Title

Veridical Deep Learning in Hard Regimes: Distribution Shift, Agentic Data Science, and Data Scarcity

ISBN

9798189276552

Author

Zane, Austin Victor

Publication Date

2026-05-01

Peer reviewed|Thesis/dissertation

Veridical Deep Learning in Hard Regimes: Distribution Shift, Agentic Data Science, and Data Scarcity

by

Austin Victor Zane

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Distinguished Professor Bin Yu, Co-chair
Professor Peter Bartlett, Co-chair
Associate Professor Avi Feller

Spring 2026

Veridical Deep Learning in Hard Regimes: Distribution Shift, Agentic Data Science, and Data Scarcity

Abstract

Veridical Deep Learning in Hard Regimes: Distribution Shift, Agentic Data Science, and Data Scarcity

by

Austin Victor Zane

Doctor of Philosophy in Statistics

University of California, Berkeley

Distinguished Professor Bin Yu, Co-chair

Professor Peter Bartlett, Co-chair

Modern deep learning has succeeded in regimes where data is abundant and high-quality, the training and deployment distributions match, and the analyst keeps close control of the modeling pipeline. Real-world deployments rarely satisfy all of these conditions. This dissertation studies three such hard regimes for deep learning, distribution shift, agentic data science, and data scarcity, through the lens of the Predictability–Computability–Stability (PCS) framework for veridical data science.

The first project addresses distribution shift in the simplest tractable setting. We prove the first non-asymptotic excess risk bounds for benignly-overfit minimum-norm linear interpolators evaluated on a target distribution that differs from the source. From these bounds we propose a taxonomy of beneficial and malignant covariate shifts parameterized by the degree of overparameterization, and corroborate the taxonomy empirically on real image data and on fully-connected neural networks.

The second project tackles agentic data science, where a large language model executes the entire analytical workflow and the analyst can no longer audit the intermediate decisions. We propose a pair of lightweight sanity checks that apply targeted perturbations to the data and the analytic prompt and ask whether the agent's conclusion survives. The checks are validated on synthetic data with controlled signal-to-noise ratios and applied to eleven real-world datasets, identifying six on which an affirmative agentic conclusion is not well-supported.

The third project develops a deep learning pipeline for pediatric focused assessment with sonography for trauma (FAST), a bedside ultrasound exam used to detect hemorrhage from intra-abdominal injury (H-IAI) in injured children. The pipeline first localizes Morison's

pouch and then classifies frames for hemorrhage. We evaluate on a pediatric cohort of 207 exams, considerably larger than previous pediatric FAST datasets but containing only 17 H-IAI positive cases. We use Monte Carlo cross-validation to report metrics stable under this small positive count.

To my family.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my advisor, Bin Yu. Over the past five years, Bin has shaped not only how I think about research, but also how I think about what it means to be a mentor. She genuinely cares about her students as researchers and as people, and I have been deeply grateful for her guidance through both the technical and personal parts of the Ph.D. She taught me to hold my work to a high standard and to think critically. I feel very fortunate to have learned from her and I will carry her example of rigor, generosity, and mentorship with me long after graduate school.

I am also grateful to Peter Bartlett for serving as my co-advisor and for his support during my Ph.D. I appreciate his role in shaping the broader intellectual environment at Berkeley, especially through his leadership at the Simons Institute during my early years in the program. The workshops, programs, and summer schools were an important part of my graduate experience.

I would also like to thank Avi Feller for serving on my committee. Avi served as a mentor both during my first industry experience at Adobe and many important moments in my Ph.D. He provided advice in a way that nobody else could, and I feel very fortunate that he was so genuinely invested in my future.

Before Berkeley, I was fortunate to have several undergraduate mentors who set me on this path. My academic journey began in Boris Hanin's office hours during my first year of college, where a question about deep learning turned into a relationship that would span the rest of my undergraduate years and beyond. Boris served as my de facto undergraduate advisor and our conversations shaped both my decision to pursue a Ph.D. and my preparation for doing so. I would not be here without him.

I had my first exposure to research with Shuying Sun during a summer at Texas State University, where she taught me what rigor and attention to detail look like in practice and guided me through my first published paper. Huiyan Sang was also an important mentor during my undergraduate years, both in the classroom and in research, and her confidence in me when I was applying to graduate school meant a great deal.

The work in this thesis would not have been possible without my excellent co-authors Neil Mallinar, Spencer Frei, Zachary Rewolinski, Hao Huang, Chandan Singh, Chenglong Wang, Jianfeng Gao, Aaron Kornblith, Nathan McNaughton, Maytal Firnberg, Newton Addo, Naina Sabbineni, and Kush Narang.

The Yu Group served as the ideal environment for research and collaboration. My desire to work with Bin was cemented after getting to know all of the kind and intelligent students in the group, including Abhineet Agarwal, Luiz Chamon, James Duncan, Corrine Elliott, Soufiane Hayou, Jakob Heiss, Yaxuan Huang, Anthony Ozerov, Zachary Rewolinski, Omer Ronen, Chandan Singh, Jingfeng Wu, and Chengzhong Ye. These students and post-docs have served as friends, mentors, collaborators, and everything in between.

Outside of the Yu Group, I would like to thank the lifelong friends I made at Berkeley, especially Andy Shen and Zachary Rewolinski. Having Melody Huang and Tiffany Tang as my officemates during the first two years of my Ph.D. was a stroke of luck and had a

surprisingly large impact on my Berkeley experience. I doubt I will ever experience a better office environment. From my cohort, I am particularly grateful to Seunghoon Paik, Licong Lin, Tiffany Ding, Karissa Huang, Yaxuan Huang, Hyunsuk Kim, and Florica Constantine, who have been a central part of my life at Berkeley from the first-year classes onward. I couldn't ask for a better group to experience a Ph.D. with. I would also like to acknowledge the Order, for reasons they will understand.

I owe everything to my family. I could not ask for better role models than my mother and father, and whatever is good in me started with them. My mother has been the first person I go to for advice on essentially anything. I've benefited greatly from her keen social intelligence, and she has shown by example what lifelong learning actually looks like. Whatever steadiness I have, in work and out of it, I learned from her. My father is one of the deepest thinkers I know, and it is no coincidence that his two oldest children ended up as Ph.D. students. He raised the four of us to think for ourselves, to work hard, and to have no patience for self-pity, and has led by example on all three.

My siblings Isabella, Eva, and Jackson have been and always will be my closest friends. Isabella and her husband James, who were my roommates, classmates, and fellow math majors, made my time at Texas A&M particularly memorable, from late nights in the math lounge to birdwatching at Northgate. I am grateful to Eva for guaranteeing that every trip home is a good time. She is the most considerate of us all and has always gone out of her way for the rest of us. Jack ending up at Stanford during my Ph.D. has been a very fortunate coincidence. He has been generous with both his time and his space, taking the long train ride up to Berkeley on the weekends and putting me up whenever I am in the South Bay.

Finally, my fiancée Chelsea has been the most important person in my life since high school. She has read more of my drafts, listened to more of my practice talks, and tolerated more of my deadline weeks than anyone should have to. When I was stressed about grad school, the job search, or the future, none of it ever felt quite as bad knowing she would always be there. I owe her more than I can say here.

Chapter 1

Introduction

Modern deep learning has produced impressive results in regimes where data is abundant, the training and deployment distributions match, and the analyst keeps close control over the modeling pipeline. Real-world deployments rarely satisfy all of these conditions. Models trained on one population are often asked to perform on another. Rare events leave only a handful of positive labels even after extensive data collection. And increasingly, the analyst is not a person at all but an autonomous system whose modeling decisions are difficult to audit. In such settings, the standard pipeline of training a model, holding out a test set, and reporting a number can fail silently, producing a metric that looks reasonable but does not reflect deployment behavior.

This dissertation considers three such hard regimes. *Distribution shift* asks what happens when the training and target distributions differ. Even in the simplest linear models, the in-distribution intuition for when overparameterization is harmless can flip in surprising ways. *Agentic data science* asks how to assess the trustworthiness of a statistical conclusion when a large language model executes the entire analysis end-to-end and the analyst can no longer inspect the intermediate decisions. *Data scarcity* asks how to evaluate a deep learning model in clinical settings where positive cases are rare despite substantial data collection effort, so that single-split estimates of accuracy or calibration are dominated by sampling noise.

The methodological response to all three regimes shares a common shape, captured by the *Predictability–Computability–Stability* (PCS) framework for veridical data science [143, 142, 101]. Where the standard evaluation, whether empirical or theoretical, cannot be trusted on its own, one performs a *reality check* on whether the analytical setup reflects deployment, together with a battery of *stability checks*. The stability checks take the form of perturbations of the training data, of the theoretical model, of the modeling decisions, or of the analytical pipeline itself, and conclusions are reported only when they survive the perturbations. Where exact answers are out of reach, one structures the analysis to be explicit, reproducible, and inexpensive enough to run these checks repeatedly. These three components are instances of, respectively, the predictability, stability, and computability principles of PCS.

This thesis presents three projects, each tackling one of these hard regimes.

Distribution Shift

In-distribution research on high-dimensional linear regression has identified the phenomenon of *benign overfitting*, in which an interpolator of noisy training labels nonetheless generalizes well, provided the source covariance matrix and input dimension satisfy specific spectral conditions. In Chapter 2 we extend this analysis to the transfer learning setting and prove the first non-asymptotic excess risk bounds for benignly-overfit linear interpolators evaluated on a target distribution that differs from the source. This extension serves both as a reality check on whether the benign-overfitting prediction reflects deployment behavior and as a stability analysis of the theoretical conclusion under perturbation of the "true model" assumption that source and target distributions match. From these bounds we propose a taxonomy of *beneficial* and *malignant* covariate shifts, parameterized by the degree of overparameterization, and corroborate the taxonomy empirically for linear models on real image data and for fully-connected neural networks in the high-dimensional regime. This work was joint with Neil Mallinar, Spencer Frei, and Bin Yu.

Agentic Data Science

Agentic data-science (ADS) pipelines such as OpenAI Codex and Claude Code can now ingest a dataset and a natural-language query and produce a polished analysis end-to-end, but their conclusions can be wrong in ways that are difficult for an analyst to detect. In Chapter 3 we propose a pair of lightweight sanity checks for ADS outputs that are grounded directly in PCS. The first is a stability check that perturbs the data and analytic prompt and asks whether the agent's conclusion survives. The second is a reality check that strips the data of signal via a null-defining perturbation and asks whether the agent's affirmative conclusion would have been reached on noise alone. Together they characterize whether an output reflects stable signal, is responding to noise, or is sensitive to incidental aspects of the input. We validate the checks on synthetic data with controlled signal-to-noise ratios and demonstrate them on eleven real-world datasets, finding that in six of them an affirmative ADS conclusion is not well-supported even though a single run of the agent may suggest otherwise. The agent's self-reported confidence is poorly calibrated to the *stability* of its conclusions under perturbation, which is what the stability check measures directly. This work was joint with Zachary Rewolinski, Hao Huang, Chandan Singh, Chenglong Wang, Jianfeng Gao, and Bin Yu.

Data Scarcity

Pediatric focused assessment with sonography for trauma (FAST) is a bedside, radiation-free ultrasound exam used to detect hemorrhage from intra-abdominal injury (H-IAI) in injured children. Its accuracy is limited by clinician inexperience with both landmark acquisition and image interpretation, which leaves a clinically meaningful gap between the exam's potential

and its routine performance. In Chapter 4 we develop a two-step deep learning pipeline that first localizes Morison's pouch via a segmentation model and then classifies frames for H-IAI. Training and evaluation are performed on a pediatric cohort of 207 exams of which only 17 are H-IAI positive, a regime in which single-split estimates of sensitivity and area under the receiver operating characteristic curve (AUROC) have standard errors comparable to the metrics themselves. The operative data scarcity here is in positive cases rather than in raw exams, with the rare-positive count driving the effective sample size. The methodological design is therefore organized around repeated cohort splits, with reported metrics averaged over the splits in which each exam appeared as a test case so that no individual allocation of the 17 positives drives the headline numbers. This procedure serves both as a reality check on deployment-time performance, by averaging over many cohort splits rather than reporting a single noisy estimate, and as a stability analysis of the conclusions under perturbation of the train-test partition. This work was joint with Nathan McNaughton, Newton Addo, Naina Sabbineni, Kush Narang, Maytal Firnberg, Bin Yu, and Aaron Kornblith.

Summary

Each of the three regimes violates a different precondition for trustworthy single-number evaluation. Distribution shift breaks the assumption that source and target distributions match. Agentic data science removes the analyst's ability to inspect and audit the intermediate modeling decisions. Data scarcity drives the effective sample size below what a single train-test split can reliably estimate. What unifies them is the PCS framework's response to such violations, which is to substitute reality checks and stability analyses for the broken assumption and to keep the analytical pipeline computable enough that those checks can be performed in the first place.

Chapter 2

Minimum-Norm Interpolation Under Covariate Shift

2.1 Introduction

Practical deployments of machine learning models are almost always in a transfer learning setting, where models trained on a *source data distribution* with noisy labels are expected to perform well on a different *target data distribution*, referred to as the "out-of-distribution" (OOD) dataset [93, 22]. There have been many experimental works on transfer learning with complex models and datasets [100, 55, 86, 45, 134, 71], but remarkably fewer attempts to study it theoretically, even in the simplest case of linear models which have been of great interest in recent years [26, 6, 44, 128].

There has been an extensive "in-distribution" (ID) theoretical interest in high-dimensional linear regression and specifically *interpolation*, meaning a model reaches zero training loss [10, 12]. Frameworks such as "benign overfitting", or "harmless interpolation" [6, 89] emerged as an attempt to explain why interpolating neural networks often do not overfit catastrophically [144]. They found that, in specific cases, overfitting can be "benign", meaning that a model interpolates noisy training labels and yet has vanishing excess risk. In linear regression, this occurs if and only if the training (source) covariance matrix satisfies very specific conditions. Under these conditions, the minimum-norm interpolator (MNI) approximately acts like a ridge regression solution.

This sparked an initial wave of in-distribution theoretical research into benign overfitting in high-dimensional linear models [17, 128, 16], kernel regression [99, 42, 7, 11], and even some shallow neural networks [32, 62, 61, 139]. The study of benign overfitting was originally motivated by overparameterized neural networks. However, recent works have shown that overparameterization alone is not sufficient for the phenomenon, with overfitting failing to be benign in many practical settings where the input dimension is small relative to the parameter count [80, 42, 63]. Thus, deeper investigations into the generalization behavior of overfit models are warranted.

Given the increasing prevalence of overparameterized models, it is natural to ask how such models perform in the transfer learning setting. There have been some efforts to answer this in the theoretically tractable cases of linear regression and random feature and kernel regression [95, 131]. However, these works either provide asymptotic bounds that require the training sample size and data dimension to go to infinity at the same rate [126], study minimax settings which only consider worst-case risk [64], or focus on augmented gradient-based training algorithms, like importance weighting [132].

Summary of contributions. In this chapter, we investigate the generalization behavior of the minimum ℓ_2-norm linear interpolator (MNI) under distribution shifts when the source distribution satisfies the conditions necessary for benign overfitting. We summarize our main contributions as follows.

- We provide the first non-asymptotic, instance-wise risk bounds for covariate shifts in interpolating linear regression when the source covariance matrix satisfies benign overfitting conditions and commutes with the target covariance matrix.
- We use our risk bounds to propose a taxonomy of covariate shifts for the MNI. We show how the ratio of target eigenvalues to source eigenvalues and the degree of overparameterization affect whether a shift is *beneficial* or *malignant*, meaning OOD risk is better or worse than ID risk, respectively. The degree of overparameterization is determined by the eigenspectrum's head and tail properties.
- We empirically show that our taxonomy of shifts holds: (1) for the MNI on real image data under natural shifts like blur (a beneficial shift) and noise (a malignant shift), underscoring the significance of our findings beyond the idealized source and target covariances for which our theory is applicable; (2) for neural networks in settings where the input data dimension is larger than the training sample size, showing that our findings for the MNI are also reflective of the behavior of more complex models.

Prior Work and Comparisons to this Work

Excess risk analysis under distribution shifts: Tripuraneni, Adlam, and Pennington [126] give an asymptotic analysis of high-dimensional random feature regression in covariate shift. They require the number of samples, n, data dimension, p, and random feature dimension to go to ∞ at the same rate. In contrast, our non-asymptotic analysis considers finite sample cases and differing rates. This allows us to draw new conclusions about how the *degree of overparameterization* changes the way in which interpolating linear models exhibit out-of-distribution (OOD) generalization. Additionally, our bounds let us analyze any sequence of eigenvalues for the target feature covariance matrix, which is not possible within their framework.

Lei, Hu, and Lee [64] study linear regression under distribution shifts in the minimax setting. Their minimax bounds consider the worst-case risk over an ℓ_2-ball of target models, whereas

we compute risk bounds specific to any model instantiation, with no restriction on the target model class. Furthermore, their experimental results only consider the underparameterized regime.

Several other works study OOD generalization in more distant settings. Wang et al. [132] study linear interpolators for classification, when trained with gradient descent and importance weighting, whereas we consider the closed-form MNI for linear regression. Simchowitz et al. [112] study covariate shifts when the target function class is the sum of two other function classes, and shifts are defined with regard to metric entropy between classes, whereas we focus on well-specified linear models. Pathak, Ma, and Wainwright [95], Ma, Pathak, and Wainwright [78], and Feng et al. [30] consider covariate shift in kernel regression based on likelihood ("importance") ratios between source and target distributions while we consider source and target eigenvalue ratios which offer granular insights into feature scale changes whereas likelihood ratios capture shifts that affect the global data distribution. Pathak, Ma, and Wainwright [95] and Ma, Pathak, and Wainwright [78] also analyze worst-case, minimax risk for nonparametric function classes. Kausik, Srivastava, and Sonthalia [51] work in the proportional asymptotic regime and consider the error in variables setting with noisy features and clean labels, while our work focuses on the linear regression setting with clean features and noisy labels. Finally, we note that risk bounds in these prior works do not sufficiently account for the behavior of the high-rank covariance tail that benign overfitting requires.

Experimental work on distribution shifts: Hendrycks and Dietterich [46] propose the CIFAR-10C dataset as an OOD counterpart to CIFAR-10, featuring test set images corrupted by visual filters like blurs and noises. Koh et al. [55] present benchmarks on more realistic datasets with modern models that can be seen "in-the-wild". Miller et al. [86] experimentally show a linear relationship between ID accuracy and OOD accuracy for a wide range of modern neural networks and datasets, though their results show ID accuracy is almost always better than OOD accuracy. On a subset of CIFAR-10C, we find settings in which OOD accuracy is *better* than ID accuracy for linear interpolators.

Benign overfitting "in-distribution": Bartlett et al. [6] propose benign overfitting, give a non-asymptotic analysis of the MNI, and show specific, necessary conditions under which the MNI achieves zero excess risk in-distribution. Tsigler and Bartlett [128] extend this work by considering benign overfitting in the case of ridge regression. Our proof techniques follow most closely to the ideas presented in these two papers for the in-distribution setting. Frei, Chatterji, and Bartlett [32] show benign overfitting in shallow non-linear MLPs trained with gradient descent on the logistic loss if the data dimension grows faster than the number of training samples. Mallinar et al. [80] experimentally show that interpolating neural networks do not benignly overfit due to the low input data dimension. Our experiments build on this by looking at settings in which $n < p$ and $n > p$ where n is the training sample size and p is the input data dimension. Other works study benign overfitting under a variety of conditions [62, 16, 34].

2.2 Preliminaries

We extend notations in Bartlett et al. [6] and Tsigler and Bartlett [128] to the transfer learning setting with OOD generalization risk as our performance metric. Appendix A.1 formalizes our setting of linear regression under distribution shift, and we provide necessary details here.

Linear Models for Source and Target Data

Let $\mathcal{D}_{\mathsf{s}}$ and $\mathcal{D}_{\mathsf{t}}$ be source and target distributions over $(x, y) \in \mathbb{R}^p \times \mathbb{R}$. We consider linear regression problems defined as follows.

Definition 2.1 (Linear regression)**.** Let the training dataset be comprised of n i.i.d. pairs $(\boldsymbol{x}^i, y^i)_{i=1}^n \sim \mathcal{D}_{\mathsf{s}}^n$ concatenated into a data matrix $X \in \mathbb{R}^{n \times p}$ and a response vector $\boldsymbol{y}_{\mathsf{s}} \in \mathbb{R}^n$, where $n < p$. We define

1. the covariance matrix $\Sigma_{\mathsf{s}} = \mathbb{E}_{\mathcal{D}_{\mathsf{s}}}[\boldsymbol{x}\boldsymbol{x}^\top]$,
2. (*centered rows*) $\mathbb{E}_{\mathcal{D}_{\mathsf{s}}}[\boldsymbol{x}] = 0$,
3. (well-specified) the optimal parameter vector $\theta_{\mathsf{s}}^* \in \mathbb{R}^p$ such that

$$y = \boldsymbol{x}^T \theta_{\mathsf{s}}^* + \varepsilon_{\mathsf{s}}$$

for $(\boldsymbol{x}, y) \sim \mathcal{D}_{\mathsf{s}}$, where $\boldsymbol{\varepsilon}_{\mathsf{s}}$ is a centered random variable with variance $v_{\varepsilon_s^2}$ and $\mathbb{E}_{\mathcal{D}_{\mathsf{s}}}[y|\boldsymbol{x}] = \boldsymbol{x}^T \theta_{\mathsf{s}}^*$.

We test on $\mathcal{D}_{\mathsf{t}}$ with Σ_{t}, θ_{t}^*, ε_{t} defined in the same way. Note that $(\boldsymbol{x}, y)$ is used to denote single observation pairs for both source and target data. We will differentiate between the two by explicitly denoting the distribution from which the pair is drawn.

To facilitate our analysis, we introduce the following assumptions on the covariance matrices and the distribution of the data.

Assumption 2.2. *For linear regression problems (Def. 2.1), with source and target covariance matrices Σ_{s} and $\Sigma_{\mathsf{t}} \in \mathbb{R}^{p \times p}$, we assume:*

1. *(simultaneous diagonalizability) Σ_{s} and Σ_{t} commute; that is, there exists an orthogonal matrix $V \in \mathbb{R}^{p \times p}$ such that $V^\top \Sigma_{\mathsf{s}} V$ and $V^\top \Sigma_{\mathsf{t}} V$ are both diagonal:*

$$\Sigma_{\mathsf{s}} = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathsf{s}}}[\boldsymbol{x}\boldsymbol{x}^T] = diag(\lambda_1, \lambda_2, ..., \lambda_p),$$
$$\Sigma_{\mathsf{t}} = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathsf{t}}}[\boldsymbol{x}\boldsymbol{x}^T] = diag(\tilde{\lambda}_1, \tilde{\lambda}_2, ..., \tilde{\lambda}_p),$$

where $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p$ and $\tilde{\lambda}_i \lambda_i \geq 0$ for all i;

2. *(subgaussianity) the whitened observations* $\boldsymbol{z} = \boldsymbol{x}^T \Sigma_{\mathsf{s}}^{-1/2}$ *are centered i.i.d. vectors with independent coordinates and subgaussian norm* σ_x*; that is, for all* $\gamma \in \mathbb{R}^p$*,*

$$\mathbb{E}[\exp(\gamma^\top \boldsymbol{z})] \leq \exp(\sigma_x^2 ||\gamma||^2/2)$$

Simultaneous diagonalizability is a common assumption in recent studies of high-dimensional linear regression [64, 51, 65] and we show in Section 2.4 with experiments that our results hold even when this is violated. Subgaussianity is also frequently used in statistical learning theory research and encompasses a wide array of distributions of interest [130].

Min-Norm Interpolator and Target Excess Risk

Given a source data matrix X, the minimum-norm interpolator (MNI) for any vector $\xi \in \mathbb{R}^n$ is defined as

$$\begin{aligned}\widehat{\theta}(\xi) &:= \operatorname{argmin} \left\{ \|\theta\|^2 : X\theta = \xi \right\} \\ &= X^T (XX^T)^{-1} \xi.\end{aligned}$$

If we consider $\xi = \boldsymbol{y}_{\mathsf{s}}$, then we recover the MNI for the labels given by the response model, but our analysis will also involve implicit MNIs for different label vectors in $\mathbb{R}^n$.

The quantity that we seek to bound is the excess risk on the target distribution, which we define for an estimator $\theta \in \mathbb{R}^p$ as,

$$R(\theta, \mathcal{D}_{\mathsf{t}}) := \mathop{\mathbb{E}}_{\mathcal{D}_{\mathsf{t}}} \left[\left(y - \boldsymbol{x}^T \theta \right)^2 - \left(y - \boldsymbol{x}^T \theta_{\mathsf{t}}^* \right)^2 \right]. \tag{2.1}$$

We now derive bounds for the target excess risk and its expectation over the source response noise. The proof of the following can be found in Appendix A.3.

Theorem 2.3. *(Target excess risk decomposition) The excess risk of the MNI trained on the source data, when evaluated on the target distribution, satisfies*

$$R(\widehat{\theta}(\boldsymbol{y}_{\mathsf{s}}), \mathcal{D}_{\mathsf{t}}) \leq 4B_1 + 4B_2 + 2V_{\boldsymbol{\varepsilon}_{\mathsf{s}}}, \tag{2.2}$$

and

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_{\mathsf{s}}} R(\widehat{\theta}(\boldsymbol{y}_{\mathsf{s}}), \mathcal{D}_{\mathsf{t}}) = B_1 + B_2 + \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_{\mathsf{s}}} V_{\boldsymbol{\varepsilon}_{\mathsf{s}}} \\ + 2(\theta_{\mathsf{t}}^* - \theta_{\mathsf{s}}^*)^\top \Sigma_{\mathsf{t}} (\theta_{\mathsf{s}}^* - \widehat{\theta}(X\theta_{\mathsf{s}}^*)),\end{aligned}$$

where we define

$$B_1 := \|\theta_{\mathsf{s}}^* - \theta_{\mathsf{t}}^*\|_{\Sigma_{\mathsf{t}}}^2, \tag{2.3}$$

$$B_2 := \|\theta_{\mathsf{s}}^* - \widehat{\theta}(X\theta_{\mathsf{s}}^*)\|_{\Sigma_{\mathsf{t}}}^2, \tag{2.4}$$

$$V_{\boldsymbol{\varepsilon}_{\mathsf{s}}} := \|\widehat{\theta}(\boldsymbol{\varepsilon}_{\mathsf{s}})\|_{\Sigma_{\mathsf{t}}}^2, \tag{2.5}$$

and $\|\boldsymbol{x}\|_M^2 := \boldsymbol{x}^\top M \boldsymbol{x}$.

We observe that B_1 is a deterministic model shift term and that no further analysis can improve its dependency on θ^*_{s}, θ^*_{t}, or Σ_{t}. The cross-term, $(\theta^*_{\mathsf{t}} - \theta^*_{\mathsf{s}})^\top \Sigma_{\mathsf{t}} (\theta^*_{\mathsf{s}} - \widehat{\theta}(X\theta^*_{\mathsf{s}}))$, is dominated by the bias and variance as evidenced by the upper bound. Therefore we focus our analysis on B_2 and $V_{\boldsymbol{\varepsilon}_{\mathsf{s}}}$. A useful normalized version of $V_{\boldsymbol{\varepsilon}_{\mathsf{s}}}$ is defined by

$$V = \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_{\mathsf{s}}} \left[V_{\boldsymbol{\varepsilon}_{\mathsf{s}}} / v^2_{\boldsymbol{\varepsilon}_{\mathsf{s}}} \right]. \tag{2.6}$$

Note that B_2, V are reminiscent of the ID bias and variance in prior work [6, 128].

Separation of Components and Effective Ranks

For an index k, we define the following quantities related to the effective rank of the tail of Σ_{s} [128]:

$$\rho_k = \frac{\sum_{i>k} \lambda_i}{n\lambda_{k+1}}, \qquad R_k = \frac{(\sum_{i>k} \lambda_i)^2}{\sum_{i>k} \lambda_i^2}.$$

ρ_k measures the ratio of the energy of the source covariance tail to the number of training data observations, after normalizing the tail eigenvalues. R_k measures the quantity of noisy features and how evenly distributed their eigenvalues are. It is minimized when there is only one nonzero eigenvalue and maximized when there are many equal eigenvalues.

Benign overfitting occurs if the MNI is overfit to noisy training labels and yet ID excess risk decays to zero. The central finding of Bartlett et al. [6] is that the only way benign overfitting happens for the MNI is if the following occurs: (1) there exists a $k^* = \min\{k \ : \ \rho_k \geq b\}$ for a universal constant $b > 1$, meaning that the last $p - k^*$ components of Σ_{s} have a high effective rank relative to the number of training samples, n; (2) the magnitudes of the bottom $p - k^*$ eigenvalues are small relative to the top k^*; and (3) $k^* \ll n$. More formally, consider quantities $p = p(n)$, a sequence of source covariance matrices $\Sigma_n = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$, $k^* = k^*(n)$ as defined above, $R_{k^*} = R_{k*}(\Sigma_n)$, and $\rho_k = \rho_k(\Sigma_n)$. A sufficient condition for benign overfitting is,

$$\lim_{n\to\infty} \rho_0 = \lim_{n\to\infty} k^*/n = \lim_{n\to\infty} n/R_{k^*} = 0. \tag{2.7}$$

If this occurs, then the MNI behaves similarly to an estimator with two components. One component has variance similar to the ordinary least squares (OLS) estimator in k^* dimensions and bias similar to the ridge regression solution with ridge parameter proportional to $\sum_{i>k} \lambda_i$, a sort of data-induced regularization. The other component is a high-dimensional component, which has vanishing variance when the data is sufficiently high-dimensional and a bias which is proportional to $\sum_{i>k} \lambda_i (\theta^*_{\mathsf{s}})^2_i$ [128]. From these conditions, we see that the top k^* components are like "signal" components of the data and the bottom $p - k^*$ components are "noise" components.

Spiked Covariance Models

We will consider a special case of the (k, ϵ)-spike model, a canonical covariance structure that exhibits benign overfitting for the MNI [17, 16], to experimentally show properties of interest.

