f}^{n_{\mathrm{ho}}}_{\mathrm{platt}}(u) - \hat{f}_{\mathrm{ang}}(u; \hat{\theta}) \right| \to 0.$$

Result 3: The Surprising Power of Platt Scaling

Theorem (Li and S. '25+)

If link function is approximately probit $(x) = \Phi(ax + b)$ *for some* $a \neq 0$ *and* $b \in \mathbb{R}$*, then we have under suitable technical conditions, as* $n_{\mathrm{ho}} \to \infty$*,*

$$\sup_{u \in \mathbb{R}} \left| \hat{f}^{n_{\mathrm{ho}}}_{\mathrm{platt}}(u) - \hat{f}_{\mathrm{ang}}(u; \hat{\theta}) \right| \to 0.$$

- Informally, if σ is a probit link function (or approximated by one up to an affine transformation—e.g., $\mathrm{sigmoid}(x) \approx \Phi(\sqrt{\pi/8}\, x)$), this simple decades old heuristic recovers the optimal "angular" correction for larger and larger holdout sets.

Result 3: The Surprising Power of Platt Scaling

Theorem (Li and S. ’25+)

If link function is approximately probit $(x) = \Phi(ax + b)$ *for some* $a \neq 0$ *and* $b \in \mathbb{R}$, *then we have under suitable technical conditions, as* $n_{\mathrm{ho}} \to \infty$,

$$\sup_{u \in \mathbb{R}} \left| \hat{f}^{n_{\mathrm{ho}}}_{\mathrm{platt}}(u) - \hat{f}_{\mathrm{ang}}(u; \hat{\theta}) \right| \to 0.$$

- Informally, if σ is a probit link function (or approximated by one up to an affine transformation—e.g., $\mathrm{sigmoid}(x) \approx \Phi(\sqrt{\pi/8}\, x)$), this simple decades old heuristic recovers the optimal “angular” correction for larger and larger holdout sets.
- This provides the first theoretical justification for Platt scaling in high dimensions.

Result 3: The Surprising Power of Platt Scaling

Theorem (Li and S. ’25+)

If link function is approximately probit $(x) = \Phi(ax + b)$ *for some* $a \neq 0$ *and* $b \in \mathbb{R}$, *then we have under suitable technical conditions, as* $n_{\mathrm{ho}} \to \infty$,

$$\sup_{u \in \mathbb{R}} \left| \hat{f}^{n_{\mathrm{ho}}}_{\mathrm{platt}}(u) - \hat{f}_{\mathrm{ang}}(u; \hat{\theta}) \right| \to 0.$$

- Informally, if σ is a probit link function (or approximated by one up to an affine transformation—e.g., $\mathrm{sigmoid}(x) \approx \Phi(\sqrt{\pi/8}\, x)$), this simple decades old heuristic recovers the optimal “angular” correction for larger and larger holdout sets.
- This provides the first theoretical justification for Platt scaling in high dimensions.

Platt scaling is provably calibrated and Bregman optimal in high dimensions when using large holdout sets.

Simulation

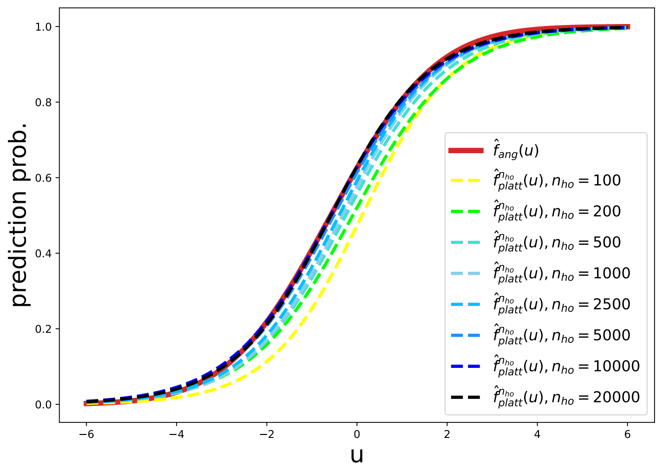


Figure 1: Platt scaling of a logistic ridge predictor converges to our angular predictor, as holdout set size increases. We set $\Sigma_{kl} = 0.5^{|k-l|}, \forall k, l \in [d]$ with $n = 1000, d = 2000$.

Proof steps

(i) By basic properties of Gaussian cdf, show that angular predictor is contained in search space of Platt scaling, i.e., for suitable $A_\star$, $B_\star$,

$$\hat{f}_{\mathrm{ang}}(u; \theta_*) = \sigma(A_* \cdot u + B_*)$$

(ii) Show the loss function $\hat{\ell}_{n_{\mathrm{ho}}}(F)$ converges to KL divergence uniformly (as $n_{\mathrm{ho}} \to \infty$)

$$\ell^\star(F) = \mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[D_{\mathrm{KL}}\left(\begin{pmatrix} (\sigma(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}) \\ 1 - \sigma(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}) \end{pmatrix} \middle\| \begin{pmatrix} (F(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}) \\ 1 - F(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}})) \end{pmatrix} \right) \right]$$

Proof steps

(i) By basic properties of Gaussian cdf, show that angular predictor is contained in search space of Platt scaling, i.e., for suitable $A_\star, B_\star$,

$$\hat{f}_{\mathrm{ang}}(u; \theta_*) = \sigma(A_* \cdot u + B_*)$$

(ii) Show the loss function $\hat{\ell}_{n_{\mathrm{ho}}}(F)$ converges to KL divergence uniformly (as $n_{\mathrm{ho}} \to \infty$)

$$\ell^\star(F) = \mathbb{E}_{\mathbf{x}_{\mathrm{new}}}\left[D_{\mathrm{KL}}\left(\begin{pmatrix} (\sigma(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}) \\ 1 - \sigma(\mathbf{w}_\star^\top \mathbf{x}_{\mathrm{new}}) \end{pmatrix} \middle\| \begin{pmatrix} (F(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}}) \\ 1 - F(\hat{\mathbf{w}}^\top \mathbf{x}_{\mathrm{new}})) \end{pmatrix}\right)\right]$$

(iii) Angular predictor minimizes the limit while Platt scaling minimizes $\hat{\ell}_{n_{\mathrm{ho}}}(F)$, convergence follows under usual minimizer consistency conditions

Illustration: Reliability Plots

1. **Predict:** Run calibrated classifier on test set to obtain predicted probabilities for each sample.
2. **Bin the Data:** Divide the range $[0, 1]$ of *predicted probabilities* into a fixed number of bins (e.g., 10 equally spaced bins).

Illustration: Reliability Plots

1. **Predict:** Run calibrated classifier on test set to obtain predicted probabilities for each sample.
2. **Bin the Data:** Divide the range $[0, 1]$ of *predicted probabilities* into a fixed number of bins (e.g., 10 equally spaced bins).
3. **Compute Averages:** For each bin, calculate *empirical frequency* (fraction of positive labels) for samples in that bin.
4. **Plot the Results:** Plot each bin's average predicted probability (x-axis) against its empirical frequency (y-axis). Overlay the ideal calibration line (the diagonal $y = x$).

Points that lie close to the 45° diagonal indicate good calibration.

Experiments

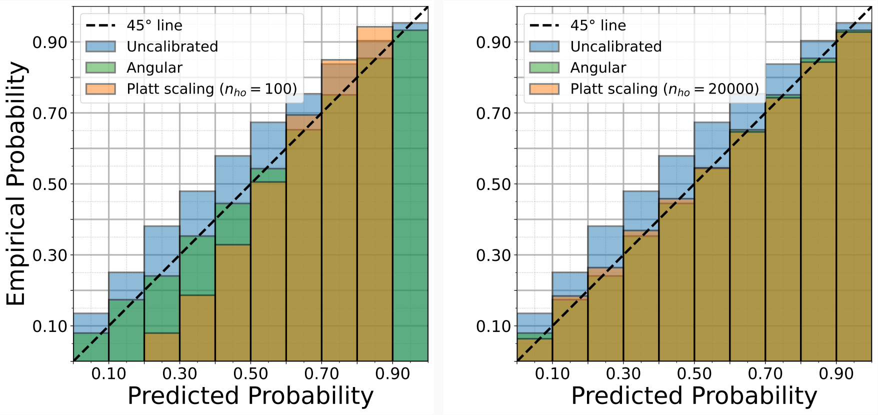


Figure 2: Reliability plots for logistic ridge predictor. Left panel uses a small holdout set for Platt scaling with $n_{\text{ho}} = 100$; Right panel uses a large holdout set with $n_{\text{ho}} = 2000$.

Universality? NonGaussian designs

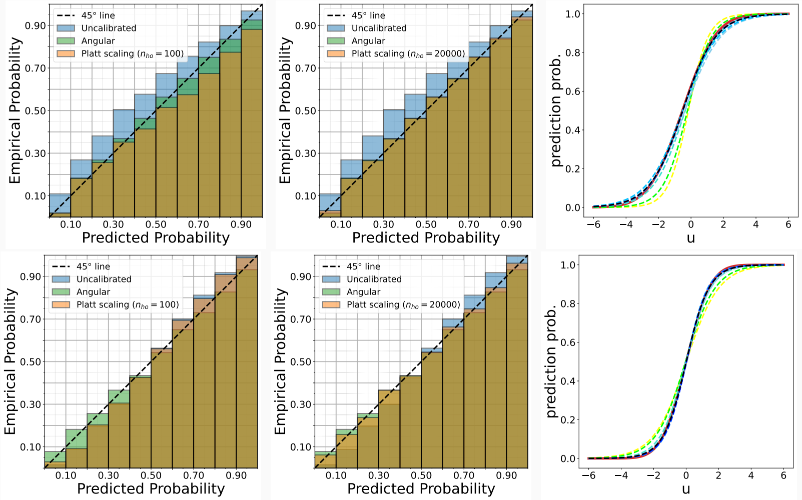


Figure 3: Rademacher entries. Upper Row: sigmoid. Bottom Row: clipped relu.

Universality? NonGaussian designs

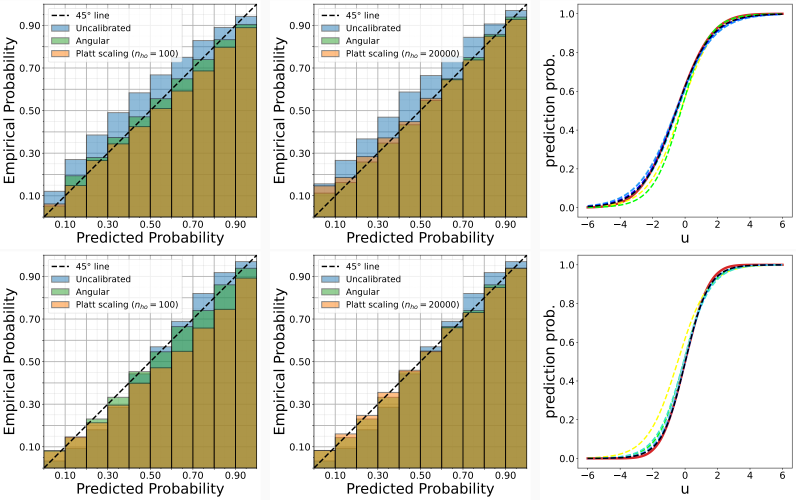


Figure 4: Unif[-1,1] entries. Upper Row: sigmoid. Bottom Row: clipped relu.

Remark on extension

Base proofs naturally extend to multi-index models:

Fix $K \geq 2$ and let $\mathbf{W}_\star = [\mathbf{w}_{\star 1}, \ldots, \mathbf{w}_{\star K}] \in \mathbb{R}^{d \times K}$. Define

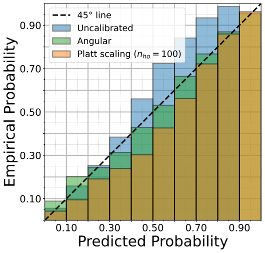
$$\mathbf{G} := \mathbf{W}_\star^\top \mathbf{x}_{\mathrm{new}} \in \mathbb{R}^K, \quad \mathrm{Prob}[y_{\mathrm{new}} = 1 | \mathbf{x}_{\mathrm{new}}] = \sigma(\mathbf{G}),$$

where σ is now generalized link (vector- or scalar-valued). This setup covers:

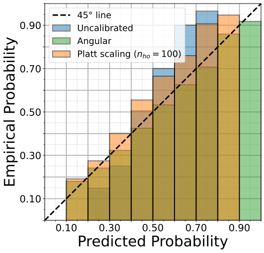
- Two-layer nets (frozen outer layer): $\sigma(\mathbf{u}) = \sum_k a_k \, \sigma(u_k)$
- Multi-class logistic models
- Additive index model: $\sigma(\mathbf{u}) = \sum_k f_k(u_k)$
- Interaction index model: $\sigma(\mathbf{u}) = \sum_k f_k(u_k) + \sum_{k<\ell} h_{k\ell}(u_k, u_\ell)$

Bottleneck: Angle estimation, needs case-by-case algorithms

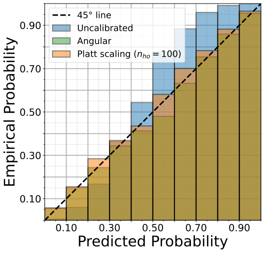
Semi-real experiments



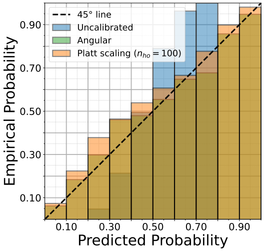
CIFAR-10



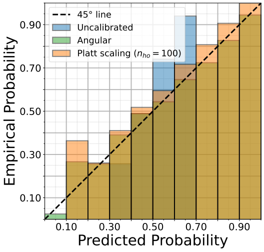
20 Newsgroups



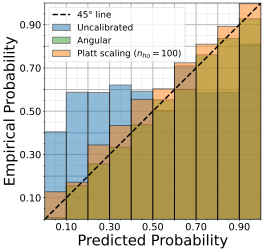
Tox21



Communities & Crime



DNA



Madelon

Conclusion: Key Takeaways

- Calibration is not just an empirical "hack"; it can be understood from first principles, even in challenging high-dimensional settings.

- The geometry of estimation, specifically the angle $\angle(\hat{w}, w^{\star})$, governs the quality of calibration in single-index odels.

- Our **Angular Calibration** method is simple, intuitive, and provably achieves both calibration and Bregman-optimality.

- Classical **Platt scaling** has surprising theoretical depth: it implicitly performs this optimal geometric correction in high dimensions.

