Data Integration: Challenges and Opportunities for Interpolation Learning under Distribution Shifts

Pragya Sur, Dept. of Statistics, Harvard University

Mathematical Statistics in the Information Age, Vienna, Sept 2025

[Joint work with Kenny Gu (Harvard → Stanford), Yanke Song (Harvard → Apple), Sohom Bhattacharya (UFL)]

Why Data Integration?

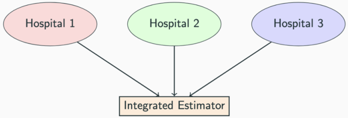
- Modern data sources are heterogeneous: multi-source, multi-sensor, multi-modal.
- Integrating diverse datasets enables improved, robust, generalizable models.
- Critical in machine learning as well as modern science; Examples :

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



What Makes Distribution Shift Challenging?

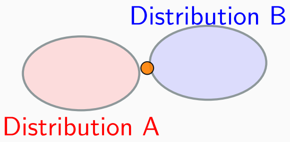
- Data come from sources with different noise, bias, and covariate structures.
- Distributional heterogeneity, e.g., covariate or label shift or mismatched features

What Makes Distribution Shift Challenging?

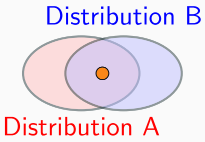
- Data come from sources with different noise, bias, and covariate structures.
- Distributional heterogeneity, e.g., covariate or label shift or mismatched features
- Theoretical understanding under-developed compared to "single distribution" statistics/machine learning.

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.



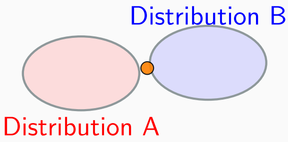
Scenario A: Integration useless



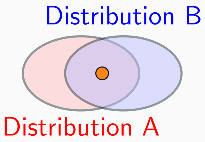
Scenario B: Integration useful

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 fundamental phase transitions in 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

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

- $\mathbf{X}^{(k)} \in \mathbb{R}^{n_k \times p}$: $\boldsymbol{X}^{(k)} = \boldsymbol{Z}^{(k)}(\boldsymbol{\Sigma}^{(k)})^{1/2}$, $\boldsymbol{Z}^{(k)}$ 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

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

- $\mathbf{X}^{(k)} \in \mathbb{R}^{n_k \times p}$: $\boldsymbol{X}^{(k)} = \boldsymbol{Z}^{(k)}(\boldsymbol{\Sigma}^{(k)})^{1/2}$, $\boldsymbol{Z}^{(k)}$ entries i.i.d. mean 0, variance 1; $\boldsymbol{\Sigma}^{(k)}$ bounded eigenvalues
- n_k: Number of samples (typically $n_1 \gg n_2$); $n_1 + n_2 =: n$.
- $\epsilon^{(k)}$ 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 ($p > n_1 + n_2$) through the lens of minimum norm (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
  $\rightarrow$ alg. properties at convergence

Utility?

- Often yields new insights on algorithms
- Common recipe: Study the implicit regularized limit
  $\rightarrow$ 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 $t \geq T(n, p, SNR)$

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

- *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
  $\rightarrow$ 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 $t \geq T(n, p, SNR)$

$$\textit{Misclassification Error}(\hat{\theta}^t) \approx \mathbb{P}\left(c_1^\star YZ_1 + c_2^\star Z_2 < 0\right), \quad a.s.$$

- *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
  $\rightarrow$ 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 $t \geq T(n, p, SNR)$

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

- *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; Significantly improves upon 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
  $\rightarrow$ 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 $t \geq T(n, p, SNR)$

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

- *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; Significantly improves upon 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, Bunea et al. '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, Bunea et al. '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:*

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

This represents both early and intermediate fusion

(will mention other estimators briefly later)

Goal

- Characterize its out-of-sample prediction error, i.e.,

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

  where $\boldsymbol{x}_0 \sim \mathbb{P}_{\boldsymbol{x}^{(2)}}$
- Guarantees will be w.h.p. over distribution of covariates

Main Results

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* $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$

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

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* $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$

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

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

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

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* $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$

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

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

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

- Involved trade-offs between target SNR, degree of shift, p, 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* $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$

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

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

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

- Involved trade-offs between target SNR, degree of shift, p, n_1, n_2
- Trade-off even between first two coefficients

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 $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$