Definition 2.4 ((k, δ, ϵ)-spike model)**.** For a source distribution $\mathcal{D}_{\mathtt{s}}$, $\delta > 0$ and $\epsilon > 0$ such that $\delta \gg \epsilon$, let

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathtt{s}}}[\boldsymbol{x}\boldsymbol{x}^T] = \mathrm{diag}(\underbrace{\lambda_1, \cdots, \lambda_k}_{=\delta}, \underbrace{\lambda_{k+1}, \cdots, \lambda_p}_{=\epsilon}).$$

In this simplified setting, there are k high-energy "signal" directions and $p - k$ low-energy "noise" directions. For a target distribution $\mathcal{D}_{\mathtt{t}}$, we use different hyperparameters $\tilde{k}, \tilde{\delta}, \tilde{\epsilon}$ to similarly characterize a shifted covariance matrix.

2.3 Main Theorems

This section provides upper and lower bounds for the variance and bias terms in Equation 2.6 and Equation 2.4, respectively. Appendix A.4 gives a high-level overview of our proof techniques. Subsequent appendices provide complete proofs. The variance bounds are adapted from Bartlett et al. [6], while the bias lower bound is derived from Tsigler and Bartlett [128]. Our contributions include a novel bias upper bound and a unique characterization of overparameterization degrees. We start with the bounds for the variance term. Appendix A.5 contains a proof of the following theorem.

Theorem 2.5. *(Upper and lower bounds for the variance term) There exist universal constants $b, c_1 > 1$ given in Lemma A.2, a universal constant c_2 given in Lemma A.8 and a constant $c > 1$ that only depends on σ_x, c_1, c_2, such that for $k \in (0, n/c)$, with probability at least $1 - 10e^{-n/c}$,*

$$V \geq \frac{1}{cn} \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \min\left(1, \frac{\lambda_i^2}{\lambda_{k+1}^2 (\rho_k + 1)^2}\right) =: \underline{V}. \tag{2.8}$$

If in addition $\rho_k \geq b$, with probability at least $1 - 7e^{-n/c}$,

$$V/c \leq \frac{1}{n} \sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} + n \frac{\sum_{i>k} \tilde{\lambda}_i \lambda_i}{(\sum_{i>k} \lambda_i)^2} =: \overline{V}/c. \tag{2.9}$$

We first note that the variance lower bound does not depend on $\rho_k \geq b$ and so it holds for any interpolating linear model, even when benign source conditions are not satisfied. However, we will see that if $\rho_k \geq b$ for some k, then the upper and lower bounds are tight. In the case where $\Sigma_{\mathtt{t}} = \Sigma_{\mathtt{s}}$, these bounds reduce to their in-distribution counterparts [6].

Our variance bounds show that the excess risk contribution of each feature is scaled by the ratio of the target and source eigenvalues, $\tilde{\lambda}_i/\lambda_i$. We immediately see that scaling down the target eigenvalues will lessen the overall contribution to variance and that scaling up the target eigenvalues will increase the contribution. We investigate these scaling factors and the separation of the first k components and last $p-k$ components in Section 2.3.

We now state upper and lower bounds for the bias term, B_2, given in Equation 2.4. The proof of the following theorem can be found in Appendix A.6.

Theorem 2.6. *(Upper and lower bounds for the bias term) For the lower bound only, assume that random models $\overline{\theta}$ are obtained from the underlying θ^*_{s} as $(\overline{\theta})_i = \gamma_i(\theta^*_{\mathsf{s}})_i$, where each γ_i is an independent Rademacher random variable. There exists a universal constant $b > 1$, constants c, C that depend only on b and σ_x, and $k < n/C$ such that if $\rho_k \geq b$, then with probability at least $1 - 10e^{-n/c}$,*

$$\mathop{\mathbb{E}}_{\bar{\theta}}[B_2] \geq \frac{1}{c}\left(\sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i (\theta^*_{\mathsf{s}})_i^2}{(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k})^2} + \sum_{i>k} \tilde{\lambda}_i (\theta^*_{\mathsf{s}})_i^2\right) =: \underline{B_2}.$$

If we assume that p is at most exponential in n, then with probability at least $1 - 5e^{-n/c}$,

$$B_2/c \leq \|\theta^*_{\mathsf{s}}\|^2 \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i}{\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)} =: \overline{B_2}/c.$$

Note that while the lower bound is in expectation over the random models $\bar{\theta}$, the resulting expression only depends on the ground-truth θ^*_{s}. This Bayesian approach also appears in prior work, i.e., Tsigler and Bartlett [128]. In studying the bias lower bound, we observe a similar separation of signal and noise components and dependence on eigenvalue ratios as in the variance bounds.

To show tightness of our bounds, we assume there exists a k such that $\rho_k \geq b$ for some universal constant $b > 1$. When this condition is satisfied, the variance bounds are tight up to constant factors. The bias bounds leave a model-dependent and source covariance-dependent gap, which we discuss in the proof overview in Appendix A.4 and in the complete proof found in Appendix A.7.

Theorem 2.7. *(Tightness of variance and bias bounds) Let the lower bound and upper bound of V be given by $\underline{V}$ and $\overline{V}$, respectively. There exists a universal constant $b \geq 1$, and constant c as defined in Theorem 2.5, and $k \in (0, n/c)$ such that if $\rho_k \geq b$, then*

$$\underline{V}/\overline{V} \in \left[b^{-2}(1+b)^{-2}/c^2, 1\right].$$

Let the lower bound and upper bound of B_2 be given by $\underline{B_2}$ and $\overline{B_2}$, respectively, and the assumptions of Theorem 2.6 be satisfied. Then

$$\underline{B_2}/\overline{B_2} \in \left[\frac{\min_i \{(\theta^*_{\mathsf{s}})_i^2 \ : \ (\theta^*_{\mathsf{s}})_i \neq 0\}}{\|\theta^*_{\mathsf{s}}\|^2 \left(1 + b^{-1}\frac{\lambda_1}{\lambda_{k+1}}\right)}, 1\right].$$

Note that the gap between our bias upper and lower bounds is independent of the target distribution.

A Taxonomy of Shifts

We now present a taxonomy of covariate shifts on the target distribution inspired by our prior analysis. We first consider OOD generalization and formally categorize shifts as *beneficial* or *malignant*.

Definition 2.8 (Beneficial and Malignant shifts)**.** For a source distribution, $\mathcal{D}_s$, a target distribution, $\mathcal{D}_t$, excess risk, R, and MNI, $\widehat{\theta}$, we say that a shift is

1. **beneficial** if $R(\widehat{\theta}, \mathcal{D}_s) > R(\widehat{\theta}, \mathcal{D}_t)$,
2. **malignant** if $R(\widehat{\theta}, \mathcal{D}_s) < R(\widehat{\theta}, \mathcal{D}_t)$.

We define these shifts for excess risk and note in Appendix A.10 that, empirically, the variance is the primary contributor to excess risk and the bias contributions are negligible when $\Sigma_{\mathtt{s}}$ satisfies benign overfitting conditions. This is in keeping with prior literature that focuses on studying variance in interpolating methods [6]. We will thus focus on variance in the following discussion.

Prior work shows that if $n, p \to \infty$ at the same rate, $\operatorname{tr}(\Sigma_{\mathtt{s}}) < \operatorname{tr}(\Sigma_{\mathtt{t}})$ results in malignant shifts on excess risk and $\operatorname{tr}(\Sigma_{\mathtt{s}}) > \operatorname{tr}(\Sigma_{\mathtt{t}})$ results in beneficial shifts on excess risk [126]. In this section we generalize these conditions by considering differing rates of $n, p \to \infty$ and measuring overparameterization by the modified "effective rank" measure R_k/n rather than p. This leads us to a novel characterization of the role of overparameterization in covariate shifts. For completeness, we describe their trace conditions in terms of our shifts in Appendix A.8.

We first consider a simplified example with separate multiplicative shifts on the signal and noise components to facilitate intuition. While this is a special case, it provides valuable insights into the dynamics of overparameterization and covariate shift that are relevant to practice. Our general results, which allow for arbitrary multiplicative shifts in every direction, are presented in Appendix A.8.

Let $\Sigma_{\mathtt{s}}$ be a covariance matrix that satisfies benign source conditions and denote the ID variance term by V_{id}. Define $\Sigma_{\mathtt{t}}$ by $\tilde{\lambda}_i = \alpha\lambda_i$ for $i \le k$, and $\tilde{\lambda}_i = \beta\lambda_i$ for $i > k$ with $\alpha, \beta \ge 0$. Let V_{ood} denote the OOD variance term. By Theorem 2.5, up to constants,

$$V_{id} \approx k/n + n/R_k, \tag{2.10}$$

$$V_{ood} \approx \alpha(k/n) + \beta(n/R_k), \tag{2.11}$$

where $R_k = (\sum_{i>k} \lambda_i)^2/(\sum_{i>k} \lambda_i^2)$.

It is clear that if $V_{ood} - V_{id} > 0$ then we have a malignant shift on the variance, and if $V_{ood} - V_{id} < 0$ then we have a beneficial shift on the variance. Observe that,

$$V_{ood} - V_{id} \approx (\alpha - 1)(k/n) + (\beta - 1)(n/R_k). \tag{2.12}$$

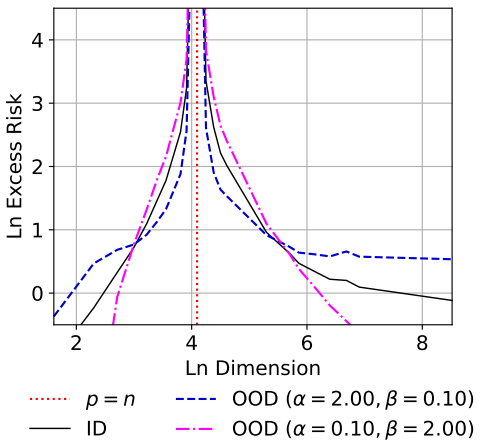
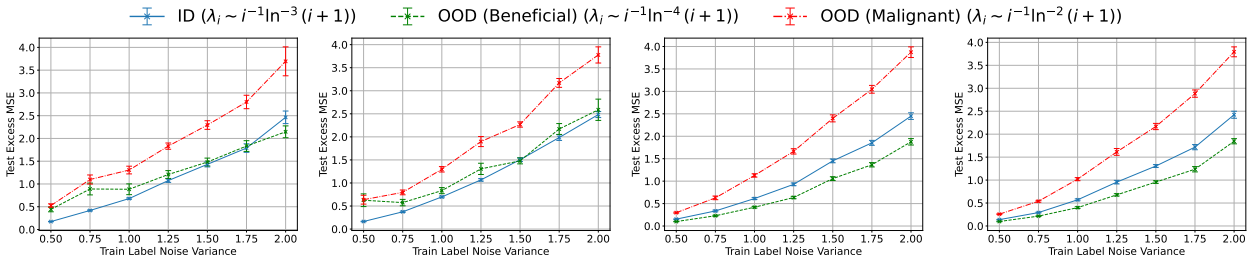


Figure 2.1: We experiment with the (k, δ, ϵ) spiked covariance models and examine conditions for beneficial and malignant shifts as given in Theorem 2.10. We take $n = 60, k = 10, \delta = 1.0, \epsilon = 1e^{-6}, \tilde{\delta} = 2.0, \tilde{\epsilon} = 1e^{-7}$, and vary p. We see a cross-over from mild to severe overparameterization on the right side of $p = n$ where both OOD shifts swap between beneficial and malignant. For both ID and OOD curves, we observe that excess risk is a decreasing function of input dimension. Curves are averaged over 100 independent runs.

In this expression, we see that the scales of signal and noise shifts, α and β, are important, as is the relationship between k/n (the "classical" rate) and n/R_k (the "high-dimensional" rate). The quantity n/R_k can be interpreted as an inverse measure of overparameterization, where smaller values correspond to higher levels of overparameterization. The rate of overparameterization relative to the classical rate of k/n determines whether the shift on the first k components (α) or the shift on the last $p - k$ components (β) contributes more to the difference in excess risk.

Based on this intuition, we define two regimes of overparameterization: *mild* and *severe*.



(a) $n = 200, p = 20, h = 512$ (b) $n = 200, p = 20, h = 2048$ (c) $n = 200, p = 2k, h = 512$ (d) $n = 200, p = 2k, h = 2048$

Figure 2.2: We train 3 layer ReLU dense neural networks with hidden width, h, on n samples from p-dimensional Gaussians. ID test data is sampled from the same distribution and OOD test sets are constructed based on beneficial and malignant covariate shifts in our theory. Ground truth models are sampled as $\theta^*_{\mathtt{s}} \sim \mathcal{S}^{p-1}$, no model shift is invoked. For training data, X, train labels are given by $\boldsymbol{y}_{\mathtt{s}} = X\theta^*_{\mathtt{s}} + \boldsymbol{\varepsilon}_{\mathtt{s}}$ with label noise $\boldsymbol{\varepsilon}_{\mathtt{s}} \sim \mathcal{N}(0, \sigma^2)$. All runs reach train loss $< 5e^{-6}$. Points are averaged over 20 independent runs with standard error bars reported.

Definition 2.9 (Mild and severe overparameterization for multiplicative shifts)**.** Let $\Sigma_{\mathtt{s}}$ be a source covariance that satisfies benign source conditions and let $k \leq n$. Define $\Sigma_{\mathtt{t}}$ as, $\tilde{\lambda}_i = \alpha\lambda_i$ for $i \leq k$ and $\tilde{\lambda}_i = \beta\lambda_i$ for $i > k$, with $\alpha, \beta \geq 0$. Let $C_{\alpha\beta} := \left|\frac{\alpha-1}{1-\beta}\right|$.

We are in the **mildly overparameterized** regime if

$$n/R_k = \omega\left(C_{\alpha\beta} \cdot k/n\right).$$

We are in the **severely overparameterized** regime if

$$n/R_k = o\left(C_{\alpha\beta} \cdot k/n\right).$$

For $\beta = 1$ we define $C_{\alpha\beta} = \infty$ and thus are effectively in the severely overparameterized regime with regard to the types of shift we observe.

It is clear that k is important in defining regimes of overparameterization. The aforementioned definitions hold for any $k < n$, however we derive our taxonomy of shifts in the case in which $\exists\ k < n$ such that $\rho_k \geq b$ for a universal constant $b > 1$. We note that even for non-linear models or settings that do not exhibit benign overfitting we can still think about a notion of a "k" akin to the intrinsic dimension of the data. We empirically show in Figures A.3 and A.4 that our taxonomy of shifts is reflective of shift behavior in realistic settings by heuristically taking k small enough to sufficiently capture the low-dimensional signal in the data.

In Equation 2.12, we see that the limit of the severe overparameterization regime would take $R_k \to \infty$ first, while holding other problem parameters fixed. In this case, we are only left

with α shifts on the top k components, as any shift on the bottom components is suppressed by the high rank covariance tail. This leads to behaviors akin to classical intuitions for an underparameterized linear regression estimator where $k = p < n$. In this case, $\alpha > 1$ leads to more variance and thus harder learning, whereas $\alpha < 1$ leads to less variance and thus easier learning. These notions of hard vs. easy learning naturally correspond to $\mathrm{tr}(\Sigma_{\mathsf{t}}) > \mathrm{tr}(\Sigma_{\mathsf{s}})$ and $\mathrm{tr}(\Sigma_{\mathsf{t}}) < \mathrm{tr}(\Sigma_{\mathsf{s}})$, respectively. This is shown experimentally in Figs. 2.1 and A.4 by looking at the left and right sides of the figures.

On the other hand, in the mildly overparameterized regime covariance tail shifts are not sufficiently suppressed and lead to non-negligible interactions with shifts on the signal components. An increase in energy in the signal components can be counteracted by a decrease in energy in the noise components, effectively increasing the contrast between signal and noise in favor of the signal. Similarly, a decrease in energy in the signal components can be harmfully counteracted by an increase in energy in the noise components. This is visible in Figs. 2.1 and A.4 just above the threshold of interpolation. Interestingly, in the mildly overparameterized regime we can also define settings in which $\mathrm{tr}(\Sigma_{\mathsf{t}}) > \mathrm{tr}(\Sigma_{\mathsf{s}})$ and yet still obtain a beneficial shift, and settings in which $\mathrm{tr}(\Sigma_{\mathsf{t}}) < \mathrm{tr}(\Sigma_{\mathsf{s}})$ and yet still obtain malignant shifts.

We formalize these observations in the following theorem, the proof of which can be found in Appendix A.8.

Theorem 2.10. *(Beneficial and Malignant Multiplicative Shifts on Variance) Let Σ_{s} be a source covariance that satisfies benign source conditions. That is, $\exists\ k$ such that $\rho_k \geq b$ for a universal constant $b > 1$. Define Σ_{t} as $\tilde{\lambda}_i = \alpha\lambda_i$ for $i \leq k$ and $\tilde{\lambda}_i = \beta\lambda_i$ for $i > k$, with $\alpha, \beta \geq 0$.*

1. *If $\alpha < 1, \beta \leq 1$ or $\alpha \leq 1, \beta < 1$ then we obtain a beneficial shift in variance.*
2. *If $\alpha > 1, \beta \geq 1$ or $\alpha \geq 1, \beta > 1$ then we obtain a malignant shift in variance.*
3. *If we are in the mildly overparameterized regime:*
   - *$\alpha > 1$ and $\beta < 1$ leads to beneficial shifts;*
   - *$\alpha < 1$ and $\beta > 1$ leads to malignant shifts.*
4. *If we are in the severely overparameterized regime:*
   - *$\alpha > 1$ and $\beta < 1$ leads to malignant shifts;*
   - *$\alpha < 1$ and $\beta > 1$ leads to beneficial shifts.*

Figs. 2.1 and A.2 demonstrate the relationship between the n/R_k and k/n rates in the case of $C_{\alpha\beta} = 1.11, C_{\alpha\beta} = 1$, respectively, for spiked covariance models. In both, we clearly see a cross-over from beneficial to malignant shifts when we transition from mild to severely overparameterized.

Overparameterization improves OOD robustness Focusing on just the target excess risk, let $\alpha = \alpha(n)$ and $\beta = \beta(n, p)$. We say that the benignly-overfit MNI is robust if its excess risk decays to zero despite the presence of multiplicative covariate shifts. In order for the variance upper bound to decay to 0, it is sufficient to have the shifts in the signal and noise components satisfy, $\alpha = o(n/k),\ \beta = o(R_k/n)$. The condition $\beta = o(R_k/n)$ allows the shift strength to increase at a rate determined by the level of overparameterization, so we conclude that increasing the amount of overparameterization improves robustness to multiplicative distribution shifts. Note that α has no dependence on R_k and so robustness to shifts on the signal components is independent of the degree of overparameterization.

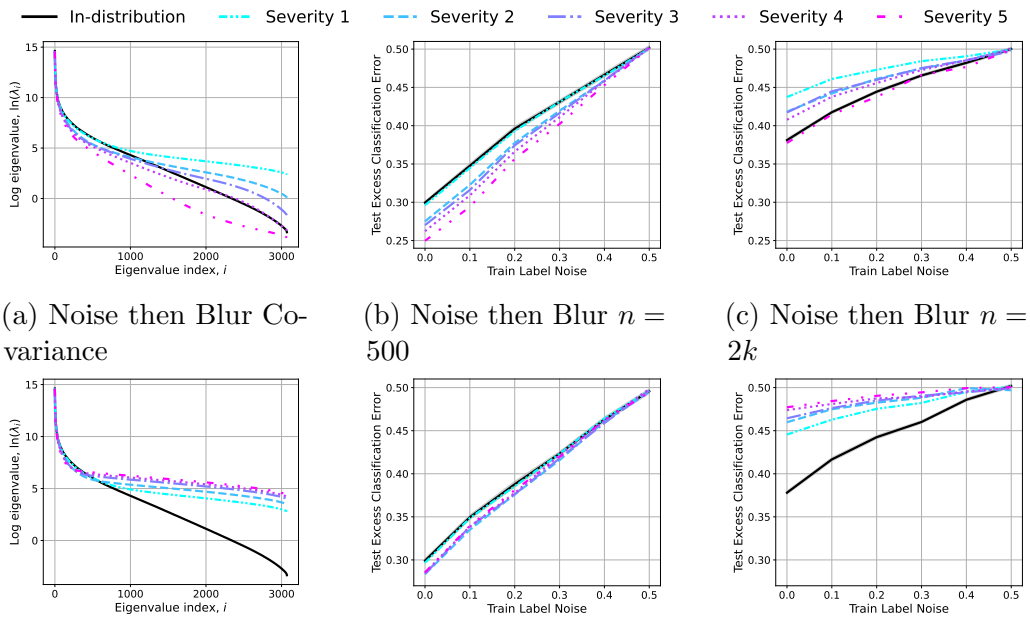
2.4 Experiments

Our theoretical results have provided insight into distribution shifts in high-dimensional linear regression. We now present experiments with linear models and neural networks, relaxing many of the assumptions used for theoretical results. Specifically, we empirically: (1) observe beneficial and malignant shifts on synthetic and real data for linear models (benignly overfit and otherwise) and even high-dimensional dense neural networks; (2) show the benefit of overparameterization in covariate shift for interpolating linear estimators; (3) validate that our findings hold when the source and target covariance matrices are not simultaneously diagonalizable, as well as under model misspecification; (4) provide experimental insight that high-dimensional neural networks, i.e., when the input data dimension is large relative to the training sample size, act similarly to the MNI whereas low-dimensional neural networks do not, regardless of the level of overparameterization. Details of experimental setup, data, and models are given in Appendix A.9. We now discuss key observations and takeaways from the experiments.

Synthetic Data Experiments

Fig. 2.1 shows excess risk vs. input dimension for data sampled from the (k, δ, ϵ)-spike covariance model with $k = 10$, $\delta = 1.0$, and $\epsilon = 1e^{-6}$. Beneficial and malignant shifts are seen in the setting of Theorem 2.10 with $\alpha = 2.0, \beta = 0.1$. That is, we see *two* cross-over points: one in the underparameterized regime and one in the overparameterized regime (going from mild to severe). This suggests that non-negligible covariance tail effects are a property of shifting when a model is in a region around the double descent peak. The further we are from the double descent peak, the more "classical" our behavior is, in that the top k components are the only ones that influence shift and the bottom $p - k$ components either don't exist or have negligible effects. Appendix A.10 explores this setup for different values of α, β and interpolating linear models for eigendecay rates that lead to harmful interpolation, i.e., *tempered* or *catastrophic* overfitting [80].

Figs. 2.2 and A.5 show similar results for 3-layer dense ReLU neural networks trained until near-interpolation (train MSE $< 5e^{-6}$) on synthetic data with benign overfitting eigendecay



(a) Noise then Blur Covariance

(b) Noise then Blur $n = 500$

(c) Noise then Blur $n = 2k$

(d) Blur then Noise Covariance

(e) Blur then Noise $n = 500$

(f) Blur then Noise $n = 2k$

Figure 2.3: We experiment with a custom variant of CIFAR-10C in which we apply the blur and noise image filters directly to the test set images of CIFAR-10 at each severity level, e.g., Severity 1 means that we add a small amount of noise *and* a small amount of blurring to the image. In the top row we first use the noise filter and then the blur filter. In the bottom row we first use the blur filter and then the noise filter. In (a) and (d), we observe that the eigenvalue decay of the shifts are non-monotonic and mirror the $\alpha < 1, \beta > 1$ setting in our taxonomy. Indeed, we also see in (b) and (e) that when we are severely overparameterized the noisy tail effects appear to be suppressed and we still obtain beneficial shifts. On the other hand, in (c) and (f) we are in the mildly overparameterized regime and observe that the noisy tail effects hurt generalization, even for severity 4 in the top row which only adds a small amount of noise in the tail. These results are exactly in keeping with our taxonomy for the $\alpha < 1, \beta > 1$ case. All curves are averaged over 50 independent runs.

rates [6]. For the neural network, we consider p to be the dimension of the input data, rather than the number of network parameters. From Fig. 2.2 we observe similar trends predicted by our theory for beneficial and malignant shifts when $p > n$, indicating that while our theory is developed for linear models we are able to extrapolate to more complex models. In both the $p > n$ and $p < n$ experiments, our results are agnostic to the hidden width of the network, further suggesting that overparameterization is qualitatively different from high-dimensionality. When $p > n$, a neural network appears to act like the interpolating MNI under distribution shifts. For $p < n$ the interpolating dense net does not exhibit the

properties of an interpolating MNI under distribution shift and the ID excess error is better than both "beneficial" and "malignant" OOD excess errors. However, the relative difference between beneficial and malignant shifts is still preserved. Note that we observe the exact same behavior in Fig. A.9 when training ResNets to interpolation on CIFAR-10 and testing on CIFAR-10C blur and noise corruptions [46].

CIFAR-10 Experiments

Next, we consider experiments with linear interpolators on a binarized CIFAR-10 and CIFAR-10C with Gaussian noise and blur corruptions at varying levels of corruption severity. For details, see Appendix A.9. This experiment breaks the assumption of simultaneous diagonalizability, and the well-specified assumption as the labels for CIFAR are not obtained by a ground-truth linear model.

Fig. A.6 shows empirical results on the MNI fit to this problem. We plot the eigenspectra of the blur and noise covariances from CIFAR-10C compared to the eigenspectra of CIFAR-10 in Figs. A.6a, A.6b. We identified these two shifts due to their eigenspectra reflecting what we expect to lead to beneficial and malignant shifts based on Theorem 2.10. We observe a tight relationship between changes in the eigenspectra of the target data and excess classification error when evaluated by the MNI. We notice that blurs reduce covariance energy with increased blurring, like a "denoising"-style operation. Experimentally this leads to improved OOD accuracy. In contrast, noise corruptions add energy to the covariance tail and lead to worsened OOD accuracy. Fig. A.8 also shows that further overparameterization in this setting leads to improved behavior of the MNI on both corruptions.

Fig. 2.3 extends these results to a more realistic setting with custom variants of CIFAR-10C that involves applying both noise and blur filters on test set images. Using both blur and noise filters together leads to covariate shifts that feature non-monotonic behaviors when comparing source to target eigenvalues. This experiment highlights the $\alpha < 1, \beta > 1$ case and we see that the OOD accuracy matches the predictions of our taxonomy based on whether we are in the mildly overparameterized regime ($n = 500$) or the severely overparameterized regime ($n = 2k$). In Fig. A.7 we show plots for an artificially constructed version of this same experiment in which Gaussian noise is injected into the high variance directions of the blur test set, simulating the equivalent of $\alpha > 1, \beta < 1$ in our taxonomy. We similarly find the OOD accuracy to match the predictions of our taxonomy.

2.5 Conclusion and Future Work

Our work provides the first finite-sample, instance-wise analysis of the MNI under transfer learning with high-dimensional linear models. We show a taxonomy of beneficial and malignant covariate shifts depending on whether we are in a *mildly* or *severely* overparameterized regime. In the mildly overparameterized regime, variance contributions on the top k components interact with that of the bottom $p - k$ components in non-negligible ways, leading to non-

standard shifts. In the severely overparameterized regime, the high-rank covariance tail suppresses variance contributions in the bottom $p-k$ components and so OOD generalization acts more "classical", akin to underparameterized linear regression where $k = p < n$.

Benign overfitting literature commonly claims to be motivated by "overparameterized" neural networks, referring to the number of parameters in the network rather than the dimension of the data. However, recent works have challenged this, suggesting the role of the ambient dimension and source covariance is more important than parameter count in determining whether overfitting is benign or catastrophic in neural networks [33, 61]. Prior work has also shown that gradient descent on 2-layer neural networks has an implicit bias towards linear decision boundaries when $p \gg n$, independent of the degree of overparameterization [35].

Our experiments further support the view that high-dimensional neural networks behave similarly to high-dimensional linear models, whereas low-dimensional neural networks do not. They provide a new and important perspective on the difference between high-dimensionality and overparameterization in neural networks in the case of distribution shift, which has yet to be appreciated in the literature. While dimensionality and degree of overparameterization are inextricably linked in linear regression, practical deep learning tends to operate in the overparameterized setting, not the high-dimensional one.

An important future direction is to investigate the extent to which our results hold for distribution shifts on more complex high-dimensional datasets. It is also of interest to extend our finite-sample theoretical analysis to shallow ReLU neural networks, other nonlinear models, and learning algorithms that overfit in a *tempered* manner [80]. Finally, future work might seek to extend our understanding of neural networks by carefully studying the interplay between the data dimension, number of network parameters, number of training samples, and the optimization algorithm and loss function, and how this interplay can affect ID & OOD generalization.

Another important future direction is to relax key assumptions in this work. These results rely on a simultaneous diagonalizability assumption on the source and target covariance matrices which frequently appears in related works on high-dimensional linear regression [64, 51, 65]. This enables us to highlight the different effects of covariate shifts in the "signal" components vs. in the "noise" components. In contrast, prior works (including those which can tolerate target covariances which do not satisfy simultaneous-diagonalizability [126, 64]) all rely on averages or traces over matrices involving the target covariance. Exploring techniques that can effectively handle non-simultaneously diagonalizable covariance matrices while preserving the insights gained from separating the impact of covariate shifts in signal and noise components is a promising avenue for future work.

Chapter 3

Sanity Checks for Agentic Data Science

3.1 Introduction

Agentic data-science (ADS) pipelines are becoming increasingly powerful [41, 90], driven in part by the rapid spread of autonomous systems such as OpenAI Codex and Claude Code. These pipelines are capable of taking as input a natural-language query and a relevant dataset and executing the full analytical workflow to produce an answer, enabling end-to-end "vibe data science." Any user can obtain a polished analysis, complete with model selection, hypothesis tests, and a written summary. However, the conclusions produced by ADS pipelines can be wrong in ways that are difficult to detect. An agent may make an unreasonable data-cleaning decision, choose an inappropriate model, or overlook confounders, but the resulting report will still appear polished and defensible [13]. Worse, for difficult datasets, a handful of runs can yield an entirely different conclusion than a full stability analysis, underscoring the need for systematic checks (Fig. 3.3). Thus, the rapid adoption of ADS pipelines poses an urgent risk of flooding science with unreliable findings [123].