Open Questions

Do our results admit universality? Should work analogous to existing proofs for training and generalization errors

Performance of other popular calibration algorithms? Do other methods such as temperature scaling and expectation consistency admit similar optimality properties?

Open Questions

Do our results admit universality? Should work analogous to existing proofs for training and generalization errors

Performance of other popular calibration algorithms? Do other methods such as temperature scaling and expectation consistency admit similar optimality properties?

When can calibration and optimality co-exist? Are other notions of optimality beyond minimizing Bregman divergence compatible with calibration?

Part II: Data Integration (Distribution Shift)

Why Data Integration?

- Modern data sources are heterogeneous: multi-source, multi-sensor, multi-modal.

Why Data Integration?

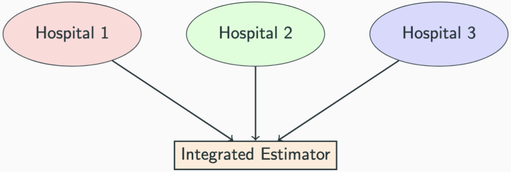
- Modern data sources are heterogeneous: multi-source, multi-sensor, multi-modal.
- Integrating diverse datasets enables improved, robust, generalizable models.

Why Data Integration?

- Modern data sources are heterogeneous: multi-source, multi-sensor, multi-modal.
- Integrating diverse datasets enables improved, robust, generalizable models.
- Critical in machine learning as well as modern science; Examples :

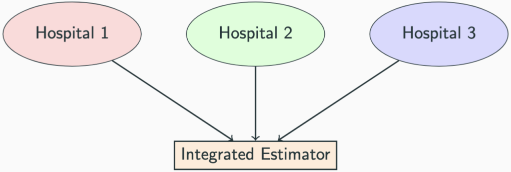
Why Data Integration?

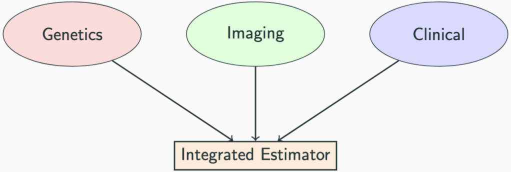
- Modern data sources are heterogeneous: multi-source, multi-sensor, multi-modal.
- Integrating diverse datasets enables improved, robust, generalizable models.
- Critical in machine learning as well as modern science; Examples :
    - **Healthcare:** Multiple hospitals measuring the same outcome but populations differ



Why Data Integration?

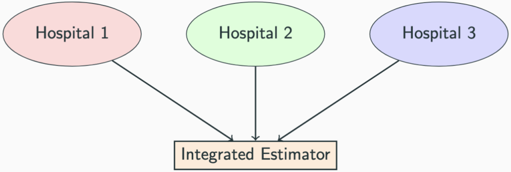
- Modern data sources are heterogeneous: multi-source, multi-sensor, multi-modal.
- Integrating diverse datasets enables improved, robust, generalizable models.
- Critical in machine learning as well as modern science; Examples :
  - **Healthcare:** Multiple hospitals measuring the same outcome but populations differ Or combining different modes, e.g., Imaging + Genetics + Clinical Measurement

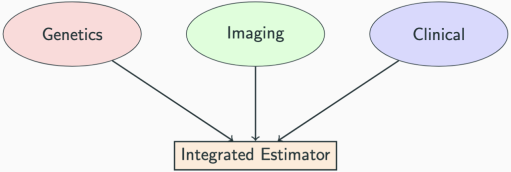




Why Data Integration?

- Modern data sources are heterogeneous: multi-source, multi-sensor, multi-modal.
- Integrating diverse datasets enables improved, robust, generalizable models.
- Critical in machine learning as well as modern science; Examples :
  - **Healthcare:** Multiple hospitals measuring the same outcome but populations differ Or combining different modes, e.g., Imaging + Genetics + Clinical Measurement

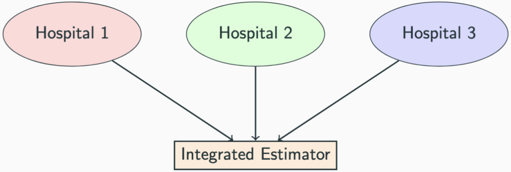




  - **Social Media:** Text + Images + Networks, **Sensor Fusion:** Radar + Cameras in autonomous driving, ...

Why Data Integration?

- Modern data sources are heterogeneous: multi-source, multi-sensor, multi-modal.
- Integrating diverse datasets enables improved, robust, generalizable models.
- Critical in machine learning as well as modern science; Examples :
  - **Healthcare:** Multiple hospitals measuring the same outcome but populations differ Or combining different modes, e.g., Imaging + Genetics + Clinical Measurement



  - **Social Media:** Text + Images + Networks, **Sensor Fusion:** Radar + Cameras in autonomous driving, ...

What Makes Distribution Shift Challenging?

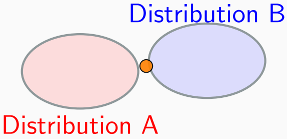
- Data come from sources with different noise, bias, and covariate structures.
- Distributional heterogeneity, e.g., covariate or label shift or mismatched features

What Makes Distribution Shift Challenging?

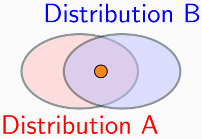
- Data come from sources with different noise, bias, and covariate structures.
- Distributional heterogeneity, e.g., covariate or label shift or mismatched features
- Theoretical understanding under-developed compared to “single distribution” statistics/machine learning.

What Makes Distribution Shift Challenging?

- Data come from sources with different noise, bias, and covariate structures.
- Distributional heterogeneity, e.g., covariate or label shift or mismatched features
- Theoretical understanding under-developed compared to “single distribution” statistics/machine learning.



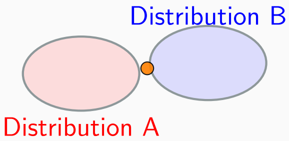
Scenario A: Integration useless



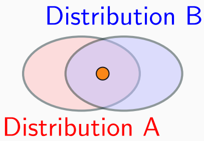
Scenario B: Integration useful

What Makes Distribution Shift Challenging?

- Data come from sources with different noise, bias, and covariate structures.
- Distributional heterogeneity, e.g., covariate or label shift or mismatched features
- Theoretical understanding under-developed compared to “single distribution” statistics/machine learning.



Scenario A: Integration useless



Scenario B: Integration useful

How do we identify in a data-adaptive manner whether we are in Scenario A or B?

Classical Approaches

- **Concatenation:** Merge features, train a joint model.
- **Ensemble Methods:** Train models separately, then combine predictions.
- **Transfer/Domain Adaptation:** Transfer knowledge from one “source” to another “target” domain.
- **Statistical Data Fusion:** Methods to combine inferences from parallel studies, e.g. Bayesian hierarchical models, meta-analysis, etc.

Classical Approaches

- **Concatenation:** Merge features, train a joint model.
- **Ensemble Methods:** Train models separately, then combine predictions.
- **Transfer/Domain Adaptation:** Transfer knowledge from one “source” to another “target” domain.
- **Statistical Data Fusion:** Methods to combine inferences from parallel studies, e.g. Bayesian hierarchical models, meta-analysis, etc.

Challenges remain in understanding the fundamentals of the problem
e.g., where does data fusion help vs hurt?

Our interest: Multi-source data integration

M datasets, from possibly different distributions. Samples i.i.d. in each.

For simplicity, assume linear models and $M = 2$. So we observe $(\mathbf{y}^{(k)}, \mathbf{X}^{(k)})$ with

$$\mathbf{y}^{(k)} = \mathbf{X}^{(k)}\boldsymbol{\theta}^{(k)} + \boldsymbol{\epsilon}^{(k)}, \quad k = 1, 2$$

Our interest: Multi-source data integration

M datasets, from possibly different distributions. Samples i.i.d. in each.

For simplicity, assume linear models and $M = 2$. So we observe $(\mathbf{y}^{(k)}, \mathbf{X}^{(k)})$ with

$$\mathbf{y}^{(k)} = \mathbf{X}^{(k)}\boldsymbol{\theta}^{(k)} + \boldsymbol{\epsilon}^{(k)}, \quad k = 1, 2$$

- $\mathbf{X}^{(k)} \in \mathbb{R}^{n_k \times d}$: $\boldsymbol{X}^{(k)} = \boldsymbol{Z}^{(k)}(\boldsymbol{\Sigma}^{(k)})^{1/2}$, $\boldsymbol{Z}^{(k)}$ entries i.i.d. mean 0, variance 1; $\boldsymbol{\Sigma}^{(k)}$ bounded eigenvalues
- n_k: Number of samples (typically $n_1 \gg n_2$); $n_1 + n_2 =: n$.
- $\boldsymbol{\epsilon}^{(k)}$ i.i.d. mean 0, finite variance σ^2

Our interest: Multi-source data integration

M datasets, from possibly different distributions. Samples i.i.d. in each.

For simplicity, assume linear models and $M = 2$. So we observe $(\mathbf{y}^{(k)}, \mathbf{X}^{(k)})$ with

$$\mathbf{y}^{(k)} = \mathbf{X}^{(k)}\boldsymbol{\theta}^{(k)} + \boldsymbol{\epsilon}^{(k)}, \quad k = 1, 2$$

- $\mathbf{X}^{(k)} \in \mathbb{R}^{n_k \times d}$: $\boldsymbol{X}^{(k)} = \boldsymbol{Z}^{(k)}(\boldsymbol{\Sigma}^{(k)})^{1/2}$, $\boldsymbol{Z}^{(k)}$ entries i.i.d. mean 0, variance 1; $\boldsymbol{\Sigma}^{(k)}$ bounded eigenvalues
- n_k: Number of samples (typically $n_1 \gg n_2$); $n_1 + n_2 =: n$.
- $\boldsymbol{\epsilon}^{(k)}$ i.i.d. mean 0, finite variance σ^2

Distribution Shift:

- Concept Shift: $\boldsymbol{\theta}^{(1)} \neq \boldsymbol{\theta}^{(2)}$.
- Covariate Shift: $\boldsymbol{\Sigma}^{(1)} \neq \boldsymbol{\Sigma}^{(2)}$

Goal

Predict in target that has low sample size with better accuracy by using source samples rather than using target only data.

Goal

Predict in target that has low sample size with better accuracy by using source samples rather than using target only data.

Question: How do we leverage the source data in a principled way to improve prediction accuracy?

Goal

Predict in target that has low sample size with better accuracy by using source samples rather than using target only data.

Question: How do we leverage the source data in a principled way to improve prediction accuracy?

This talk: Study in an overparametrized regime $(d > n_1 + n_2)$ through the lens of min-norm interpolation: one of the most commonly seen implicit regularized limits in the ML literature

Quick Detour: Implicit Regularization and Min-norm Interpolation

Implicit Regularization

With suitable initialization, step size, etc. modern ML algorithms show implicit regularization to special prediction rules/classifiers

Implicit Regularization

With suitable initialization, step size, etc. modern ML algorithms show implicit regularization to special prediction rules/classifiers

Examples abound:

– One of earliest example: AdaBoost (Zhang and Yu '05)

– GD (suitable initialization...) on overparametrized unregularized logistic loss (Soudry et al. '18)

– GD on linear convolutional neural networks (Gunasekar '18)

– GD training a self-attention layer, i.e. a stylized version of a transformer (Tarzanagh et al '23; Vasudeva et al '24)

Utility?

- Often yields new insights on algorithms
- Common recipe: Study the implicit regularized limit
  → alg. properties at convergence

Utility?

- Often yields new insights on algorithms
- Common recipe: Study the implicit regularized limit

  → alg. properties at convergence

Theorem (An example result: Liang and S. AoS '22)
In binary classification, with proper (non-vanishing) stepsize, Adaboost iterates $\hat{\theta}^t$ *satisfy for all* $t \geq T(n, d, SNR)$

$$\textit{Misclassification Error}(\hat{\theta}^t) \approx \mathbb{P}\left(c_1^\star Y Z_1 + c_2^\star Z_2 < 0\right), \quad a.s.$$

- *Precise characterization of* (Y, Z_1, Z_2) *and* $(c_1^\star, c_2^\star, s^\star)$

Utility?

- Often yields new insights on algorithms
- Common recipe: Study the implicit regularized limit

  → alg. properties at convergence

Theorem (An example result: Liang and S. AoS '22)
In binary classification, with proper (non-vanishing) stepsize, Adaboost iterates $\hat{\theta}^t$ *satisfy for all* $t \geq T(n, d, SNR)$

$$\textit{Misclassification Error}(\hat{\theta}^t) \approx \mathbb{P}\left(c_1^\star Y Z_1 + c_2^\star Z_2 < 0\right), \quad a.s.$$

- *Precise characterization of* (Y, Z_1, Z_2) *and* $(c_1^\star, c_2^\star, s^\star)$
- *Approach: Characterize prediction error of the limiting min-*ℓ_1*-norm interpolator and use connection with AdaBoost;*

Utility?

- Often yields new insights on algorithms
- Common recipe: Study the implicit regularized limit
  → alg. properties at convergence

Theorem (An example result: Liang and S. AoS '22)
In binary classification, with proper (non-vanishing) stepsize, Adaboost iterates $\hat{\theta}^t$ *satisfy for all* $t \geq T(n, d, SNR)$

$$\textit{Misclassification Error}(\hat{\theta}^t) \approx \mathbb{P}\left(c_1^\star Y Z_1 + c_2^\star Z_2 < 0\right), \quad a.s.$$

- *Precise characterization of* (Y, Z_1, Z_2) *and* $(c_1^\star, c_2^\star, s^\star)$
- *Approach: Characterize prediction error of the limiting min-*ℓ_1*-norm interpolator and use connection with AdaBoost; Complements classical bounds by Schapire et al '98, Koltchinskii and Panchenko '05;*

Utility?