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

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

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

- Involved trade-offs between target SNR, degree of shift, p, n_1, n_2
- Trade-off even between first two coefficients

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* $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$

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

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

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

- Involved trade-offs between target SNR, degree of shift, p, 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* $\mathrm{SNR} \leq \frac{\mathrm{p}^2}{(\mathrm{p}-\mathrm{n})(\mathrm{p}-\mathrm{n}_2)}$, *then*

$$R(\hat{\boldsymbol{\theta}}_{target}) \leq R(\hat{\boldsymbol{\theta}}_{pool}) + o(1) \tag{1}$$

2. *Else, define* $\rho := \frac{p-n}{p-n_1} - \frac{p^2}{(p-n_1)(p-n_2)} \cdot \frac{1}{\mathrm{SNR}}$. *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* $\mathrm{SNR} \leq \frac{\mathrm{p}^2}{(\mathrm{p}-\mathrm{n})(\mathrm{p}-\mathrm{n}_2)}$, *then*
$$R(\hat{\boldsymbol{\theta}}_{target}) \leq R(\hat{\boldsymbol{\theta}}_{pool}) + o(1) \tag{1}$$

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

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

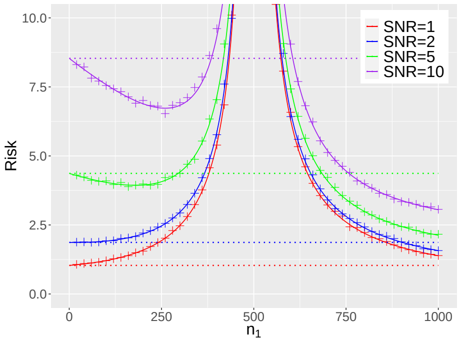
1. *If* $\mathrm{SNR} \leq \frac{\mathrm{p}^2}{(\mathrm{p}-\mathrm{n})(\mathrm{p}-\mathrm{n}_2)}$, *then*
$$R(\hat{\boldsymbol{\theta}}_{target}) \leq R(\hat{\boldsymbol{\theta}}_{pool}) + o(1) \tag{1}$$

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

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

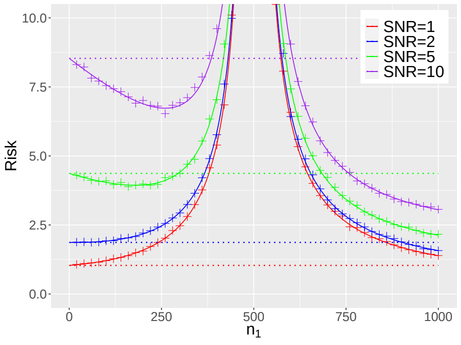
(ii) If SNR is large transfer gain depends on the degree of shift

Effects of SNR



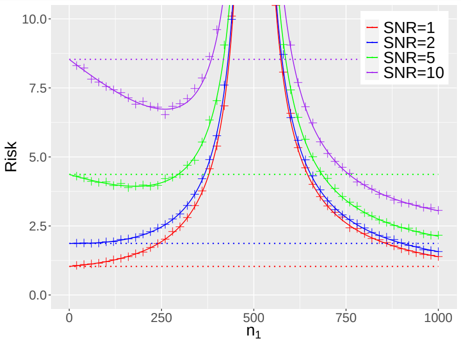
- SNR $= \|\boldsymbol{\theta}^{(2)}\|^2/\sigma^2$
- $n_2 = 100$, $p = 600$, Shift-to-signal ratio (SSR)$= \|\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}\|^2/\|\boldsymbol{\theta}^{(2)}\|^2 = 0.2$

Effects of SNR



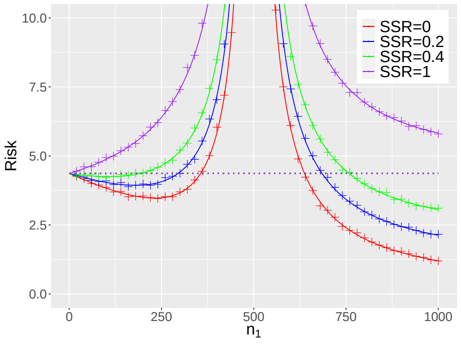
- SNR $= \|\boldsymbol{\theta}^{(2)}\|^2/\sigma^2$
- $n_2 = 100$, $p = 600$, Shift-to-signal ratio (SSR)$= \|\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}\|^2/\|\boldsymbol{\theta}^{(2)}\|^2 = 0.2$
- Takeaways: For low SNR, pooling does not help