Existing solutions to tackle this problem have limitations. Classical tests of statistical significance generally fail in agentic settings. For example, we observe that agents handle statistical evidence poorly in practice, over-interpreting estimates that lack significance and "p-hacking" by anchoring on whichever test is most favorable (Sec. 3.4). Alternatively, simply asking an agent to assess the uncertainty of an analysis also does not succeed, as agent-reported confidence is poorly calibrated to the actual stability of an underlying result (Sec. 3.5). Finally, "multiverse" analyses enable quantifying the impact of different choices in an ADS pipeline, but their outputs quickly become too large for human review, creating a need for actionable checks [13, 102].

To tackle this challenge, we make use of the Predictability-Computability-Stability (PCS) framework for veridical data science [143, 142, 101]. It establishes that trustworthy, reality-checked conclusions must be stable under reasonable perturbations across different data

science stages, such as problem specification, data cleaning, and model choices. In this spirit, we generate small, reasonable changes, which we call *PCS perturbations*. These perturbations probe agentic failure modes which would not impact a competent human analyst's conclusions. If an agent's conclusions change dramatically under these perturbations, they are not trustworthy.

Consistency under these perturbations is necessary but not sufficient, since an agent could also arrive at a stable but biased conclusion. Since ground-truth conclusions are rarely known in data science, we use a *null-defining perturbation* to put the agent in a setting where the answer is known and testable. For example, we may remove all the signal in the data (e.g., independently shuffling each column) to establish a negative control against which we can measure the agent's capabilities. We then run the agent many times to generate two empirical distributions: one where both null-defining and PCS perturbations are applied, which we call the *null distribution*, and another where only PCS perturbations are applied, which we call the *alternative distribution*. Using these two distributions, we apply two complementary sanity checks (see Fig. 3.1):

(1) *The Yes check*: a one-sample bootstrap test of whether the ADS mean response on the original data is significantly above the midpoint of the scale (i.e., in the "Yes" range), thereby testing whether a positive conclusion was reached by chance.

(2) *The Overlap check*: quantifies the similarity between the null and alternative distributions, thereby testing for bias in the ADS conclusion that does not come from signal in the data.

In experiments using OpenAI Codex, we first validate the approach on simulations with controlled signal-to-noise ratios, confirming that the checks track ground-truth signal strength (Sec. 3.4). We then apply them to eleven real-world datasets from the BLADE benchmark, where the checks characterize each dataset's signal regime and reveal one case in which the agent's affirmative conclusion is not well-separated from its behavior on pure noise (Sec. 3.4). We further analyze failure modes of ADS systems (Sec. 3.5), showing that ADS self-reported confidence is poorly calibrated to the empirical stability of its conclusions and characterizing how different steps fail using the BLADE agentic harness (an alternative to Codex). We view this work as a first step toward evaluation standards for agentic data science, grounded in the core principles of the PCS framework for veridical data science.

3.2 Related Work

Agentic data science (ADS). The risks of AI automation in the scientific workflows have been documented in fields adjacent to data science, including LLM-hacking for annotation [9, 127] and broader stages of the research pipeline [77]. More directly relevant, Asher et al. [3] demonstrates that guardrails are necessary to prevent humans from p-hacking using LLM-driven data science, while other work highlights the value of transparency in intermediate data-science artifacts as a partial remedy [115, 90]. Despite these warning signs, much of the ADS literature has focused on capability and benchmarking, including evaluation suites

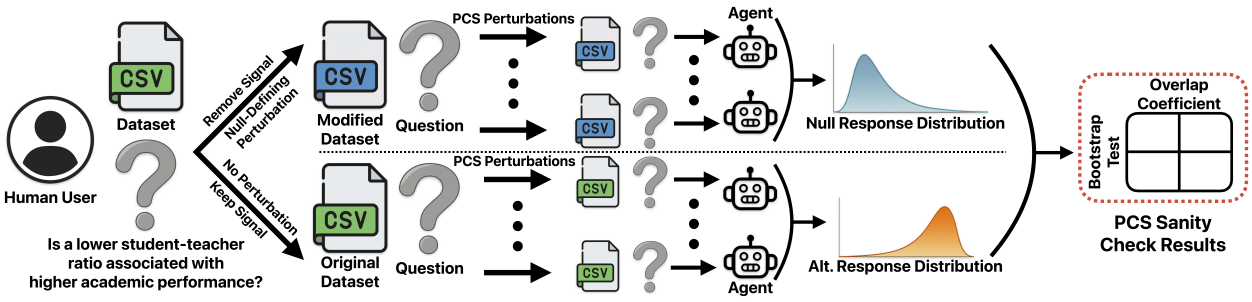


Figure 3.1: Pipeline for computing the PCS sanity checks. The user supplies a dataset and question. The pipeline then generates perturbed variants of these under both null-defining and PCS perturbations. The coding agent analyzes each independently, producing null and alternative response distributions. A bootstrap test & overlap coefficient then inform whether the agent found a data-grounded, stable positive conclusion.

such as DSGym [91], ScienceAgentBench [18], IDA-Bench [67] and others [118, 69], alongside accompanying agentic frameworks [41]. Comparatively, little attention has been paid to whether ADS conclusions can be trusted, motivating the need for intuitive and actionable sanity checks.

Perturbation-based analysis. The PCS framework [143, 142] and its use of generative AI [101] has been shown to work well in solving scientific problems [8, 1, 133, 125, 136, 2] and generalizes various perturbations or viewpoints common in the literature, e.g., the multiverse analysis [119, 23], sensitivity analysis [20], and others [31], along with accompanying software [74, 104, 25]. Recent work has proposed the promising use of multiverse analyses for reporting ADS conclusions [13], but the resulting multiverse is often too large for human review, and thus its nature is more descriptive than actionable. This creates a need for actionable sanity checks which evaluate conclusion *stability*.

Uncertainty quantification for language models. Many frameworks for uncertainty quantification in traditional statistical learning have been configured for language model applications. This includes conformal inference [98, 92], confidence intervals [137], and Bayesian approaches [103]. It also includes more holistic approaches, such as automatic evaluation of an LLM-generated plan [108] and participatory refinement [24]. However, these approaches have not been tailored to data science, which contains a compounding variety of judgment calls made by the user [36, 143].

3.3 Methods

We consider the setting where a user supplies an agentic system with a dataset and a yes or no question. This setup can produce qualitative insights for a wide range of questions.

Generating response distributions using perturbations

As discussed in Sec. 3.1, PCS posits that trustworthy conclusions must be stable under reasonable perturbations to the data and the problem formulation. We define two perturbation types: one that modifies the data's signal, and one that leaves it unchanged. Combined, they provide a useful sanity check on the trustworthiness of an agent's conclusion.

Null-defining perturbations. Evaluating ADS conclusions is difficult without ground truth. We address this by examining responses in settings where the answer is known by removing signal from the data, similar to negative control methods [109]. By breaking the association between features and outcomes, any inferred relationship should disappear, creating a distribution of conclusions under known null conditions. We call such modifications "null-defining perturbations". Examples include shuffling feature values independently or replacing outcomes with random draws. Running the agent on the perturbed data yields the empirical *null distribution*. If no null-defining perturbation is applied, we call the resulting distribution the *alternative distribution*.

Probing failure modes with PCS perturbations. To understand the response variation, we probe ADS failure modes using PCS-inspired perturbations to metadata or prompts. These "PCS perturbations" (Table 3.1) leave data signal unchanged and should not affect experienced data scientists. We apply these PCS perturbations when computing both for the null and alternative distributions to assess stability.

Eliciting continuous responses. We prompt the agent to respond on a scale from 0 (strong "No") to 100 (strong "Yes"). While binary outputs still permit stability analysis, restricting the agent's output space discards information about the strength of its conviction.

Two PCS sanity checks

Given the null and alternative distributions, we want to answer two questions. First, does the agent consistently affirm the research question on the real data, or could its positive response be a fluke? Second, does the agent behave differently on real data than on noise, or does it produce similar outputs regardless of whether signal is present? Each question targets a failure mode that the other misses.

***The Yes Check*. Is the agent consistently affirmative?** A single high score could be a fluke. We therefore test whether the agent's mean response on the original data consistently exceeds the scale midpoint (50) across repeated runs and perturbations. A mean above 50 indicates net agreement with the research question, while a mean at or below 50 suggests a lack of evidence for a "Yes" answer. Importantly, by aggregating over many runs, the test guards against drawing conclusions from a single lucky (or unlucky) pipeline execution. To formalize this, we test the one-sided hypothesis $H_0 \colon \mu_{\text{alt}} = 50$ against $H_1 \colon \mu_{\text{alt}} > 50$, where μ_{alt} is defined as the population mean, drawn from the alternative distribution. We draw B bootstrap samples from the observed alternative responses, computing the p-value as the proportion of bootstrap means that fall at or below 50, applying the Phipson and Smyth [96] correction to avoid p-values of exactly zero. We also report bootstrap percentile confidence

Table 3.1: Perturbations used in the PCS sanity checks. The null-defining distribution breaks associations between the features and the outcome. The PCS perturbations should not affect an experienced data scientist or agentic data science system, although some perturbations (e.g., *positive leading statement*) may mislead human data scientists due to confirmation bias.

Perturbation	**Type**	**Description**	**Failure Mode Tested**
Shuffle Feature Values	Null-Defining	Shuffles each column independently to break associations in the data.	Incorrect data analysis conclusions.
Add Non-Signal Features	PCS	Appends features irrelevant to & uncorrelated with the task of interest.	Distraction by irrelevant covariates.
Anonymize Feature Names	PCS	Replaces all column names with generic identifiers (e.g., `feature1`, `feature2`).	Reliance on semantic variable names from pre-training data.
Shuffle Feature Names	PCS	Applies a random permutation of existing column names while leaving data values and feature descriptions in metadata unchanged.	Robustness to misleading variable names.
Positive Leading Statement	PCS	Prepends the question with a strong prior belief that the answer is "Yes."	Sycophancy toward user-stated beliefs.
Negative Leading Statement	PCS	Prepends the question with a strong prior belief that the answer is "No."	Sycophancy toward user-stated beliefs.

intervals, providing a range of plausible average responses. This check treats responses across PCS perturbation types as identically distributed. We verify this in Appendix B.1.

The Overlap check. **Can the agent distinguish signal from noise?** An agent's average response could be well above 50, yet still produce a response distribution that overlaps heavily with the null, suggesting it cannot distinguish signal from noise. To capture this, we compute the overlap coefficient (OVL) between the alternative & null densities (f_{alt} & f_{null}):

$$\mathrm{OVL} = \int \min\big(f_{\mathrm{alt}}(x),\ f_{\mathrm{null}}(x)\big)\, dx. \tag{3.1}$$

We estimate densities via Gaussian kernel density estimation with Scott's rule [107] for bandwidth selection (the default in most standard scientific computing libraries), and the integral is evaluated numerically over $[0, 100]$. An OVL near 0 indicates well-separated distributions, while an OVL near 1 indicates that the agent cannot tell real data from noise.

Table 3.2: Interpretations of the four PCS sanity-check regimes.

	***Overlap check* passed ✓**	***Overlap check* failed ✗**
***Yes check* passed ✓**	The positive conclusion is stable.	The positive conclusion may not be grounded in the data.
***Yes check* failed ✗**	There is some positive signal, but not enough for a "Yes".	No evidence supporting a positive conclusion.

Interpreting the PCS sanity checks. Combining both checks produces a 2×2 classification of each dataset, summarized in Table 3.2. We can binarize each sanity check using a threshold, e.g., for the *Yes check* we assess whether the bootstrap p-value is below significance level α, and for the *Overlap check* whether OVL $< \tau$ for a user-chosen threshold τ. Note from Table 3.2 that we obtain different information when one check fails but the other succeeds.

3.4 Results

In this section, we demonstrate that the PCS sanity checks effectively measure simulated signal in the data (Sec. 3.4) and report results for each of the BLADE datasets (Sec. 3.4).

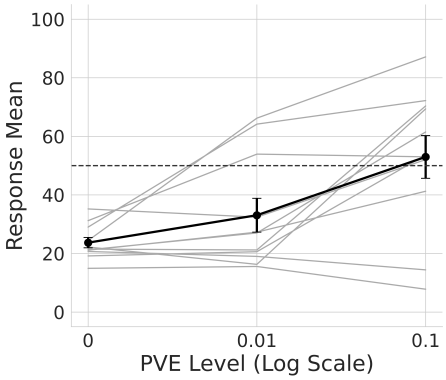
Experimental Setup

Datasets. We evaluate our approach on eleven datasets from the BLADE benchmark [40], a collection of tabular datasets paired with binary yes-no questions. Further details regarding the datasets and their corresponding tasks are provided in Appendix B.2.

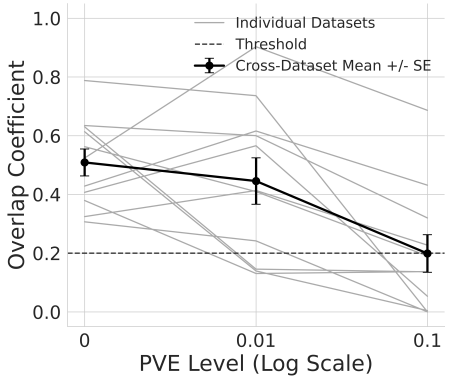
Perturbations. To define the null distribution, we break all data associations by shuffling feature values independently. We apply five PCS perturbations (Table 3.1) 20 times each, creating 100 observations for both the null and alternative distributions per dataset.

Agent & prompt. We use GPT-5.2-Codex as our agent, set to "high" reasoning effort. Prompted as "an expert data scientist", the agent provides: (1) an integer Likert score, where 0 indicates strong "No" and 100 indicates strong "Yes", and (2) a text explanation. The exact prompt is in Appendix B.3. We also use the BLADE agentic system [40] for a step-level analysis of where instability enters (Sec. 3.5).

Sanity check details. The bootstrap test (level $\alpha = 0.05$) evaluates if the alternative mean exceeds 50 (i.e., is in the "Yes" range) using $B = 10{,}000$ resamples. The overlap coefficient threshold separating low from high distributional overlap is set to $\tau = 0.2$.



(a) The agent's response increases as we add signal to the data. The two datasets that decrease, `affairs` and `reading`, ask directional questions, with the signal going in the opposite direction.



(b) The higher levels of signal in the data cause the agent to respond more positively. This creates lower overlap with the null distribution, since it does not contain signal.

Figure 3.2: The PCS sanity checks behave as expected. As more signal is introduced to the data, the agent's responses become more positive, leading to less similarity with the null.

PCS Sanity Checks Correlate with Signal Strength in Simulations

Lacking ground-truth for real-world data, we use simulations to determine the signal needed for an agent to distinguish it from noise, using the signal-to-noise ratio as a tuning knob.

Setup. For each BLADE dataset, we identify the dependent variable Y and (one-hot encoded) independent variables X, which includes an intercept. We regress Y on X to fit an OLS model $\widehat{Y} = X\beta$. We then construct a synthetic dependent variable $Z = \widehat{Y} + \varepsilon$, where the noise variance is calibrated so that the signal explains a target fraction of the total variance. Specifically, we set $\sigma_\varepsilon = \sqrt{\mathrm{Var}(\widehat{Y}) \cdot \frac{1-\mathrm{PVE}}{\mathrm{PVE}}}$, where $\mathrm{PVE} = \mathrm{Var}\left(\mathbb{E}[Y \mid X]\right)/\mathrm{Var}(Y)$ represents the proportion of variance explained. At the boundary, $\mathrm{PVE} = 0$ represents pure noise ($Z \sim N(\bar{Y}, \sigma_Y)$), while $\mathrm{PVE} = 1$ represents pure signal ($Z = \widehat{Y}$). Responses on data with $\mathrm{PVE} \in \{0.0, 0.01, 0.1\}$ serve as the alternative group. The null group is as characterized in Sec. 3.4. The PCS perturbations from Table 3.1 are applied five times each, creating an alternative distribution with 25 observations.

Results. We present the results of the two stability checks in Fig. 3.2. We see in Fig. 3.2a that the agent's responses become more positive as we introduce signal into the data. These responses then cause a decrease in overlap with the null distribution, as shown in Fig. 3.2b.

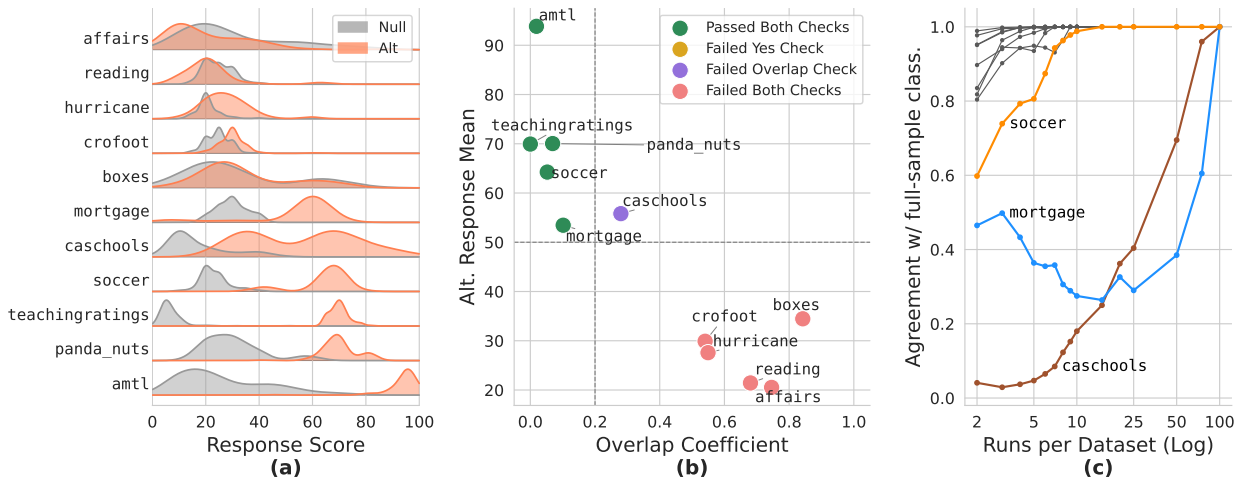


Figure 3.3: Sanity check results on BLADE datasets. **(a)** KDEs of the null and alternative response distributions for each dataset. **(b)** Each dataset's OVL and alternative response mean, colored by category (Table 3.2). Vertical and horizontal dashed lines mark the OVL decision boundary and the scale midpoint, respectively. **(c)** Agreement between subsampled and full-sample classifications as a function of the number of alternative runs (Appendix B.5).

Main result: Sanity checks on benchmark data highlight unreliable conclusions

We evaluate the PCS sanity checks on 11 BLADE datasets, following the experimental setup described in Sec. 3.4. Summary statistics for the distributions and their corresponding PCS sanity check results are reported in Table B.3. Fig. 3.3 shows the full score distributions, each dataset's classification according to Table 3.2, and a convergence analysis showing how classification stability varies with the number of runs.

Datasets passing the checks. 5 datasets pass both sanity checks: `amtl`, `panda_nuts`, `soccer`, `teachingratings`, and `mortgage`. Their bootstrap p-values are well below $\alpha = 0.05$ with overlap coefficients below $\tau = 0.2$, indicating that the agent consistently affirms the research question in a manner clearly distinguishable from noise. The effect sizes in this group, however, span a wide range. While `teachingratings` shows near-perfect separation (OVL ≈ 0), `mortgage` shows a more moderate separation (OVL $= 0.102$), suggesting more variable confidence levels. These five cases represent settings where a practitioner can place reasonable trust in the agent's affirmative conclusion.

Datasets with no signal. Five datasets fail both sanity checks: `affairs`, `boxes`, `crofoot`, `hurricane`, and `reading`. These datasets have bootstrap p-values at or near 1.0, since the one-sided test observes that the agent's mean response on the real data is in the "No" range (<50). High OVLs for `boxes` (0.842) and `affairs` (0.746) confirm that the agent behaves similarly on real data and noise. The overlap coefficients reinforce this picture.

Moderate OVLs for `crofoot` (0.54) and `hurricane` (0.55) suggest the agent may detect a faint difference, but not enough to push it toward an affirmative conclusion. These cases are the least ambiguous, as the agent does not find compelling evidence for a "Yes" response.

An analysis failure. One dataset passes the *Yes check* but fails the *Overlap check*: `caschools`. Although the agent leans affirmative ($55.8, p = 0.002$), the overlap coefficient of 0.280 exceeds the threshold. The large standard deviation of the alternative responses (19.9) reflects wide confidence swings. In this case, relying on the mean response alone would give a misleading sense of reliability. The sanity check reveals this by requiring the agent not only to respond "Yes", but to do so in a manner distinguishable from its behavior on noise.

None of the eleven datasets fail only the *Yes check*. This is unsurprising, as benchmark datasets are typically chosen to exhibit a clear answer, making weak-but-detectable effects unlikely to appear. However, it does occur in our simulation experiments at intermediate signal strengths (Sec. 3.5, Appendix B.4), confirming the framework can detect this regime.

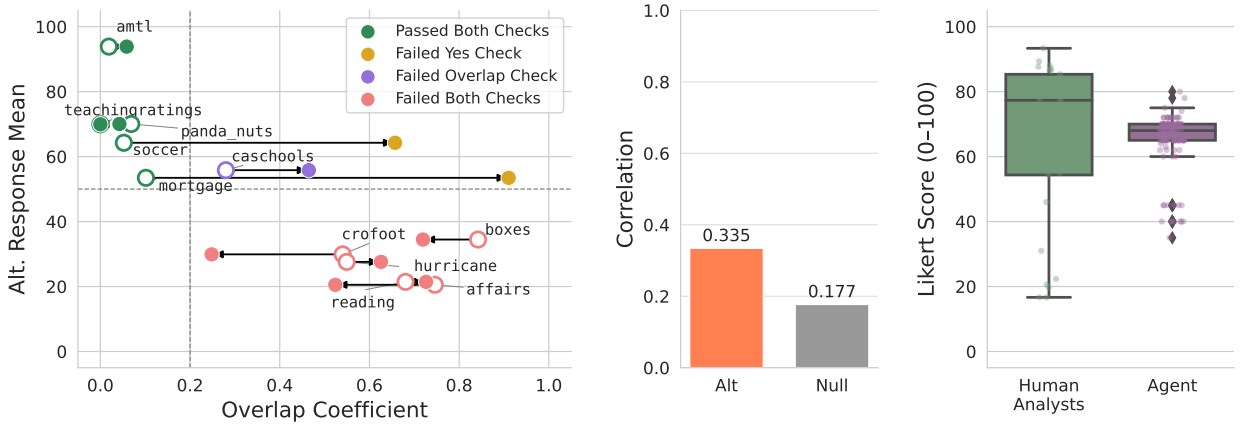
Comparing the two checks. The value of the two-dimensional classification becomes clear when contrasting adjacent cases. For example, `mortgage` and `caschools` have similar alternative means (53.5 vs. 55.8) and significant p-values. To a test that only examines response significance, these two datasets would look comparable. The overlap coefficient is needed to reveal the difference: `mortgage` is grounded (OVL= 0.102), while `caschools` overlaps heavily with its null (OVL= 0.280). This contrast confirms that both PCS sanity checks are needed; the bootstrap test ensures the agent is saying something affirmative, and the overlap coefficient ensures that the affirmation is distinguishable from behavior on noise.

How many runs are needed? Fig. 3.3c shows that most classifications stabilize within five runs, but the three most ambiguous cases (`caschools`, `mortgage`, `soccer`) require far more, suggesting that a small number of runs may not suffice when the underlying question is difficult for the agent to resolve (Appendix B.5). When taking only a single run, we find that it is possible to obtain an affirmative result for all the datasets (Fig. 3.3a), despite the fact that 6 fail at least one of the sanity checks (Fig. 3.3b).

Why do null hypotheses sometimes give strong 'yes' answers? Even after breaking all associations in the dataset with the null-defining perturbation, the agent sometimes concludes that a strong relationship exists. Inspecting these analyses reveals two distinct failure modes, where the agent either: (1) misunderstands statistical significance, or (2) tries multiple statistical approaches and anchors its answer on the test which yields the lowest p-value. Examples of both failure modes are included in Appendix B.7.

3.5 Analysis

In this section, we investigate questions regarding failure modes of ADS pipelines.



(a) When PCS sanity checks are applied using a null with PVE=0.01, we see less "Passed both checks".

(b) Agent-reported confidence is only weakly correlated with empirical exceedance, especially under the null.

(c) The conclusions formed by human analysts exhibit greater variance than those from agents.

Figure 3.4: Multi-faceted analysis of agentic instability.

How Do Perturbations Affect Individual Data Analysis Steps?

To understand why PCS sanity checks are necessary, it helps to investigate where instability enters the agentic pipeline. By pinpointing the most volatile stages of analysis, we can better understand how small, reasonable perturbations can compound into untrustworthy conclusions. To do so, we use the BLADE harness [40] rather than treating the pipeline as a black box (as in Sec. 3.4). We decompose each analysis into four steps: independent variable selection, control variable selection, model specification, and conclusion. Using the PCS perturbations from Table 3.1 with five replicates, an LLM judge scored the similarity between perturbed and unperturbed outputs. Full details are in Appendix B.8.

A clear stability gradient emerges across the pipeline. Independent variable selection is the most stable, as research questions often dictate these choices. Volatility increases during control variable selection and model specification, and peaks at the conclusion step, where the agent weighs conflicting evidence and commits to a qualitative judgment. The perturbation which shuffles feature names tends to produce the largest deviations at the conclusion step. Notably, within-perturbation and baseline-vs-perturbation heatmaps are nearly identical in structure, suggesting that perturbations increase volatility without introducing specific biases. This compounding of instability reinforces the need for PCS sanity checks.

Evaluation with a More Precise Null Hypothesis

Throughout the paper, we choose the null-defining perturbation to be shuffling all feature values independently (Sec. 3.4). This is a useful default choice as it destroys all relationships in the data. However, a user may choose a different null-defining perturbation if it better suits their needs; here, we demonstrate one such choice.

Motivation. Practical applications often require a specific effect size to justify action. For example, in the `caschools` dataset, a school might need a minimum impact from student-to-teacher ratios to justify hiring costs. In such cases, a "no relationship" null is not contextually aligned; it is better to ensure the signal meaningfully exceeds a user-selected criterion. This allows the PCS sanity checks to act as a customizable filter for real-world decision-making.

Setup. We apply PCS sanity checks using distributions where the PVE of each dataset is set to 0.01, as described in Sec. 3.4. Since large datasets can yield statistically significant results even at PVE= 0.01, we modify our scaffolding: a "Failed the Overlap check" classification now requires the null mean to fall within the "No" range. This ensures that high-overlap distributions in the "Yes" range are correctly labeled "Failed the Yes check".

Results. Fig. 3.4a shows that requiring the explanatory variable to capture more than 1% of variance leads to fewer "Passed both checks" conclusions. Specifically, the `soccer` and `mortgage` datasets are classified as "Failed the Yes check" because their null means shifted into the "Yes" range. The full distributions are provided in Appendix B.6.

Can Agents Approximate the Uncertainty of Their Conclusions?

A naïve way to assess pipeline stability is to have an agent estimate uncertainty. If agents could accurately self-assess stability, empirical sanity checks might be redundant. Thus, we evaluate agentic self-assessment by having a supervisor agent review the responses and code from Sec. 3.4, finding that self-reported confidence is nearly uninformative.

Setup. The supervising agent provides a number akin to an exceedance probability. In particular, we ask: "If you were to reconduct this analysis 100 times with slightly different reasonable decisions in the data science pipeline, how many times would you expect to get an answer more positive (larger on the Likert scale) than seen in the conclusion?"

Results. If accurate, the agent's stated confidence should track the empirical exceedance rate, i.e., the actual fraction of peer runs that produced a higher response. Fig. 3.4b shows that under the alternative distribution, where real signal is present, confidence and empirical exceedance are only moderately correlated (Spearman $\rho = 0.34$). Under the null distribution, where no signal exists, the correlation drops to $\rho = 0.18$, making confidence scores nearly uninformative. In both cases, scores cluster between 30–70 rather than spanning the full scale, indicating that the agent hedges rather than providing calibrated estimates. This miscalibration peaks in the null setting, where practitioners most need to identify instability. Self-assessment is therefore not a substitute for empirical sanity checks.

Comparison to Human Variations

Our final question is whether this instability is a unique flaw of LLMs or an inherent challenge of data science. To contextualize agentic uncertainty, we examine variability in human-performed studies where independent analysts reach different conclusions on identical data. We focus on the dataset and task from Silberzahn et al. [111] (part of the BLADE benchmark), which asks if soccer players with darker skin tones receive more red cards. To compare human & agentic variation, three data scientists scored the 29 original human reports with the prompt from Sec. 3.4. We then averaged these scores for each report.

Results. Fig. 3.4c compares human- and agent-performed analyses (using the `soccer` alternative distribution). We observe that 76% of human and 86% of agent analyses yielded a "Yes" conclusion (score > 50). Notably, agentic conclusions are bimodal, with tight clusters around 70 (moderate "Yes") and 40 (weak "No"), while human conclusions are more uniformly spread. Inter-rater agreement and further details are in Appendix B.9. The high human variation suggests that the 'problem' isn't just the LLM; it is the inherent ambiguity of the data science task. Thus, as LLMs continue to scale and inherit these human-like failure modes, PCS sanity checks will be essential for assessing conclusion trustworthiness.

3.6 Discussion

Limitations. While they are diverse, the BLADE datasets we study here are already present in the literature and may have been seen by LLMs during pretraining, potentially biasing some of their responses. Our evaluation also largely relies on a single null-defining perturbation and one primary agent (OpenAI Codex), due to computational cost constraints. These same cost constraints may also limit casual use of the sanity checks here, though these concerns may be alleviated over time as the cost of LLM calls continues to decrease.

Future work. This work opens many avenues for future research. Some promising directions include exploring null-defining perturbations which preserve realistic covariate structure, early stopping rules to reduce the number of agent calls required, and evaluation across a broader set of agents and benchmarks. A natural future direction is automating the perturbation approach itself, the data science analog of moving from fault analysis to site reliability engineering. The sanity checks here could also be generalized beyond data science to other settings where perturbations can be defined. For example, some fitting broader scopes are the more general problems of generating novel hypotheses [140, 110], conducting human-in-the-loop data science [38, 105, 122, 29], building better agentic tools for explaining data [145, 114, 106], and fully autonomous science [76, 124, 87].