- Often yields new insights on algorithms
- Common recipe: Study the implicit regularized limit
  → alg. properties at convergence

Theorem (An example result: Liang and S. AoS '22)
In binary classification, with proper (non-vanishing) stepsize, Adaboost iterates $\hat{\theta}^t$ satisfy for all $t \geq T(n, d, SNR)$

$$\textit{Misclassification Error}(\hat{\theta}^t) \approx \mathbb{P}\left(c_1^\star Y Z_1 + c_2^\star Z_2 < 0\right), \quad a.s.$$

- *Precise characterization of (Y, Z_1, Z_2) and $(c_1^\star, c_2^\star, s^\star)$*
- *Approach: Characterize prediction error of the limiting min-ℓ_1-norm interpolator and use connection with AdaBoost; Complements classical bounds by Schapire et al '98, Koltchinskii and Panchenko '05; similar characterization possible for any algorithm converging to these interpolators*

Min-norm interpolators

For i.i.d. data $(y_i, \boldsymbol{x}_i)$ that can be perfectly interpolated, define the min-ℓ_q-norm interpolator as

$$\hat{\boldsymbol{\theta}}_q = \arg\min \|\boldsymbol{\theta}\|_q \quad \text{s.t.} \quad y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}, y_i \in \mathbb{R} \quad \text{or} \quad y_i \boldsymbol{x}_i^\top \boldsymbol{\theta} \geq 0, y_i \in \{1, -1\}$$

- Important class–arises as implicit regularized limits of many algs

Min-norm interpolators

For i.i.d. data $(y_i, \boldsymbol{x}_i)$ that can be perfectly interpolated, define the min-ℓ_q-norm interpolator as

$$\hat{\boldsymbol{\theta}}_q = \arg\min \|\boldsymbol{\theta}\|_q \quad \text{s.t.} \quad y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}, y_i \in \mathbb{R} \quad \text{or} \quad y_i \boldsymbol{x}_i^\top \boldsymbol{\theta} \geq 0, y_i \in \{1, -1\}$$

- Important class–arises as implicit regularized limits of many algs

Min-norm interpolators

For i.i.d. data $(y_i, \boldsymbol{x}_i)$ that can be perfectly interpolated, define the min-ℓ_q-norm interpolator as

$$\hat{\boldsymbol{\theta}}_q = \arg\min \|\boldsymbol{\theta}\|_q \quad \text{s.t.} \quad y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}, y_i \in \mathbb{R} \quad \text{or} \quad y_i \boldsymbol{x}_i^\top \boldsymbol{\theta} \geq 0, y_i \in \{1, -1\}$$

- Important class–arises as implicit regularized limits of many algs
- Extensively studied under single distribution overparametrized models (Montanari et al. '19, Deng et al. '19, Liang and S. '20, Chatterji et al. '20 Donhauser et al. '21, Zhou et al. '21, '22)

Min-norm interpolators

For i.i.d. data $(y_i, \boldsymbol{x}_i)$ that can be perfectly interpolated, define the min-ℓ_q-norm interpolator as

$$\hat{\boldsymbol{\theta}}_q = \arg\min \|\boldsymbol{\theta}\|_q \quad \text{s.t.} \quad y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}, y_i \in \mathbb{R} \quad \text{or} \quad y_i \boldsymbol{x}_i^\top \boldsymbol{\theta} \geq 0, y_i \in \{1, -1\}$$

- Important class–arises as implicit regularized limits of many algs
- Extensively studied under single distribution overparametrized models (Montanari et al. '19, Deng et al. '19, Liang and S. '20, Chatterji et al. '20 Donhauser et al. '21, Zhou et al. '21, '22)
- Under-explored in presence of distributions shifts; Mallinar et al. '24, Patil et al. '24 study out-of-distribution settings with no target data during training

Natural analogue of min-norm interpolators under distribution shifts?

- Start from simplest: $q = 2$

Natural analogue of min-norm interpolators under distribution shifts?

- Start from simplest: $q = 2$
- How do we think about the analogue for distribution shift settings?

Natural analogue of min-norm interpolators under distribution shifts?

- Start from simplest: $q = 2$
- How do we think about the analogue for distribution shift settings?
- Revisit single training data results

Natural analogue of min-norm interpolators under distribution shifts?

– Start from simplest: $q = 2$

– How do we think about the analogue for distribution shift settings?

– Revisit single training data results

Different formulations

- Min-ℓ_2-norm interpolator: $\arg\min \|\boldsymbol{\theta}\|_2$ s.t. $y_i = \boldsymbol{x}_i^\top \theta$ for all i
- Alternate (Hastie et al. '22): Ridgeless or $\lambda \to 0^+$ limit of solution to

$$\hat{\boldsymbol{\theta}}_\lambda = \arg\min_{\boldsymbol{\theta}} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|^2$$

Segue from regularized regression

- Regularized regression extensively studied for transfer learning (Yang et al. '20, Bastani '21, Cai et al. '21 Li et al. '22, Zhang et al. '22, Tian and Feng '23, Zhou et al. '24, new synthetic correlated data models proposed in Gerace et al. '22)

Segue from regularized regression

- Regularized regression extensively studied for transfer learning (Yang et al. '20, Bastani '21, Cai et al. '21 Li et al. '22, Zhang et al. '22, Tian and Feng '23, Zhou et al. '24, new synthetic correlated data models proposed in Gerace et al. '22)

- Natural regularized loss: for suitable weights $w_1, w_2 \geq 0$,

$$\arg\min_{\boldsymbol{\theta}} \left\{ \frac{w_1}{n} \|\boldsymbol{y}^{(1)} - \boldsymbol{X}^{(1)}\boldsymbol{\theta}\|_2^2 + \frac{w_2}{n} \|\boldsymbol{y}^{(2)} - \boldsymbol{X}^{(2)}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2 \right\}$$

Segue from regularized regression

- Regularized regression extensively studied for transfer learning (Yang et al. '20, Bastani '21, Cai et al. '21 Li et al. '22, Zhang et al. '22, Tian and Feng '23, Zhou et al. '24, new synthetic correlated data models proposed in Gerace et al. '22)
- Natural regularized loss: for suitable weights $w_1, w_2 \geq 0$,

$$\arg\min_{\boldsymbol{\theta}} \left\{ \frac{w_1}{n} \|\boldsymbol{y}^{(1)} - \boldsymbol{X}^{(1)}\boldsymbol{\theta}\|_2^2 + \frac{w_2}{n} \|\boldsymbol{y}^{(2)} - \boldsymbol{X}^{(2)}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2 \right\}$$

- Ridgeless limit for any w_1, w_2 is a *pooled min-ℓ_2-norm interpolator:*

$$\hat{\boldsymbol{\theta}}_{\text{pool}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_2 \quad \text{s.t.} \quad y_i^{(k)} = \boldsymbol{x}_i^{(k)\top}\boldsymbol{\theta} \quad \text{for all i,k}$$

In some sense, this is both early and intermediate fusion estimator

Goal

- Characterize its out-of-sample prediction error—dependence on dimensionality, level of shifts, SNRs, etc.

- Formally, for $\boldsymbol{x}_0 \sim \mathbb{P}_{\boldsymbol{x}^{(2)}}$, characterize out-of-sample prediction risk on target distribution

$$\text{Risk} = R(\hat{\boldsymbol{\theta}}_{\text{pool}}) = \mathbb{E}[(\boldsymbol{x}_0^\top \hat{\boldsymbol{\theta}}_{\text{pool}} - \boldsymbol{x}_0^\top \boldsymbol{\theta}^{(2)})^2 | \boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}]$$

- Guarantees will be w.h.p. over distribution of covariates;

Risk under Concept Shift

Theorem (Song, Bhattacharya, S. '24+)
Assume $\boldsymbol{\Sigma}^{(1)} = \boldsymbol{\Sigma}^{(2)} = \boldsymbol{I}, \boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$ *Gaussian. With high probability over randomness of* $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$

$$R(\hat{\boldsymbol{\theta}}_{pool}) = \frac{n}{d-n}\sigma^2 + \frac{d-n}{d}||\boldsymbol{\theta}^{(2)}||_2^2 + \frac{n_1(d-n_1)}{d(d-n)}||\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}||_2^2 + o(1)$$

Risk under Concept Shift

Theorem (Song, Bhattacharya, S. '24+)
Assume $\boldsymbol{\Sigma}^{(1)} = \boldsymbol{\Sigma}^{(2)} = \boldsymbol{I}, \boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$ *Gaussian. With high probability over randomness of* $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$

$$R(\hat{\boldsymbol{\theta}}_{pool}) = \frac{n}{d-n}\sigma^2 + \frac{d-n}{d}||\boldsymbol{\theta}^{(2)}||_2^2 + \frac{n_1(d-n_1)}{d(d-n)}||\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}||_2^2 + o(1)$$

- For target-only interpolator, with high probability (Hastie et al 2022),

$$R(\hat{\boldsymbol{\theta}}^{(2)}) = \frac{n_2}{d-n_2}\sigma^2 + \frac{d-n_2}{p}||\boldsymbol{\theta}^{(2)}||_2^2 + o(1)$$

Risk under Concept Shift

Theorem (Song, Bhattacharya, S. '24+)
Assume $\boldsymbol{\Sigma}^{(1)} = \boldsymbol{\Sigma}^{(2)} = \boldsymbol{I}, \boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$ Gaussian. With high probability over randomness of $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$

$$R(\hat{\boldsymbol{\theta}}_{pool}) = \frac{n}{d-n}\sigma^2 + \frac{d-n}{d}||\boldsymbol{\theta}^{(2)}||_2^2 + \frac{n_1(d-n_1)}{d(d-n)}||\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}||_2^2 + o(1)$$

- For target-only interpolator, with high probability (Hastie et al 2022),

$$R(\hat{\boldsymbol{\theta}}^{(2)}) = \frac{n_2}{d-n_2}\sigma^2 + \frac{d-n_2}{p}||\boldsymbol{\theta}^{(2)}||_2^2 + o(1)$$

- Involved trade-offs between target SNR, degree of shift, d, n_1, n_2

Risk under Concept Shift

Theorem (Song, Bhattacharya, S. '24+)
Assume $\boldsymbol{\Sigma}^{(1)} = \boldsymbol{\Sigma}^{(2)} = \boldsymbol{I}, \boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$ *Gaussian. With high probability over randomness of* $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$

$$R(\hat{\boldsymbol{\theta}}_{pool}) = \frac{n}{d-n}\sigma^2 + \frac{d-n}{d}||\boldsymbol{\theta}^{(2)}||_2^2 + \frac{n_1(d-n_1)}{d(d-n)}||\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}||_2^2 + o(1)$$

- For target-only interpolator, with high probability (Hastie et al 2022),

$$R(\hat{\boldsymbol{\theta}}^{(2)}) = \frac{n_2}{d-n_2}\sigma^2 + \frac{d-n_2}{p}||\boldsymbol{\theta}^{(2)}||_2^2 + o(1)$$

- Involved trade-offs between target SNR, degree of shift, d, n_1, n_2
- Trade-off even between first two coefficients
- Universality results ongoing with Kenny Gu

Risk under Concept Shift

Theorem (Song, Bhattacharya, S. '24+)
Assume $\boldsymbol{\Sigma}^{(1)} = \boldsymbol{\Sigma}^{(2)} = \boldsymbol{I}, \boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$ Gaussian. With high probability over randomness of $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$

$$R(\hat{\boldsymbol{\theta}}_{pool}) = \frac{n}{d-n}\sigma^2 + \frac{d-n}{d}||\boldsymbol{\theta}^{(2)}||_2^2 + \frac{n_1(d-n_1)}{d(d-n)}||\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}||_2^2 + o(1)$$

- For target-only interpolator, with high probability (Hastie et al 2022),

$$R(\hat{\boldsymbol{\theta}}^{(2)}) = \frac{n_2}{d-n_2}\sigma^2 + \frac{d-n_2}{p}||\boldsymbol{\theta}^{(2)}||_2^2 + o(1)$$

- Involved trade-offs between target SNR, degree of shift, d, n_1, n_2
- Trade-off even between first two coefficients
- Universality results ongoing with Kenny Gu

Concept Shift: Corollary

SNR (Signal-to-noise ratio) $:= \frac{||\boldsymbol{\theta}^{(2)}||_2^2}{\sigma^2}$, SSR (Shift-to-signal ratio) $:= \frac{||\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}||_2^2}{||\boldsymbol{\theta}^{(2)}||_2^2}$

Concept Shift: Corollary

SNR (Signal-to-noise ratio) $:= \frac{||\boldsymbol{\theta}^{(2)}||_2^2}{\sigma^2}$, SSR (Shift-to-signal ratio) $:= \frac{||\boldsymbol{\theta}^{(1)}-\boldsymbol{\theta}^{(2)}||_2^2}{||\boldsymbol{\theta}^{(2)}||_2^2}$

Theorem (Song, Bhattacharya, S. '24+)
Under model shift assumptions

1. *If* $\mathrm{SNR} \leq \frac{\mathrm{d}^2}{(\mathrm{d}-\mathrm{n})(\mathrm{d}-\mathrm{n}_2)}$, *then*
$$R(\hat{\boldsymbol{\theta}}^{(2)}) \leq R(\hat{\boldsymbol{\theta}}_{pool}) + o(1) \tag{1}$$

2. *Else, define* $\rho := \frac{d-n}{d-n_1} - \frac{d^2}{(d-n_1)(d-n_2)} \cdot \frac{1}{\mathrm{SNR}}$. *When* SSR $\geq \rho$, *then (1) holds;*

Concept Shift: Corollary

SNR (Signal-to-noise ratio) $:= \frac{||\boldsymbol{\theta}^{(2)}||_2^2}{\sigma^2}$, SSR (Shift-to-signal ratio) $:= \frac{||\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}||_2^2}{||\boldsymbol{\theta}^{(2)}||_2^2}$

Theorem (Song, Bhattacharya, S. '24+)
Under model shift assumptions

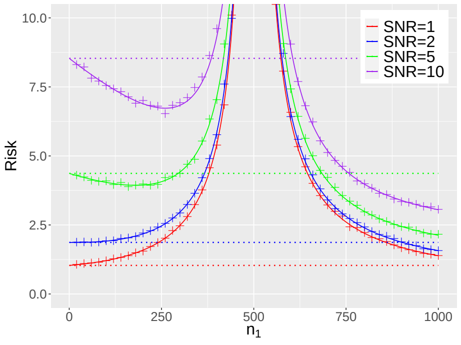
1. *If* $\mathrm{SNR} \leq \frac{\mathrm{d}^2}{(\mathrm{d}-\mathrm{n})(\mathrm{d}-\mathrm{n}_2)}$, *then*
$$R(\hat{\boldsymbol{\theta}}^{(2)}) \leq R(\hat{\boldsymbol{\theta}}_{pool}) + o(1) \tag{1}$$

2. *Else, define* $\rho := \frac{d-n}{d-n_1} - \frac{d^2}{(d-n_1)(d-n_2)} \cdot \frac{1}{\mathrm{SNR}}$. *When* SSR $\geq \rho$, *then (1) holds; when* SSR $< \rho$, *then*
$$R(\hat{\boldsymbol{\theta}}_{pool}) \leq R(\hat{\boldsymbol{\theta}}^{(2)}) + o(1)$$

Takeaways: (i) When the SNR of target is small, pooling always hurts, increases noise