Effects of SNR



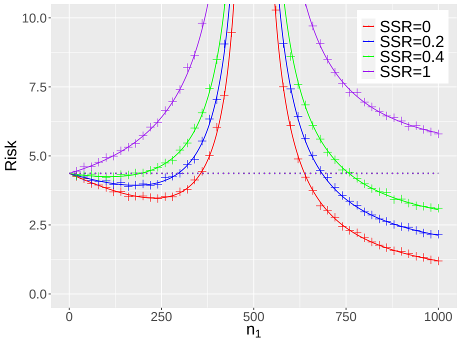
- SNR $= \|\boldsymbol{\theta}^{(2)}\|^2/\sigma^2$
- $n_2 = 100$, $p = 600$, Shift-to-signal ratio (SSR)$= \|\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}\|^2/\|\boldsymbol{\theta}^{(2)}\|^2 = 0.2$
- Takeaways: For low SNR, pooling does not help
- For higher SNR it does till n_1 below a threshold

Effects of SSR



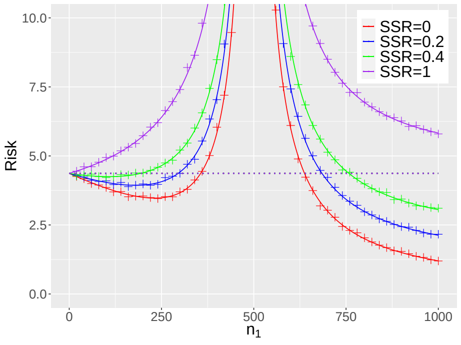
- $n_2 = 100, \; p = 600, \; \mathrm{SNR} = 5$

Effects of SSR



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

Effects of SSR



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

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

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):

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

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

Risk under Covariate Shift

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

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

- *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 $p/n = \gamma$; $n = n_1 + n_2$

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

- *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 $p/n = \gamma$; $n = n_1 + n_2$

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

- *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 (p even), $\lambda^{(1)}_{p+1-i} = 1/\lambda^{(1)}_i$ for $i = 1, ..., p/2$
- 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 (p even), $\lambda_{p+1-i}^{(1)} = 1/\lambda_i^{(1)}$ for $i = 1, \ldots, p/2$
- 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* $n_1 < \min\{p/2, p - n_2\}$, *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 (p even), $\lambda^{(1)}_{p+1-i} = 1/\lambda^{(1)}_i$ for $i = 1, ..., p/2$
- 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* $n_1 < \min\{p/2, p - n_2\}$, *then*

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

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

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

Risk monotonicity in eigenvalue

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

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

Risk monotonicity in eigenvalue

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

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

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

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

(iii) *If* $n_1 = \min\{p/2, p - n_2\}$, *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):* $\lambda^{(1)}_{p+1-i} = 1/\lambda^{(1)}_i = \frac{1}{\kappa}$ *for* $i = 1, \ldots, p/2$
- *Let* $\boldsymbol{M}(\kappa)$ *denote such a diagonal matrix*
- $\boldsymbol{\Sigma}^{(2)} = I$

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

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

   (iii) *If* $n_1 = \min\{p/2, p - n_2\}$, *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):* $\lambda^{(1)}_{p+1-i} = 1/\lambda^{(1)}_i = \frac{1}{\kappa}$ *for* $i = 1, ..., p/2$
- *Let* $\boldsymbol{M}(\kappa)$ *denote such a diagonal matrix*
- $\mathbf{\Sigma}^{(2)} = I$

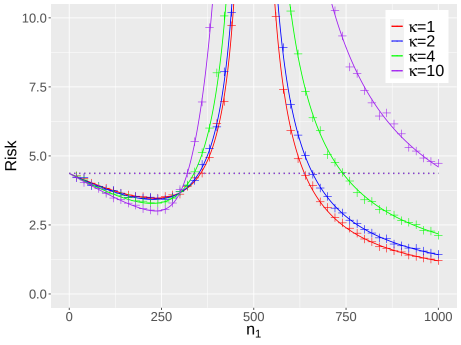
(i) *When* $n_1 < \min\{p/2, p - n_2\}$, $\hat{R}(\boldsymbol{M}(\kappa_1)) \leq \hat{R}(\boldsymbol{M}(\kappa_2)) + o(1)$ *for any* $\kappa_1 > \kappa_2 > 1$