Chapter 4

Pediatric FAST under Data Scarcity: A Deep Learning Pipeline for H-IAI Detection

4.1 Introduction

Traumatic injury is the leading cause of death in children in the United States, and hemorrhage (bleeding) from intra-abdominal injury (H-IAI) following blunt trauma accounts for roughly 30% of pediatric trauma deaths [135, 52]. The reference standard for diagnosis of H-IAI in the emergency department (ED) is the computed tomography scan (CT) [84], which has two major drawbacks in children. The first is ionizing radiation, which increases the lifetime risk of radiation-induced cancer in children [85, 117]. The second is time and resource use. CT requires transport out of the trauma bay and coordination with radiology, which delays the evaluation and therapeutic treatment of injured children. Despite a substantial rise in pediatric CT use for injured children, there is not a proportional decrease in mortality in children after blunt trauma [57, 15]. This suggests that many CT scans performed on injured children may not be necessary.

Focused assessment with sonography for trauma (FAST) is a point-of-care ultrasound examination used to detect H-IAI at the bedside within several minutes of the child's arrival to the ED. It is radiation-free, inexpensive, portable, and performed by the treating clinician without requiring a radiologist or ultrasound technician. FAST has been shown in injured adults to decrease time to operative care [83]. The FAST is a protocolized examination that evaluates a series of anatomic landmarks. The pediatric FAST protocol has been standardized through expert consensus [59], and Morison's pouch, the potential space between the liver and right kidney in the right upper quadrant, is one of the primary anatomic landmarks because free fluid can accumulate there in the supine patient. Despite the promise of FAST in improving care, many clinicians remain under-trained in pediatric FAST, resulting in significant diagnostic inaccuracy and unreliability [70, 120, 39, 146]. Hence, many clinicians

are reluctant to act on FAST results, and opt to obtain CT scans instead.

A clinician performing a FAST must complete two operator-dependent tasks, each of which is difficult for clinicians without substantial pediatric FAST experience. The first is anatomic landmark acquisition: positioning the ultrasound probe to produce a view in which the relevant anatomy is clearly visible. The second is interpretation: judging whether H-IAI is present within that view. Diagnostic accuracy of pediatric FAST is variable across operators, and training curves are long, with skill continuing to improve after tens or hundreds of supervised examinations [39, 146]. A recent meta-analysis found pediatric FAST sensitivity for intra-abdominal injury may be insufficient to support the examination as a standalone rule-out test in children [70]. However, when FAST is used in a diagnostic strategy, using physical examination findings or laboratory testing, it may be able to reduce unnecessary CT scan use [60].

Both tasks, anatomic landmark identification and H-IAI interpretation, are computer vision problems of the kind that deep learning has handled well in adjacent medical imaging domains. We develop two models for these tasks, a segmentation model that localizes Morison's pouch on a single ultrasound frame and a classification model that scores the same frame for H-IAI. The two models are trained and deployed independently, and mirror the two-step clinical reading of a FAST frame. A FAST examination is not complete until each anatomical landmark has been identified and examined, so localizing the pouch is not merely an auxiliary to classification but a requirement of the FAST itself.

Existing machine learning (ML) work on FAST is divided by task and by population. For classification of H-IAI, a 2021 study evaluated FAST on a balanced adult cohort of 324 patients and reported frame-level sensitivity and specificity of 0.985 and 0.913 on a held-out test set [19]. The same study also trained a separate binary classifier that scored each frame as qualified or non-qualified based on whether at least one-third of the right kidney was visible. Another adult FAST study trained an automated machine learning pipeline on 20 exams and directly segmented H-IAI when present [116]. That study used image-level cross-validation rather than exam-level splitting, so frames from the same exam could contribute to both training and evaluation folds. Both studies focused on adult FAST and did not evaluate performance in injured children.

For localization, prior FAST ML work has focused on adjacent organs or H-IAI itself rather than direct localization of Morison's pouch. In [72], a U-Net was trained to segment the liver and right kidney on adult FAST, using adjacent organs as a proxy for the pouch. In [66], hemoperitoneum was localized with YOLOv3 bounding boxes in adult right-upper-quadrant FAST examinations. These studies leave two gaps that are central to pediatric FAST decision support. First, pathology-based localization is not available on negative exams, which constitute most pediatric trauma evaluations. Second, organ or pathology segmentation does not directly answer whether the clinician has localized Morison's pouch, the landmark needed to perform the right upper quadrant FAST view.

Pediatric FAST has received less attention in the ML community. However, early work in this area shows promise. The approach in [58] trained a classifier to identify which of the four standard FAST views a given frame depicts, establishing feasibility for deep learning on

pediatric FAST but addressing a different task than ours. To our knowledge, no prior work has built an H-IAI classifier or a landmark segmenter for pediatric FAST.

The PECARN clinical decision rule ([47, 48]) is used in clinical practice for pediatric H-IAI risk stratification and achieves 100% sensitivity and 47.2% specificity on pediatric blunt trauma using tabular clinical inputs such as mechanism of injury, vital signs, and examination findings. The PECARN CDR and our imaging models operate on disjoint inputs, since the CDR uses structured clinical observations and our models use only ultrasound frames.

We frame this work within the predictability, computability, and stability (PCS) framework for veridical data science [143, 142, 101]. PCS shapes the methodology by requiring the modeling pipeline to be explicit, reproducible, and checked for stability across the choices most likely to affect the conclusions. In this study, those choices include exam-level splitting, repeated resampling, exam-level aggregation, and threshold selection under positive-label scarcity.

This chapter makes three main contributions.

- **Direct segmentation of the FAST anatomic landmark, Morison's pouch.** We develop a segmentation model that targets Morison's pouch itself, rather than an adjacent organ or the fluid when present. We evaluate the model using both pixel-overlap metrics, benchmarked against a clinician inter-rater reference, and a blinded clinician review of practical localization usefulness.
- **An H-IAI classifier for pediatric FAST.** We develop a frame-level classifier for H-IAI detection on pediatric FAST ultrasound. The classifier is trained independently from the segmentation model and produces exam-level predictions through aggregation of frame-level scores.
- **A PCS-guided evaluation protocol calibrated to the data-scarcity regime.** The protocol combines Monte Carlo cross-validation, exam-level bootstrap confidence intervals, and a procedure for selecting the operating threshold when positive counts are low. It is applicable to other medical imaging tasks with similar label scarcity.

4.2 Materials and Methods

Study design and data sources

We conducted a retrospective analysis of 207 pediatric FAST ultrasound exams from UCSF Benioff Children's Hospital Oakland and UCSF Benioff Children's Hospital San Francisco, along with a separate adult cohort of 123 exams from Zuckerberg San Francisco General Hospital used only for model pretraining. The pediatric cohort included 17 H-IAI-positive exams, reflecting the rare-event nature of pediatric FAST evaluation. All data were obtained

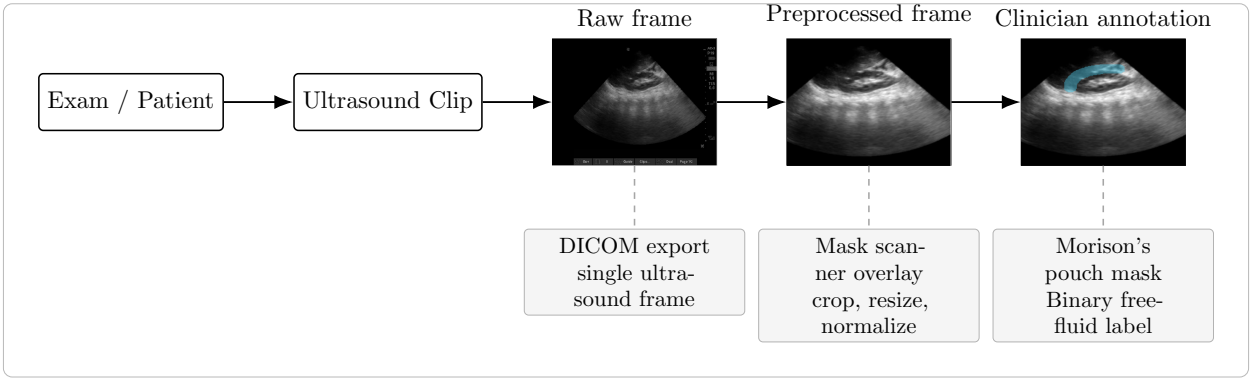


Figure 4.1: Dataset processing workflow. Each pediatric FAST study consists of one or more ultrasound clips, which are decomposed into raw frames. Frames are preprocessed by masking scanner overlays, cropping, resizing, and normalizing. Clinicians then annotate visible Morison's pouch frames with a segmentation mask and binary free-fluid label.

from hospital quality assurance archives under an Institutional Review Board-approved protocol with waived informed consent, and were fully de-identified prior to analysis. Pediatric exams were acquired on Sonosite Edge II and X-Porte point-of-care ultrasound systems as part of routine trauma evaluation in the ED, and exported as DICOM.

Dataset and annotation

Each FAST exam is organized hierarchically. An exam corresponds to one trauma evaluation of one patient and contains one or more ultrasound clips, with each clip composed of individual frames. The exam is the independent sampling unit throughout. All train and test splits operate at the exam level, so frames from a given exam never appear on both sides of a split.

Annotations were produced on the MD.ai platform (MD.ai, New York, NY, USA) by pediatric emergency medicine clinicians experienced in point-of-care ultrasound. For each exam, the annotator reviewed all clips, identified frames in which Morison's pouch was visible, and applied two labels to a subset of such frames, (1) a pixel-wise segmentation mask of the pouch and (2) a binary indicator of H-IAI. Frames in which Morison's pouch was not visible were not annotated. For some exams, a clinician identified the first and last frame of a clip in which the pouch was visible and a trained non-clinician annotator interpolated labels on the intervening frames. This extends coverage without sacrificing expert oversight, since the clinical judgments of pouch visibility and fluid presence are made by the clinician at the endpoints.

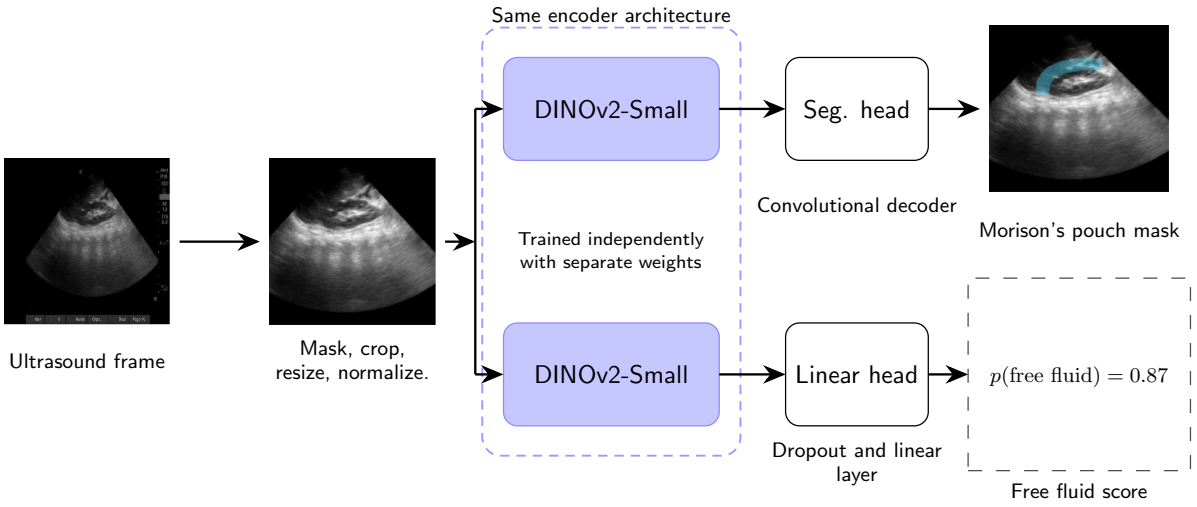


Figure 4.2: **Training Pipeline.** A raw ultrasound frame passes through the fixed preprocessing pipeline and is then processed by a DINOv2-Small encoder. The segmentation and classification models use the same encoder architecture but are trained independently, with separate weights for each task. The segmentation model uses a convolutional decoder to produce a Morison's pouch mask. The classification model uses a linear head to produce a free fluid score.

Two-step pipeline and training

Our system addresses the two operator-dependent tasks of pediatric FAST interpretation with two deep learning models, a segmentation model that localizes Morison's pouch and a classification model that detects H-IAI. Both models take a single raw FAST frame as input and produce their outputs in parallel. One alternative is a sequential design in which the segmentation output is used to restrict the classifier input to the predicted Morison's pouch region. We evaluated this variant during development and found that restricting the classifier's field of view to the predicted Morison's pouch region did not improve performance over full-frame classification, so the two models remained independent.

Both models use DINOv2-Small, a ViT-S/14 encoder with 22.1M parameters, as the frame encoder [94]. For both tasks, the encoder was initialized from the publicly released facebook/dinov2-small checkpoint and trained independently for segmentation and classification. We selected the Small variant of the DINOv2 family because the pediatric fine-tuning cohort contains too few positive exams to support a larger parameter count without overfitting. The only architectural difference between the two models is the task-specific head.

Segmentation model (Morison's Pouch). A lightweight convolutional decoder upsamples the backbone's patch embeddings to a 224x224 single-channel probability map of landmark membership. During inference, a predicted mask is produced independently for each frame. The loss is an equally weighted sum of Dice loss with Laplace smoothing and

binary cross-entropy. Full decoder architecture is given in Sec. C.1.

Classification model (H-IAI). The classification head is a dropout layer followed by a linear projection to a scalar logit, trained with binary cross-entropy. At inference time, the classifier produces an independent score for every frame in an exam.

Preprocessing and augmentation. Raw frames are masked and cropped to remove scanner overlay and GUI artifacts, resized to 224x224 via bilinear interpolation, replicated from one grayscale channel to three, and normalized with ImageNet statistics. Training augmentation consists of horizontal flips, small shift-scale-rotate perturbations, and mild brightness and contrast jitter. No augmentation is applied at inference. The same preprocessing and augmentation pipeline is used for both models and for both adult pretraining and pediatric fine-tuning. Full details are in Sec. C.1.

Training procedure. Each model is trained in two stages, a one-time pretraining run on the adult cohort followed by 50 independent fine-tuning runs on the pediatric cohort as part of the Monte Carlo cross-validation (MCCV) protocol (Sec. 4.2). Prior work on cross-age transfer has found that adult-trained models require pediatric-specific fine-tuning and evaluation [53], which motivates this separation. Both stages use AdamW with linear warmup and cosine decay, mixed precision on a single NVIDIA A100 GPU, and a freeze-then-unfreeze schedule in which the backbone is held frozen while the task head trains alone and the last two transformer blocks are unfrozen partway through training at a reduced learning rate. Stochastic weight averaging [49] is applied over the later portion of training to produce the final checkpoint. These choices are each motivated by the risk of overfitting to a pediatric fine-tuning set that contains as few as 13 positive exams per split. Pediatric fine-tuning initializes from the adult-pretrained checkpoint and uses a further reduced backbone learning rate to limit drift from the pretrained representation. Full hyperparameters for both pretraining and fine-tuning are in Sec. C.2.

Sampling. As shown in Table 4.1, the pediatric cohort is imbalanced both between classes and within classes, since positive exams are fewer but contain more frames per exam than negative exams. Each frame is assigned a weight of $0.5/(S_{\text{class}} \cdot n_{\text{frames}})$, where S_{class} is the number of exams in the frame's class and n_{frames} is the number of frames in the frame's exam. This produces balanced 50/50 class sampling across mini-batches while also ensuring that each exam within a class contributes equal expected weight regardless of frame count. The same sampler is used for both models and for both training stages.

Methodological hyperparameter locking. Architecture, preprocessing, augmentation, learning rates, freeze and unfreeze schedule, SWA schedule, epoch budgets, and top-k aggregation were selected using the adult cohort before pediatric evaluation. These choices were then held fixed for all pediatric MCCV splits. Pediatric data were used only for fine-tuning and held-out evaluation within the prespecified MCCV protocol.

Monte Carlo cross-validation

Of the 207 exams in the pediatric cohort, 17 are positive. At this label count, a single fixed train and test split is not a credible basis for reporting performance, since any allocation of

three or four positive exams to a test set produces metric estimates with standard errors comparable to the metrics themselves. A conventional five or ten-fold cross-validation produces exactly one test prediction per exam and yields noisy per-fold sensitivity estimates for the same reason. We therefore use Monte Carlo cross-validation (MCCV), in which the cohort is repeatedly partitioned into training and test sets and a separate model is trained on each partition [97, 138].

We perform 50 iterations of stratified random splitting at the exam level. Each iteration partitions the 207 pediatric exams 75/25 into training and test sets, stratifying by the exam-level H-IAI label so that each split contains approximately 13 positive and 142 negative training exams against 4 positive and 48 negative test exams. The 50 splits are generated once and reused identically for both tasks, so any comparison between segmentation and classification is made on matched data partitions.

For each split, we fine-tune the adult-pretrained checkpoint on that split's training exams and apply the resulting model to the held-out test exams. Each exam appears in the test set of approximately 12 of the 50 splits, so every exam receives multiple out-of-fold predictions from models that were never trained on it. To produce a single out-of-fold score for each exam, we average its per-exam scores across all splits in which it appeared as a test case. For segmentation, Dice and IoU are first computed at the frame level, then averaged within each held-out exam to produce one score per exam per split. These per-exam scores are then averaged across all splits in which the exam appeared as a test case. For classification, frame-level H-IAI scores are reduced to one exam-level score using top-k mean aggregation, and those exam-level scores are then averaged across held-out appearances. The resulting 207 pooled per-exam scores are the unit of analysis for all reported metrics.

The adult cohort was used to set the modeling recipe before pediatric evaluation. Within each pediatric MCCV split, the model was fine-tuned on that split's training exams and evaluated on its held-out test exams using the fixed recipe described above. We did not adapt architecture, preprocessing, augmentation, learning rates, training duration, or exam-level aggregation to pediatric test performance. This choice limits optimization for the pediatric cohort but avoids letting held-out pediatric data influence the modeling pipeline.

Evaluation

Segmentation metrics. Segmentation quality is evaluated using the Dice similarity coefficient between the predicted binary mask and the clinician-annotated reference mask, with predictions binarized at a threshold of 0.5. The Dice coefficient is twice the area of overlap between the predicted and reference masks divided by the sum of their areas, ranging from 0 for no overlap to 1 for perfect agreement. Intersection over Union (IoU) is reported as a secondary metric. Dice and IoU are computed at the frame level and aggregated to the exam level as described in Sec. 4.2, consistent with the exam as the independent sampling unit. Dice and IoU measure pixel-level overlap against a fixed reference mask, which penalizes small boundary disagreements on a landmark whose edges are poorly defined. A high Dice score implies a useful prediction, but the converse does not necessarily hold, since a mask

that differs from the reference at the margins may still identify the same anatomic region. Documented physician requirements for AI-augmented point-of-care ultrasound emphasize localization utility over pixel-level agreement [66], so we complement the geometric metrics with a clinician review described below. In this review, two pediatric emergency medicine clinicians viewed raw ultrasound frames with predicted masks overlaid and judged whether each mask helped localize Morison's pouch.

Classification metrics. During inference, the classifier produces a H-IAI score for each frame. To reduce these frame-level scores to a single exam-level score, we sort the frame scores in descending order and average the top ten. This top-k mean aggregation reflects the clinical interpretation task, in which an exam is considered positive if there is compelling evidence of H-IAI in any view, and it provides robustness against individual noisy predictions without being drowned out by uninformative frames. The overall classification performance is reported using exam-level area under the receiver operating characteristic curve (AUROC) computed on the pooled per-exam scores, together with sensitivity and specificity at a clinically-motivated operating point. We select the operating point to target 100% sensitivity, reflecting the priority that a missed hemorrhage is substantially more costly than a false positive, which can be resolved by clinician review or confirmatory imaging.

Threshold selection. Selecting the operating threshold on the same data used to evaluate sensitivity and specificity at that threshold produces optimistically biased estimates. We therefore select the threshold via a leave-one-positive-out (LOPO) procedure on the 207 pooled per-exam scores. For each of the 17 positive exams, we hold out that exam, compute the smallest threshold at which the remaining 16 positives all receive a positive classification, evaluate whether the held-out positive is detected at that threshold, and record the specificity achieved on the 190 negatives. Repeating this for all 17 positives yields an honest leave-one-out sensitivity estimate and a specificity estimate that is not contaminated by the threshold selection procedure. The selected threshold is highly stable across the 17 folds, reflecting that removing any single positive from a pool of 17 barely changes the smallest threshold at which the remaining exams are all above the decision boundary.

Clinician review. To assess whether predicted segmentation masks are clinically useful, two pediatric emergency medicine clinicians reviewed model outputs in a blinded evaluation on MD.ai. We sampled all 17 positive pediatric exams and 34 randomly selected negative exams, drawing up to five frames per exam for a total of 239 frames across 51 exams. For each frame, the clinician saw the raw ultrasound frame with the predicted mask overlaid, and answered two questions. The first asked whether the predicted mask helps localize Morison's Pouch. The second asked how the predicted mask compares to how they would place an annotation, with response options of exactly the same, model does less, or model does more. The first question captures clinical utility as a landmark indicator regardless of pixel-level agreement, and the second captures the direction of any disagreement between the model and clinician expertise. Eleven frames were excluded from analysis because one or both clinicians marked the target anatomy as not clearly visible, leaving 228 frames for the reported results. Inter-annotator agreement on frames reviewed by both clinicians was 94.3% (33 of 35). Full protocol details, including exact question wording, are in Sec. C.3.

Statistical analysis

All confidence intervals are computed at the exam level, which is the independent sampling unit. Methods were chosen according to the estimand.

For aggregate per-exam metrics whose sampling distribution has no clean closed form, namely pooled segmentation Dice, IoU, and clinical usefulness rates, we use nonparametric bootstrap resampling of the 207 per-exam scores with 10,000 resamples. We report bias-corrected and accelerated (BCa) intervals [27], which adjust for bias and skewness in the bootstrap distribution and give better coverage than percentile intervals when the underlying distribution is skewed. For pooled AUROC we use the same bootstrap approach but resample positives and negatives independently, which preserves the marginal class proportions across resamples and is a common approach for AUROC at small positive counts.

For sensitivity and specificity under the leave-one-positive-out procedure, we report Wilson 95% intervals for the corresponding exam-level proportions. At a fixed operating threshold, each exam contributes a binary classification outcome, so a binomial proportion interval is appropriate. When the observed rate is 100%, we instead report a one-sided 95% Clopper-Pearson lower bound.

4.3 Results

We evaluated Morison's pouch segmentation in three ways. First, we measured pixel overlap against clinician reference masks using Dice and IoU. Second, we compared model performance with a preliminary clinician inter-rater reference. Third, we asked pediatric emergency medicine clinicians whether predicted masks helped localize Morison's pouch when overlaid on raw ultrasound frames.

Cohort composition

The full dataset included 330 FAST exams and 10,912 annotated frames. The adult cohort contained 123 exams, including 43 positive and 80 negative exams, totaling 3,862 frames. The pediatric cohort contained 207 exams, including 17 positive and 190 negative exams, totaling 7,050 frames. Frame counts differed by exam label, with more frames available from positive exams than negative exams in both cohorts. Pediatric positive exams averaged 148 frames per exam compared with 24 for negative exams, while adult positive exams averaged 74 frames per exam compared with 9 for negative exams.

Table 4.1: Dataset composition. Counts are shown at the exam level and the frame level for each cohort, broken down by H-IAI label. The adult cohort is used exclusively for model pretraining. The pediatric cohort is used for fine-tuning and for all reported performance estimates. An exam corresponds to one FAST examination of one patient and is the independent sampling unit for all train and test splits.

	Positive Studies	Frames	Negative Studies	Frames	Total Studies	Frames
Pediatric	17	2,471	190	4,579	207	7,050
Adult	43	3,164	80	698	123	3,862
Total	60	5,635	270	5,277	330	10,912

Segmentation performance

Across 50 MCCV splits, the segmentation model achieved a pooled exam-level Dice of 0.664 [0.646, 0.679] and IoU of 0.518 [0.501, 0.535]. Performance was modestly lower on positive exams (Dice 0.609 [0.541, 0.661], N=17) than on negative exams (Dice 0.668 [0.650, 0.684], N=190), with a parallel gap in IoU. The direction of this gap is consistent with the task, since H-IAI in Morison's pouch distorts the hepatorenal interface that anchors the reference annotation, so the landmark the model must segment against is itself less stable on positive frames than on negative ones.

Table 4.2: Segmentation performance on the pediatric cohort, evaluated via 50 iterations of Monte Carlo cross-validation. Per-exam Dice and IoU scores are averaged across the splits in which each exam appeared as a test case, and reported with 95% BCa bootstrap confidence intervals over the per-exam scores. N denotes the number of exams in each subgroup.

	Positive (n=17)	Negative (n=190)	Overall (n=207)
Dice	0.609 [0.541, 0.661]	0.668 [0.650, 0.684]	0.664 [0.646, 0.679]
IoU	0.459 [0.395, 0.511]	0.523 [0.504, 0.540]	0.518 [0.501, 0.535]

A pooled Dice of 0.664 is best read against a reference point for what agreement looks like between trained human annotators on the same task. In a preliminary inter-rater experiment conducted as part of this project, two clinicians independently annotating Morison's pouch on 220 pediatric FAST frames agreed at a mean Dice of 0.76. This figure was obtained under conditions more favorable than those the model operates in, including more consistent reference annotations and higher-quality source frames. It therefore represents an optimistic ceiling on human-human agreement rather than a matched comparison, and the gap between

0.76 and the model's 0.664 on the full MCCV evaluation is correspondingly smaller than it appears.

Even with the ceiling in view, geometric overlap against a fixed reference is not the question a clinician cares about. The operationally meaningful question is whether a predicted mask helps locate Morison's Pouch in the frame in front of them. To answer it, two pediatric emergency medicine clinicians followed the manual evaluation procedure described in Sec. 4.2, consisting of 228 frames sampled from 51 pediatric exams.

Clinician evaluations are reported in Table 4.3. On the localization question, clinicians judged that the predicted mask helped localize Morison's Pouch on 100% of the 51 reviewed exams, with a one-sided 95% Clopper-Pearson lower bound of 94.3%. On the overlap question, clinicians rated the predicted mask as exactly matching the reference on 84.9% [75.9%, 90.2%] of frames, with the remaining frames split between model-does-less (6.6% [3.4%, 12.5%]) and model-does-more (8.6% [4.3%, 17.2%]).

The results demonstrate that a pooled exam-level Dice of 0.664 coexists with clinicians finding the predicted mask useful for localization on every reviewed exam and rating it as exactly matching the reference on roughly 85% of frames. The geometric metric understates the operational value of the model at the point of care.

Table 4.3: Clinician-rated segmentation overlap, stratified by exam label. Rates are per-exam averages over reviewed frames, with 95% BCa bootstrap confidence intervals. N denotes the number of exams in each subgroup.

	Positive (n=17)	Negative (n=34)	Overall (n=51)
Helps localize Morison's Pouch	100% [≥86.2%]	100% [≥91.8%]	100% [≥94.3%]
Exactly same	82.2% [68.4%, 90.4%]	86.2% [74.5%, 93.0%]	84.9% [75.9%, 90.2%]
Model does less	13.2% [5.9%, 27.4%]	3.2% [1.2%, 8.5%]	6.6% [3.4%, 12.5%]
Model does more	4.6% [1.2%, 11.5%]	10.5% [4.3%, 23.1%]	8.6% [4.3%, 17.2%]

Classification performance

Pooled across 50 MCCV splits, the classifier achieved an exam-level AUROC of 0.933 [0.878, 0.976], with stratified bootstrap resampling positives and negatives independently. At the leave-one-positive-out operating point targeting 100% sensitivity on the training positives, LOPO sensitivity was 16 of 17 (94.1%) [73.0%, 99.0%] and LOPO specificity was 135 of 190 (71.1%) [64.2%, 77.0%], with Wilson intervals on both proportions. The selected threshold was stable across folds, reflecting that removing any single positive from a pool of 17 barely shifts the smallest threshold at which the remaining positives are all above the decision boundary.

One positive exam was not detected under the LOPO procedure, with a pooled score falling below the threshold set by the remaining 16 positives. The wide sensitivity interval reflects the fundamental constraint of evaluating on 17 positive exams rather than instability in the model itself.

During development, we also evaluated UltraFedFM, an ultrasound-specific foundation model, as an alternative to the DINOv2-Small backbone used throughout this chapter. UltraFedFM is substantially larger, but on the pediatric cohort its classification performance was comparable to DINOv2-Small, with differences within the run-to-run variability of the MCCV protocol. We selected the smaller DINOv2-Small model on the grounds of deployment simplicity and training cost.

Table 4.4: Classification performance on the pediatric cohort, evaluated via 50 iterations of Monte Carlo cross-validation. Pooled AUROC is computed on per-exam scores averaged across the splits in which each exam appeared as a test case, with a 95% stratified bootstrap confidence interval resampling positives and negatives independently. Sensitivity and specificity are reported at a leave-one-positive-out (LOPO) operating threshold, with 95% Wilson intervals on the underlying binomial proportions.

Pooled AUROC	LOPO sensitivity	LOPO specificity
0.933 [0.878, 0.976]	94.1% [73.0%, 99.0%]	71.1% [64.2%, 77.0%]

4.4 Discussion

This study demonstrates that a clinically structured two-stage approach can support pediatric FAST interpretation by mirroring how clinicians perform the examination, first identifying a relevant anatomic landmark and then assessing for H-IAI. In this initial implementation, the models achieved performance approaching expert agreement for Morison's pouch localization and high sensitivity for H-IAI classification, suggesting potential value as a decision-support and training tool for novice clinicians. More broadly, these findings support a staged, anatomically grounded approach for rare, high-stakes outcome detection in low signal-to-noise ultrasound data.