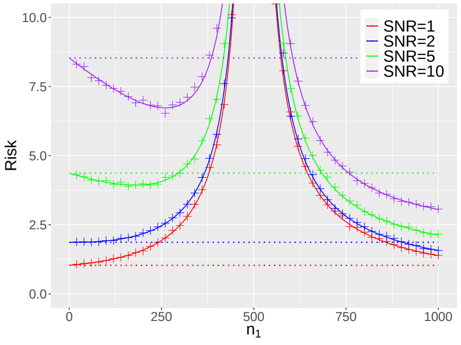
(ii) If SNR is large transfer gain depends on the degree of shift

Effects of SNR



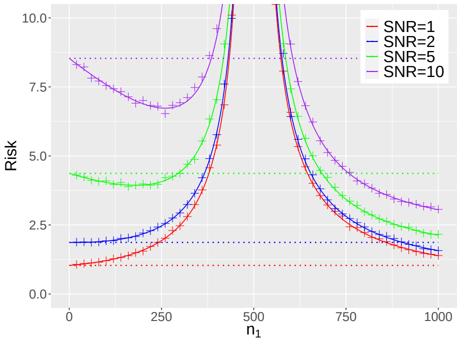
- $\text{SNR} = \|\boldsymbol{\theta}^{(2)}\|^2/\sigma^2$
- $n_2 = 100$, $d = 600$, Shift-to-signal ratio (SSR)$= \|\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}\|^2/\|\boldsymbol{\theta}^{(2)}\|^2 = 0.2$

Effects of SNR



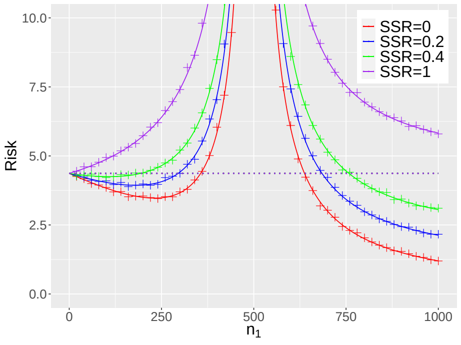
- SNR $= \|\boldsymbol{\theta}^{(2)}\|^2/\sigma^2$
- $n_2 = 100$, $d = 600$, Shift-to-signal ratio (SSR)$= \|\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}\|^2/\|\boldsymbol{\theta}^{(2)}\|^2 = 0.2$
- Takeaways: For low SNR, pooling does not help

Effects of SNR



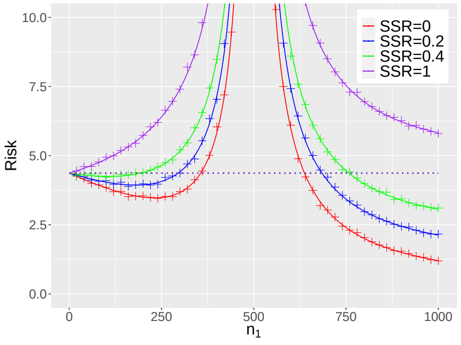
- SNR $= \|\boldsymbol{\theta}^{(2)}\|^2/\sigma^2$
- $n_2 = 100$, $d = 600$, Shift-to-signal ratio (SSR)$= \|\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}\|^2/\|\boldsymbol{\theta}^{(2)}\|^2 = 0.2$
- Takeaways: For low SNR, pooling does not help
- For higher SNR it does till n_1 below a threshold

Effects of SSR



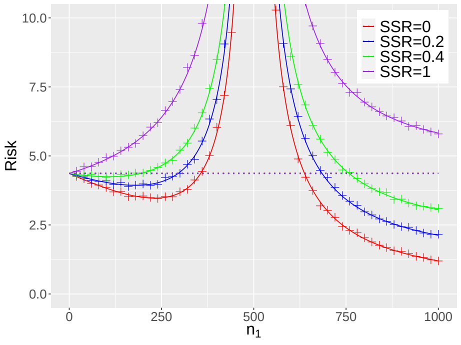
- $n_2 = 100,\ d = 600,\ \mathrm{SNR} = 5$

Effects of SSR



- $n_2 = 100$, $d = 600$, $\mathrm{SNR} = 5$
- Transfer helps for low SSR but not higher SSR

Effects of SSR



- $n_2 = 100$, $d = 600$, $\text{SNR} = 5$
- Transfer helps for low SSR but not higher SSR
- Key: Data-driven SNR, SSR estimators in paper
  Useful to decide to pool or not to pool

Optimal target sample size

Theorem (Song, Bhattacharya, S. '24+)
In the setting of our previous theorem, the following target sample size optimizes the generalization error:

$$n_{2,opt} = \arg\min_{n_2 \in \mathbb{N}} R(\hat{\theta}_{pool}) = \left(d - n_1 - \sqrt{\frac{d^2}{\mathrm{SNR}} + n_1 \mathrm{SSR}} \right)_+$$

- Consistent SNR, SSR estimators in the paper, producing consistent $\hat{n}_{2,opt}$

Optimal target sample size

Theorem (Song, Bhattacharya, S. '24+)
In the setting of our previous theorem, the following target sample size optimizes the generalization error:

$$n_{2,opt} = \arg\min_{n_2 \in \mathbb{N}} R(\hat{\theta}_{pool}) = \left(d - n_1 - \sqrt{\frac{d^2}{\mathrm{SNR}} + n_1 \mathrm{SSR}} \right)_+$$

- Consistent SNR, SSR estimators in the paper, producing consistent $\hat{n}_{2,opt}$
- Similar results can be derived for optimal source sample size for generalization

Optimal target sample size

Theorem (Song, Bhattacharya, S. '24+)
In the setting of our previous theorem, the following target sample size optimizes the generalization error:

$$n_{2,opt} = \arg\min_{n_2 \in \mathbb{N}} R(\hat{\theta}_{pool}) = \left(d - n_1 - \sqrt{\frac{d^2}{\mathrm{SNR}} + n_1 \mathrm{SSR}} \right)_+$$

- Consistent SNR, SSR estimators in the paper, producing consistent $\hat{n}_{2,opt}$
- Similar results can be derived for optimal source sample size for generalization
- Key takeaway: Optimal sample size involves SNR, SSR, etc.

Optimal target sample size

Theorem (Song, Bhattacharya, S. ’24+)
In the setting of our previous theorem, the following target sample size optimizes the generalization error:

$$n_{2,opt} = \arg\min_{n_2 \in \mathbb{N}} R(\hat{\theta}_{pool}) = \left(d - n_1 - \sqrt{\frac{d^2}{\mathrm{SNR}} + n_1 \mathrm{SSR}} \right)_+$$

- Consistent SNR, SSR estimators in the paper, producing consistent $\hat{n}_{2,opt}$
- Similar results can be derived for optimal source sample size for generalization
- Key takeaway: Optimal sample size involves SNR, SSR, etc.
- Estimate of RHS provide principled guidance for how many samples to include

Optimal target sample size

Theorem (Song, Bhattacharya, S. '24+)
In the setting of our previous theorem, the following target sample size optimizes the generalization error:

$$n_{2,opt} = \arg\min_{n_2 \in \mathbb{N}} R(\hat{\theta}_{pool}) = \left(d - n_1 - \sqrt{\frac{d^2}{\mathrm{SNR}} + n_1 \mathrm{SSR}} \right)_+$$

- Consistent SNR, SSR estimators in the paper, producing consistent $\hat{n}_{2,opt}$
- Similar results can be derived for optimal source sample size for generalization
- Key takeaway: Optimal sample size involves SNR, SSR, etc.
- Estimate of RHS provide principled guidance for how many samples to include

Including too much can hurt performance: contrary to traditional statistical wisdom!

Covariate shift: Setting

- Recall $\boldsymbol{y}^{(k)} = \boldsymbol{X}^{(k)}\theta^{(k)} + \varepsilon^{(k)}$; $k = 1$ source, $k = 2$ target
- $\boldsymbol{X}^{(k)} = \boldsymbol{Z}^{(k)}(\boldsymbol{\Sigma}^{(k)})^{1/2}$, $\boldsymbol{Z}^{(k)}$ entries i.i.d. mean 0, variance 1

Covariate shift: Setting

- Recall $\boldsymbol{y}^{(k)} = \boldsymbol{X}^{(k)}\boldsymbol{\theta}^{(k)} + \varepsilon^{(k)}$; $k = 1$ source, $k = 2$ target
- $\boldsymbol{X}^{(k)} = \boldsymbol{Z}^{(k)}(\boldsymbol{\Sigma}^{(k)})^{1/2}$, $\boldsymbol{Z}^{(k)}$ entries i.i.d. mean 0, variance 1
- Assume $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(2)}$, $(\boldsymbol{\Sigma}^{(1)}, \boldsymbol{\Sigma}^{(2)}) = \boldsymbol{V}(\boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)})\boldsymbol{V}^\top$ (Simultaneous diagonalizability)

Covariate shift: Setting

- Recall $\boldsymbol{y}^{(k)} = \boldsymbol{X}^{(k)}\boldsymbol{\theta}^{(k)} + \varepsilon^{(k)}$; $k = 1$ source, $k = 2$ target
- $\boldsymbol{X}^{(k)} = \boldsymbol{Z}^{(k)}(\boldsymbol{\Sigma}^{(k)})^{1/2}$, $\boldsymbol{Z}^{(k)}$ entries i.i.d. mean 0, variance 1
- Assume $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(2)}$, $(\boldsymbol{\Sigma}^{(1)}, \boldsymbol{\Sigma}^{(2)}) = \boldsymbol{V}(\boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)})\boldsymbol{V}^\top$ (Simultaneous diagonalizability)
- Relevant distributions (also appear in Hastie et al '22):

$$(i)\hat{H}_d(a,b) := \frac{1}{d}\sum_{i=1}^{d} 1_{\{(a,b)=(\lambda_i^{(1)},\lambda_i^{(2)})\}},$$

$$(ii)\hat{G}_d(a,b) := \frac{1}{||\boldsymbol{\theta}^{(2)}||_2^2}\sum_{i=1}^{d}\langle\boldsymbol{\theta}^{(2)}, \boldsymbol{v}_i\rangle^2 1_{\{(a,b)=(\lambda_i^{(1)},\lambda_i^{(2)})\}}$$

Risk under Covariate Shift

Theorem (Song, Bhattacharya, S. '24+)
Error variance: σ^2, dimension to total sample size ratio $d/n = \gamma$; $n = n_1 + n_2$

$$R(\hat{\boldsymbol{\theta}}_{pool}) = -\sigma^2\gamma \int \frac{\lambda^{(2)}(\tilde{a}_3\lambda^{(1)} + \tilde{a}_4\lambda^{(2)})}{(\tilde{a}_1\lambda^{(1)} + \tilde{a}_2\lambda^{(2)} + 1)^2} d\hat{H}_d(\lambda^{(1)}, \lambda^{(2)})$$

$$+ ||\boldsymbol{\theta}^{(2)}||_2^2 \cdot \int \frac{\tilde{b}_3\lambda^{(1)} + (\tilde{b}_4 + 1)\lambda^{(2)}}{(\tilde{b}_1\lambda^{(1)} + \tilde{b}_2\lambda^{(2)} + 1)^2} d\hat{G}_d(\lambda^{(1)}, \lambda^{(2)}) + o(1)$$

- *Precise description of constants $\tilde{a}_i, \tilde{b}_i$ in paper*

Risk under Covariate Shift

Theorem (Song, Bhattacharya, S. '24+)
Error variance: σ^2, dimension to total sample size ratio $d/n = \gamma$; $n = n_1 + n_2$

$$R(\hat{\boldsymbol{\theta}}_{pool}) = -\,\sigma^2\gamma \int \frac{\lambda^{(2)}(\tilde{a}_3\lambda^{(1)} + \tilde{a}_4\lambda^{(2)})}{(\tilde{a}_1\lambda^{(1)} + \tilde{a}_2\lambda^{(2)} + 1)^2} d\hat{H}_d(\lambda^{(1)}, \lambda^{(2)}) + ||\boldsymbol{\theta}^{(2)}||_2^2 \cdot \int \frac{\tilde{b}_3\lambda^{(1)} + (\tilde{b}_4 + 1)\lambda^{(2)}}{(\tilde{b}_1\lambda^{(1)} + \tilde{b}_2\lambda^{(2)} + 1)^2} d\hat{G}_d(\lambda^{(1)}, \lambda^{(2)}) + o(1)$$

- *Precise description of constants $\tilde{a}_i, \tilde{b}_i$ in paper*
- *Depends only on $\lambda^{(i)}$'s not $\boldsymbol{v}_i'$s*

Risk under Covariate Shift

Theorem (Song, Bhattacharya, S. ’24+)
Error variance: σ^2, dimension to total sample size ratio $d/n = \gamma$; $n = n_1 + n_2$

$$R(\hat{\boldsymbol{\theta}}_{pool}) = -\sigma^2\gamma \int \frac{\lambda^{(2)}(\tilde{a}_3\lambda^{(1)} + \tilde{a}_4\lambda^{(2)})}{(\tilde{a}_1\lambda^{(1)} + \tilde{a}_2\lambda^{(2)} + 1)^2} d\hat{H}_d(\lambda^{(1)}, \lambda^{(2)}) + ||\boldsymbol{\theta}^{(2)}||_2^2 \cdot \int \frac{\tilde{b}_3\lambda^{(1)} + (\tilde{b}_4 + 1)\lambda^{(2)}}{(\tilde{b}_1\lambda^{(1)} + \tilde{b}_2\lambda^{(2)} + 1)^2} d\hat{G}_d(\lambda^{(1)}, \lambda^{(2)}) + o(1)$$

- *Precise description of constants $\tilde{a}_i, \tilde{b}_i$ in paper*
- *Depends only on $\lambda^{(i)}$'s not $\boldsymbol{v}_i'$s*
- *More involved to study transfer versus target only performance*

Example (Does covariate shift help?)

- Setup: Define $\boldsymbol{M}$ to be diagonal with reciprocal eigenvalues (d even), $\lambda^{(1)}_{d+1-i} = 1/\lambda^{(1)}_i$ for $i = 1, ..., d/2$
- Define $\hat{R}(\boldsymbol{M}) := R(\hat{\boldsymbol{\theta}}_{\text{pool}} | \boldsymbol{\Sigma}^{(1)} = \boldsymbol{M}, \boldsymbol{\Sigma}^{(2)} = \boldsymbol{I})$
- So $\hat{R}(\boldsymbol{I})$ denotes the no-covariate shift case

Example (Does covariate shift help?)