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

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

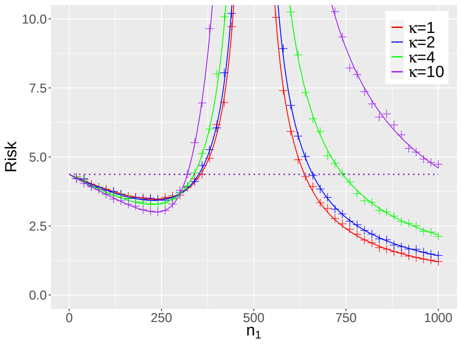
– Takeaway: Under sufficient overparametrization, the more the covariate shift, the less the risk and vice versa

Illustration



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

Illustration



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

Extensions

1. **Other Estimators:** Pre-training/fine-tuning type estimators (start with an initial pooled estimator–biased, fine-tune using the target data), regularization based estimators, late fusion estimators

Extensions

1. **Other Estimators:** Pre-training/fine-tuning type estimators (start with an initial pooled estimator–biased, fine-tune using the target data), regularization based estimators, late fusion estimators

2. **Non-linear models:** Random features regression (Gaussian equivalence trick: Hu and Lu ('20), Liang and S. ('22)) or nonparametric models with basis expansions (Equivalent Parametrization trick: Lahiry and S. ('24))

Extensions

1. **Other Estimators:** Pre-training/fine-tuning type estimators (start with an initial pooled estimator–biased, fine-tune using the target data), regularization based estimators, late fusion estimators

2. **Non-linear models:** Random features regression (Gaussian equivalence trick: Hu and Lu ('20), Liang and S. ('22)) or nonparametric models with basis expansions (Equivalent Parametrization trick: Lahiry and S. ('24))

3. **Beyond simultaneous diagonalizability:** Feasible but requires novel developments in random matrix theory (will discuss later)

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.

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."

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."
- Results form a starting point—many extensions possible.

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.”
- Results form a starting point—many extensions possible.
- We significantly advance Random Matrix Theory for this work.

Technical detour: Our Contribution to Random Matrix Theory

Review: Basic Setup

- Let $\mathbf{Z} \in \mathbb{R}^{n \times p}$ 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 $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, define $\mathbf{X} = \mathbf{Z}\Sigma^{1/2}$

Review: Basic Setup

- Let $\mathbf{Z} \in \mathbb{R}^{n \times p}$ 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 $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, define $\mathbf{X} = \mathbf{Z}\Sigma^{1/2}$
- Suppose that $p/n \to \gamma \in (0, \infty)$
- Consider the scaled sample covariance matrix $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{X}^\top \mathbf{X}$

Review: Basic Setup

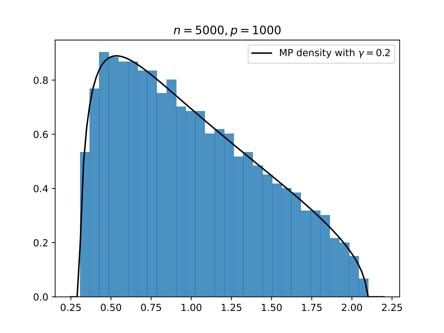
- Let $\mathbf{Z} \in \mathbb{R}^{n \times p}$ be a matrix with i.i.d. entries satisfying $\mathbb{E}[Z_{ij}] = 0$, $\mathrm{Var}(Z_{ij}) = 1$, and necessary moment conditions
- For some $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$, define $\mathbf{X} = \mathbf{Z}\Sigma^{1/2}$
- Suppose that $p/n \to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

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

A global law

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



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

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

But, true in quite some generality.

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

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

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

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

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

- 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

Applying the global law

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

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

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

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

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

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

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

Ridgeless regression

- In the overparameterized regime ($p > n$), 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 ($p > n$), 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

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

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

Ridgeless regression

- In the overparameterized regime ($p > n$), 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 $\hat{\mathbf{\Sigma}}^\dagger = \lim_{\lambda\to 0^+}(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1}$

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

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 quantitative result that allows for $|z| \to 0$ as $n \to \infty$.