The purpose of this work is not to exceed expert performance, but to develop a clinically useful assistive tool that approaches expert-level localization and H-IAI detection, particularly for novice or low-volume FAST users. Pediatric FAST is underused in part because landmark acquisition and interpretation both have long training curves, and a system that assists with either step may lower the skill threshold for obtaining reliable information from the exam. The classifier's high-sensitivity operating point reflects the asymmetry of the underlying decision, in which a missed hemorrhage carries substantially more cost than a false positive that prompts closer bedside review or confirmatory imaging. In this setting, the model would serve as an additional set of eyes during clinician-performed FAST rather than as

an autonomous triage tool. For context, the PECARN clinical decision rule, which is the current standard tool for risk-stratifying pediatric blunt abdominal trauma, operates at 100% sensitivity and 47.2% specificity [48]. Our classifier uses a different input modality, so we report these figures alongside PECARN rather than as a direct comparison.

A second contribution of this work is the separation of the two operator-dependent parts of pediatric FAST, landmark localization and H-IAI classification, into parallel model outputs. Dice and IoU measure pixel-level overlap against a fixed reference annotation, which can penalize small boundary disagreements on a potential-space landmark whose edges are poorly defined. In our clinician review, the same masks that produced modest geometric overlap were judged useful for localizing Morison's pouch on all reviewed exams. This supports pairing pixel-overlap metrics with clinical utility review for anatomic landmarks in point-of-care ultrasound. It also suggests that the first stage can provide human-aligned localization even when it does not improve the second-stage classifier as an input mask. In development, restricting the classifier to the predicted Morison's pouch region did not improve performance, likely because masking can remove contextual information or propagate segmentation error into classification.

This work also differs from prior FAST AI studies in both population and task structure. Prior work on H-IAI classification has largely focused on adult FAST, including frame-level H-IAI classification and automated approaches that segment H-IAI when present [19, 116, 66]. Other work has used adjacent organs or pathology as substitutes for the FAST landmark rather than directly targeting Morison's pouch [72, 73]. Pediatric FAST has received less ML attention, and existing pediatric work has focused on view classification rather than H-IAI detection or direct landmark segmentation [58]. Our results extend this literature to injured children and suggest that adult ultrasound pretraining followed by pediatric fine-tuning may help address pediatric label scarcity, while still requiring pediatric-specific evaluation.

There are several important limitations. First, the pediatric cohort contains a small number of H-IAI-positive FAST exams, reflecting the core challenge of rare-event modeling in pediatric FAST. H-IAI-positive exams are uncommon but clinically high-stakes, which makes both model development and performance estimation difficult in low signal-to-noise ultrasound data. The Monte Carlo cross-validation and leave-one-positive-out procedures were designed to report performance with honest uncertainty quantification, but they cannot remove the uncertainty created by 17 positive exams. Second, the pipeline operates on single frames, whereas clinical FAST interpretation integrates information across frames as the probe moves. Third, the study is restricted to the right upper quadrant view, while a complete FAST exam includes additional views with their own landmarks and failure modes. External validation on a multi-site pediatric cohort, extension to the remaining FAST views, and clip-level modeling are the natural next steps before prospective evaluation at the bedside.

Bibliography

[1] Reza Abbasi-Asl, Yuansi Chen, Adam Bloniarz, Michael Oliver, Ben DB Willmore, Jack L Gallant, and Bin Yu. "The DeepTune framework for modeling and characterizing neurons in visual cortex area V4". In: *bioRxiv* (2018), p. 465534.

[2] Richard Antonello, Chandan Singh, Shailee Jain, Aliyah Hsu, Sihang Guo, Jianfeng Gao, Bin Yu, and Alexander Huth. "Generative causal testing to bridge data-driven models and scientific theories in language neuroscience". In: *arXiv preprint arXiv:2410.00812* (2024).

[3] Samuel G. Z. Asher, Janet Malzahn, Jessica M. Persano, Elliot J. Paschal, Andrew C. W. Myers, and Andrew B. Hall. *Do Claude Code and Codex P-Hack? Sycophancy and Statistical Analysis in Large Language Models*. Preprint. 2026.

[4] Katrin Auspurg and Josef Brüderl. "Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the "Many Analysts, One Data Set" Project". In: *Socius* 7 (2021), p. 23780231211024421.

[5] Laura A. Bakkensen and William Larson. "Population matters when modeling hurricane fatalities". In: *Proceedings of the National Academy of Sciences* 111.50 (2014), E5331–E5332.

[6] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. "Benign Overfitting in Linear Regression". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30063–30070.

[7] Daniel Barzilai and Ohad Shamir. "Generalization in Kernel Regression Under Realistic Assumptions". In: *Preprint, arXiv:2312.15995* (2023).

[8] Sumanta Basu, Karl Kumbier, James B. Brown, and Bin Yu. "Iterative random forests to discover predictive and stable high-order interactions". en. In: *Proceedings of the National Academy of Sciences* 115.8 (2018), pp. 1943–1948. ISSN: 0027-8424, 1091-6490.

[9] Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza-del Arco, Johannes B Gruber, and Dirk Hovy. "Large language model hacking: Quantifying the hidden risks of using llms for text annotation". In: *arXiv preprint arXiv:2509.08825* (2025).

[10] Mikhail Belkin, Daniel J Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine-learning practice and the classical bias–variance trade-off". In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854.

[11] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. "Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018.

[12] Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. "Does data interpolation contradict statistical optimality?" In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2019.

[13] Martin Bertran, Riccardo Fogliato, and Zhiwei Steven Wu. "Many AI Analysts, One Dataset: Navigating the Agentic Data Science Multiverse". In: *arXiv preprint arXiv:2602.18710* (2026).

[14] Christophe Boesch, Daša Bombjaková, Amelia Meier, and Roger Mundry. "Learning curves and teaching when acquiring nut-cracking in humans and chimpanzees". In: *Scientific Reports* 9.1 (2019), p. 1515. ISSN: 2045-2322.

[15] Joshua Broder, Lynn Ansley Fordham, and David M. Warshauer. "Increasing utilization of computed tomography in the pediatric emergency department, 2000-2006". eng. In: *Emergency Radiology* 14.4 (Sept. 2007), pp. 227–232. ISSN: 1070-3004.

[16] Niladri S. Chatterji and Philip M. Long. "Deep Linear Networks can Benignly Overfit when Shallow Ones Do". In: *Journal of Machine Learning Research* 24.117 (2023), pp. 1–39.

[17] Niladri S. Chatterji, Philip M. Long, and Peter L. Bartlett. "The Interplay Between Implicit Bias and Benign Overfitting in Two-Layer Linear Networks". In: *Journal of Machine Learning Research* 23.263 (2022), pp. 1–48.

[18] Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. "Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery". In: *arXiv preprint arXiv:2410.05080* (2024).

[19] Chi-Yung Cheng, I.-Min Chiu, Ming-Ya Hsu, Hsiu-Yung Pan, Chih-Min Tsai, and Chun-Hung Richard Lin. "Deep Learning Assisted Detection of Abdominal Free Fluid in Morison's Pouch During Focused Assessment With Sonography in Trauma". English. In: *Frontiers in Medicine* 8 (Sept. 2021). ISSN: 2296-858X.

[20] H Christopher Frey and Sumeet R Patil. "Identification and review of sensitivity analysis methods". In: *Risk analysis* 22.3 (2002), pp. 553–578.

[21] Margaret C. Crofoot, Ian C. Gilby, Martin C. Wikelski, and Roland W. Kays. "Interaction location outweighs the competitive advantage of numerical superiority in ¡i¿Cebus capucinus¡/i¿ intergroup contests". In: *Proceedings of the National Academy of Sciences* 105.2 (2008), pp. 577–581.

[22] Alexander D'Amour et al. "Underspecification Presents Challenges for Credibility in Modern Machine Learning". In: *Journal of Machine Learning Research* 23.226 (2022), pp. 1–61.

[23] Marco Del Giudice and Steven W Gangestad. "A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions". In: *Advances in Methods and Practices in Psychological Science* 4.1 (2021), p. 2515245920954925.

[24] Sylvie Delacroix, Diana Robinson, Umang Bhatt, Jacopo Domenicucci, Jessica Montgomery, Gaël Varoquaux, Carl Henrik Ek, Vincent Fortuin, Yulan He, Tom Diethe, et al. "Beyond quantification: Navigating uncertainty in professional AI systems". In: *RSS: Data Science and Artificial Intelligence* 1.1 (2025), udaf002.

[25] James Duncan, Rush Kapoor, Abhineet Agarwal, Chandan Singh, and Bin Yu. "VeridicalFlow: a Python package for building trustworthy data science pipelines with PCS". In: *Journal of Open Source Software* 7.69 (2022), p. 3895.

[26] Raaz Dwivedi, Chandan Singh, Bin Yu, and Martin J. Wainwright. "Revisiting minimum description length complexity in overparameterized models". In: *Preprint, arXiv:2006.10189* (2020).

[27] Bradley Efron and Robert Tibshirani. *An introduction to the bootstrap*. Monographs on statistics and applied probability 57. New York: Chapman & Hall, 1993. ISBN: 978-0-412-04231-7.

[28] Ray C. Fair. "A Theory of Extramarital Affairs". In: *Journal of Political Economy* 86.1 (1978), pp. 45–61. ISSN: 00223808, 1537534X.

[29] Jean Feng, Avni Kothari, Patrick Vossler, Andrew Bishara, Lucas Zier, Newton Addo, Aaron Kornblith, Yan Shuo Tan, and Chandan Singh. "Human-AI Co-design for Clinical Prediction Models". In: *arXiv preprint arXiv:2601.09072* (2026).

[30] Xingdong Feng, Xin He, Caixing Wang, Chao Wang, and Jingnan Zhang. "Towards a Unified Analysis of Kernel-based Methods Under Covariate Shift". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2023.

[31] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously". In: *Journal of Machine Learning Research* 20.177 (2019), pp. 1–81.

[32] Spencer Frei, Niladri S Chatterji, and Peter Bartlett. "Benign Overfitting without Linearity: Neural Network Classifiers Trained by Gradient Descent for Noisy Linear Data". In: *Conference on Learning Theory (COLT)*. 2022.

[33] Spencer Frei, Niladri S. Chatterji, and Peter L. Bartlett. "Random Feature Amplification: Feature Learning and Generalization in Neural Networks". In: *Journal of Machine Learning Research* 24.303 (2023), pp. 1–49.

[34] Spencer Frei, Gal Vardi, Peter L. Bartlett, and Nathan Srebro. "Benign Overfitting in Linear Classifiers and Leaky ReLU Networks from KKT Conditions for Margin Maximization". In: *Conference on Learning Theory (COLT)*. 2023.

[35] Spencer Frei, Gal Vardi, Peter L. Bartlett, Nathan Srebro, and Wei Hu. "Implicit Bias in Leaky ReLU Networks Trained on High-Dimensional Data". In: *International Conference on Learning Representations (ICLR)*. 2022.

[36] Andrew Gelman and Eric Loken. "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time". In: *Department of Statistics, Columbia University* 348.1-17 (2013), p. 3.

[37] Cassandra C. Gilmore. "A comparison of antemortem tooth loss in human hunter-gatherers and non-human catarrhines: Implications for the identification of behavioral evolution in the human fossil record". In: *American Journal of Physical Anthropology* 151.2 (2013), pp. 252–264.

[38] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. "Towards an AI co-scientist". In: *arXiv preprint arXiv:2502.18864* (2025).

[39] V. H. Gracias, H. L. Frankel, R. Gupta, J. Malcynski, R. Gandhi, L. Collazzo, H. Nisenbaum, and C. W. Schwab. "Defining the learning curve for the Focused Abdominal Sonogram for Trauma (FAST) examination: implications for credentialing". eng. In: *The American Surgeon* 67.4 (Apr. 2001), pp. 364–368. ISSN: 0003-1348.

[40] Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, et al. "Blade: Benchmarking language model agents for data-driven science". In: *arXiv preprint arXiv:2408.09667* (2024).

[41] Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. "Ds-agent: Automated data science by empowering large language models with case-based reasoning". In: *arXiv preprint arXiv:2402.17453* (2024).

[42] Moritz Haas, David Holzmüller, Ulrike von Luxburg, and Ingo Steinwart. "Mind the spikes: Benign overfitting of kernels and neural networks in fixed dimension". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2023.

[43] Daniel S. Hamermesh and Amy Parker. "Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity". In: *Economics of Education Review* 24.4 (2005), pp. 369–376. ISSN: 0272-7757.

[44] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. "Surprises in High-Dimensional Ridgeless Least Squares Interpolation". In: *Annals of Statistics* 50.2 (2022), pp. 949–986.

[45] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. "The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization". In: *International Conference on Computer Vision (ICCV)*. 2021.

[46] Dan Hendrycks and Thomas G. Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations". In: *International Conference on Learning Representations (ICLR)*. 2019.

[47] James F. Holmes, Kathleen Lillis, David Monroe, Dominic Borgialli, Benjamin T. Kerrey, Prashant Mahajan, Kathleen Adelgais, Angela M. Ellison, Kenneth Yen, Shireen Atabaki, Jay Menaker, Bema Bonsu, Kimberly S. Quayle, Madelyn Garcia, Alexander Rogers, Stephen Blumberg, Lois Lee, Michael Tunik, Joshua Kooistra, Maria Kwok, Lawrence J. Cook, J. Michael Dean, Peter E. Sokolove, David H. Wisner, Peter Ehrlich, Arthur Cooper, Peter S. Dayan, Sandra Wootton-Gorges, Nathan Kuppermann, and Pediatric Emergency Care Applied Research Network (PECARN). "Identifying children at very low risk of clinically important blunt abdominal injuries". eng. In: *Annals of Emergency Medicine* 62.2 (Aug. 2013), 107–116.e2. ISSN: 1097-6760.

[48] James F Holmes, Kenneth Yen, Irma T Ugalde, Paul Ishimine, Pradip P Chaudhari, Nisa Atigapramoj, Mohamed Badawy, Kevan A McCarten-Gibbs, Donovan Nielsen, Allyson C Sage, Grant Tatro, Jeffrey S Upperman, P David Adelson, Daniel J Tancredi, and Nathan Kuppermann. "PECARN prediction rules for CT imaging of children presenting to the emergency department with blunt abdominal or minor head trauma: a multicentre prospective validation study". en. In: *The Lancet Child & Adolescent Health* 8.5 (May 2024), pp. 339–347. ISSN: 23524642.

[49] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. *Averaging Weights Leads to Wider Optima and Better Generalization*. arXiv:1803.05407 [cs]. Feb. 2019.

[50] Kiju Jung, Sharon Shavitt, Madhu Viswanathan, and Joseph M. Hilbe. "Female hurricanes are deadlier than male hurricanes". In: *Proceedings of the National Academy of Sciences* 111.24 (2014), pp. 8782–8787.

[51] Chinmay Kausik, Kshitij Srivastava, and Rahul Sonthalia. "Generalization Error Without Independence: Denoising, Linear Regression, and Transfer Learning". In: *arXiv preprint arXiv:2305.17297* (2023).

[52] Mary Ella Kenefake, Matthew Swarm, and Jennifer Walthall. "Nuances in pediatric trauma". eng. In: *Emergency Medicine Clinics of North America* 31.3 (Aug. 2013), pp. 627–652. ISSN: 1558-0539.

[53] Tristan Kirscher, Sylvain Faisan, Xavier Coubez, Loris Barrier, and Philippe Meyer. *PSAT: Pediatric Segmentation Approaches via Adult Augmentations and Transfer Learning*. en. Sept. 2025.

[54] Christian Kleiber and Achim Zeileis. *Applied econometrics with R*. Springer Science & Business Media, 2008.

[55] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard L. Phillips, Ian Gao, et al. "Wilds: A benchmark of in-the-wild distribution shifts". In: *International Conference on Machine Learning (ICML)*. 2021.

[56] Lyle W. Konigsberg and Susan R. Frankenberg. "Bayes in biological anthropology". In: *American Journal of Physical Anthropology* 152.S57 (2013), pp. 153–184.

[57] Frederick Kofi Korley, Julius Cuong Pham, and Thomas Dean Kirsch. "Use of Advanced Radiology During Visits to US Emergency Departments for Injury-Related Conditions, 1998-2007". In: *JAMA* 304.13 (Oct. 2010), pp. 1465–1471. ISSN: 0098-7484.

[58] Aaron E. Kornblith, Newton Addo, Ruolei Dong, Robert Rogers, Jacqueline Grupp-Phelan, Atul Butte, Pavan Gupta, Rachael A. Callcut, and Rima Arnaout. "Development and Validation of a Deep Learning Strategy for Automated View Classification of Pediatric Focused Assessment With Sonography for Trauma". en. In: *Journal of Ultrasound in Medicine* 41.8 (Aug. 2022), pp. 1915–1924. ISSN: 0278-4297, 1550-9613.

[59] Aaron E. Kornblith, Newton Addo, Monica Plasencia, Ashkon Shaahinfar, Margaret Lin-Martore, Naina Sabbineni, Delia Gold, Lily Bellman, Ron Berant, Kelly R. Bergmann, Timothy E. Brenkert, Aaron Chen, Erika Constantine, J. Kate Deanehan, Almaz Dessie, Marsha Elkhunovich, Jason Fischer, Cynthia A. Gravel, Sig Kharasch, Charisse W. Kwan, Samuel H. F. Lam, Jeffrey T. Neal, Kathyrn H. Pade, Rachel Rempell, Allan E. Shefrin, Adam Sivitz, Peter J. Snelling, Mark O. Tessaro, and William White. "Development of a Consensus-Based Definition of Focused Assessment With Sonography for Trauma in Children". eng. In: *JAMA network open* 5.3 (Mar. 2022), e222922. ISSN: 2574-3805.

[60] Aaron E. Kornblith, Jahanara Graf, Newton Addo, Christopher Newton, Rachael Callcut, Jacqueline Grupp-Phelan, and David M. Jaffe. "The Utility of Focused Assessment With Sonography for Trauma Enhanced Physical Examination in Children With Blunt Torso Trauma". en. In: *Academic Emergency Medicine* 27.9 (Sept. 2020). Ed. by Timothy B. Jang, pp. 866–875. ISSN: 1069-6563, 1553-2712.

[61] Guy Kornowski, Gilad Yehudai, and Ohad Shamir. "From Tempered to Benign Overfitting in ReLU Neural Networks". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2023.

[62] Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. "Benign Overfitting in Two-layer ReLU Convolutional Neural Networks". In: *International Conference on Machine Learning (ICML)*. 2023.

[63] Jianfa Lai, Zixiong Yu, Songtao Tian, and Qian Lin. "Generalization Ability of Wide Residual Networks". In: *Preprint, arXiv:2305.18506* (2023).

[64] Qi Lei, Wei Hu, and Jason D. Lee. "Near-Optimal Linear Regression under Distribution Shift". In: *International Conference on Machine Learning (ICML)*. 2021.

[65] Daniel LeJeune, Jiayu Liu, and Reinhard Heckel. "Monotonic Risk Relationships under Distribution Shifts for Regularized Risk Minimization". In: *Journal of Machine Learning Research* 25.54 (2024), pp. 1–37.

[66] Megan M. Leo, Ilkay Yildiz Potter, Mohsen Zahiri, Ashkan Vaziri, Christine F. Jung, and James A. Feldman. "Using Deep Learning to Detect the Presence and Location of Hemoperitoneum on the Focused Assessment with Sonography in Trauma (FAST) Examination in Adults". In: *Journal of Digital Imaging* 36.5 (Oct. 2023), pp. 2035–2050. ISSN: 0897-1889.

[67] Hanyu Li, Haoyu Liu, Tingyu Zhu, Tianyu Guo, Zeyu Zheng, Xiaotie Deng, and Michael I Jordan. "IDA-Bench: Evaluating LLMs on Interactive Guided Data Analysis". In: *arXiv preprint arXiv:2505.18223* (2025).

[68] Qisheng Li, Meredith Ringel Morris, Adam Fourney, Kevin Larson, and Katharina Reinecke. "The Impact of Web Browser Reader Views on Reading Speed and User Experience". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, 1–12. ISBN: 9781450359702.

[69] Yifei Li, Hanane Nour Moussa, Ziru Chen, Shijie Chen, Botao Yu, Mingyi Xue, Benjamin Burns, Tzu-Yao Chiu, Vishal Dey, Zitong Lu, et al. "AutoSDT: Scaling Data-Driven Discovery Tasks Toward Open Co-Scientists". In: *arXiv preprint arXiv:2506.08140* (2025).

[70] Tian Liang, Eric Roseman, Melanie Gao, and Richard Sinert. "The Utility of the Focused Assessment With Sonography in Trauma Examination in Pediatric Blunt Abdominal Trauma: A Systematic Review and Meta-Analysis". eng. In: *Pediatric Emergency Care* 37.2 (Feb. 2021), pp. 108–118. ISSN: 1535-1815.

[71] Weixin Liang, Yining Mao, Yongchan Kwon, Xinyu Yang, and James Y. Zou. "Accuracy on the Curve: On the Nonlinear Correlation of ML Performance Between Data Subpopulations". In: *International Conference on Machine Learning (ICML)*. 2023.

[72] P. Lin and C.-Y. Tsai. "334 Novel Application of Artificial Intelligence in Ultrasound of Solid Organ Adjacent Morison's Pouch by Object Segmentation". English. In: *Annals of Emergency Medicine* 82.4 (Oct. 2023), S147. ISSN: 0196-0644, 1097-6760.

[73] Zhanye Lin, Zhengyi Li, Peng Cao, Yingying Lin, Fengting Liang, Jiajun He, and Libing Huang. "Deep learning for emergency ascites diagnosis using ultrasonography images". In: *Journal of Applied Clinical Medical Physics* 23.7 (June 2022), e13695. ISSN: 1526-9914.

[74] Yang Liu, Alex Kale, Tim Althoff, and Jeffrey Heer. "Boba: Authoring and visualizing multiverse analyses". In: *IEEE Transactions on Visualization and Computer Graphics* 27.2 (2020), pp. 1753–1763.

[75] J Scott Long and Jeremy Freese. *Regression models for categorical dependent variables using Stata*. Vol. 7. Stata press, 2006.

[76] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. "The ai scientist: Towards fully automated open-ended scientific discovery". In: *arXiv preprint arXiv:2408.06292* (2024).

[77] Ziming Luo, Atoosa Kasirzadeh, and Nihar B Shah. "The More You Automate, the Less You See: Hidden Pitfalls of AI Scientist Systems". In: *arXiv preprint arXiv:2509.08713* (2025).

[78] Cong Ma, Reese Pathak, and Martin J. Wainwright. "Optimally tackling covariate shift in RKHS-based nonparametric regression". In: *Annals of Statistics* 51.2 (2023), pp. 738–761.

[79] Steve Maley. "Statistics show no evidence of gender bias in the public's hurricane preparedness". In: *Proceedings of the National Academy of Sciences* 111.37 (2014), E3834–E3834.

[80] Neil Mallinar, James Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. "Benign, Tempered, or Catastrophic: Toward a Refined Taxonomy of Overfitting". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2022.

[81] Daniel Malter. "Female hurricanes are not deadlier than male hurricanes". In: *Proceedings of the National Academy of Sciences* 111.34 (2014), E3496–E3496.

[82] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2018.

[83] Lawrence A. Melniker, Evan Leibner, Mark G. McKenney, Peter Lopez, William M. Briggs, and Carol A. Mancuso. "Randomized Controlled Clinical Trial of Point-of-Care, Limited Ultrasonography for Trauma in the Emergency Department: The First Sonography Outcomes Assessment Program Trial". English. In: *Annals of Emergency Medicine* 48.3 (Sept. 2006), pp. 227–235. ISSN: 0196-0644, 1097-6760.

[84] James A. Meltzer, Melvin E. Stone Jr, Srinivas H. Reddy, and Ellen J. Silver. "Association of Whole-Body Computed Tomography With Mortality Risk in Children With Blunt Trauma". In: *JAMA Pediatrics* 172.6 (June 2018), pp. 542–549. ISSN: 2168-6203.

[85] Diana L. Miglioretti, Eric Johnson, Andrew Williams, Robert T. Greenlee, Sheila Weinmann, Leif I. Solberg, Heather Spencer Feigelson, Douglas Roblin, Michael J. Flynn, Nicholas Vanneman, and Rebecca Smith-Bindman. "The Use of Computed Tomography in Pediatrics and the Associated Radiation Exposure and Estimated Cancer Risk". In: *JAMA Pediatrics* 167.8 (Aug. 2013), pp. 700–707. ISSN: 2168-6203.

[86] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. "Accuracy on the Line: on the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization". In: *International Conference on Machine Learning (ICML)*. 2021.

[87] Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C Landsness, Daniel L Barabasi, Siddharth Narayanan, Nicky Evans, et al. "Kosmos: An ai scientist for autonomous discovery". In: *arXiv preprint arXiv:2511.02824* (2025).

[88] Alicia H Munnell, Geoffrey M.B Tootell, Lynn E Browne, and James McEneaney. "Mortgage lending in Boston: interpreting HMDA data". eng. In: *The American economic review* 86.1 (1996), pp. 25 –53. ISSN: 0002-8282.

[89] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. "Harmless interpolation of noisy data in regression". In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 67–83.

[90] Jaehyun Nam, Jinsung Yoon, Jiefeng Chen, and Tomas Pfister. "DS-STAR: Data Science Agent via Iterative Planning and Verification". In: *arXiv preprint arXiv:2509.21825* (2025).

[91] Fan Nie, Junlin Wang, Harper Hua, Federico Bianchi, Yongchan Kwon, Zhenting Qi, Owen Queen, Shang Zhu, and James Zou. "DSGym: A Holistic Framework for Evaluating and Training Data Science Agents". In: *arXiv preprint arXiv:2601.16344* (2026).

[92] Sima Noorani, Shayan Kiyani, George J. Pappas, and Hamed Hassani. "Conformal Prediction Beyond the Seen: A Missing Mass Perspective for Uncertainty Quantification in Generative Models". In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2025.

[93] Dino Oglic, Zoran Cvetkovic, Peter Sollich, Steve Renals, and Bin Yu. "Towards Robust Waveform-Based Acoustic Models". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 1977–1992.

[94] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. "DINOv2: Learning Robust Visual Features without Supervision". In: *Transactions on Machine Learning Research* (2024). Version Number: 2.

[95] Reese Pathak, Cong Ma, and Martin Wainwright. "A new similarity measure for covariate shift with applications to nonparametric regression". In: *International Conference on Machine Learning (ICML)*. 2022.

[96] Belinda Phipson and Gordon K. Smyth. "Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn". In: *Statistical Applications in Genetics and Molecular Biology* 9.1 (2010), Article 39.

[97] Richard R. Picard and R. Dennis Cook. "Cross-Validation of Regression Models". In: *Journal of the American Statistical Association* 79.387 (1984), pp. 575–583. ISSN: 0162-1459.

[98] Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi Jaakkola, and Regina Barzilay. "Conformal Language Modeling". In: *International Conference on Learning Representations*. Ed. by B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun. Vol. 2024. 2024, pp. 11654–11681.

[99] Alexander Rakhlin and Xiyu Zhai. "Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon". In: *Conference on Learning Theory (COLT)*. 2019.

[100] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. "Do ImageNet Classifiers Generalize to ImageNet?" In: *International Conference on Machine Learning (ICML)*. 2019.

[101] Zachary T Rewolinski and Bin Yu. "PCS Workflow for Veridical Data Science in the Age of AI". In: *arXiv preprint arXiv:2508.00835* (2025).

[102] Julia M. Rohrer, Jessica Hullman, and Andrew Gelman. *What's a Multiverse Good for Anyway?* 2026.

[103] Brendan Leigh Ross, Noël Vouitsis, Atiyeh Ashari Ghomi, Rasa Hosseinzadeh, Ji Xin, Zhaoyan Liu, Yi Sui, Shiyi Hou, Kin Kwan Leung, Gabriel Loaiza-Ganem, and Jesse C. Cresswell. *Textual Bayes: Quantifying Uncertainty in LLM-Based Systems*. 2025. arXiv: `2506.10060 [cs.LG]`.

[104] Abhraneel Sarma, Kyle Hwang, Jessica Hullman, and Matthew Kay. "Milliways: Taming multiverses through principled evaluation of data analysis paths". In: *Proceedings of the 2024 CHI conference on human factors in computing systems*. 2024, pp. 1–15.

[105] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. "Agent laboratory: Using llm agents as research assistants". In: *Findings of the Association for Computational Linguistics: EMNLP 2025* (2025), pp. 5977–6043.

[106] Lisa Schut, Nenad Tomasev, Tom McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. "Bridging the Human-AI Knowledge Gap: Concept Discovery and Transfer in AlphaZero". In: *arXiv preprint arXiv:2310.16410* (2023).

[107] David W Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.

[108] SeungWon Seo, SooBin Lim, SeongRae Noh, Haneul Kim, and HyeongYeop Kang. *From Assumptions to Actions: Turning LLM Reasoning into Uncertainty-Aware Planning for Embodied Agents*. 2026. arXiv: 2602.04326 [cs.AI].

[109] Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. "A selective review of negative control methods in epidemiology". In: *Current epidemiology reports* 7.4 (2020), pp. 190–202.

[110] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. "Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers". In: *arXiv preprint arXiv:2409.04109* (2024).

[111] Raphael Silberzahn, Eric L Uhlmann, Daniel P Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Štěpán Bahník, Feng Bai, Colin Bannard, Evelina Bonnier, et al. "Many analysts, one data set: Making transparent how variations in analytic choices affect results". In: *Advances in methods and practices in psychological science* 1.3 (2018), pp. 337–356.

[112] Max Simchowitz, Anurag Ajay, Pulkit Agrawal, and Akshay Krishnamurthy. "Statistical Learning under Heterogeneous Distribution Shift". In: *International Conference on Machine Learning (ICML)*. 2023.

[113] Uri Simonsohn, Joseph P Simmons, and Leif D Nelson. "Specification curve analysis". In: *Nature human behaviour* 4.11 (2020), pp. 1208–1214.

[114] Chandan Singh, John X. Morris, Jyoti Aneja, Alexander M. Rush, and Jianfeng Gao. *Explaining Patterns in Data with Language Models via Interpretable Autoprompting*. 2023. arXiv: 2210.01848 [cs.LG].