- Setup: Define $\boldsymbol{M}$ to be diagonal with reciprocal eigenvalues (d even), $\lambda^{(1)}_{d+1-i} = 1/\lambda^{(1)}_i$ for $i = 1, ..., d/2$
- Define $\hat{R}(\boldsymbol{M}) := R(\hat{\boldsymbol{\theta}}_{\text{pool}} | \boldsymbol{\Sigma}^{(1)} = \boldsymbol{M}, \boldsymbol{\Sigma}^{(2)} = \boldsymbol{I})$
- So $\hat{R}(\boldsymbol{I})$ denotes the no-covariate shift case

Theorem (Song, Bhattacharya, S. '24+)

1. *When* $n_1 < \min\{d/2, p - n_2\}$, *then*

$$\hat{R}(\boldsymbol{M}) < \hat{R}(\boldsymbol{I}) + o(1)$$

Example (Does covariate shift help?)

- Setup: Define $\boldsymbol{M}$ to be diagonal with reciprocal eigenvalues (d even), $\lambda^{(1)}_{d+1-i} = 1/\lambda^{(1)}_i$ for $i = 1, ..., d/2$
- Define $\hat{R}(\boldsymbol{M}) := R(\hat{\boldsymbol{\theta}}_{\text{pool}} | \boldsymbol{\Sigma}^{(1)} = \boldsymbol{M}, \boldsymbol{\Sigma}^{(2)} = \boldsymbol{I})$
- So $\hat{R}(\boldsymbol{I})$ denotes the no-covariate shift case

Theorem (Song, Bhattacharya, S. '24+)

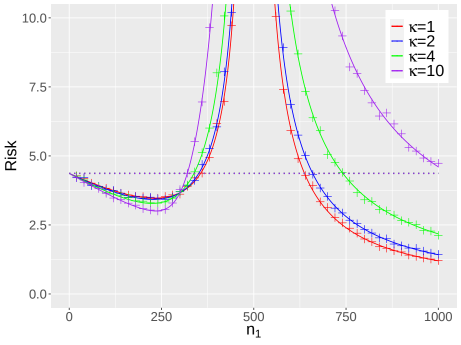
1. *When* $n_1 < \min\{d/2, p - n_2\}$, *then*

$$\hat{R}(\boldsymbol{M}) < \hat{R}(\boldsymbol{I}) + o(1)$$

2. *When* $d/2 \leq n_1 < d - n_2$, *then,*

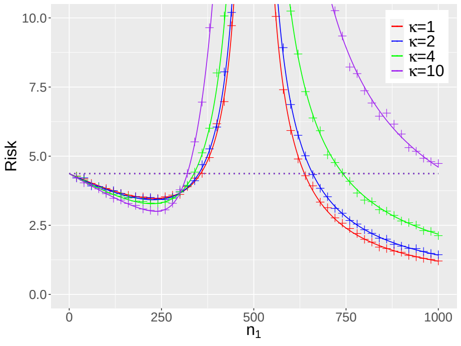
$$\hat{R}(\boldsymbol{M}) \geq \hat{R}(\boldsymbol{I}) + o(1)$$

Illustration



- $\lambda^{(1)}_{d+1-i} = 1/\lambda^{(1)}_i = \frac{1}{\kappa}$ for $i = 1, \ldots, d/2$, and $\Sigma^{(2)} = I$, $d = 600$, $n_2 = 100$

Illustration



- $\lambda^{(1)}_{d+1-i} = 1/\lambda^{(1)}_i = \frac{1}{\kappa}$ for $i = 1, \ldots, d/2$, and $\Sigma^{(2)} = I$, $d = 600$, $n_2 = 100$
- $\kappa = 1$ (red) gives risk curve for no cov. shift
- The crossing point on left is $n_1 = d/2$. Below, cov. shift helps

Risk monotonicity in eigenvalue

Theorem (Song, Bhattacharya, S. '24+)

- $\boldsymbol{\Sigma}^{(1)}$ *has two eigenvalues (previous plot setting):* $\lambda^{(1)}_{d+1-i} = 1/\lambda^{(1)}_{i} = \frac{1}{\kappa}$ *for* $i = 1, \ldots, d/2$
- *Let* $\boldsymbol{M}(\kappa)$ *denote such a diagonal matrix*
- $\boldsymbol{\Sigma}^{(2)} = I$

Risk monotonicity in eigenvalue

Theorem (Song, Bhattacharya, S. '24+)

- $\mathbf{\Sigma}^{(1)}$ *has two eigenvalues (previous plot setting):* $\lambda_{d+1-i}^{(1)} = 1/\lambda_i^{(1)} = \frac{1}{\kappa}$ *for* $i = 1, ..., d/2$
- *Let* $\boldsymbol{M}(\kappa)$ *denote such a diagonal matrix*
- $\mathbf{\Sigma}^{(2)} = I$

(i) *When* $n_1 < \min\{d/2, d - n_2\}$, $\hat{R}(\boldsymbol{M}(\kappa_1)) \leq \hat{R}(\boldsymbol{M}(\kappa_2)) + o(1)$ *for any* $\kappa_1 > \kappa_2 > 1$

(ii) *When* $d/2 < n_1 < d - n_2$, $\hat{R}(\boldsymbol{M}(\kappa_1)) \geq \hat{R}(\boldsymbol{M}(\kappa_2)) + o(1)$ *for any* $\kappa_1 > \kappa_2 > 1$

(iii) *If* $n_1 = \min\{d/2, d - n_2\}$, *then* $\hat{R}(\boldsymbol{M}(\kappa))$ *does not depend on* $\kappa \geq 1$

Risk monotonicity in eigenvalue

Theorem (Song, Bhattacharya, S. '24+)

- $\mathbf{\Sigma}^{(1)}$ *has two eigenvalues (previous plot setting):* $\lambda^{(1)}_{d+1-i} = 1/\lambda^{(1)}_i = \frac{1}{\kappa}$ *for* $i = 1, \ldots, d/2$
- *Let* $\boldsymbol{M}(\kappa)$ *denote such a diagonal matrix*
- $\mathbf{\Sigma}^{(2)} = I$

(i) *When* $n_1 < \min\{d/2, d - n_2\}$, $\hat{R}(\boldsymbol{M}(\kappa_1)) \leq \hat{R}(\boldsymbol{M}(\kappa_2)) + o(1)$ *for any* $\kappa_1 > \kappa_2 > 1$

(ii) *When* $d/2 < n_1 < d - n_2$, $\hat{R}(\boldsymbol{M}(\kappa_1)) \geq \hat{R}(\boldsymbol{M}(\kappa_2)) + o(1)$ *for any* $\kappa_1 > \kappa_2 > 1$

(iii) *If* $n_1 = \min\{d/2, d - n_2\}$, *then* $\hat{R}(\boldsymbol{M}(\kappa))$ *does not depend on* $\kappa \geq 1$

Risk monotonicity in eigenvalue

Theorem (Song, Bhattacharya, S. '24+)

- $\boldsymbol{\Sigma}^{(1)}$ *has two eigenvalues (previous plot setting):* $\lambda^{(1)}_{d+1-i} = 1/\lambda^{(1)}_i = \frac{1}{\kappa}$ *for* $i = 1, ..., d/2$
- *Let* $\boldsymbol{M}(\kappa)$ *denote such a diagonal matrix*
- $\boldsymbol{\Sigma}^{(2)} = I$

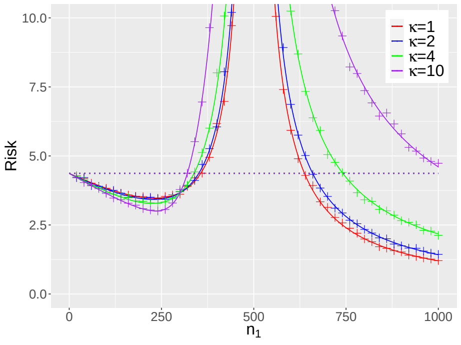
(i) *When* $n_1 < \min\{d/2, d - n_2\}$, $\hat{R}(\boldsymbol{M}(\kappa_1)) \leq \hat{R}(\boldsymbol{M}(\kappa_2)) + o(1)$ *for any* $\kappa_1 > \kappa_2 > 1$

(ii) *When* $d/2 < n_1 < d - n_2$, $\hat{R}(\boldsymbol{M}(\kappa_1)) \geq \hat{R}(\boldsymbol{M}(\kappa_2)) + o(1)$ *for any* $\kappa_1 > \kappa_2 > 1$

(iii) *If* $n_1 = \min\{d/2, d - n_2\}$, *then* $\hat{R}(\boldsymbol{M}(\kappa))$ *does not depend on* $\kappa \geq 1$

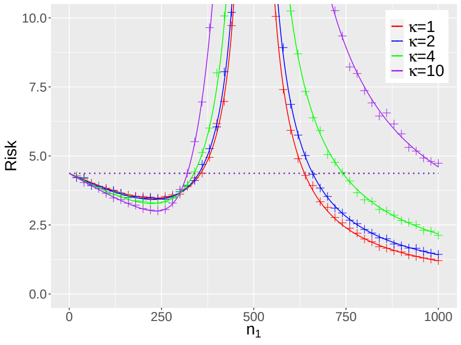
– *Takeaway: Under sufficient overparametrization, the more the covariate shift, the less the risk and vice versa*

Illustration



- $\lambda^{(1)}_{d+1-i} = 1/\lambda^{(1)}_i = \frac{1}{\kappa}$ for $i = 1, \ldots, d/2$, and $\mathbf{\Sigma}^{(2)} = I$
- $\kappa = 1$ (red) gives risk curve for no covariate shift
- The crossing point on left is $n_1 = d/2, d = 600, n_2 = 100$; all curves cross red curve

Illustration



- $\lambda^{(1)}_{d+1-i} = 1/\lambda^{(1)}_i = \frac{1}{\kappa}$ for $i = 1, \ldots, d/2$, and $\mathbf{\Sigma}^{(2)} = I$
- $\kappa = 1$ (red) gives risk curve for no covariate shift
- The crossing point on left is $n_1 = d/2, d = 600, n_2 = 100$; all curves cross red curve monotonicity pattern between κ's changes

Conclusion: Key Takeaways

- Heterogeneity: opportunity *and* risk.
- Distribution shift in the interpolating regime can be rigorously analyzed.
- We provide the first precise, analytic formulas for the generalization error of pooled min-norm interpolator under concept and covariate shift.

Conclusion: Key Takeaways

- Heterogeneity: opportunity *and* risk.
- Distribution shift in the interpolating regime can be rigorously analyzed.
- We provide the first precise, analytic formulas for the generalization error of pooled min-norm interpolator under concept and covariate shift.
- Our results reveal sharp phase transitions thresholds for positive vs. negative transfer, quantifying when to share, when to "keep separate."

Conclusion: Key Takeaways

- Heterogeneity: opportunity *and* risk.
- Distribution shift in the interpolating regime can be rigorously analyzed.
- We provide the first precise, analytic formulas for the generalization error of pooled min-norm interpolator under concept and covariate shift.
- Our results reveal sharp phase transitions thresholds for positive vs. negative transfer, quantifying when to share, when to “keep separate.”
- The analysis required developing novel tools in Random Matrix Theory, specifically a new anisotropic local law.

Future Directions

Deeper dive into covariate shift: Understand covariate shift phenomena for broader class of distributions?

Other Estimators:

(i) *Late fusion*: Average estimators trained on separate datasets

(ii) Compare with other interpolators or explicit regularized strategies

Beyond Linear models: Do these insights persist for classification problems, or more complex models, e.g., multi-index or random features regression?

Recall Key Takeaways

- Heterogeneity: opportunity *and* risk.
- Distribution shift in the interpolating regime can be rigorously analyzed.
- We provided the first precise, analytic formulas for the generalization error of pooled min-norm interpolator under concept and covariate shift.
- Our results reveal sharp phase transitions thresholds for positive vs. negative transfer, quantifying when to share, when to "keep separate."
- The analysis required developing novel tools in Random Matrix Theory, specifically a new anisotropic local law.

Technical detour: Anisotropic local laws and their utility in data integration theory

Basic Setup

- Let $\mathbf{Z} \in \mathbb{R}^{n \times d}$ be a matrix with i.i.d. entries satisfying $\mathbb{E}[Z_{ij}] = 0$, $\mathrm{Var}(Z_{ij}) = 1$, and necessary moment conditions
- For some $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$, define $\mathbf{X} = \mathbf{Z}\Sigma^{1/2}$

Basic Setup

- Let $\mathbf{Z} \in \mathbb{R}^{n\times d}$ be a matrix with i.i.d. entries satisfying $\mathbb{E}[Z_{ij}] = 0$, $\mathrm{Var}(Z_{ij}) = 1$, and necessary moment conditions
- For some $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, define $\mathbf{X} = \mathbf{Z}\Sigma^{1/2}$
- Suppose that $d/n \to \gamma \in (0, \infty)$
- Consider the scaled sample covariance matrix $\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}^\top \mathbf{X}$

Basic Setup

- Let $\mathbf{Z} \in \mathbb{R}^{n \times d}$ be a matrix with i.i.d. entries satisfying $\mathbb{E}[Z_{ij}] = 0$, $\mathrm{Var}(Z_{ij}) = 1$, and necessary moment conditions
- For some $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$, define $\mathbf{X} = \mathbf{Z}\Sigma^{1/2}$
- Suppose that $d/n \to \gamma \in (0, \infty)$
- Consider the scaled sample covariance matrix $\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}^\top \mathbf{X}$

Wish to understand the *behavior* of the empirical spectral distribution (ESD) of $\hat{\mathbf{\Sigma}}$:

$$\mu_{\hat{\mathbf{\Sigma}}} = \frac{1}{d} \sum_{i \leq d} \delta_{\lambda_i(\hat{\mathbf{\Sigma}})}$$

The resolvent–why study it?

Useful to study the ESD of the resolvent

$$R(z) = (\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}$$

The resolvent–why study it?

Useful to study the ESD of the resolvent

$$R(z) = (\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}$$

and its normalized trace:

$$m_{\hat{\mathbf{\Sigma}}}(z) = \frac{1}{d}\mathrm{Tr}[R(z)] = \frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}] = \frac{1}{d}\sum_{i \le d} \frac{1}{\lambda_i(\hat{\mathbf{\Sigma}}) - z} = \int \frac{\mathrm{d}\mu_{\hat{\mathbf{\Sigma}}}(\lambda)}{\lambda - z}$$

which is precisely the Stieltjes transform of $\mu_{\hat{\mathbf{\Sigma}}}$.

The resolvent–why study it?