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 quantitative result 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 $n^{-1+\epsilon} \le \eta$ and $|z| \ge \epsilon$, 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 $\mathbf{v}, \mathbf{w} \in \mathbb{C}^p$ of fixed norm.

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 quantitative result 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 $\mathbf{v}, \mathbf{w} \in \mathbb{C}^p$ 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.

Initiated in Erdős et al. (2009), this line of work called results of this form local laws

Application: high-dimensional ridgeless regression, revisited

Recall that for fixed λ, the risk calculation required

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

Application: high-dimensional ridgeless regression, revisited

Recall that for fixed λ, the risk calculation required

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

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

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

Application: high-dimensional ridgeless regression, revisited

Recall that for fixed λ, the risk calculation required

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

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

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

and further

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

These assumed $\mathbf{\Sigma}=\mathbf{I}$ but anisotropic versions exist (Knowles and Yin '16).

Distribution shift problems

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

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

with $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$ differing in distribution.

Distribution shift problems

- In the covariate shift setting, the 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. Free probability theory provides global law (Voiculescu ('91), Speicher ('93); Book length treatments: Nica and Speicher ('06), Bai and Silverstein ('10), Mingo and Speicher ('17), Erdős and Yau ('17), Potters and Bouchaud ('20))

Distribution shift problems

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

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

with $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$ differing in distribution. Free probability theory provides global law (Voiculescu ('91), Speicher ('93); Book length treatments: Nica and Speicher ('06), Bai and Silverstein ('10), Mingo and Speicher ('17), Erdős and Yau ('17), Potters and Bouchaud ('20))

- We establish a new **anisotropic local law** for the resolvent of such sums.

Distribution shift problems

- In the covariate shift setting, the 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. Free probability theory provides global law (Voiculescu ('91), Speicher ('93); Book length treatments: Nica and Speicher ('06), Bai and Silverstein ('10), Mingo and Speicher ('17), Erdős and Yau ('17), Potters and Bouchaud ('20))
- We establish a new **anisotropic local law** for the resolvent of such sums.
- Allows to characterize risk of the interpolator by tracking λ-dependent quantities.

Distribution shift problems

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

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

  with $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$ differing in distribution. Free probability theory provides global law (Voiculescu ('91), Speicher ('93); Book length treatments: Nica and Speicher ('06), Bai and Silverstein ('10), Mingo and Speicher ('17), Erdős and Yau ('17), Potters and Bouchaud ('20))
- We establish a new **anisotropic local law** for the resolvent of such sums.
- Allows to characterize risk of the interpolator by tracking λ-dependent quantities.
- Recent: A double application of our local law allows to relax assumptions such as simultaneous diagonalizability! (joint with Kenny Gu)

Widespread Utility Across Modern ML Problems

Examples

- Knowledge Distillation and Weak-to-strong Generalization: Teacher-student scenario, two kinds of model training, rich to small or small to rich. Similar sum of sample covariances of different distributions arise (initial work: Ildiz et al. 2024, interesting statistical questions outstanding: ongoing with Radu Lecoui and Debarghya Mukherjee)

Examples

- Knowledge Distillation and Weak-to-strong Generalization: Teacher-student scenario, two kinds of model training, rich to small or small to rich. Similar sum of sample covariances of different distributions arise (initial work: Ildiz et al. 2024, interesting statistical questions outstanding: ongoing with Radu Lecoui and Debarghya Mukherjee)
- Multi-objective optimization for economics problems e.g., to understand incumbent/entrant market dynamics for AI companies (Jagadeesan et al. '24)

Examples

- Knowledge Distillation and Weak-to-strong Generalization: Teacher-student scenario, two kinds of model training, rich to small or small to rich. Similar sum of sample covariances of different distributions arise (initial work: Ildiz et al. 2024, interesting statistical questions outstanding: ongoing with Radu Lecoui and Debarghya Mukherjee)
- Multi-objective optimization for economics problems e.g., to understand incumbent/entrant market dynamics for AI companies (Jagadeesan et al. '24)
- Boosting generalization performance by mixing real data with synthetic data generated from AI models (extensively studied experimentally or in low dimensions: Dohmatob et al. '24, Gerstgrasser et al. '24, Dey and Donoho '24, He et al. '25)