[115] Venkatesh Sivaraman, Patrick Vossler, Adam Perer, Julian Hong, and Jean Feng. *More Than "Means to an End": Supporting Reasoning with Transparently Designed AI Data Science Processes*. 2026. arXiv: 2603.24877 [cs.HC].

[116] Anna R. Sjogren, Megan M. Leo, James Feldman, and Joseph T. Gwin. "Image Segmentation and Machine Learning for Detection of Abdominal Free Fluid in Focused Assessment With Sonography for Trauma Examinations: A Pilot Study". eng. In: *Journal of Ultrasound in Medicine: Official Journal of the American Institute of Ultrasound in Medicine* 35.11 (Nov. 2016), pp. 2501–2509. ISSN: 1550-9613.

[117] Rebecca Smith-Bindman, Jafi Lipson, Ralph Marcus, Kwang-Pyo Kim, Mahadevappa Mahesh, Robert Gould, Amy Berrington de González, and Diana L. Miglioretti. "Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer". eng. In: *Archives of Internal Medicine* 169.22 (Dec. 2009), pp. 2078–2086. ISSN: 1538-3679.

[118] Zhangde Song, Jieyu Lu, Yuanqi Du, Botao Yu, Thomas M Pruyn, Yue Huang, Kehan Guo, Xiuzhe Luo, Yuanhao Qu, Yi Qu, et al. "Evaluating large language models in scientific discovery". In: *arXiv preprint arXiv:2512.15567* (2025).

[119] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. "Increasing transparency through a multiverse analysis". In: *Perspectives on Psychological Science* 11.5 (2016), pp. 702–712.

[120] Susan Steinemann and Mayumi Fernandez. "Variation in training and use of the focused assessment with sonography in trauma (FAST)". en. In: *The American Journal of Surgery* 215.2 (Feb. 2018), pp. 255–258. ISSN: 00029610.

[121] James H Stock and Mark W Watson. *Introduction to econometrics*. 4th. Pearson, 2020.

[122] Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. "The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies". In: *Nature* 646.8085 (2025), pp. 716–723.

[123] Chenhao Tan and Haokun Liu. *The Mirage of Autonomous AI Scientists*. Preprint, `https://chenhaot.com/papers/mirage_ai_scientist.pdf`. Accessed: 2026-02-24. 2026.

[124] Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. "AI-Researcher: Autonomous Scientific Innovation". In: *arXiv preprint arXiv:2505.18705* (2025).

[125] Tiffany M Tang, Yuping Zhang, Ana M Kenney, Cassie Xie, Lanbo Xiao, Javed Siddiqui, Sudhir Srivastava, Martin G Sanda, John T Wei, Ziding Feng, et al. "A simplified MyProstateScore2. 0 for high-grade prostate cancer". In: *Cancer Biomarkers* 42.1 (2025), p. 18758592241308755.

[126] Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. "Overparameterization Improves Robustness to Covariate Shift in High Dimensions". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2021.

[127] Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. "Evaluating Large Language Models as Expert Annotators". In: *arXiv preprint arXiv:2508.07827* (2025).

[128] Alexander Tsigler and Peter L. Bartlett. "Benign overfitting in ridge regression". In: *Journal of Machine Learning Research* 24.123 (2023), pp. 1–76.

[129] Edwin JC Van Leeuwen, Emma Cohen, Emma Collier-Baker, Christian J Rapold, Marie Schäfer, Sebastian Schütte, and Daniel BM Haun. "The development of human social learning across seven societies". In: *Nature Communications* 9.1 (2018), p. 2076.

[130] Roman Vershynin. *High-dimensional probability*. Cambridge University Press, 2018.

[131] Kaizheng Wang. "Pseudo-Labeling for Kernel Ridge Regression under Covariate Shift". In: *Preprint, arXiv:2302.10160* (2023).

[132] Ke Alexander Wang, Niladri S Chatterji, Saminul Haque, and Tatsunori Hashimoto. "Is Importance Weighting Incompatible with Interpolating Classifiers?" In: *International Conference on Learning Representations (ICLR)*. 2022.

[133] Qianru Wang, Tiffany M Tang, Nathan Youlton, Chad S Weldy, Ana M Kenney, Omer Ronen, J Weston Hughes, Elizabeth T Chin, Shirley C Sutton, Abhineet Agarwal, et al. "Epistasis regulates genetic control of cardiac hypertrophy". In: *medRxiv* (2024), pp. 2023–11.

[134] F. Wenzel, Andrea Dittadi, Peter Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, Bernhard Scholkopf, and Francesco Locatello. "Assaying Out-Of-Distribution Generalization in Transfer Learning". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2022.

[135] Steven H. Woolf, Elizabeth R. Wolf, and Frederick P. Rivara. "The New Crisis of Increasing All-Cause Mortality in US Children and Adolescents". In: *JAMA* 329.12 (Mar. 2023), pp. 975–976. ISSN: 0098-7484.

[136] Siqi Wu, Antony Joseph, Ann S Hammonds, Susan E Celniker, Bin Yu, and Erwin Frise. "Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks". In: *Proceedings of the National Academy of Sciences* 113.16 (2016), pp. 4290–4295.

[137] Skyler Wu, Yash Nair, and Emmanuel J. Candès. *Efficient Evaluation of LLM Performance with Statistical Guarantees*. 2026. arXiv: `2601.20251 [stat.ML]`.

[138] Qing-Song Xu and Yi-Zeng Liang. "Monte Carlo cross validation". In: *Chemometrics and Intelligent Laboratory Systems* 56.1 (Apr. 2001), pp. 1–11. ISSN: 0169-7439.

[139] Zhiwei Xu, Yutong Wang, Spencer Frei, Gal Vardi, and Wei Hu. "Benign Overfitting and Grokking in ReLU Networks for XOR Cluster Data". In: *International Conference on Learning Representations (ICLR)*. 2024.

[140] Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. "Large language models for automated open-domain scientific hypotheses discovery". In: *Findings of the Association for Computational Linguistics: ACL 2024*. 2024, pp. 13545–13565.

[141] Cristobal Young and Katherine Holsteen. "Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis". In: *Sociological Methods & Research* 46.1 (2017), pp. 3–40.

[142] Bin Yu and Rebecca L. Barter. *Veridical Data Science: The Practice of Responsible Data Analysis and Decision Making*. MIT Press, 2024.

[143] Bin Yu and Karl Kumbier. "Veridical data science". In: *Proceedings of the National Academy of Sciences of the United States of America* 117.8 (2020), pp. 3920–3929.

[144] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization". In: *International Conference on Learning Representations (ICLR)*. 2017.

[145] Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. "Goal driven discovery of distributional differences via language descriptions". In: *arXiv preprint arXiv:2302.14233* (2023).

[146] Markus Tyler Ziesmann, Jason Park, Bertram J. Unger, Andrew W. Kirkpatrick, Ashley Vergis, Sarvesh Logsetty, Chau Pham, David Kirschner, and Lawrence M. Gillman. "Validation of the quality of ultrasound imaging and competence (QUICk) score as an objective assessment tool for the FAST examination". eng. In: *The Journal of Trauma and Acute Care Surgery* 78.5 (May 2015), pp. 1008–1013. ISSN: 2163-0763.

Appendix A

Content Deferred from Chapter 2

A.1 Formal Assumptions

Definition A.1 (Linear regression under distribution shift)**.** We consider a training dataset comprised of n i.i.d. pairs $(\boldsymbol{x}^i, y^i)_{i=1}^n \sim \mathcal{D}_\mathsf{s}^n$ concatenated into a data matrix $X \in \mathbb{R}^{n \times p}$ and a response vector $\boldsymbol{y}_\mathsf{s} \in \mathbb{R}^n$. The setting is overparameterized, meaning we have more input features than training samples, or $n < p$.

We define

1. the covariance matrix $\Sigma_\mathsf{s} = \mathbb{E}_{\mathcal{D}_\mathsf{s}}[\boldsymbol{x}\boldsymbol{x}^\top]$,
2. the optimal parameter vector $\theta^*_{src} \in \mathbb{R}^p$, satisfying

$$\mathop{\mathbb{E}}_{\mathcal{D}_\mathsf{s}}\left[(y - x^\top \theta^*_{src})^2\right] = \min_\theta \mathop{\mathbb{E}}_{\mathcal{D}_\mathsf{s}}\left[(y - x^\top \theta)^2\right].$$

We test on the distribution $\mathcal{D}_\mathsf{t}$ with Σ_t and θ^*_t defined in the same way. We assume

1. (*centered rows*) $\mathbb{E}_{\mathcal{D}_\mathsf{s}}[\boldsymbol{x}] = 0$;
2. (*well-specified - source*) For $(X, \boldsymbol{y}) \subseteq \mathcal{D}_\mathsf{s}$, $\boldsymbol{y} = X\theta^*_\mathsf{s} + \boldsymbol{\varepsilon}_\mathsf{s}$. We assume that the components of the source noise vector $\boldsymbol{\varepsilon}_\mathsf{s}$ are i.i.d. centered random variables with positive variance $v_{\varepsilon_s^2}$ and that $\mathbb{E}_{\mathcal{D}_\mathsf{s}}[y|x] = x^T\theta^*_\mathsf{s}$;
3. (*well-specified - target*) For $(X, \boldsymbol{y}) \subseteq \mathcal{D}_\mathsf{t}$, $\boldsymbol{y} = X\theta^*_\mathsf{t} + \boldsymbol{\varepsilon}_\mathsf{t}$. We assume that the components of the target noise vector $\boldsymbol{\varepsilon}_\mathsf{t}$ are i.i.d. centered random variables with noise variance, $v_{\varepsilon_t^2}$, and that $\mathbb{E}_{\mathcal{D}_\mathsf{t}}[y|x] = x^T\theta^*_\mathsf{t}$;
4. (*simultaneously diagonalizability*) Σ_s and Σ_t commute; that is, there exists an orthogonal matrix $V \in \mathbb{R}^p$ such that $V^\top \Sigma_\mathsf{s} V$ and $V^\top \Sigma_\mathsf{t} V$ are both diagonal. This allows us to fix

an orthonormal basis in which we can express the covariance matrices as

$$\Sigma_{\mathsf{s}} = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{s}}}[xx^T] = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_p),$$
$$\Sigma_{\mathsf{t}} = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}[xx^T] = \mathrm{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2, ..., \tilde{\lambda}_p),$$

where the source eigenvalues are a non-increasing sequence, $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p$. Note that we do not require the target eigenvalues to be a non-increasing sequence, however we require that $\tilde{\lambda}_i \lambda_i \geq 0$ for all i;

5. (*subgaussianity*) the whitened data matrix, denoted $Z = X\Sigma_{\mathsf{s}}^{-1/2}$, has centered i.i.d. row vectors with independent coordinates. We assume that the rows are subgaussian with subgaussian norm σ_x; that is, for all $\gamma \in \mathbb{R}^p$,

$$\mathbb{E}[\exp(\gamma^\top z)] \leq \exp(\sigma_x^2 ||\gamma||^2/2).$$

A.2 Key results from prior work and technical lemmas

For ease on the reader, we replicate some key lemma statements from Bartlett et al. [6] and Tsigler and Bartlett [128] and provide new lemmas and corollaries that we use in our work.

Recall that $\rho_k = \frac{1}{n\lambda_{k+1}} \sum_{i>k} \lambda_i$, $X \in \mathbb{R}^{n\times p}$, and $\Sigma_{\mathsf{s}} \in \mathbb{R}^{p\times p} = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$. Let $X_{0:k} \in \mathbb{R}^{n\times k}$ denote the matrix comprised of the first k feature columns. Similarly, $X_{k:p} \in \mathbb{R}^{n\times(p-k)}$ denote the matrix of the last $p-k$ feature columns. The Gram matrix of the data, denoted here by

$$A = XX^T,$$

plays a central role in the investigation of high-dimensional linear regression. Analogous to the above, we express $A_{0:k} = X_{0:k}X_{0:k}^T \in \mathbb{R}^{n\times n}$ and similarly for $A_{k:p} \in \mathbb{R}^{n\times n}$.

Letting $Z = X\Sigma_{\mathsf{s}}^{-1/2} \in \mathbb{R}^{n\times p}$ and denoting the independent column vectors of Z by $z_i \in \mathbb{R}^n$, we have the following expressions:

$$A = \sum_i \lambda_i z_i z_i^T, \qquad A_{-i} = \sum_{j\neq i} \lambda_j z_j z_j^T, \qquad A_k = \sum_{i>k} \lambda_i z_i z_i^T.$$

The following lemma from Bartlett et al. [6] is key in controlling the largest and smallest eigenvalues of the data Gram matrix and its variants A_{-i} and A_k. Importantly, it also shows that if the energy in the bottom $p-k$ components of the covariance matrix is sufficiently large (ρ_k is lower bounded by a constant), then the largest and smallest eigenvalues of A_k are equal up to constants.

Lemma A.2 (Lemma 5 from Bartlett et al. [6])**.** *There are constants $b, c \geq 1$ such that for any $k \geq 0$, with probability at least $1 - 2e^{-n/c}$,*

1. *for all* $i \geq 1$,

$$\mu_{k+1}(A_{-i}) \leq \mu_{k+1}(A) \leq \mu_1(A_k) \leq c\left(\sum_{j>k} \lambda_j + \lambda_{k+1} n\right),$$

2. *for all* $1 \leq i \leq k$,

$$\mu_n(A) \geq \mu_n(A_{-i}) \geq \mu_n(A_k) \geq \frac{1}{c}\sum_{j>k} \lambda_j - c\lambda_{k+1} n,$$

3. *if* $\rho_k \geq b$, *then*

$$\frac{1}{c}\lambda_{k+1}\rho_k n \leq \mu_n(A_k) \leq \mu_1(A_k) \leq c\lambda_{k+1}\rho_k n.$$

A consequence of the prior eigenvalue bounds is that when ρ_k is lower bounded by a constant, the condition number of A_k is upper bounded by a constant. Therefore even as problem parameters such as training sample size and input dimension grow to ∞, A_k is still well-conditioned. This is important as non-benign overfitting occurs when the condition number bound on A_k grows with problem parameters. This would happen if the lower bound on the smallest eigenvalue of A_k decays to zero too quickly which would cause the condition number of A_k to diverge. If this occurs then the excess risk of the MNI would be lower bounded. This is shown for the in-distribution case in Bartlett et al. [6].

Corollary A.3. *Following from Lemma A.2, there are constants* $b, c \geq 1$ *such that for any* $k \geq 0$, *with probability at least* $1 - 2e^{-n/c}$, *if* $\rho_k \geq b$ *then*

$$\frac{\mu_1(A_k)}{\mu_n(A_k)} \leq c^2 \tag{A.1}$$

which is the equivalent of the assumption CondNum$(k, 2e^{-n/c}, c^2)$ *as defined in Tsigler and Bartlett [128].*

The following definition and lemma omit all references to *NonCritReg* and the ridge parameter in Tsigler and Bartlett [128].

Definition A.4 (StableLowerEig(k, δ, L) from Tsigler and Bartlett [128])**.** Assume that for any $j \in \{1, 2, \cdots, p\}$ with probability (separate for every j) at least $1 - \delta$,

$$\mu_n(A_{-j}) \geq \mu_n(\mathbb{E}\, A_k)/L = (\sum_{i>k} \lambda_i)/L. \tag{A.2}$$

We now state key assumptions that are necessary in order to obtain an explicit bias lower bound. Exchangeable coordinates (*ExchCoord*) is a weaker assumption than independent components of the data vector. It is used in Tsigler and Bartlett [128] instead of independent components. We assume that components of Z are independent and so we immediately satisfy the *ExchCoord*, which we define here.

Definition A.5 (ExchCoord)**.** Assume the sequence of coordinates of $\Sigma_{\mathsf{s}}^{-1/2}x$, for any $x \in X$, is exchangable (any deterministic permutation of the coordinates of whitened data vectors doesn't change their distribution).

The *PriorSigns* assumption is necessary to obtain lower bounds on the bias term. It allows us to use bounds on the expectation of a quadratic form, $\mathbb{E}_v[v^T M v]$, in order to separately analyze the contributions of v and M. As the bias takes the form $\theta_{\mathsf{s}}^{*T}(I - X^T A^{-1} X)\Sigma_{\mathsf{t}}(I - X^T A^{-1} X)\theta_{\mathsf{s}}^*$ we see that such a bound would separate the contributions of the model from that of data-dependent matrix expressions.

Definition A.6 (PriorSigns)**.** Assume that θ^* is sampled from a prior distribution in the following way: one starts with vector $\overline{\theta}$ and flips signs of all its coordinates with probability 0.5 independently.

Under *PriorSigns*, the random model vector is obtained by flipping signs on the components of the ground-truth model vector. This does not affect our bounds as we see in Theorem 2.6 that our bias lower bound only relies on squared components of the random model vector which are equivalent to the squared components of the ground truth model.

An important consequence of having a bounded condition number and independent coordinates is that with high probability the smallest eigenvalue of A_{-i} for all $i \geq 1$ is lower bounded by $n\lambda_{k+1}\rho_k$ up to constants. These assumptions allow Bartlett et al. [6] to prove Lemma A.2, which in turn allows us to derive the *StableLowerEig* condition. This is a simple consequence of B.1 and we provide details here for completeness.

Corollary A.7 (Our variant of StableLowerEig from Tsigler and Bartlett [128])**.** *For all $i \geq 1$, with probability at least $1 - 2e^{-n/c_2}$*

$$\mu_n(A_{-i}) \geq \frac{1}{c_2}\mu_n(\mathbb{E}\, A_k) = \frac{1}{c_2}\sum_{j>k}\lambda_j.$$

Proof. By Lemma A.2, for some absolute constant $c_1 \geq 1$ with probability at least $1 - 2e^{-n/c_1}$

$$\mu_n(A_k) \geq \frac{1}{c_1}\sum_{i>k}\lambda_i - c_1\lambda_{k+1}n.$$

The assumption $\rho_k \geq b$ for some $b \geq 1$ gives us

$$\begin{aligned}\frac{1}{c_1}\sum_{i>k}\lambda_i - c_1\lambda_{k+1}n &= \frac{1}{c_1}\lambda_{k+1}n\rho_k - c_1\lambda_{k+1}n \\ &\geq \left(\frac{1}{c_1} - \frac{c_1}{b}\right)\lambda_{k+1}n\rho_k \\ &= \left(\frac{1}{c_1} - \frac{c_1}{b}\right)\sum_{i>k}\lambda_i.\end{aligned}$$

Choosing $b > c_1^2$ and $c_2 = \max\{c_1, (1/c_1 - c_1/b)^{-1}\}$, we get that with probability at least $1 - 2e^{-n/c_2}$

$$\mu_n(A_k) \geq \frac{1}{c_2} \sum_{i>k} \lambda_i.$$

The next step is to extend this result to A_{-i} for all i.

For $i \leq k$, observe that $A_{-i} \succeq A_k$ gives us $\mu_n(A_{-i}) \geq \mu_n(A_k)$. For the case of $i > k$, we have

$$\begin{aligned} A_{-i} &= \sum_{j \neq i} \lambda_j z_j z_j^\top \\ &= \sum_{j \leq k} \lambda_j z_j z_j^\top + \sum_{j>k, j\neq i} \lambda_j z_j z_j^\top \\ &\succeq \lambda_1 z_1 z_1^\top + \sum_{j>k, j\neq i} \lambda_j z_j z_j^\top \\ &\succeq \lambda_i z_1 z_1^\top + \sum_{j>k, j\neq i} \lambda_j z_j z_j^\top. \end{aligned}$$

We assume that the features are independent and z_i is centered and whitened, so $\lambda_i z_1 z_1^\top + \sum_{j>k, j\neq i} \lambda_j z_j z_j^\top$ has the same distribution as $A_k = \sum_{j>k} \lambda_j z_j z_j^\top$. Therefore,

$$\begin{aligned} &\mathbb{P}\left(\mu_n(A_{-i}) \geq \frac{1}{c_2} \sum_{i>k} \lambda_i \right) \\ \geq &\mathbb{P}\left(\mu_n\left(\lambda_i z_1 z_1^\top + \sum_{j>k, j\neq i} \lambda_j z_j z_j^\top \right) \geq \frac{1}{c_2} \sum_{i>k} \lambda_i \right) \\ = &\mathbb{P}\left(\mu_n(A_k) \geq \frac{1}{c_2} \sum_{i>k} \lambda_i \right) \\ \geq &1 - 2e^{-n/c}. \end{aligned}$$

□

The following corollaries provide high-probability bounds on random subgaussian vectors with independent coordinates.

Corollary A.8 (Corollary 1 from Bartlett et al. [6])**.** *There is a universal constant c such that for any centered random vector $z \in \mathbb{R}^n$ with independent σ^2-subgaussian coordinates with unit variances, any random subspace $\mathscr{L}$ of $\mathbb{R}^n$ of codimension k that is independent of z, and any $t > 0$, with probability at least $1 - 3e^{-t}$,*

$$\begin{aligned} \|z\|^2 &\leq n + c\sigma^2(t + \sqrt{nt}), \\ \|\Pi_{\mathscr{L}} z\|^2 &\geq n - c\sigma^2(k + t + \sqrt{nt}), \end{aligned}$$

where $\Pi_{\mathscr{L}}$ is the orthogonal projection on $\mathscr{L}$.

In our proofs, we will need to control the norm of z_i for all $i \leq p$ on the same high-probability event. In these cases we need to apply a union bound over the events in the summation. The following corollary shows how to invoke a union bound over ℓ of these events in such a way that the probability over the union of all such events holds with high probability that depends n.

Corollary A.9. *There is a universal constant c as defined in Corollary A.8. Let $z \in \mathbb{R}^n$ be a centered random vector with σ^2-subgaussian coordinates and unit variances. Let $\mathscr{L}$ be a random subspace of $\mathbb{R}^n$ of codimension k that is independent of z.*

For $0 < t < n/c_0$ and $k \in (0, n/c_1)$ for $c_1 > c_0$ with c_0 sufficiently large, with probability $1 - 3e^{-t}$,

$$\|z\|^2 \leq c_2 n$$
$$\|\Pi_{\mathscr{L}} z\|^2 \geq n/c_3$$

where c_2, c_3 only depends on c, c_0, σ.

We obtain a union bound over the intersection of ℓ of these events so long as $\ln(\ell) \leq n/c_0 \Rightarrow \ell \leq e^{n/c_0}$. Then for $k \in (0, n/c_1)$ for $c_1 > c_0$ with c_0 sufficiently large, if $\ell \leq e^{n/c_0}$, with probability at least $1 - 3e^{-n/c_0}$, ℓ of the above events independently hold.

Proof. Let Corollary A.8 hold with universal constant c. Then, with probability $1 - 3e^{-t}$ for $t > 0$,

$$\|z\|^2 \leq n + c\sigma^2(t + \sqrt{nt})$$
$$\|\Pi_{\mathscr{L}} z\|^2 \geq n - c\sigma^2(k + t + \sqrt{nt}).$$

Let $t \leq \frac{n}{c_0}$. Then we have that,

$$-\frac{n}{c_0} \leq -t \qquad -\frac{n}{\sqrt{c_0}} \leq -\sqrt{nt}.$$

Plugging in for $\|z\|^2$,

$$\begin{aligned}\|z\|^2 &\leq n + c\sigma^2(t + \sqrt{nt}) \\ &\leq n + c\sigma^2(\frac{n}{c_0} + \frac{n}{\sqrt{c_0}}) \\ &= n(1 + c\sigma^2(c_0^{-1} + c_0^{-1/2})) \\ &= c_1 n\end{aligned}$$

for c_1 only dependent on c, c_0, σ. Now, plugging in for $\|\Pi_{\mathscr{L}} z\|^2$,

$$\begin{aligned}\|\Pi_{\mathscr{L}} z\|^2 &\geq n - c\sigma^2(k + t + \sqrt{nt}) \\ &\geq n - c\sigma^2(k + \frac{n}{c_0} + \frac{n}{\sqrt{c_0}}) \\ &= n(1 - c\sigma^2(\frac{k}{n} + c_0^{-1} + c_0^{-1/2})).\end{aligned}$$

Let $k < \frac{n}{c_2}$ for $c_2 > c_0$. Then it is clear that $-\frac{k}{n} > -\frac{1}{c_2}$ and,

$$\begin{aligned} n(1 - c\sigma^2(\frac{k}{n} + c_0^{-1} + c_0^{-1/2})) &\geq n(1 - c\sigma^2(c_2^{-1} + c_0^{-1} + c_0^{-1/2})) \\ &= n/c_3 \end{aligned}$$

for constant c_3 that only depends on c, σ^2, c_0. We finally require that $1 - c\sigma^2(c_1^{-1} + c_0^{-1} + c_0^{-1/2}) > 0$ which we can achieve by taking c_0 sufficiently large.

We now proceed to bound the union of ℓ of the complement events, in order to obtain a bound over the intersection of ℓ of these events.

For multiple z_i's, define by A_i the events shown above, that $\|z_i\|^2 \leq c_2 n$ and $\|\Pi_{\mathscr{L}_i} z_i\|^2 \geq n/c_3$ where z_i and $\mathscr{L}_i$ are defined analogous to $z, \mathscr{L}$ above. Then

$$\begin{aligned} P(\cup_{i=1}^{\ell}(A_i)^c) &\leq \sum_{i=1}^{\ell} P((A_i)^c) \\ &\leq \sum_{i=1}^{\ell} 3e^{-t} \\ &= 3\ell e^{-t}. \end{aligned}$$

Then $P(\cap_{i=1}^{\ell} A_i) \geq 1 - 3\ell e^{-t}$. Observing that $3\ell e^{-t} = 3e^{\ln(\ell)} e^{-t} = 3e^{-t+\ln(\ell)} = 3e^{-(t-\ln(\ell))}$ we can set the per-event t accordingly and obtain the necessary bound. We want $0 < t - \ln(\ell) \leq n/c_0$ to complete the bound. Therefore, we need that, per-event, $\ln(\ell) < t \leq n/c_0 + \ln(\ell)$. If $\ln(\ell) \leq n/c_0$ then this reduces to needing $\ln(\ell) < t \leq 2n/c_0$. Since each event is defined for $t \in (0, n/c_0]$ the union bound proof is complete by taking $t = n/c_0$ and requiring that $\ln(\ell) \leq n/c_0$.

□

The following lemma is necessary in order to extend a summation over random variables, each lower bounded by a real number with equal probability, to a unified lower bound over the entire summation.

Lemma A.10 (Lemma 9 from Bartlett et al. [6])**.** *Suppose $n \leq \infty$ and $\{\eta_i\}_{i=1}^n$ is a sequence of non-negative random variables, $\{t_i\}_{i=1}^n$ is a sequence of non-negative real numbers (at least one of which is strictly positive) such that for some $\delta \in (0, 1)$ and any $i \leq n$, $P(\eta_i > t_i) \geq 1 - \delta$. Then,*

$$P\left(\sum_{i=1}^{n} \eta_i \geq \frac{1}{2}\sum_{i=1}^{n} t_i\right) \geq 1 - 2\delta.$$

We now provide a minor generalization of Corollary S.6 in Bartlett et al. [6] that comes from replacing a_1 in a non-increasing sequence of non-negative numbers $\{a_i\}_{i=1}^p$ with $\max_i a_i$ and only requiring that $\{a_i\}_{i=1}^p$ is a sequence of non-negative numbers.

Corollary A.11. *There is a universal constant c such that for any sequence $\{a_i\}_{i=1}^p$ of non-negative numbers such that $\sum_{i=1}^p a_i < \infty$, and any independent, centered, σ-subexponential random variables $\{\xi_i\}_{i=1}^p$, and any $x > 0$, with probability at least $1 - 2e^{-cx}$,*

$$|\sum_i a_i \xi_i| \leq \sigma \max\left(x \max_i a_i, \sqrt{x \sum_{i=1}^p a_i^2}\right).$$

Lastly, the following identity will allow us to use the *PriorSigns* assumption to derive a new form for the bias term, which will be used for the proof of the lower bound.

Lemma A.12 (Identity for expectation of a quadratic form)**.** *Assume $M \in \mathbb{R}^{p\times p}$ is a symmetric matrix. For a random vector $x \in \mathbb{R}^p$ with mean $\mathbb{E}[x]$ and covariance $\mathrm{Cov}(x)$,*

$$\mathop{\mathbb{E}}_x[x^T M x] = \mathbb{E}[x]^T M\, \mathbb{E}[x] + \mathrm{tr}(M\, Cov(x)).$$

Proof.

$$\begin{aligned}
\mathop{\mathbb{E}}_x[x^T M x] &= \mathbb{E}[\mathrm{tr}(x^T M x)] \\
&= \mathbb{E}[\mathrm{tr}(M x x^T)] \\
&= \mathrm{tr}(M\, \mathbb{E}[x x^T]) \\
&= \mathrm{tr}(M \mathrm{Cov}(x) + M\, \mathbb{E}[x]\, \mathbb{E}[x]^T) \\
&= \mathrm{tr}(M \mathrm{Cov}(x)) + \mathrm{tr}(\mathbb{E}[x] M\, \mathbb{E}[x]^T) \\
&= \mathrm{tr}(M \mathrm{Cov}(x)) + \mathbb{E}[x] M\, \mathbb{E}[x]^T.
\end{aligned}$$

□

A.3 Proof of excess risk bound

We start by restating Theorem 2.3.