Useful to study the ESD of the resolvent

$$R(z) = (\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}$$

and its normalized trace:

$$m_{\hat{\mathbf{\Sigma}}}(z) = \frac{1}{d}\mathrm{Tr}[R(z)] = \frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}] = \frac{1}{d}\sum_{i \leq d} \frac{1}{\lambda_i(\hat{\mathbf{\Sigma}}) - z} = \int \frac{\mathrm{d}\mu_{\hat{\mathbf{\Sigma}}}(\lambda)}{\lambda - z}$$

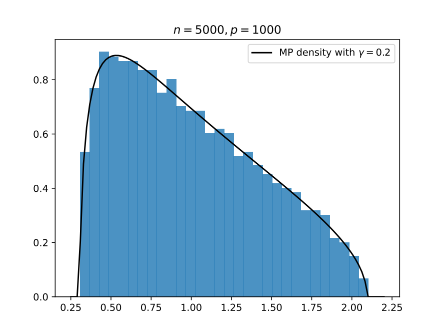
which is precisely the Stieltjes transform of $\mu_{\hat{\mathbf{\Sigma}}}$.

The resolvent is fundamental for statistical inference questions, e.g.,

$$\text{Ridge regression yields } \hat{\beta}_{\text{ridge}} = (\mathbf{X}^\top\mathbf{X} + n\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} = \frac{1}{n}R(-\lambda)\mathbf{X}^\top\mathbf{y}$$

A global law

Assume that $\mathbf{\Sigma} = \mathbf{I}$. Recall that $\mu_{\hat{\mathbf{\Sigma}}}$ converges weakly to the Marchenko-Pastur law μ_γ.



In particular, the Stieltjes transform of $\mu_{\hat{\mathbf{\Sigma}}}$ converges to the Stieltjes transform of μ_γ:

$$\underbrace{\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}]}_{\text{Stieltjes transform of ESD}} = \frac{1}{d}\sum_{i \leq d} \frac{1}{\lambda_i(\hat{\mathbf{\Sigma}}) - z} \to \int \frac{\mathrm{d}\mu_\gamma(t)}{t - z} = m_\gamma(z)$$

A more general global law

Note $\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}] = \langle \mathbf{v}, (\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}\mathbf{v}\rangle$ for $\mathbf{v} = \mathbf{1}/\sqrt{d}$. What about other $\mathbf{v} \in \mathbb{C}^d$?

A more general global law

Note $\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}] = \langle \mathbf{v}, (\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}\mathbf{v}\rangle$ for $\mathbf{v} = \mathbf{1}/\sqrt{d}$. What about other $\mathbf{v} \in \mathbb{C}^d$?

Theorem (Theorem 10.16, Bai and Silverstein)
For any $\mathbf{v} \in \mathbb{C}^d$ with $\|\mathbf{v}\|_2 = 1$, we have $\langle \mathbf{v}, (\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}\mathbf{v}\rangle \to m_\gamma(z)$.

By a polarization identity, we immediately have the corollary:

$$\begin{aligned}\langle \mathbf{v}, (\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}\mathbf{w}\rangle = \frac{1}{4}\big(&\langle \mathbf{v}+\mathbf{w}, (\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}(\mathbf{v}+\mathbf{w})\rangle \\ &- \langle \mathbf{v}-\mathbf{w}, (\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}(\mathbf{v}-\mathbf{w})\rangle \\ &- i\langle \mathbf{v}+i\mathbf{w}, (\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}(\mathbf{v}+i\mathbf{w})\rangle \\ &+ i\langle \mathbf{v}-i\mathbf{w}, (\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}(\mathbf{v}-i\mathbf{w})\rangle\big) \\ \to m_\gamma(z)\langle \mathbf{v}, \mathbf{w}\rangle&\end{aligned}$$

for $\mathbf{v}, \mathbf{w} \in \mathbb{C}^d$ with $\|\mathbf{v}\|_2 = \|\mathbf{w}\|_2 = 1$.

Application: high-dimensional ridge(less) regression
Hastie et al. (2020)

- Suppose $\mathbf{y} = \mathbf{X}\beta + \epsilon$ for fixed β and i.i.d. noise ϵ.
- Consider the ridge estimator

$$\hat{\beta}_\lambda = \operatorname*{argmin}_{\mathbf{b} \in \mathbb{R}^p} \{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + n\lambda\|\mathbf{b}\|_2^2\} = \frac{1}{n}(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1}\mathbf{X}^\top \mathbf{y}$$

Application: high-dimensional ridge(less) regression Hastie et al. (2020)

- Suppose $\mathbf{y} = \mathbf{X}\beta + \epsilon$ for fixed β and i.i.d. noise ϵ.
- Consider the ridge estimator

$$\hat{\beta}_\lambda = \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^p}\{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + n\lambda\|\mathbf{b}\|_2^2\} = \frac{1}{n}(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

- The bias and variance expressions (conditional on $\mathbf{X}$) are

$$B_{\mathbf{X}}(\hat{\beta}_\lambda, \beta) := \|\mathbb{E}[\hat{\beta}_\lambda - \beta \mid \mathbf{X}]\|_{\mathbf{\Sigma}}^2 = \lambda^2\beta^\top(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1}\mathbf{\Sigma}(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1}\beta$$

$$V_{\mathbf{X}}(\hat{\beta}_\lambda, \beta) := \frac{\mathrm{Var}(\epsilon_1)}{n}\mathrm{Tr}[\mathbf{\Sigma}\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-2}]$$

which we can study using variants of these global laws

Application: high-dimensional ridge(less) regression
Hastie et al. (2020)

- Suppose $\mathbf{y} = \mathbf{X}\beta + \epsilon$ for fixed β and i.i.d. noise ϵ.
- Consider the ridge estimator

$$\hat{\beta}_\lambda = \underset{\mathbf{b}\in\mathbb{R}^p}{\operatorname{argmin}}\{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + n\lambda\|\mathbf{b}\|_2^2\} = \frac{1}{n}(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

- The bias and variance expressions (conditional on $\mathbf{X}$) are

$$B_{\mathbf{X}}(\hat{\beta}_\lambda, \beta) := \|\mathbb{E}[\hat{\beta}_\lambda - \beta \mid \mathbf{X}]\|_{\mathbf{\Sigma}}^2 = \lambda^2\beta^\top(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1}\mathbf{\Sigma}(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1}\beta$$

$$V_{\mathbf{X}}(\hat{\beta}_\lambda, \beta) := \frac{\mathrm{Var}(\epsilon_1)}{n}\mathrm{Tr}[\mathbf{\Sigma}\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-2}]$$

which we can study using variants of these global laws

Applying the global law

Continue to assume $\mathbf{\Sigma} = \mathbf{I}$ (the anisotropic global laws are similar).

Then, to analyze $V_{\mathbf{X}}(\hat{\beta}_\lambda, \beta)$, it suffices to understand

$$\begin{aligned}\lim_{d\to\infty} \frac{1}{d}\mathrm{Tr}[\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-2}] &= \lim_{d\to\infty}\left(\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1}] - \frac{\lambda}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-2}]\right) \\ &= \lim_{d\to\infty}\left(\underbrace{\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1}]}_{\to m_\gamma(-\lambda)} - \lambda \cdot \underbrace{\frac{\partial}{\partial \lambda}\left[\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1}]\right]}_{\text{Claim: } \to \frac{\partial}{\partial\lambda} m_\gamma(-\lambda)}\right)\end{aligned}$$

Applying the global law

Continue to assume $\mathbf{\Sigma} = \mathbf{I}$ (the anisotropic global laws are similar).

Then, to analyze $V_{\mathbf{X}}(\hat{\beta}_\lambda, \beta)$, it suffices to understand

$$\begin{aligned}\lim_{d\to\infty} \frac{1}{d}\mathrm{Tr}[\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-2}] &= \lim_{d\to\infty}\left(\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1}] - \frac{\lambda}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-2}]\right) \\ &= \lim_{d\to\infty}\left(\underbrace{\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1}]}_{\to m_\gamma(-\lambda)} - \lambda \cdot \underbrace{\frac{\partial}{\partial\lambda}\left[\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1}]\right]}_{\text{Claim: } \to \frac{\partial}{\partial\lambda} m_\gamma(-\lambda)}\right)\end{aligned}$$

$\lambda \mapsto (\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1}$ is analytic and **uniformly bounded** for λ bounded away from 0

Uniform convergence in a compact set around λ, which allows us to exchange $\lim_{d\to\infty}$ and $\frac{\partial}{\partial\lambda}$.

Ridgeless regression

- In the overparameterized regime ($d > n$), for the min-norm interpolator

$$\begin{aligned}\hat{\beta} &= \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^p}\{\|\mathbf{b}\|_2 : \mathbf{X}\mathbf{b} = \mathbf{y}\} \\ &= (\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{X}^\top\mathbf{y} \\ &= \lim_{\lambda\to 0^+}(\mathbf{X}^\top\mathbf{X} + n\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}\end{aligned}$$

Ridgeless regression

- In the overparameterized regime ($d > n$), for the min-norm interpolator

$$\begin{aligned}\hat{\beta} &= \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^p}\{\|\mathbf{b}\|_2 : \mathbf{X}\mathbf{b} = \mathbf{y}\} \\ &= (\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{X}^\top\mathbf{y} \\ &= \lim_{\lambda\to 0^+}(\mathbf{X}^\top\mathbf{X} + n\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}\end{aligned}$$

- The bias and variance expressions (conditional on $\mathbf{X}$) are

$$B_{\mathbf{X}}(\hat{\beta};\beta) = \beta^\top(I - \hat{\mathbf{\Sigma}}^\dagger\hat{\mathbf{\Sigma}})\mathbf{\Sigma}(I - \hat{\mathbf{\Sigma}}^\dagger\hat{\mathbf{\Sigma}})\beta, \quad V_{\mathbf{X}}(\hat{\beta};\beta) = \frac{\mathrm{Var}(\epsilon_1)}{n}\mathrm{Tr}[\hat{\mathbf{\Sigma}}^\dagger\mathbf{\Sigma}]$$

where $\hat{\mathbf{\Sigma}}^\dagger = \lim_{\lambda\to 0^+}(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1}$

Ridgeless regression

- In the overparameterized regime ($d > n$), for the min-norm interpolator

$$\begin{aligned}\hat{\beta} &= \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^p}\{\|\mathbf{b}\|_2 : \mathbf{X}\mathbf{b} = \mathbf{y}\} \\ &= (\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{X}^\top\mathbf{y} \\ &= \lim_{\lambda\to 0^+}(\mathbf{X}^\top\mathbf{X} + n\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}\end{aligned}$$

- The bias and variance expressions (conditional on $\mathbf{X}$) are

$$B_{\mathbf{X}}(\hat{\beta};\beta) = \beta^\top(I - \hat{\boldsymbol{\Sigma}}^\dagger\hat{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}(I - \hat{\boldsymbol{\Sigma}}^\dagger\hat{\boldsymbol{\Sigma}})\beta, \quad V_{\mathbf{X}}(\hat{\beta};\beta) = \frac{\mathrm{Var}(\epsilon_1)}{n}\mathrm{Tr}[\hat{\boldsymbol{\Sigma}}^\dagger\boldsymbol{\Sigma}]$$

where $\hat{\boldsymbol{\Sigma}}^\dagger = \lim_{\lambda\to 0^+}(\hat{\boldsymbol{\Sigma}} + \lambda\mathbf{I})^{-1}$

- Previous argument fails since we lose uniform boundedness; need new tools!

Local laws

Isotropic local laws

Global laws establish control of the spectrum of $\hat{\boldsymbol{\Sigma}}$ on an average sense.

- To probe eigenvalue behavior on a finer scale, we need a **local law** that allows for $|z| \to 0$ as $n \to \infty$.

Isotropic local laws

Global laws establish control of the spectrum of $\hat{\boldsymbol{\Sigma}}$ on an average sense.

- To probe eigenvalue behavior on a finer scale, we need a **local law** that allows for $|z| \to 0$ as $n \to \infty$.

Theorem (Bloemendal et al., 2014, Theorem 2.4, roughly)
For sufficiently small ϵ, if $z = E + i\eta$ satisfies $n^{-1+\epsilon} \leq \eta$ and $|z| \geq \epsilon$, then

$$|\langle \mathbf{v}, (\hat{\boldsymbol{\Sigma}} - z\mathbf{I})^{-1}\mathbf{w}\rangle - m_\gamma(z)\langle \mathbf{v}, \mathbf{w}\rangle| \prec \sqrt{\frac{\operatorname{Im} m_\gamma(z)}{n\eta}} + \frac{1}{n\eta}$$

for deterministic vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^d$ of fixed norm.

Isotropic local laws

Global laws establish control of the spectrum of $\hat{\mathbf{\Sigma}}$ on an average sense.

- To probe eigenvalue behavior on a finer scale, we need a **local law** that allows for $|z| \to 0$ as $n \to \infty$.

Theorem (Bloemendal et al., 2014, Theorem 2.4, roughly)
For sufficiently small ϵ, if $z = E + i\eta$ satisfies $n^{-1+\epsilon} \leq \eta$ and $|z| \geq \epsilon$, then

$$|\langle \mathbf{v}, (\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}\mathbf{w}\rangle - m_\gamma(z)\langle \mathbf{v}, \mathbf{w}\rangle| \prec \sqrt{\frac{\operatorname{Im} m_\gamma(z)}{n\eta}} + \frac{1}{n\eta}$$

for deterministic vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^d$ of fixed norm.

Morally, says $(\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1} \approx m_\gamma(z)\mathbf{I}$ in a much stronger sense than a global law.