Examples

- Knowledge Distillation and Weak-to-strong Generalization: Teacher-student scenario, two kinds of model training, rich to small or small to rich. Similar sum of sample covariances of different distributions arise (initial work: Ildiz et al. 2024, interesting statistical questions outstanding: ongoing with Radu Lecoui and Debarghya Mukherjee)
- Multi-objective optimization for economics problems e.g., to understand incumbent/entrant market dynamics for AI companies (Jagadeesan et al. '24)
- Boosting generalization performance by mixing real data with synthetic data generated from AI models (extensively studied experimentally or in low dimensions: Dohmatob et al. '24, Gerstgrasser et al. '24, Dey and Donoho '24, He et al. '25)

All of these problems provide rich test beds for our RMT advances

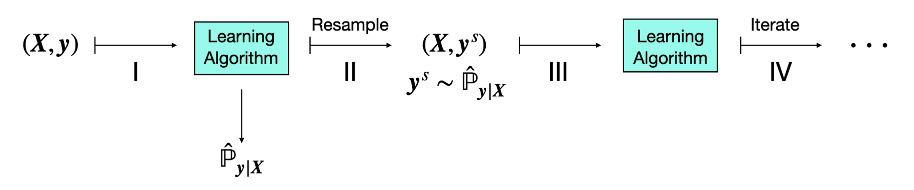
Examples

- Knowledge Distillation and Weak-to-strong Generalization: Teacher-student scenario, two kinds of model training, rich to small or small to rich. Similar sum of sample covariances of different distributions arise (initial work: Ildiz et al. 2024, interesting statistical questions outstanding: ongoing with Radu Lecoui and Debarghya Mukherjee)
- Multi-objective optimization for economics problems e.g., to understand incumbent/entrant market dynamics for AI companies (Jagadeesan et al. '24)
- Boosting generalization performance by mixing real data with synthetic data generated from AI models (extensively studied experimentally or in low dimensions: Dohmatob et al. '24, Gerstgrasser et al. '24, Dey and Donoho '24, He et al. '25)

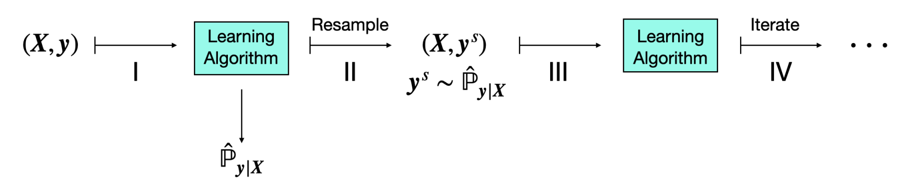
  Q. When and how can model collapse be prevented under overparametrization? (Ongoing with Anvit Garg and Sohom Bhattacharya)

All of these problems provide rich test beds for our RMT advances

Synthetic data and model collapse

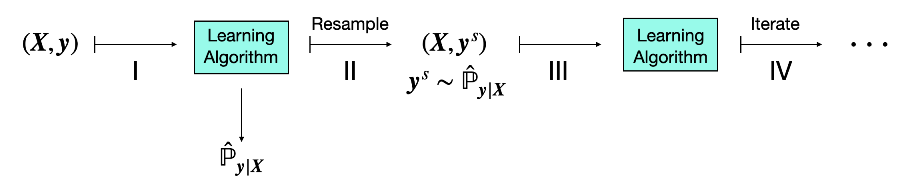


Synthetic data and model collapse



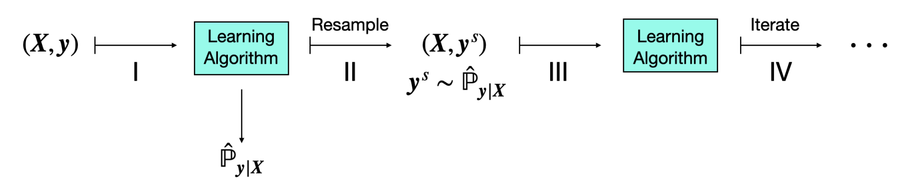
- Common learning paradigm in modern ML (Anaby-Tavor et al. '19, Huang et al. '22), improves performance often