Theorem 2.3. *(Target excess risk decomposition) The excess risk of the MNI trained on the source data, when evaluated on the target distribution, satisfies*

$$R(\widehat{\theta}(\boldsymbol{y}_\mathsf{s}), \mathcal{D}_\mathsf{t}) \leq 4B_1 + 4B_2 + 2V_{\boldsymbol{\varepsilon}_\mathsf{s}}, \tag{2.2}$$

and

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_\mathsf{s}} R(\widehat{\theta}(\boldsymbol{y}_\mathsf{s}), \mathcal{D}_\mathsf{t}) = B_1 + B_2 + \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_\mathsf{s}} V_{\boldsymbol{\varepsilon}_\mathsf{s}} \\
+ 2(\theta^*_\mathsf{t} - \theta^*_\mathsf{s})^\top \Sigma_\mathsf{t} (\theta^*_\mathsf{s} - \widehat{\theta}(X\theta^*_\mathsf{s})),
\end{aligned}$$

where we define

$$B_1 := \|\theta^*_{\mathsf{s}} - \theta^*_{\mathsf{t}}\|^2_{\Sigma_{\mathsf{t}}}, \tag{2.3}$$

$$B_2 := \|\theta^*_{\mathsf{s}} - \widehat{\theta}(X\theta^*_{\mathsf{s}})\|^2_{\Sigma_{\mathsf{t}}}, \tag{2.4}$$

$$V_{\boldsymbol{\varepsilon}_{\mathsf{s}}} := \|\widehat{\theta}(\boldsymbol{\varepsilon}_{\mathsf{s}})\|^2_{\Sigma_{\mathsf{t}}}, \tag{2.5}$$

and $\|\boldsymbol{x}\|^2_M := \boldsymbol{x}^\top M \boldsymbol{x}$.

Proof. Let us begin by noting that the excess risk of any θ is given by,

$$\begin{aligned}
R(\theta) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\mathsf{t}}}\left[\left(y - x^\top\theta\right)^2\right] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\mathsf{t}}}\left[\left(y - x^\top\theta^*_{\mathsf{t}}\right)^2\right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\mathsf{t}}}\left[\left(y - x^\top\theta^*_{\mathsf{t}} + x^\top\theta^*_{\mathsf{t}} - x^\top\theta\right)^2\right] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\mathsf{t}}}\left[\left(y - x^\top\theta^*_{\mathsf{t}}\right)^2\right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\mathsf{t}}}\left[\left(x^\top\theta^*_{\mathsf{t}} - x^\top\theta\right)^2\right] + 2\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\mathsf{t}}}\left[\left(y - x^\top\theta^*_{\mathsf{t}}\right)\left(x^\top\theta^*_{\mathsf{t}} - x^\top\theta\right)\right] \\
&\overset{(i)}{=} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[\left(x^\top\theta^*_{\mathsf{t}} - x^\top\theta\right)^2\right].
\end{aligned} \tag{A.3}$$

Equality (i) uses that, conditional on x, $y - x^\top\theta^*_{\mathsf{t}}|x$ is mean-zero, which is given in Assumption 3 (*well-specified - target*). So that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\mathsf{t}}}\left[\left(y - x^\top\theta^*_{\mathsf{t}}\right)\left(x^\top\theta^*_{\mathsf{t}} - x^\top\theta\right)\right] = \mathbb{E}\left[\left(x^\top\theta^*_{\mathsf{t}} - x^\top\theta\right)\mathbb{E}\left[\left(y - x^\top\theta^*_{\mathsf{t}}\right)|x\right]\right] = 0.$$

We now note that the source-data MNI can be decomposed as follows,

$$\begin{aligned}
\widehat{\theta}(\boldsymbol{y}_{\mathsf{s}}) &= X^\top(X_{\mathsf{s}}X_{\mathsf{s}}^\top)^{-1}\boldsymbol{y}_{\mathsf{s}} \\
&= X^\top(X_{\mathsf{s}}X_{\mathsf{s}}^\top)^{-1}(X\theta^*_{\mathsf{s}} + \boldsymbol{\varepsilon}_{\mathsf{s}}) \\
&= \widehat{\theta}(X\theta^*_{\mathsf{s}}) + \widehat{\theta}(\boldsymbol{\varepsilon}_{\mathsf{s}})
\end{aligned}$$

We can thus continue from equation A.3 to characterize the excess risk of the source-data MNI as

$$\begin{aligned}
R(\widehat{\theta}(\boldsymbol{y}_{\mathsf{s}})) &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[\left(x^\top\theta^*_{\mathsf{t}} - x^\top\widehat{\theta}(\boldsymbol{y}_{\mathsf{s}})\right)^2\right] \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[\left(x^\top\theta^*_{\mathsf{t}} - x^\top(\widehat{\theta}(X\theta^*_{\mathsf{s}}) + \widehat{\theta}(\boldsymbol{\varepsilon}_{\mathsf{s}}))\right)^2\right] \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[\left(x^\top(\theta^*_{\mathsf{t}} - \widehat{\theta}(X\theta^*_{\mathsf{s}})) - x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_{\mathsf{s}})\right)^2\right] \\
&\overset{(i)}{\leq} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[2\left(x^\top(\theta^*_{\mathsf{t}} - \widehat{\theta}(X\theta^*_{\mathsf{s}}))\right)^2 + 2\left(x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_{\mathsf{s}})\right)^2\right] \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[2\left(x^\top(\theta^*_{\mathsf{t}} - \theta^*_{\mathsf{s}} + \theta^*_{\mathsf{s}} - \widehat{\theta}(X\theta^*_{\mathsf{s}}))\right)^2 + 2\left(x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_{\mathsf{s}})\right)^2\right] \\
&\overset{(ii)}{\leq} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[4\left(x^\top(\theta^*_{\mathsf{t}} - \theta^*_{\mathsf{s}})\right)^2 + 4\left(x^\top(\theta^*_{\mathsf{s}} - \widehat{\theta}(X\theta^*_{\mathsf{s}}))\right)^2 + 2\left(x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_{\mathsf{s}})\right)^2\right].
\end{aligned} \tag{A.4}$$

In inequalities (i) and (ii), we have used Young's inequality, which implies $(a-b)^2 \leq 2(a-c)^2 + 2(b-c)^2$ for any $a, b, c \in \mathbb{R}$. Recalling that

$$\|x\|_M^2 := x^\top M x,$$

it is apparent that the first term is just the weighted distance between the source and target vectors,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top(\theta_\mathsf{t}^* - \theta_\mathsf{s}^*)\right)^2\right] = (\theta_\mathsf{t}^* - \theta_\mathsf{s}^*)^\top \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}\left[xx^\top\right](\theta_\mathsf{t}^* - \theta_\mathsf{s}^*) = \|\theta_\mathsf{s}^* - \theta_\mathsf{t}^*\|_{\Sigma_\mathsf{t}}^2. \tag{A.5}$$

The second term looks quite similar to the bias term, B, in Bartlett et al. [6] and Tsigler and Bartlett [128].

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top(\theta_\mathsf{s}^* - \widehat{\theta}(X\theta_\mathsf{s}^*))\right)^2\right] \\
&\quad= \left(\theta_\mathsf{s}^* - \widehat{\theta}\,(X\theta_\mathsf{s}^*)\right) \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}[xx^\top]\left(\theta_\mathsf{s}^* - \widehat{\theta}\,(X\theta_\mathsf{s}^*)\right) \\
&\quad= \|\theta_\mathsf{s}^* - \widehat{\theta}\,(X\theta_\mathsf{s}^*)\,\|_{\Sigma_\mathsf{t}}^2.
\end{aligned} \tag{A.6}$$

The key difference with the standard supervised setting is that now the quantitiy in the middle is Σ_t, not Σ_s. Equivalently, the norm on $\theta_\mathsf{s}^* - \widehat{\theta}(X\theta_\mathsf{s}^*)$ is induced by Σ_t rather than Σ_s.

And finally, the third term is similar to the variance term, C, in Bartlett et al. [6]:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})\right)^2\right] &= \widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})^\top \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}[xx^\top]\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s}) \\
&= \widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})^\top\Sigma_\mathsf{t}\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s}) \\
&= \|\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})\|_{\Sigma_\mathsf{t}}^2.
\end{aligned} \tag{A.7}$$

As in the bias term, the only difference is that the middle term is Σ_t rather than Σ_s. Equivalently, the norm on $\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})$ is induced by Σ_t rather than Σ_s.

Putting it all together, we get the following upper bound for the excess risk of the minimum-norm interpolator on the training data,

$$R(\widehat{\theta}(\boldsymbol{y}_\mathsf{s})) \leq 4\|\theta_\mathsf{s}^* - \theta_\mathsf{t}^*\|_{\Sigma_\mathsf{t}}^2 + 4\|\theta_\mathsf{s}^* - \widehat{\theta}(X\theta_\mathsf{s}^*)\|_{\Sigma_\mathsf{t}}^2 + 2\|\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})\|_{\Sigma_\mathsf{t}}^2.$$

This completes the upper bound for the risk.

For the lower bound, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_\mathsf{s}} R(\widehat{\theta}(\boldsymbol{y}_\mathsf{s})) &= \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_\mathsf{s}, x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top\theta^*_\mathsf{t} - x^\top\widehat{\theta}(\boldsymbol{y}_\mathsf{s})\right)^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_\mathsf{s}, x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top(\theta^*_\mathsf{t} - \widehat{\theta}(X\theta^*_\mathsf{s})) - x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})\right)^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_\mathsf{s}, x\sim\mathcal{D}_\mathsf{t}}\Bigg[\left(x^\top(\theta^*_\mathsf{t} - \widehat{\theta}(X\theta^*_\mathsf{s}))\right)^2 - 2\cdot x^\top(\theta^*_\mathsf{t} - \widehat{\theta}(X\theta^*_\mathsf{s}))\cdot x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s}) \\
&\qquad + \left(x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})\right)^2\Bigg] \\
&\overset{(i)}{=} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top(\theta^*_\mathsf{t} - \widehat{\theta}(X\theta^*_\mathsf{s}))\right)^2\right] + \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_\mathsf{s}, x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})\right)^2\right]
\end{aligned}$$

The equality (i) uses that, conditional on X, $\boldsymbol{\varepsilon}_\mathsf{s}$ is zero-mean. Note that the second term above is just $\mathbb{E}_{\boldsymbol{\varepsilon}_\mathsf{s}}\|\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})\|^2_{\Sigma_\mathsf{t}}$, so we need only deal with the first term. Adding and subtracting θ^*_s inside the square and expanding, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top(\theta^*_\mathsf{t} - \widehat{\theta}(X\theta^*_\mathsf{s}))\right)^2\right] \\
&\quad= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top(\theta^*_\mathsf{t} - \theta^*_\mathsf{s})\right)^2\right] + \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top(\theta^*_\mathsf{s} - \widehat{\theta}(X\theta^*_\mathsf{s}))\right)^2\right] \\
&\qquad + 2\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}\left[(\theta^*_\mathsf{t} - \theta^*_\mathsf{s})^\top xx^\top(\theta^*_\mathsf{s} - \widehat{\theta}(X\theta^*_\mathsf{s}))\right] \\
&\quad= \|\theta^*_\mathsf{t} - \theta^*_\mathsf{s}\|^2_{\Sigma_\mathsf{t}} + \|\theta^*_\mathsf{s} - \widehat{\theta}(X\theta^*_\mathsf{s})\|^2_{\Sigma_\mathsf{t}} + 2(\theta^*_\mathsf{t} - \theta^*_\mathsf{s})^\top\Sigma_\mathsf{t}(\theta^*_\mathsf{s} - \widehat{\theta}(X\theta^*_\mathsf{s})).
\end{aligned}$$

□

A.4 Overview of variance and bias proof techniques

The central pillar of both proofs is controlling the eigenvalues of A_k, which in turn provides certain bounds on the eigenvalues of A and A_{-i}. A key finding of Bartlett et al. [6] is that once ρ_k is large enough, all eigenvalues of A_k are identical up to a constant factor. Specifically,

$$z^T Az \approx n^2\lambda_{k+1}\rho_k, \qquad\qquad z^T A^{-1}z \approx n(n\lambda_{k+1}\rho_k)^{-1}.$$

Variance

Due to independence between the components of $\boldsymbol{\varepsilon}_{\mathsf{s}}$, the variance term from Eqn. 2.6 can be expressed as

$$\begin{aligned} V &= \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}}[V_{\varepsilon_{\mathsf{s}}}/v_{\varepsilon}^2] \\ &= \mathrm{tr}(A^{-1}X\tilde{\Sigma}X^{\top}A^{-1}) \\ &= \sum_{i=1}^{p} \tilde{\lambda}_i \lambda_i z_i^T A^{-2} z_i. \end{aligned}$$

Now that we are dealing with a sum of quadratic forms, we consider the first k^* signal and last $p - k^*$ noise components separately. Using the Sherman-Morrison formula the former can be written as

$$\begin{aligned} \sum_{i \le k^*} \tilde{\lambda}_i \lambda_i z_i^T A^{-2} z_i &= \sum_{i \le k^*} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i^2 z_i^T A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^T A_{-i}^{-1} z_i)^2} \\ &\approx \sum_{i \le k^*} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i^2 n (n\lambda_{k+1}\rho_k)^{-2}}{\lambda_i^2 n^2 (n\lambda_{k+1}\rho_k)^{-2}} \\ &= \sum_{i \le k^*} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{1}{n}, \end{aligned}$$

where $\lambda_i z_i^T A_{-i}^{-1} z_i$ dominates 1 for $i \le k^*$. For the sum over the noise components the 1 in the denominator dominates the other term and so we directly analyze the tail contributions as,

$$\sum_{i > k^*} \frac{\tilde{\lambda}_i}{\lambda_i} \lambda_i^2 z_i^T A^{-2} z_i \approx \sum_{i > k^*} \frac{\tilde{\lambda}_i}{\lambda_i} \lambda_i^2 n (n\lambda_{k+1}\rho_k)^{-2}.$$

The result is that the variance term is upper and lower bounded by

$$\frac{1}{n} \sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} + \sum_{i > k} \frac{\tilde{\lambda}_i}{\lambda_i} \left(\frac{\lambda_i^2}{n \lambda_{k+1}^2 \rho_k^2} \right)$$

times constant factors.

Bias

As in Eqn. 2.4, the bias term is given by

$$\begin{aligned} B_2 &= \|\theta^*_{\mathsf{s}} - X^T A^{-1} X \theta^*_{\mathsf{s}}\|^2_{\Sigma_{\mathsf{t}}} \\ &= \mathrm{tr}(\theta^{*T}_{\mathsf{s}}(I - X^T A^{-1} X)\Sigma_{\mathsf{t}}(I - X^T A^{-1} X)\theta^*_{\mathsf{s}}) \\ &\leq \mathrm{tr}(\theta^*_{\mathsf{s}}\theta^{*T}_{\mathsf{s}}) \cdot \mathrm{tr}((I - X^T A^{-1} X)\Sigma_{\mathsf{t}}(I - X^T A^{-1} X)) \\ &= \|\theta^*_{\mathsf{s}}\|^2 \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \sum_{j=1}^{p} \left(e_i[j] - \sqrt{\lambda_i \lambda_j} z_i^\top A^{-1} z_j\right)^2, \end{aligned}$$

where we use the Cauchy-Scharwz inequality to separate the parameter vector from the quadratic form. A quick application of the Sherman-Morrison formula allows us to write

$$B_2 \leq \|\theta^*_{\mathsf{s}}\|^2 \sum_{i=1}^{p} \tilde{\lambda}_i \frac{1}{1 + \lambda_i z_i^T A_{-i}^{-1} z_i}.$$

From here, we once again exert control over the eigenvalues of A_{-i} to get

$$\frac{1}{1 + \lambda_i z_i^T A_{-i}^{-1} z_i} \approx \frac{1}{1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k}},$$

which completes the upper bound proof sketch.

Note that the looseness of the bias bounds largely stems from the application of the Cauchy-Schwarz inequality. The only situations in which the bound becomes an equality are when

$$c\theta^*_{\mathsf{s}} = (I - X^T A^{-1} X)\Sigma_{\mathsf{t}}^{1/2}$$

for some scalar $c \in \mathbb{R}$ or when θ^*_{s} is the zero vector.

Between the upper and lower bounds, the latter is likely tighter due to the use of the *PriorSigns* assumption. As detailed in Appendix A.6, it allows us to write

$$B \geq \theta^{*T}_{\mathsf{s}}(I - \mathrm{diag}(X^T A^{-1} X))\Sigma_{\mathsf{t}}(I - \mathrm{diag}(X^T A^{-1} X))\theta^*_{\mathsf{s}},$$

where for a matrix $Q \in \mathbb{R}^{m \times m}$, we use $\mathrm{diag}(Q) \in \mathbb{R}^{m \times m}$ to denote zeroed off-diagonal entries. The contribution of the off-diagonal entries is non-negative and dominated by the diagonals, so they can be dropped in the lower bound while preserving tightness under the *PriorSigns* assumption. In general, non-negative terms cannot be discarded in the proof of an upper bound, so we resort to the Cauchy-Schwarz inequality in order to avoid addressing the off-diagonals directly. However, decoupling the model vector θ^*_{s} from the matrix $(I - X^T A^{-1} X)\Sigma_{\mathsf{t}}^{1/2}$ introduces another degree of looseness, contributing to the gap between our bounds. Improving our upper bound will require controlling the off-diagonals of this matrix product with a technique more appropriate than Cauchy-Schwarz.

A.5 Proof of variance bounds

Theorem 2.5. *(Upper and lower bounds for the variance term) There exist universal constants $b, c_1 > 1$ given in Lemma A.2, a universal constant c_2 given in Lemma A.8 and a constant $c > 1$ that only depends on σ_x, c_1, c_2, such that for $k \in (0, n/c)$, with probability at least $1 - 10e^{-n/c}$,*

$$V \geq \frac{1}{cn} \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \min\left(1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k + 1)^2}\right) =: \underline{V}. \tag{2.8}$$

If in addition $\rho_k \geq b$, with probability at least $1 - 7e^{-n/c}$,

$$V/c \leq \frac{1}{n} \sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} + n \frac{\sum_{i>k} \tilde{\lambda}_i \lambda_i}{(\sum_{i>k} \lambda_i)^2} =: \overline{V}/c. \tag{2.9}$$

Proof. We derive the variance terms necessary here and finish the proof of the upper bound in Appendix A.5 and the lower bound in Appendix A.5.

We follow the proof techniques in Bartlett et al. [6] and Tsigler and Bartlett [128]. Observe that we can express the variance term as follows,

$$\begin{aligned} V &= \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}}[V_{\varepsilon_{\mathsf{s}}}/v_\varepsilon^2] \\ &= \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}}[\|X^\top (XX^\top)^{-1} \varepsilon_{\mathsf{s}}\|_{\Sigma_{\mathsf{t}}}^2 / v_\varepsilon^2]. \end{aligned}$$

Defining $A = XX^\top$,

$$\begin{aligned} V &= \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}}[\|X^\top A^{-1} \varepsilon_{\mathsf{s}}\|_{\Sigma_{\mathsf{t}}}^2 / v_\varepsilon^2] \\ &= \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}}[(\varepsilon_{\mathsf{s}}^\top A^{-1} X \Sigma_{\mathsf{t}} X^\top A^{-1} \varepsilon_{\mathsf{s}}) / v_\varepsilon^2] \\ &= \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}}[\mathrm{tr}(\varepsilon_{\mathsf{s}}^\top A^{-1} X \Sigma_{\mathsf{t}} X^\top A^{-1} \varepsilon_{\mathsf{s}}) / v_\varepsilon^2]. \end{aligned}$$

Using the trace trick,

$$\begin{aligned} V &= \mathrm{tr}(A^{-1} X \Sigma_{\mathsf{t}} X^\top A^{-1} \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}}[\varepsilon_{\mathsf{s}} \varepsilon_{\mathsf{s}}^\top]) / v_\epsilon^2 \\ &= \mathrm{tr}(A^{-1} X \Sigma_{\mathsf{t}} X^\top A^{-1} v_\epsilon^2 I_n) / v_\epsilon^2 \\ &= \mathrm{tr}(A^{-1} X \Sigma_{\mathsf{t}} X^\top A^{-1}) \\ &= \mathrm{tr}(X \Sigma_{\mathsf{t}} X^\top A^{-2}) \\ &= \mathrm{tr}((\sum_{i=1}^{p} \tilde{\lambda}_i x^i (x^i)^\top) A^{-2}) \\ &= \mathrm{tr}((\sum_{i=1}^{p} \tilde{\lambda}_i \lambda_i z_i z_i^\top) A^{-2}) \end{aligned}$$

where $x^i \in \mathbb{R}^n$ and $\frac{x^i}{\sqrt{\lambda_i}} = z_i \in \mathbb{R}^n$ are columns of $X \in \mathbb{R}^{n\times p}$ and $X\Sigma_{\mathsf{s}}^{-1/2} \in \mathbb{R}^{n\times p}$, respectively. Continuing the calculation, we have that

$$\begin{aligned} V &= \sum_{i=1}^{p} \tilde{\lambda}_i \lambda_i \mathrm{tr}(z_i^T A^{-2} z_i) \\ &= \sum_{i=1}^{p} \tilde{\lambda}_i \lambda_i \mathrm{tr}(z_i^T (A_{-i} + z_i z_i^T \lambda_i)^{-2} z_i) \\ &= \sum_{i=1}^{p} \tilde{\lambda}_i \lambda_i \frac{z_i^T A_{-i}^{-2} z_i}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} \end{aligned}$$

where $A_{-i} = XX^T - \lambda_i z_i z_i^T = \sum_{j\neq i} \lambda_j z_j z_j^T$. This expression will serve as the starting point for the variance term, which we will now proceed to upper and lower bound.

□

Upper Bound

After isolating the contribution of $\frac{\tilde{\lambda}_i}{\lambda_i}$, most of the components of this proof are as given in the proof of Lemma 6 in Bartlett et al. [6]. For completeness, we replicate them here and refer the reader to their paper for further details and intuitions.

We start by separating the variance term into the top k components and the bottom $p-k$ components as follows,

$$V = \sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i^2 z_i^T A_{-i}^{-2} z_i}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} + \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} (\lambda_i^2 z_i^T A^{-2} z_i).$$

Fix constants $b, c_1 \geq 1$ as defined in Lemma A.2. Then, with probability $1 - 2e^{-n/c_1}$, if $\rho_k \geq b$ then for all $z \in \mathbb{R}^n$ and $i \in [1,k]$,

$$\begin{aligned} z_i^T A_{-i}^{-2} z_i &\leq \mu_1(A_{-i}^{-2}) \|z_i\|^2 \\ &\leq \mu_n(A_{-i})^{-2} \|z_i\|^2 \\ &\leq \frac{c_1^2 \|z_i\|^2}{(n\lambda_{k+1}\rho_k)^2} \end{aligned}$$

and on the same event,

$$\begin{aligned} z_i^T A_{-i}^{-1} z_i &\geq (\Pi_{\mathcal{L}_i} z_i)^T A_{-i}^{-1} (\Pi_{\mathcal{L}_i} z_i) \\ &\geq \mu_n(A_{-i}^{-1}) \|\Pi_{\mathcal{L}_i} z_i\|^2 \\ &\geq \mu_{k+1}(A_{-i})^{-1} \|\Pi_{\mathcal{L}_i} z_i\|^2 \\ &\geq \frac{\|\Pi_{\mathcal{L}_i} z_i\|^2}{n c_1 \lambda_{k+1} \rho_k} \end{aligned}$$

where $\Pi_{\mathcal{L}_i}$ is the orthogonal projection onto the span of the bottom $n-k$ eigenvectors of A_{-i}. It is important to use the projection onto the bottom eigenvectors of A_{-i} in lower bounding the denominator term because we have to use the fact that $\mu_n(A_{-i}^{-1}) \geq \mu_1(A_{-i})^{-1}$. When we don't do the projection, then z_i is affected by all of A_{-i} and so the largest eigenvalue that affects this expression is $\mu_1(A_{-i}) = \lambda_1$. After doing this projection, we no longer have contributions from the top k eigenvectors / eigenvalues in the summation of $z_i^T A_{-i}^{-1} z_i$. Therefore, the largest eigenvalue that affects this summation is now λ_{k+1} instead of λ_1, and so we can use this in our lower bound instead, as desired.

Putting it together, for $i \leq k$,

$$\begin{aligned}\frac{\lambda_i^2 z_i^T A_{-i}^{-2} z_i}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} &\leq \frac{z_i^T A_{-i}^{-2} z_i}{(z_i^T A_{-i}^{-1} z_i)^2} \\ &\leq c_1^4 \frac{\|z_i\|^2}{\|\Pi_{\mathcal{L}_i} z_i\|^4}.\end{aligned}$$

We now invoke Corollary A.9 with a union bound over k events. Let $t < n/c_0$ and $k \in (0, n/c)$ for $c > c_0$ and c_0 sufficiently large. Since $k < n/c$ we also satisfy the union bound condition that $\ln(k) < n/c$. Then, with probability at least $1 - 3e^{-t}$,

$$\begin{aligned}\|z_i\|^2 &\leq c_2 n \\ \|\Pi_{\mathcal{L}_i} z_i\|^2 &\geq n/c_3\end{aligned}$$

for constants c_2, c_3 that only depend on σ_x, c_0, and a universal constant c as defined in Corollary A.8.

Altogether, with probability $1 - 5e^{-n/c_0}$ for c_0 sufficiently large,

$$\begin{aligned}\sum_{i=1}^k \frac{\tilde{\lambda}_i}{\lambda_i}\left(\frac{\lambda_i^2 z_i^T A_{-i}^{-2} z_i}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2}\right) &\leq \sum_{i=1}^k \frac{\tilde{\lambda}_i}{\lambda_i} c_1^4 \frac{\|z_i\|^2}{\|\Pi_{\mathcal{L}_i} z_i\|^4} \\ &\leq \sum_{i=1}^k \frac{\tilde{\lambda}_i}{\lambda_i} c_1^4 \frac{c_2 c_3^2}{n} \\ &= c_4 \sum_{i=1}^k \frac{\tilde{\lambda}_i}{\lambda_i} \frac{1}{n}.\end{aligned}$$

On the same event we use to bound $\mu_{k+1}(A_{-i})$ via Lemma A.2, we also have that $\mu_1(A^{-2}) \leq \mu_n(A)^{-2}$. As such,

$$\sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i}(\lambda_i^2 z_i^T A^{-2} z_i) \leq \frac{c_1^2 \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \lambda_i^2 \|z_i\|^2}{(n\lambda_{k+1}\rho_k)^2}.$$

Then by Corollary A.11, there is a universal constant a such that with probability at least $1 - 2e^{-t}$ for $t < n/c_0$ and $c_0 > a^{-1}$,

$$\begin{aligned}
\sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \lambda_i^2 \|z_i\|^2 &\leq \sigma_x^2 \max(t \max_{i>k}(\tilde{\lambda}_i \lambda_i), \sqrt{t \sum_{i>k} (\tilde{\lambda}_i \lambda_i)^2}) \\
&\leq n \sum_{i>k} \tilde{\lambda}_i \lambda_i + \sigma_x^2 \max(t \max_{i>k}(\tilde{\lambda}_i \lambda_i), \sqrt{tn \sum_{i>k} (\tilde{\lambda}_i \lambda_i)^2}) \\
&\leq n \sum_{i>k} \tilde{\lambda}_i \lambda_i + \sigma_x^2 \max(t \sum_{i>k} \tilde{\lambda}_i \lambda_i, \sqrt{tn} \sum_{i>k} \tilde{\lambda}_i \lambda_i) \\
&\leq c_5 n \sum_{i>k} \tilde{\lambda}_i \lambda_i \\
&= c_5 n \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \lambda_i^2.
\end{aligned}$$

Altogether,

$$\begin{aligned}
\sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} (\lambda_i^2 z_i^T A^{-2} z_i) &\leq \frac{c_1^2 \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \lambda_i^2 \|z_i\|^2}{(n \lambda_{k+1} \rho_k)^2} \\
&\leq \frac{c_1^2 c_5 n}{(n \lambda_{k+1} \rho_k)^2} \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \lambda_i^2 \\
&= c_6 \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \left(\frac{\lambda_i^2}{n \lambda_{k+1}^2 \rho_k^2} \right).
\end{aligned}$$

By taking $c > \max(c_0, c_4, c_6)$ we have that with probability $1 - 7e^{-n/c}$,

$$\begin{aligned}
V &\leq c \left(\sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{1}{n} + \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \left(\frac{\lambda_i^2}{n \lambda_{k+1}^2 \rho_k^2} \right) \right) \\
&= \frac{1}{n} \sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} + n \frac{\sum_{i>k} \tilde{\lambda}_i \lambda_i}{(\sum_{i>k} \lambda_i)^2}.
\end{aligned}$$

Lower Bound

Recall that the variance takes the form,

$$V = \sum_{i=1}^{p} \tilde{\lambda}_i \lambda_i \frac{z_i^T A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^T A_{-i}^{-1} z_i)^2}.$$

By Cauchy-Schwartz,

$$(z_i^T A_{-i}^{-1} z_i)^2 = |\langle z_i, A_{-i}^{-1} z_i \rangle|^2 \leq \|z_i\|^2 \cdot (z_i^T A_{-i}^{-2} z_i).$$

We plug this identity into our lower bound, and further multiply by $\frac{\lambda_i}{\lambda_i}$, resulting in

$$\begin{aligned}
V &= \sum_{i=1}^{p} \tilde{\lambda}_i \lambda_i \frac{z_i^T A_{-i}^{-2} z_i}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} \\
&= \sum_{i=1}^{p} (\frac{\tilde{\lambda}_i}{\lambda_i}) \frac{\lambda_i^2 z_i^T A_{-i}^{-2} z_i}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} \\
&\geq \sum_{i=1}^{p} (\frac{\tilde{\lambda}_i}{\lambda_i}) \frac{\lambda_i^2 (z_i^T A_{-i}^{-1} z_i)^2}{||z_i||^2 (1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} \\
&= \sum_{i=1}^{p} (\frac{\tilde{\lambda}_i}{\lambda_i}) \frac{1}{||z_i||^2 (1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2 (\lambda_i z_i^T A_{-i}^{-1} z_i)^{-2}} \\
&= \sum_{i=1}^{p} (\frac{\tilde{\lambda}_i}{\lambda_i}) \frac{1}{||z_i||^2 (1+(\lambda_i z_i^T A_{-i}^{-1} z_i)^{-1})^2}.
\end{aligned}$$

Then, let $k \in (0, n)$ and $\mathscr{L}_i$ be the span of the bottom $n-k$ eigenvectors of A_{-i} and $\Pi_{\mathscr{L}_i}$ be the projection onto the orthogonal complement of $\mathscr{L}_i$. We have that

$$\begin{aligned}
z_i^T A_{-i}^{-1} z_i &\geq (\Pi_{\mathscr{L}_i} z_i)^T A_{-i}^{-1} (\Pi_{\mathscr{L}_i} z_i) \\
&\geq \|\Pi_{\mathscr{L}_i} z_i\|^2 \mu_{k+1}(A_{-i})^{-1}.
\end{aligned}$$

From Lemma A.2, there is a constant $c_1 \geq 1$, such that for any $k \geq 0$, with probability $1 - 2e^{-n/c_1}$, $\mu_{k+1}(A_{-i}) \leq c_1(\sum_{j>k} \lambda_j + \lambda_{k+1} n)$. Additionally, by Corollary A.9, let $t < n/c_3$ and $k \in (0, n/c)$ for $c > c_3$ and c_3 sufficiently large. Then, with probability at least $1 - 3e^{-t}$

$$\|\Pi_{\mathscr{L}_i} z_i\|^2 \geq n/c_4$$

where c_4 only depends on c_3, σ_x and the universal constant given in Corollary A.8.