Application: high-dimensional ridgeless regression, revisited

Recall that for fixed λ, the risk calculation required

$$\lim_{d\to\infty}\frac{1}{d}\mathrm{Tr}[\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-2}] = \lim_{d\to\infty}\left(\underbrace{\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-1}]}_{\to m_\gamma(-\lambda)} - \lambda\cdot\underbrace{\frac{\partial}{\partial\lambda}\left[\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-1}]\right]}_{\text{Claim: }\to\frac{\partial}{\partial\lambda}m_\gamma(-\lambda)}\right)$$

Application: high-dimensional ridgeless regression, revisited

Recall that for fixed λ, the risk calculation required

$$\lim_{d\to\infty} \frac{1}{d}\mathrm{Tr}[\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-2}] = \lim_{d\to\infty}\left(\underbrace{\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-1}]}_{\to m_\gamma(-\lambda)} - \lambda\cdot\underbrace{\frac{\partial}{\partial\lambda}\left[\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-1}]\right]}_{\text{Claim: }\to\frac{\partial}{\partial\lambda}m_\gamma(-\lambda)}\right)$$

By simplifying the bound in the anisotropic local law, one can show that

$$\left|\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-1}] - m_\gamma(-\lambda)\right| \lesssim \frac{1}{\mathrm{Re}(\lambda)\cdot n^{(1-\epsilon)/2}}$$

Application: high-dimensional ridgeless regression, revisited

Recall that for fixed λ, the risk calculation required

$$\lim_{d\to\infty} \frac{1}{d}\mathrm{Tr}[\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-2}] = \lim_{d\to\infty}\left(\underbrace{\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-1}]}_{\to m_\gamma(-\lambda)} - \lambda\cdot\underbrace{\frac{\partial}{\partial\lambda}\left[\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-1}]\right]}_{\text{Claim: }\to\frac{\partial}{\partial\lambda}m_\gamma(-\lambda)}\right)$$

By simplifying the bound in the anisotropic local law, one can show that

$$\left|\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-1}] - m_\gamma(-\lambda)\right| \lesssim \frac{1}{\mathrm{Re}(\lambda)\cdot n^{(1-\epsilon)/2}}$$

and further

$$\left|\frac{\partial}{\partial\lambda}\left[\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-1}] - m_\gamma(-\lambda)\right]\right| \lesssim \frac{1}{\mathrm{Re}(\lambda)^2\cdot n^{(1-\epsilon)/2}}$$

Similar bounds allow to compute the interpolator risk (note these assumed $\mathbf{\Sigma}=\mathbf{I}$).

Application: high-dimensional ridgeless regression, revisited

Recall that for fixed λ, the risk calculation required

$$\lim_{d\to\infty} \frac{1}{d}\mathrm{Tr}[\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-2}] = \lim_{d\to\infty}\left(\underbrace{\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-1}]}_{\to m_\gamma(-\lambda)} - \lambda\cdot\underbrace{\frac{\partial}{\partial\lambda}\left[\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-1}]\right]}_{\text{Claim: }\to\frac{\partial}{\partial\lambda}m_\gamma(-\lambda)}\right)$$

By simplifying the bound in the anisotropic local law, one can show that

$$\left|\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-1}] - m_\gamma(-\lambda)\right| \lesssim \frac{1}{\mathrm{Re}(\lambda)\cdot n^{(1-\epsilon)/2}}$$

and further

$$\left|\frac{\partial}{\partial\lambda}\left[\frac{1}{d}\mathrm{Tr}[(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-1}] - m_\gamma(-\lambda)\right]\right| \lesssim \frac{1}{\mathrm{Re}(\lambda)^2\cdot n^{(1-\epsilon)/2}}$$

Similar bounds allow to compute the interpolator risk (note these assumed $\mathbf{\Sigma}=\mathbf{I}$).

Anisotropic setting

Let $\mathbf{\Sigma}$ be any covariance matrix (i.e., $\mathbf{\Sigma} \neq \mathbf{I}$).

No reason to expect the resolvent $(\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}$ to behave as a multiple of $\mathbf{I}$.

Anisotropic setting

Let $\mathbf{\Sigma}$ be any covariance matrix (i.e., $\mathbf{\Sigma} \neq \mathbf{I}$).

No reason to expect the resolvent $(\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}$ to behave as a multiple of $\mathbf{I}$. If the eigenvalues of $\mathbf{\Sigma}$ converge to a distribution H, then the LSD of $\hat{\mathbf{\Sigma}}$ has a Stieltjes transform $m_{\gamma,\mathbf{\Sigma}}$ satisfying

$$m_{\gamma,\mathbf{\Sigma}}(z) = \int \frac{\mathrm{d}H(\lambda)}{z - \gamma\lambda m_{\gamma,\mathbf{\Sigma}}(z)}$$

Anisotropic setting

Let $\mathbf{\Sigma}$ be any covariance matrix (i.e., $\mathbf{\Sigma} \neq \mathbf{I}$).

No reason to expect the resolvent $(\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}$ to behave as a multiple of $\mathbf{I}$. If the eigenvalues of $\mathbf{\Sigma}$ converge to a distribution H, then the LSD of $\hat{\mathbf{\Sigma}}$ has a Stieltjes transform $m_{\gamma,\mathbf{\Sigma}}$ satisfying

$$m_{\gamma,\mathbf{\Sigma}}(z) = \int \frac{\mathrm{d}H(\lambda)}{z - \gamma\lambda m_{\gamma,\mathbf{\Sigma}}(z)}$$

Can we prove a similar *anisotropic* local law?

An anisotropic local law

Theorem (Knowles and Yin, 2016, Theorem 3.21)
For (small) ϵ, if $z = E + i\eta \in \mathbb{C}_+$ satisfies $n^{-1+\epsilon} \leq \eta$ and $|z| \geq \epsilon$, then

$$|\langle \mathbf{v}, (\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}\mathbf{w}\rangle - \langle \mathbf{v}, -(z(\mathbf{I} + m_{\gamma,\mathbf{\Sigma}}(z)\mathbf{\Sigma}))^{-1}\mathbf{w}\rangle| \prec \sqrt{\frac{\operatorname{Im} m_\gamma(z)}{n\eta}} + \frac{1}{n\eta}$$

for deterministic vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^d$ of fixed norm.

An anisotropic local law

Theorem (Knowles and Yin, 2016, Theorem 3.21)
For (small) ϵ, if $z = E + i\eta \in \mathbb{C}_+$ satisfies $n^{-1+\epsilon} \leq \eta$ and $|z| \geq \epsilon$, then

$$|\langle \mathbf{v}, (\hat{\boldsymbol{\Sigma}} - z\mathbf{I})^{-1}\mathbf{w}\rangle - \langle \mathbf{v}, -(z(\mathbf{I} + m_{\gamma,\boldsymbol{\Sigma}}(z)\boldsymbol{\Sigma}))^{-1}\mathbf{w}\rangle| \prec \sqrt{\frac{\operatorname{Im} m_\gamma(z)}{n\eta}} + \frac{1}{n\eta}$$

for deterministic vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^d$ of fixed norm.

Morally, the resolvent $(\hat{\boldsymbol{\Sigma}} - z\mathbf{I})^{-1}$ behaves as $-(z(\mathbf{I} + m_{\gamma,\boldsymbol{\Sigma}}(z)\boldsymbol{\Sigma}))^{-1}$.

An anisotropic local law

Theorem (Knowles and Yin, 2016, Theorem 3.21)
For (small) ϵ, if $z = E + i\eta \in \mathbb{C}_+$ satisfies $n^{-1+\epsilon} \leq \eta$ and $|z| \geq \epsilon$, then

$$|\langle \mathbf{v}, (\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}\mathbf{w}\rangle - \langle \mathbf{v}, -(z(\mathbf{I} + m_{\gamma,\mathbf{\Sigma}}(z)\mathbf{\Sigma}))^{-1}\mathbf{w}\rangle| \prec \sqrt{\frac{\operatorname{Im} m_\gamma(z)}{n\eta}} + \frac{1}{n\eta}$$

for deterministic vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^d$ of fixed norm.

Morally, the resolvent $(\hat{\mathbf{\Sigma}} - z\mathbf{I})^{-1}$ behaves as $-(z(\mathbf{I} + m_{\gamma,\mathbf{\Sigma}}(z)\mathbf{\Sigma}))^{-1}$.

This is already useful for stat/ML problems with a non-trivial covariance matrix,

e.g., risk of interpolators in high-dimensional regression in presence of non-trivial feature covariances.

Distribution shift problems

- In the covariate shift setting, our covariance matrix is now

  $$\hat{\mathbf{\Sigma}} = \mathbf{X}^{(1)\top}\mathbf{X}^{(1)} + \mathbf{X}^{(2)\top}\mathbf{X}^{(2)}$$

  with $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$ differing in distribution.
- Prior anisotropic local laws no longer apply.

Distribution shift problems

- In the covariate shift setting, our covariance matrix is now

  $$\hat{\mathbf{\Sigma}} = \mathbf{X}^{(1)\top}\mathbf{X}^{(1)} + \mathbf{X}^{(2)\top}\mathbf{X}^{(2)}$$

  with $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$ differing in distribution.
- Prior anisotropic local laws no longer apply.
- We establish a new **anisotropic local law** for the resolvent of sums of such sample covariance matrices.
- Allows us to characterize the risk of the interpolator by tracking λ-dependent quantities precisely through the proof.

An anisotropic local law for covariate shift

For technical reasons, we assume that $\mathbf{\Sigma}^{(1)}$ and $\mathbf{\Sigma}^{(2)}$ are *co-diagonalizable*: i.e., $\mathbf{\Sigma}^{(1)} = \mathbf{V}\mathbf{\Lambda}^{(1)}\mathbf{V}^\top$, $\mathbf{\Sigma}^{(2)} = \mathbf{V}\mathbf{\Lambda}^{(2)}\mathbf{V}^\top$ with $\mathbf{V}$ orthogonal and $\mathbf{\Lambda}^{(1)}, \mathbf{\Lambda}^{(2)}$ diagonal.

An anisotropic local law for covariate shift

For technical reasons, we assume that $\mathbf{\Sigma}^{(1)}$ and $\mathbf{\Sigma}^{(2)}$ are *co-diagonalizable*: i.e., $\mathbf{\Sigma}^{(1)} = \mathbf{V}\mathbf{\Lambda}^{(1)}\mathbf{V}^\top$, $\mathbf{\Sigma}^{(2)} = \mathbf{V}\mathbf{\Lambda}^{(2)}\mathbf{V}^\top$ with $\mathbf{V}$ orthogonal and $\mathbf{\Lambda}^{(1)}, \mathbf{\Lambda}^{(2)}$ diagonal.

Letting $\mathbf{F} = n^{-1/2}[(\mathbf{\Lambda}^{(1)})^{1/2}\mathbf{V}^\top\mathbf{Z}^{(1)\top}, (\mathbf{\Lambda}^{(2)})^{1/2}\mathbf{V}^\top\mathbf{Z}^{(2)\top}]$, with $\mathbf{X}^{(1)} = \mathbf{Z}^{(1)}\mathbf{\Sigma}^{(1)}, \mathbf{X}^{(2)} = \mathbf{Z}^{(2)}\mathbf{\Sigma}^{(2)}$, so $\mathbf{F} \in \mathbb{R}^{d+n\times d+n}$

An anisotropic local law for covariate shift

For technical reasons, we assume that $\mathbf{\Sigma}^{(1)}$ and $\mathbf{\Sigma}^{(2)}$ are *co-diagonalizable*: i.e., $\mathbf{\Sigma}^{(1)} = \mathbf{V}\mathbf{\Lambda}^{(1)}\mathbf{V}^\top$, $\mathbf{\Sigma}^{(2)} = \mathbf{V}\mathbf{\Lambda}^{(2)}\mathbf{V}^\top$ with $\mathbf{V}$ orthogonal and $\mathbf{\Lambda}^{(1)}, \mathbf{\Lambda}^{(2)}$ diagonal.

Letting $\mathbf{F} = n^{-1/2}[(\mathbf{\Lambda}^{(1)})^{1/2}\mathbf{V}^\top\mathbf{Z}^{(1)\top}, (\mathbf{\Lambda}^{(2)})^{1/2}\mathbf{V}^\top\mathbf{Z}^{(2)\top}]$, with $\mathbf{X}^{(1)} = \mathbf{Z}^{(1)}\mathbf{\Sigma}^{(1)}, \mathbf{X}^{(2)} = \mathbf{Z}^{(2)}\mathbf{\Sigma}^{(2)}$, so $\mathbf{F} \in \mathbb{R}^{d+n\times d+n}$

the following resolvent becomes important,

$$\mathbf{G}(z) = \left[\begin{pmatrix} \mathbf{0} & \mathbf{F} \\ \mathbf{F}^\top & \mathbf{0} \end{pmatrix} - \begin{pmatrix} z\mathbf{I}_d & 0 \\ 0 & \mathbf{I}_n \end{pmatrix}\right]^{-1} = \begin{pmatrix} (\mathbf{F}\mathbf{F}^\top - z\mathbf{I})^{-1} & (\mathbf{F}\mathbf{F}^\top - z\mathbf{I})^{-1}\mathbf{F} \\ \mathbf{F}^\top(\mathbf{F}\mathbf{F}^\top - z\mathbf{I})^{-1} & z(\mathbf{F}\mathbf{F}^\top - z\mathbf{I})^{-1} \end{pmatrix}$$

An anisotropic local law for covariate shift

For technical reasons, we assume that $\boldsymbol{\Sigma}^{(1)}$ and $\boldsymbol{\Sigma}^{(2)}$ are *co-diagonalizable*: i.e., $\boldsymbol{\Sigma}^{(1)} = \mathbf{V}\boldsymbol{\Lambda}^{(1)}\mathbf{V}^\top$, $\boldsymbol{\Sigma}^{(2)} = \mathbf{V}\boldsymbol{\Lambda}^{(2)}\mathbf{V}^\top$ with $\mathbf{V}$ orthogonal and $\boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}$ diagonal.