Synthetic data and model collapse



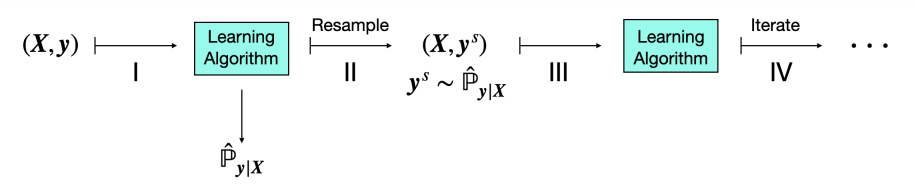
- Common learning paradigm in modern ML (Anaby-Tavor et al. '19, Huang et al. '22), improves performance often
- But, naive synthetic data-based retraining degrades performance (model collapse, Shumailov et al. '23, Hataya et al. '22)

Synthetic data and model collapse



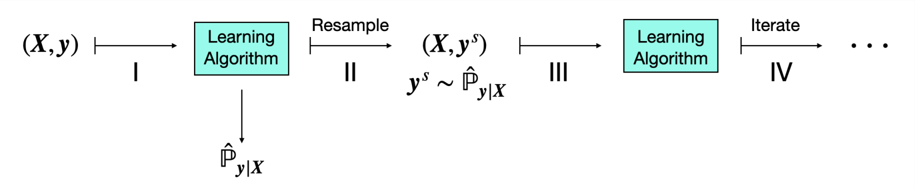
- Common learning paradigm in modern ML (Anaby-Tavor et al. '19, Huang et al. '22), improves performance often
- But, naive synthetic data-based retraining degrades performance (model collapse, Shumailov et al. '23, Hataya et al. '22)
- To prevent collapse: Mix original real data with synthetic data in Step III (Dohmatob et al. '24, Gerstgrasser et al. '24)

Mitigating model collapse under overparametrization



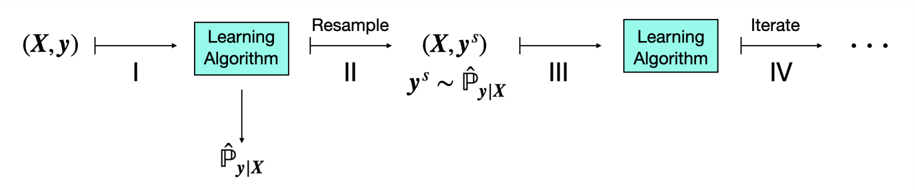
- *In what fraction should you mix the real and synthetic data to see optimal gains in prediction performance?*

Mitigating model collapse under overparametrization



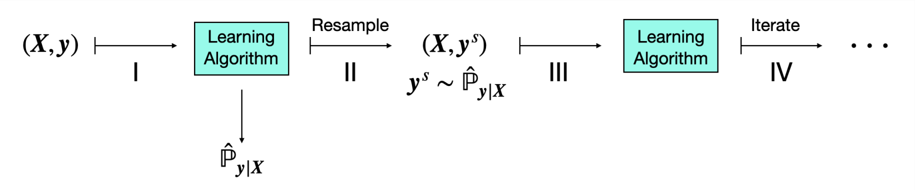
- *In what fraction should you mix the real and synthetic data to see optimal gains in prediction performance?*
- We can quantify this precisely by building upon our RMT advances

Mitigating model collapse under overparametrization



- *In what fraction should you mix the real and synthetic data to see optimal gains in prediction performance?*
- We can quantify this precisely by building upon our RMT advances
  E.g., If learning algorithm is min-ℓ_2-norm interpolator, the optimal real data fraction is reciprocal of the Golden Ratio (Garg, Bhattacharya and S. '25+)

Mitigating model collapse under overparametrization



- *In what fraction should you mix the real and synthetic data to see optimal gains in prediction performance?*
- We can quantify this precisely by building upon our RMT advances
  E.g., If learning algorithm is min-ℓ_2-norm interpolator, the optimal real data fraction is reciprocal of the Golden Ratio (Garg, Bhattacharya and S. '25+)
- Previously, rigorous understanding on how to mitigate model collapse existed only under low dimensions (Dey and Donoho '24, He et al. '25)

Thank you!

Contact: pragya@fas.harvard.edu

- Main Reference: Song, Y., Bhattacharya, S. and Sur, P., 2024+. Generalization error of min-norm interpolators in transfer learning. arXiv:2406.13944.
- See also: 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.
- Upcoming: Garg, A., Bhattacharya, S. and Sur, P., 2025+ Optimal Mixing Ratios for Preventing Model Collapse under Overparametrization, arXiv:2510.XXXX.