Letting $\mathbf{F} = n^{-1/2}[(\boldsymbol{\Lambda}^{(1)})^{1/2}\mathbf{V}^\top\mathbf{Z}^{(1)\top}, (\boldsymbol{\Lambda}^{(2)})^{1/2}\mathbf{V}^\top\mathbf{Z}^{(2)\top}]$, with $\mathbf{X}^{(1)} = \mathbf{Z}^{(1)}\boldsymbol{\Sigma}^{(1)}, \mathbf{X}^{(2)} = \mathbf{Z}^{(2)}\boldsymbol{\Sigma}^{(2)}$, so $\mathbf{F} \in \mathbb{R}^{d+n\times d+n}$

the following resolvent becomes important,

$$\mathbf{G}(z) = \left[\begin{pmatrix} \mathbf{0} & \mathbf{F} \\ \mathbf{F}^\top & \mathbf{0} \end{pmatrix} - \begin{pmatrix} z\mathbf{I}_d & 0 \\ 0 & \mathbf{I}_n \end{pmatrix}\right]^{-1} = \begin{pmatrix} (\mathbf{F}\mathbf{F}^\top - z\mathbf{I})^{-1} & (\mathbf{F}\mathbf{F}^\top - z\mathbf{I})^{-1}\mathbf{F} \\ \mathbf{F}^\top(\mathbf{F}\mathbf{F}^\top - z\mathbf{I})^{-1} & z(\mathbf{F}\mathbf{F}^\top - z\mathbf{I})^{-1} \end{pmatrix}$$

Wish to characterize the limit $\mathfrak{G}(z)$ of $\mathbf{G}(z)$ but with control of the rate as $z \to 0$

An anisotropic local law for covariate shift

The limit is

$$\mathfrak{G}(z) = \begin{pmatrix} [a_1(z)\mathbf{\Lambda}^{(1)} + a_2(z)\mathbf{\Lambda}^{(2)} - z\mathbf{I}_d]^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\frac{n}{n_1} a_1(z)\mathbf{I}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\frac{n}{n_2} a_2(z)\mathbf{I}_{n_2} \end{pmatrix}$$

where a_1, a_2 are the unique solutions to

$$a_2 - \frac{n_2}{n} + \gamma \int \frac{a_2\lambda^{(2)}}{a_1\lambda^{(1)} + a_2\lambda^{(2)} - z} d\hat{H}(\lambda^{(1)}, \lambda^{(2)}) = 0$$

$$a_1 - \frac{n_1}{n} + \gamma \int \frac{a_1\lambda^{(1)}}{a_1\lambda^{(1)} + a_2\lambda^{(2)} - z} d\hat{H}(\lambda^{(1)}, \lambda^{(2)}) = 0$$

Simultaneous diagonalizability allows for a tractable form here;

An anisotropic local law for covariate shift

The limit is

$$\mathfrak{G}(z) = \begin{pmatrix} [a_1(z)\mathbf{\Lambda}^{(1)} + a_2(z)\mathbf{\Lambda}^{(2)} - z\mathbf{I}_d]^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\frac{n}{n_1}a_1(z)\mathbf{I}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\frac{n}{n_2}a_2(z)\mathbf{I}_{n_2} \end{pmatrix}$$

where a_1, a_2 are the unique solutions to

$$a_2 - \frac{n_2}{n} + \gamma \int \frac{a_2\lambda^{(2)}}{a_1\lambda^{(1)} + a_2\lambda^{(2)} - z} d\hat{H}(\lambda^{(1)}, \lambda^{(2)}) = 0$$

$$a_1 - \frac{n_1}{n} + \gamma \int \frac{a_1\lambda^{(1)}}{a_1\lambda^{(1)} + a_2\lambda^{(2)} - z} d\hat{H}(\lambda^{(1)}, \lambda^{(2)}) = 0$$

Simultaneous diagonalizability allows for a tractable form here; $\hat{H}$ is empirical distribution of eigenvalues;

An anisotropic local law for covariate shift

The limit is

$$\mathfrak{G}(z) = \begin{pmatrix} [a_1(z)\mathbf{\Lambda}^{(1)} + a_2(z)\mathbf{\Lambda}^{(2)} - z\mathbf{I}_d]^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\frac{n}{n_1}a_1(z)\mathbf{I}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\frac{n}{n_2}a_2(z)\mathbf{I}_{n_2} \end{pmatrix}$$

where a_1, a_2 are the unique solutions to

$$a_2 - \frac{n_2}{n} + \gamma \int \frac{a_2\lambda^{(2)}}{a_1\lambda^{(1)} + a_2\lambda^{(2)} - z} d\hat{H}(\lambda^{(1)}, \lambda^{(2)}) = 0$$
$$a_1 - \frac{n_1}{n} + \gamma \int \frac{a_1\lambda^{(1)}}{a_1\lambda^{(1)} + a_2\lambda^{(2)} - z} d\hat{H}(\lambda^{(1)}, \lambda^{(2)}) = 0$$

Simultaneous diagonalizability allows for a tractable form here; $\hat{H}$ is empirical distribution of eigenvalues; second and third blocks allow to isolate coefficients a_1, a_2.

Precise form of anisotropic local law

Theorem (Song et al., 2024)
On a suitable domain $\mathbf{D}$ *for* $\lambda > d^{-1/7+\epsilon}$ *and deterministic unit vectors* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{d+n}$,

$$\sup_{z \in \mathbf{D}} |\mathbf{u}^\top [\mathbf{G}(z) - \mathfrak{G}(z)] \mathbf{v}| \prec d^{-1/2} \lambda^{-3}.$$

The desired anisotropic local law follows as a corollary by taking $\mathbf{u}, \mathbf{v}$ to only be nonzero in the first d entries.

Proof outline

Recall $\mathbf{X}^{(1)} = \mathbf{Z}^{(1)}\mathbf{\Sigma}^{(1)}, \mathbf{X}^{(2)} = \mathbf{Z}^{(2)}\mathbf{\Sigma}^{(2)}$.

(a) Establish an entrywise local law for diagonal $\mathbf{\Sigma}^{(1)}, \mathbf{\Sigma}^{(2)}$:

- $\sup_{z \in \mathbf{D}} \max_{1 \leq i,j \leq n+d} |\mathbf{G}_{ij}(z) - \mathfrak{G}_{ij}(z)| \prec d^{-1/2}\lambda^{-3}$

Proof outline

Recall $\mathbf{X}^{(1)} = \mathbf{Z}^{(1)}\mathbf{\Sigma}^{(1)}, \mathbf{X}^{(2)} = \mathbf{Z}^{(2)}\mathbf{\Sigma}^{(2)}$.

(a) Establish an entrywise local law for diagonal $\mathbf{\Sigma}^{(1)}, \mathbf{\Sigma}^{(2)}$:

- $\sup_{z\in\mathbf{D}} \max_{1\le i,j\le n+d} |\mathbf{G}_{ij}(z) - \mathfrak{G}_{ij}(z)| \prec d^{-1/2}\lambda^{-3}$

(b) Use this to establish an anisotropic local law for Gaussian $\mathbf{Z}$ and general $\mathbf{\Sigma}$ [Use rotational invariance to reduce dominant terms to diagonal case]

Proof outline

Recall $\mathbf{X}^{(1)} = \mathbf{Z}^{(1)}\mathbf{\Sigma}^{(1)}, \mathbf{X}^{(2)} = \mathbf{Z}^{(2)}\mathbf{\Sigma}^{(2)}$.

(a) Establish an entrywise local law for diagonal $\mathbf{\Sigma}^{(1)}, \mathbf{\Sigma}^{(2)}$:

- $\sup_{z \in \mathbf{D}} \max_{1 \le i,j \le n+d} \left|\mathbf{G}_{ij}(z) - \mathfrak{G}_{ij}(z)\right| \prec d^{-1/2}\lambda^{-3}$

(b) Use this to establish an anisotropic local law for Gaussian $\mathbf{Z}$ and general $\mathbf{\Sigma}$ [Use rotational invariance to reduce dominant terms to diagonal case]

(c) Interpolate between Gaussian $\mathbf{Z}$ and general $\mathbf{Z}$ (both with general $\mathbf{\Sigma}$) [Usual Lindeberg argument]

Proof outline

Recall $\mathbf{X}^{(1)} = \mathbf{Z}^{(1)}\boldsymbol{\Sigma}^{(1)}, \mathbf{X}^{(2)} = \mathbf{Z}^{(2)}\boldsymbol{\Sigma}^{(2)}$.

(a) Establish an entrywise local law for diagonal $\boldsymbol{\Sigma}^{(1)}, \boldsymbol{\Sigma}^{(2)}$:

- $\sup_{z \in \mathbf{D}} \max_{1 \le i,j \le n+d} |\mathbf{G}_{ij}(z) - \mathfrak{G}_{ij}(z)| \prec d^{-1/2}\lambda^{-3}$

(b) Use this to establish an anisotropic local law for Gaussian $\mathbf{Z}$ and general $\boldsymbol{\Sigma}$ [Use rotational invariance to reduce dominant terms to diagonal case]

(c) Interpolate between Gaussian $\mathbf{Z}$ and general $\mathbf{Z}$ (both with general $\boldsymbol{\Sigma}$) [Usual Lindeberg argument]

This is proof structure for anisotropic local law with single covariance matrix as well.

Our main difficulty is tracking the dependence on λ throughout, but especially in (a).

Proof sketch

(a) Establish an entrywise local law for diagonals

- Prove that $a_1(z)$ and $a_2(z)$ exist and are unique (contraction argument)

Proof sketch

(a) Establish an entrywise local law for diagonals

- Prove that $a_1(z)$ and $a_2(z)$ exist and are unique (contraction argument)

- Prove a stability result: if $m_1(z), m_2(z)$ satisfy scaled versions of the equations up to additive errors, then $|m_1(z) - (-\frac{n}{n_1} a_1(z))| + |m_2(z) - (-\frac{n}{n_2} a_2(z))|$ can be controlled in terms of these errors

Proof sketch

(a) Establish an entrywise local law for diagonals

- Prove that $a_1(z)$ and $a_2(z)$ exist and are unique (contraction argument)

- Prove a stability result: if $m_1(z), m_2(z)$ satisfy scaled versions of the equations up to additive errors, then $|m_1(z) - (-\frac{n}{n_1} a_1(z))| + |m_2(z) - (-\frac{n}{n_2} a_2(z))|$ can be controlled in terms of these errors

- Show that $m_1(z) = \frac{1}{n_1} \sum_{i=d+1}^{d+n_1} G_{ii}(z)$ and $m_2(z) = \frac{1}{n_2} \sum_{i=d+n_1+1}^{d+n_1+n_2} G_{ii}(z)$ satisfy such versions of the equations upto additive errors; control error terms.

Impact of such anistropic local laws in other ML problems

High-dimensional Analysis of Knowledge Distillation: Weak-to-Strong Generalization and Scaling Laws

Ildiz et al. (2024)
(Corresponds to arXiv:2410.18837)

The Problem: Knowledge distillation is a powerful technique where a small "student" model is trained on synthetic labels generated by a large, powerful "teacher" model.

Goal: What is the generalization behavior of the student? Can such student perform comparably or outperform strong teacher?

Problem Setup and Main Findings

They study a teacher-student setup for linear models in the high-dimensional limit.

- A **teacher** model is trained on n_T real data points (X_T, y_T).
- It generates n_S synthetic (”surrogate”) labels for new data X_S.
- A **student** model is trained on this surrogate data $(X_S, y_S^{\text{teacher}})$.

Problem Setup and Main Findings

They study a teacher-student setup for linear models in the high-dimensional limit.

- A **teacher** model is trained on n_T real data points (X_T, y_T).
- It generates n_S synthetic ("surrogate") labels for new data X_S.
- A **student** model is trained on this surrogate data $(X_S, y_S^{\text{teacher}})$.

Main Findings:

- Precise asymptotic formulae for the student's final test error
- The error decomposes into terms related to the teacher's error, the student's approximation error, and the number of real (n_T) vs. surrogate (n_S) samples.
- Allows to quantify when student performs similar or better than teacher.

Technical Connection to Our Work

Knowledge distillation is a form of transfer learning where knowledge is transferred via synthetic labels.

Two kinds of samples: the ones for the teacher and the ones for the student

Technical Connection to Our Work

Knowledge distillation is a form of transfer learning where knowledge is transferred via synthetic labels.

Two kinds of samples: the ones for the teacher and the ones for the student **The Mathematical Core:** The total sample covariance matrix admits the decomposition:

$$\hat{\Sigma}_{\text{total}} = \underbrace{\hat{\Sigma}_{\text{teacher}}}_{\text{Source}} + \underbrace{\hat{\Sigma}_{\text{student}}}_{\text{Target}}$$

Technical Connection to Our Work

Knowledge distillation is a form of transfer learning where knowledge is transferred via synthetic labels.

Two kinds of samples: the ones for the teacher and the ones for the student **The Mathematical Core:** The total sample covariance matrix admits the decomposition:

$$\hat{\Sigma}_{\text{total}} = \underbrace{\hat{\Sigma}_{\text{teacher}}}_{\text{Source}} + \underbrace{\hat{\Sigma}_{\text{student}}}_{\text{Target}}$$

Inherently a sum of anisotropic matrices since there are typically distribution shifts in the data.

Similar anisotropic local laws useful.

Similar other problems

- Weak-to-strong generalization: Reverse of Knowledge distillation in the way we described; when teacher is weak, but wish to use it to train stronger student with less compute than would be required to train strong student from scratch.

Similar other problems

- Weak-to-strong generalization: Reverse of Knowledge distillation in the way we described; when teacher is weak, but wish to use it to train stronger student with less compute than would be required to train strong student from scratch.
- Boosting generalization performance by mixing surrogate data with real data in settings where data collection is difficult, e.g. by appending synthetic data (Ildiz et al.2024)

Similar other problems

- Weak-to-strong generalization: Reverse of Knowledge distillation in the way we described; when teacher is weak, but wish to use it to train stronger student with less compute than would be required to train strong student from scratch.
- Boosting generalization performance by mixing surrogate data with real data in settings where data collection is difficult, e.g. by appending synthetic data (Ildiz et al.2024)
- Time series forecasting where distribution shifts are common due to seasonal changes, market shocks, etc. (Taga et al. 2025)

Similar other problems

- Weak-to-strong generalization: Reverse of Knowledge distillation in the way we described; when teacher is weak, but wish to use it to train stronger student with less compute than would be required to train strong student from scratch.
- Boosting generalization performance by mixing surrogate data with real data in settings where data collection is difficult, e.g. by appending synthetic data (Ildiz et al.2024)
- Time series forecasting where distribution shifts are common due to seasonal changes, market shocks, etc. (Taga et al. 2025)
- Multi-objective optimization for economics problems (e.g. understanding markets: Jagadeesan et al. '24)

Similar other problems

- Weak-to-strong generalization: Reverse of Knowledge distillation in the way we described; when teacher is weak, but wish to use it to train stronger student with less compute than would be required to train strong student from scratch.
- Boosting generalization performance by mixing surrogate data with real data in settings where data collection is difficult, e.g. by appending synthetic data (Ildiz et al.2024)
- Time series forecasting where distribution shifts are common due to seasonal changes, market shocks, etc. (Taga et al. 2025)
- Multi-objective optimization for economics problems (e.g. understanding markets: Jagadeesan et al. '24)
- All these problems serve as test beds for such anisotropic local laws

Part III: Beyond Distribution Shift: Multimodal learning

This part will be shared later

References

Thank you!

Contact: pragya@fas.harvard.edu

Main References:

1. Li, Y. and Sur, P., 2025. Optimal and provable calibration in high-dimensional binary classification: Angular calibration and platt scaling. arXiv preprint arXiv:2502.15131. In review, Neurips
2. Song, Y., Bhattacharya, S. and Sur, P., 2024. Generalization error of min-norm interpolators in transfer learning. arXiv preprint arXiv:2406.13944. In revision for The Annals of Statistics

See also:

1. Liang, T. and Sur, P., 2022. A precise high-dimensional asymptotic theory for boosting and minimum-ℓ_1-norm interpolated classifiers. The Annals of Statistics, 50(3), pp.1669-1695.
2. Sur, P. and Candès, E.J., 2019. A modern maximum-likelihood theory for high-dimensional logistic regression. Proceedings of the National Academy of Sciences, 116(29), pp.14516-14525.