Then, for $c \geq \max\{c_1, c_3\}$, with probability $1 - 5e^{-n/c}$,

$$\begin{aligned}
z_i^T A_{-i}^{-1} z_i &\geq \|\Pi_{\mathscr{L}_i} z_i\|^2 \mu_{k+1}(A_{-i})^{-1} \\
&\geq \frac{n}{c_4(\sum_{j>k} \lambda_j + \lambda_{k+1} n)}.
\end{aligned}$$

By again applying Corollary A.9 on the same event we have

$$\|z_i\|^2 \leq c_5 n.$$

where c_5 has the same dependencies as c_4.

Altogether, we have for each i, with probability $1 - 5e^{-n/c}$,

$$\begin{aligned}\frac{1}{||z_i||^2(1+(\lambda_i z_i^T A_{-i}^{-1} z_i)^{-1})^2} &\geq \frac{1}{c_5 n(1+(\frac{c_4(\sum_{j>k}\lambda_j + \lambda_{k+1}n)}{\lambda_i n}))^2} \\ &= \frac{1}{c_5 n(1+\frac{c_4\lambda_{k+1}}{\lambda_i}(\frac{\sum_{j>k}\lambda_j}{\lambda_{k+1}n}+1))^2} \\ &= \frac{1}{c_5 c_4^2 n(1/c_4+\frac{\lambda_{k+1}}{\lambda_i}(\rho_k+1))^2} \\ &\geq \frac{1}{c_6 n(1+\frac{\lambda_{k+1}}{\lambda_i}(\rho_k+1))^2}\end{aligned}$$

where $c_6 = c_5 c_4^2$ and $c > \max\{c_1, c_3\}$ as defined above.

Finally, we invoke Lemma A.10 and that $1/(a+b)^2 \geq \min(a^{-2}, b^{-2})/4$ to get that, with probability $1 - 10e^{-n/c}$,

$$V \geq \frac{1}{8c_6 n}\sum_{i=1}^{p}\frac{\tilde{\lambda}_i}{\lambda_i}\min(1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}).$$

For $c_7 \geq \max\{8c_6, c\}$ we have that with probability $1 - 10e^{-n/c_7}$,

$$V \geq \frac{1}{c_7 n}\sum_{i=1}^{p}\frac{\tilde{\lambda}_i}{\lambda_i}\min(1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}).$$

A.6 Proof of bias bounds

Theorem 2.6. *(Upper and lower bounds for the bias term) For the lower bound only, assume that random models $\overline{\theta}$ are obtained from the underlying θ^*_{s} as $(\overline{\theta})_i = \gamma_i(\theta^*_{\mathsf{s}})_i$, where each γ_i is an independent Rademacher random variable. There exists a universal constant $b > 1$, constants c, C that depend only on b and σ_x, and $k < n/C$ such that if $\rho_k \geq b$, then with probability at least $1 - 10e^{-n/c}$,*

$$\underset{\overline{\theta}}{\mathbb{E}}[B_2] \geq \frac{1}{c}\left(\sum_{i=1}^{k}\frac{\tilde{\lambda}_i}{\lambda_i}\frac{\lambda_i(\theta^*_{\mathsf{s}})_i^2}{(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k})^2} + \sum_{i>k}\tilde{\lambda}_i(\theta^*_{\mathsf{s}})_i^2\right) =: \underline{B_2}.$$

If we assume that p is at most exponential in n, then with probability at least $1 - 5e^{-n/c}$,

$$B_2/c \leq \|\theta^*_{\mathsf{s}}\|^2\sum_{i=1}^{p}\frac{\tilde{\lambda}_i}{\lambda_i}\frac{\lambda_i}{\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)} =: \overline{B_2}/c.$$

Upper Bound

Proof. As defined in Eqn. 2.4,

$$\begin{aligned} B_2 &= \|\theta^*_{\mathsf{s}} - \widehat{\theta}(X\theta^*_{\mathsf{s}})\|^2_{\Sigma_{\mathsf{t}}} \\ &= \|\theta^*_{\mathsf{s}} - X^T A^{-1} X \theta^*_{\mathsf{s}}\|^2_{\Sigma_{\mathsf{t}}} \\ &= \theta^{*T}_{\mathsf{s}} (I - X^T A^{-1} X) \Sigma_{\mathsf{t}} (I - X^T A^{-1} X) \theta^*_{\mathsf{s}}. \end{aligned} \tag{A.8}$$

The i^{th} row of $I_p - X^T A^{-1} X$ is given by $e_i - \sqrt{\lambda_i} z_i^T A^{-1} X$. It follows that

$$\begin{aligned} (\theta^*)^T M \theta^* &= \begin{pmatrix} \vdots \\ \sum_{j=1}^p \theta_j (e_i[j] - \sqrt{\lambda_i \lambda_j} z_i^\top A^{-1} z_j) \\ \vdots \end{pmatrix}^\top \Sigma_{\mathsf{t}} \begin{pmatrix} \vdots \\ \sum_{j=1}^p \theta_j (e_i[j] - \sqrt{\lambda_i \lambda_j} z_i^\top A^{-1} z_j) \\ \vdots \end{pmatrix} \quad i^{th} \text{ row shown} \\ &= \sum_{i=1}^p \tilde{\lambda}_i \Big(\sum_{j=1}^p \theta_j (e_i[j] - \sqrt{\lambda_i \lambda_j} z_i^\top A^{-1} z_j) \Big)^2 \\ &\le \sum_{i=1}^p \tilde{\lambda}_i \Big(\sum_{j=1}^p \theta_j^2 \Big) \sum_{j=1}^p \Big(e_i[j] - \sqrt{\lambda_i \lambda_j} z_i^\top A^{-1} z_j \Big)^2 \\ &= \|\theta^*\|^2 \sum_{i=1}^p \tilde{\lambda}_i \sum_{j=1}^p \Big(e_i[j] - \sqrt{\lambda_i \lambda_j} z_i^\top A^{-1} z_j \Big)^2. \end{aligned}$$

Next we look at i^{th} term in the outer sum.

$$\begin{aligned} \tilde{\lambda}_i \sum_{j=1}^p (e_i[j] - \sqrt{\lambda_i \lambda_j} z_i^\top A^{-1} z_j)^2 &= \tilde{\lambda}_i (1 - \lambda_i z_i^\top A^{-1} z_i)^2 + \tilde{\lambda}_i \sum_{j \neq i} \lambda_i \lambda_j (z_i^\top A^{-1} z_j)^2 \\ &= \tilde{\lambda}_i (1 - 2\lambda_i z_i^T A^{-1} z_i + \lambda_i^2 (z_i^T A^{-1} z_i)^2 + \sum_{j \neq i} \lambda_i \lambda_j (z_i^\top A^{-1} z_j)^2) \\ &= \tilde{\lambda}_i (1 - 2\lambda_i z_i^T A^{-1} z_i + \sum_{i=1}^p \lambda_i \lambda_j (z_i^\top A^{-1} z_j)^2) \\ &= \tilde{\lambda}_i (1 - 2\lambda_i z_i^T A^{-1} z_i + \lambda_i z_i^\top A^{-1} \Big(\sum_{i=1}^p \lambda_j z_j z_j^T \Big) A^{-1} z_i) \\ &= \tilde{\lambda}_i \left(1 - 2\lambda_i z_i^T A^{-1} z_i + \lambda_i z_i^\top A^{-1} A A^{-1} z_i \right) \\ &= \tilde{\lambda}_i \left(1 - 2\lambda_i z_i^T A^{-1} z_i + \lambda_i z_i^\top A^{-1} z_i \right) \\ &= \tilde{\lambda}_i \left(1 - \lambda_i z_i^T A^{-1} z_i \right). \end{aligned}$$

Using the Sherman-Morrison formula, we get that

$$\begin{aligned}
1 - \lambda_i z_i^T A^{-1} z_i &= 1 - \lambda_i z_i^T \left(A_{-i} + \lambda_i z_i z_i^T\right)^{-1} z_i \\
&= 1 - \lambda_i z_i^T \left(A_{-i}^{-1} - \lambda_i A_{-i}^{-1} z_i (1 + \lambda_i z_i^T A_{-i}^{-1} z_i)^{-1} z_i^T A_{-i}^{-1}\right) z_i \\
&= 1 - \lambda_i z_i^T A_{-i}^{-1} z_i + \frac{(\lambda_i z_i^T A_{-i}^{-1} z_i)^2}{1 + \lambda_i z_i^T A_{-i}^{-1} z_i} \\
&= \frac{1}{1 + \lambda_i z_i^T A_{-i}^{-1} z_i}.
\end{aligned}$$

We now provide an upper bound for the remaining term. Let $\Pi_{\mathcal{L}_i}$ be the orthogonal projection onto the bottom $n-k$ eigenvectors of A_{-i}. By Lemma A.2, there exist constants $b, c_0 \geq 1$ such that if $\rho_k \geq b$, then with probability at least $1 - 2e^{-n/c_0}$,

$$\mu_{k+1}(A_{-i}) \leq c_0 \lambda_{k+1} \rho_k n,$$

so we get

$$\begin{aligned}
1 + \lambda_i z_i^T A_{-i}^{-1} z_i &\geq 1 + \lambda_i (\Pi_{\mathcal{L}_i} z_i)^T A_{-i}^{-1} (\Pi_{\mathcal{L}_i} z_i) \\
&\geq 1 + \frac{\lambda_i \left\|\Pi_{\mathcal{L}_i} z_i\right\|^2}{c_0 \lambda_{k+1} n \rho_k}.
\end{aligned}$$

By Corollary A.9, there exist constants c_1 and c_2 with $c_2 > c_1$ and c_1 sufficiently large such that for $0 < k < n/c_2$, we have with probability at least $1 - 3e^{-n/c_1}$,

$$\left\|\Pi_{\mathcal{L}_i} z_i\right\|^2 \geq n/c_3,$$

where c_3 depends only on c_1 and σ.

Plugging these in gives us with probability at least $1 - 5e^{-n/c_4}$,

$$\begin{aligned}
\tilde{\lambda}_i \left(1 - \lambda_i z_i^T A^{-1} z_i\right) &\leq \frac{\tilde{\lambda}_i}{\left(1 + \frac{c_5^2 \lambda_i}{\lambda_{k+1} \rho_k}\right)} \\
&= \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i}{\left(1 + \frac{c_5^2 \lambda_i}{\lambda_{k+1} \rho_k}\right)},
\end{aligned}$$

where $c_4 = \max(c_0, c_1)$ and $c_5 = \min(c_0, c_3)$.

Therefore by union bound over the application of Corollary A.9,

$$\begin{aligned}
B &\leq \|\theta^*\|^2 \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i}{\left(1 + \frac{c_5^2 \lambda_i}{\lambda_{k+1} \rho_k}\right)} \\
&\leq \frac{1}{c_6} \|\theta^*\|^2 \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i}{\left(1 + \frac{\lambda_i}{\lambda_{k+1} \rho_k}\right)},
\end{aligned}$$

where $c_6 = \min(c_5^2, 1)$. Taking $c = \max(c_6^{-1}, c_4)$ gives us the result.

□

Lower Bound

After isolating the contribution of $\frac{\tilde{\lambda}_i}{\lambda_i}$, many of the components of this proof are as given in Tsigler and Bartlett [128]. For completeness, we replicate them here.

Proof. Assume that the vector θ^*_{s} is randomly distributed according to the PriorSigns($\overline{\theta}_{\mathsf{s}}$) assumption. Using Lemma A.12, the bias term can be rewritten as

$$\begin{aligned}
B &= \mathop{\mathbb{E}}_{\theta^*_{\mathsf{s}}}[B_{\theta^*_{\mathsf{s}}}] \\
&= \mathop{\mathbb{E}}_{\theta^*_{\mathsf{s}}}[\|\theta^*_{\mathsf{s}} - \widehat{\theta}(X\theta^*_{\mathsf{s}})\|^2_{\Sigma_{\mathsf{t}}}] \\
&= \mathop{\mathbb{E}}_{\theta^*_{\mathsf{s}}}[(\theta^*_{\mathsf{s}})^T(I_p - X^T(XX^T)^{-1}X)\Sigma_{\mathsf{t}}(I_p - X^T(XX^T)^{-1}X)\theta^*_{\mathsf{s}}] \\
&= \mathop{\mathbb{E}}_{\theta^*_{\mathsf{s}}}[(\theta^*_{\mathsf{s}})^T M\theta^*_{\mathsf{s}}] \\
&= \mathop{\mathbb{E}}_{\theta^*_{\mathsf{s}}}[\theta^*_{\mathsf{s}}]^T M \mathop{\mathbb{E}}_{\theta^*_{\mathsf{s}}}[\theta^*_{\mathsf{s}}] + \mathrm{tr}(M\mathrm{Cov}(\theta^*_{\mathsf{s}})) \\
&= \mathrm{tr}(M\mathrm{Cov}(\theta^*_{\mathsf{s}})).
\end{aligned}$$

where $M = (I_p - X^T(XX^T)^{-1}X)\Sigma_{\mathsf{t}}(I_p - X^T(XX^T)^{-1}X)$. The last equality follows from the assumption $\mathbb{E}_{\theta^*_{\mathsf{s}}}[(\theta^*_{\mathsf{s}})] = 0$. The diagonal elements of $\mathrm{Cov}(\theta^*_{\mathsf{s}})$ are the component-wise variances of θ^*_{s}, which are given by $(\theta^*_{\mathsf{s}})^2_i = (\overline{\theta}_{\mathsf{s}})^2_i$. The off-diagonal elements are 0 since the components of θ^*_{s} are independent. As such, we need only consider the diagonal elements of M.

Note that the i^{th} row of $I_p - X^T(XX^T)^{-1}X$ is equal to $e_i - \sqrt{\lambda_i}z_i^T(XX^T)^{-1}X$, where e_i is the i^{th} vector of the standard orthonormal basis. It follows that the i^{th} diagonal element of M is given by

$$\begin{aligned}
M_{ii} &= \sum_{j=1}^{p} \tilde{\lambda}_j(e_i[j] - \sqrt{\lambda_i\lambda_j}z_i^T A^{-1}z_j)^2 \\
&= \tilde{\lambda}_i(1 - \lambda_i z_i^T A^{-1}z_i)^2 + \sum_{j\neq i} \tilde{\lambda}_j\lambda_i\lambda_j(z_i^T A^{-1}z_j)^2.
\end{aligned}$$

Hence, we can express the bias term as

$$\begin{aligned}
B &= \sum_{i=1}^{p} (\overline{\theta}_{\mathsf{s}})^2_i \big[\tilde{\lambda}_i(1 - \lambda_i z_i^T A^{-1}z_i)^2 + \sum_{j\neq i} \tilde{\lambda}_j\lambda_i\lambda_j(z_i^T A^{-1}z_j)^2\big] \\
&\geq \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i}\lambda_i(\overline{\theta}_{\mathsf{s}})^2_i(1 - \lambda_i z_i^T A^{-1}z_i)^2.
\end{aligned}$$

We are able to eliminate the second term because it is non-negative. Substituting $A = A_{-i} + \lambda_i z_i z_i^T$ and using the Sherman-Morrison identity, we have that $1 - \lambda_i z_i^T A^{-1}z_i = \frac{1}{1+\lambda_i z_i^T A_{-i}^{-1}z_i}$

(see proof of bias upper bound in Appendix A.6). Then,

$$B \geq \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i (\overline{\theta}_{\mathsf{s}})_i^2}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2}$$

Let's bound each term in that sum from below with high probability. By Corollary A.7, there exist constants $b, c_0 \geq 1$ such that for any $i \geq 0$ with probability at least $1 - 2e^{-n/c_0}$, if $\rho_k \geq b$, then

$$\mu_n(A_{-i}) \geq \frac{1}{c_0} n \lambda_{k+1} \rho_k.$$

Next,

$$\frac{\lambda_i}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} \geq \frac{\lambda_i}{(1+\lambda_i \mu_n(A_{-i})^{-1} \left\|z_i\right\|^2)^2}.$$

By Corollary A.9, for constants c_1, c_2 such that $k < n/c_2$ with $c_2 > c_1$ for sufficiently large c_1 with probability at least $1 - 3e^{-n/c_1}$ we have $\left\|z_i\right\|^2 \leq c_3 n$, where c_3 depends only on c_1 and σ.

We obtain that w.p. at least $1 - 5e^{-n/c_4}$,

$$\frac{\lambda_i \bar{\theta}_i^2}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} \geq \frac{\lambda_i \bar{\theta}_i^2}{\left(1 + \frac{c_4^2 \lambda_i}{\lambda_{k+1} \rho_k}\right)^2},$$

where $c_4 = \max(c_0, c_1, c_3)$. All the terms are non-negative so Lemma A.10 provides a lower bound on their sum. With probability at least $1 - 10e^{-n/c_4}$,

$$\begin{aligned} B &\geq \frac{1}{2} \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i \bar{\theta}_i^2}{(1 + \frac{c_4^2 \lambda_i}{\lambda_{k+1} \rho_k})^2} \\ &\geq \frac{1}{c_5} \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i \bar{\theta}_i^2}{(1 + \frac{\lambda_i}{\lambda_{k+1} \rho_k})^2}, \end{aligned}$$

where $c_5 = 2\max(c_4^2, 1)$.

Finally, we notice that on $i > k$ we have $\rho_k \geq b > 1$ and $\lambda_i \leq \lambda_{k+1}$ giving us,

$$\begin{aligned} \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i \bar{\theta}_i^2}{(1 + \frac{\lambda_i}{\lambda_{k+1} \rho_k})^2} &\geq \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i \bar{\theta}_i^2}{(1 + \frac{\lambda_i}{\lambda k+1})^2} \\ &\geq \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i \bar{\theta}_i^2}{4} \\ &= \frac{1}{4} \sum_{i>k} \tilde{\lambda}_i \bar{\theta}_i^2. \end{aligned}$$

Letting $c = 4\max(c_4, c_5)$ gives us the result.

□

A.7 Proof of tightness of bounds

Theorem 2.7. *(Tightness of variance and bias bounds) Let the lower bound and upper bound of V be given by $\underline{V}$ and $\overline{V}$, respectively. There exists a universal constant $b \geq 1$, and constant c as defined in Theorem 2.5, and $k \in (0, n/c)$ such that if $\rho_k \geq b$, then*

$$\underline{V}/\overline{V} \in \left[b^{-2}(1+b)^{-2}/c^2, 1\right].$$

Let the lower bound and upper bound of B_2 be given by $\underline{B_2}$ and $\overline{B_2}$, respectively, and the assumptions of Theorem 2.6 be satisfied. Then

$$\underline{B_2}/\overline{B_2} \in \left[\frac{\min_i \left\{(\theta^*_{\mathsf{s}})_i^2 \ : \ (\theta^*_{\mathsf{s}})_i \neq 0\right\}}{\|\theta^*_{\mathsf{s}}\|^2 \left(1 + b^{-1}\frac{\lambda_1}{\lambda_{k+1}}\right)}, 1\right].$$

Proof. We split the proof into the variance proof in Appendix A.7 and the bias proof in Appendix A.7. □

Variance Proof

Proof. Recall that

$$\underline{V} = \frac{1}{8c_6 n}\sum_{i=1}^{p}\frac{\tilde{\lambda}_i}{\lambda_i}\min\left(1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right)$$
$$\overline{V} = c\left(\sum_{i=1}^{k}\frac{\tilde{\lambda}_i}{\lambda_i}\frac{1}{n} + \sum_{i>k}\frac{\tilde{\lambda}_i}{\lambda_i}\left(\frac{\lambda_i^2}{n\lambda_{k+1}^2\rho_k^2}\right)\right).$$

Since k is the smallest ℓ such that $\rho_\ell \geq b$, it is clear by definition that $\rho_{k-1} < b$. Then we observe that

$$\rho_{k-1} = \frac{1}{n\lambda_k}\sum_{j>k-1}\lambda_j = \frac{\lambda_k + \sum_{j>k}\lambda_j}{n\lambda_k} = \frac{\lambda_k + n\lambda_{k+1}\rho_k}{n\lambda_k} < b$$
$$\therefore \lambda_k + n\lambda_{k+1}\rho_k < nb\lambda_k \Rightarrow \lambda_k > \frac{\lambda_k + n\lambda_{k+1}\rho_k}{nb} > \frac{n\lambda_{k+1}\rho_k}{nb} = \frac{\lambda_{k+1}\rho_k}{b}.$$

On $i \leq k$,

$$\underline{V} : \overline{V} = \frac{1}{8c_6 n}\sum_{i=1}^{k}\frac{\tilde{\lambda}_i}{\lambda_i}\min\left(1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right) : \frac{\tilde{\lambda}_i}{\lambda_i}\frac{c}{n}$$
$$\geq \frac{1}{8c_6 c}\sum_{i=1}^{k}\min\left(1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right) : 1.$$

If the min is 1 then we are okay otherwise, using the identity above and that fact that $\lambda_i \geq \lambda_k$, we have that

$$\begin{aligned}\frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2} &> \frac{(\lambda_{k+1}\rho_k)^2}{b^2\lambda_{k+1}^2(\rho_k+1)^2} \\ &= \frac{\rho_k^2}{b^2(\rho_k+1)^2}.\end{aligned}$$

Examining the ρ_k terms:

$$\begin{aligned}\frac{\rho_k^2}{(\rho_k+1)^2} &= \frac{1}{\rho_k^{-2}(\rho_k+1)^2} \\ &= \frac{1}{(1+\rho_k^{-1})^2}.\end{aligned}$$

As $\rho_k \geq b$ we have that $\rho_k^{-1} \leq b \Rightarrow 1+\rho_k^{-1} \leq 1+b \Rightarrow (1+\rho_k^{-1})^{-2} \geq (1+b)^{-2}$.

Putting it together we get that

$$\begin{aligned}\frac{1}{8c_6c}\sum_{i=1}^{k}\frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2} &\geq \frac{1}{8c_6c}\sum_{i=1}^{k}\frac{1}{b^2(1+b)^2} \\ &= \frac{k}{8c_6c\cdot b^2(1+b)^2} \\ &\geq \frac{1}{8c_6c\cdot b^2(1+b)^2}.\end{aligned}$$

On $i > k$, it is clear that the min is always given by the second term, as $\lambda_i \leq \lambda_{k+1}$, so we get

$$\begin{aligned}\underline{V} : \overline{V} &= \frac{1}{8c_6n}\sum_{i>k}\frac{\tilde{\lambda}_i}{\lambda_i}\min\left(1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right) : c\frac{\tilde{\lambda}_i}{\lambda_i}\frac{\lambda_i^2}{n\lambda_{k+1}^2\rho_k^2} \\ &= \frac{1}{8c_6c}\sum_{i>k}\frac{\rho_k^2}{(\rho_k+1)^2} \\ &\geq \frac{1}{8c_6c}\sum_{i>k}\frac{1}{(1+b)^2} = \frac{1}{8c_6c}\frac{p-k}{(1+b)^2} > \frac{1}{8c_6c(1+b)^2}.\end{aligned}$$

Finally we note that for $b \geq 1$ it is clear that $\min(b^{-2}(1+b)^{-2}, (1+b)^{-2}) = b^{-2}(1+b)^{-2}$. Therefore,

$$\underline{V} : \overline{V} \geq \frac{1}{8c_6c}b^{-2}(1+b)^{-2}.$$

By setting c in the upper bound such that $c > 8c_6$, we get

$$\underline{V} : \overline{V} \geq \frac{1}{c^2}b^{-2}(1+b)^{-2}.$$

□

Bias Proof

Proof. We will bound the ratio of the lower and upper bounds by bounding the ratios of the corresponding terms in each sum. Observe that for all i, the ratio of the terms is equal to

$$\frac{(\theta_i^*)^2}{\|\theta^*\|^2} \cdot \frac{1}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)}.$$

On $i \leq k$,

$$\begin{aligned}&\frac{(\theta_i^*)^2}{\|\theta^*\|^2} \cdot \frac{1}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)} \\ &\geq \min_i \frac{(\theta_i^*)^2}{\|\theta^*\|^2} \cdot \frac{1}{\left(1 + \frac{\lambda_1}{\lambda_{k+1}} b^{-1}\right)}.\end{aligned}$$

On $i > k$, we have $\lambda_i/\lambda_{k+1} \leq 1$, so

$$\begin{aligned}&\frac{(\theta_i^*)^2}{\|\theta^*\|^2} \cdot \frac{1}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)} \\ &\geq \min_i \frac{(\theta_i^*)^2}{\|\theta^*\|^2} \cdot \frac{1}{(1 + b^{-1})}.\end{aligned}$$

Unfortunately, the looseness in the top k components coming from the gap λ_1/λ_{k+1} dominates the tighter ratios in the bottom $p - k$ components which only contain a model-dependent gap, $\min_i \theta_i^2/\|\theta\|^2$. Future work would seek to resolve this and provide tight upper and lower bounds for the bias terms.

□

A.8 Proof of beneficial and malignant shifts

Trace Conditions for Simple Shifts

Let Σ_{s} be any source covariance and define Σ_{t} as $\tilde{\lambda}_i = \alpha\lambda_i$ for $i \leq k$ and $\tilde{\lambda}_i = \beta\lambda_i$ for $i > k$ with $\alpha, \beta \geq 0$.

Then $\operatorname{tr}(\Sigma_{\mathsf{s}}) = \sum_{i=1}^k \lambda_i + \sum_{i>k} \lambda_i$ and $\operatorname{tr}(\Sigma_{\mathsf{t}}) = \alpha(\sum_{i=1}^k \tilde{\lambda}_i) + \beta(\sum_{i>k} \tilde{\lambda}_i)$.

For $\alpha > 1, \beta < 1$, if

$$\frac{\sum_{i>k} \lambda_i}{\sum_{i=1}^k \lambda_i} < \frac{\alpha - 1}{1 - \beta}$$

then we have that $\text{tr}(\Sigma_{\mathsf{s}}) < \text{tr}(\Sigma_{\mathsf{t}})$ and if the inequality is flipped then we obtain $\text{tr}(\Sigma_{\mathsf{s}}) > \text{tr}(\Sigma_{\mathsf{t}})$.

For $\alpha < 1, \beta > 1$, if

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i>k} \lambda_i} < \frac{\beta - 1}{1 - \alpha}$$

then we have that $\text{tr}(\Sigma_{\mathsf{s}}) < \text{tr}(\Sigma_{\mathsf{t}})$ and if the inequality is flipped then we obtain $\text{tr}(\Sigma_{\mathsf{s}}) > \text{tr}(\Sigma_{\mathsf{t}})$.

Proof of Beneficial and Malignant Shifts for Simple Shifts

We restate the theorem for ease.

Theorem 2.10. *(Beneficial and Malignant Multiplicative Shifts on Variance) Let Σ_{s} be a source covariance that satisfies benign source conditions. That is, $\exists\ k$ such that $\rho_k \geq b$ for a universal constant $b > 1$. Define Σ_{t} as $\tilde{\lambda}_i = \alpha\lambda_i$ for $i \leq k$ and $\tilde{\lambda}_i = \beta\lambda_i$ for $i > k$, with $\alpha, \beta \geq 0$.*

1. *If $\alpha < 1, \beta \leq 1$ or $\alpha \leq 1, \beta < 1$ then we obtain a beneficial shift in variance.*
2. *If $\alpha > 1, \beta \geq 1$ or $\alpha \geq 1, \beta > 1$ then we obtain a malignant shift in variance.*
3. *If we are in the mildly overparameterized regime:*
   - *$\alpha > 1$ and $\beta < 1$ leads to beneficial shifts;*
   - *$\alpha < 1$ and $\beta > 1$ leads to malignant shifts.*
4. *If we are in the severely overparameterized regime:*
   - *$\alpha > 1$ and $\beta < 1$ leads to malignant shifts;*
   - *$\alpha < 1$ and $\beta > 1$ leads to beneficial shifts.*

Proof. From Theorem 2.5 and Theorem 2.7, we have that for a universal constant $b > 1$ if $\rho_k \geq b$ we get the following upper and lower bounds on the out-of-distribution variance for some constants c_1, c_2,

$$V_{ood} \leq c_1 \left(\frac{1}{n} \sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} + n \frac{\sum_{i>k} \tilde{\lambda}_i \lambda_i}{(\sum_{i>k} \lambda_i)^2} \right)$$
$$V_{ood} \geq c_2 \left(\frac{1}{n} \sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} + n \frac{\sum_{i>k} \tilde{\lambda}_i \lambda_i}{(\sum_{i>k} \lambda_i)^2} \right).$$

Analogously, the in-distribution variance is upper and lower bounded by,

$$V_{id} \leq c_1 \left(\frac{k}{n} + \frac{n}{R_k} \right)$$
$$V_{id} \geq c_2 \left(\frac{k}{n} + \frac{n}{R_k} \right)$$

where $R_k = (\sum_{i>k} \lambda_i)^2 / \sum_{i>k} \lambda_i^2$.

Let Σ_{s} be any source covariance model that satisfies benign source conditions. Define Σ_{t} by,

$$\tilde{\lambda}_i = \begin{cases} \alpha \lambda_i, & i \leq k \\ \beta \lambda_i, & i > k \end{cases}$$

for $\alpha, \beta \geq 0$.

Beneficial shifts. We use the upper bound to specify requirements for the beneficial shifts.

$$\begin{aligned} V_{ood} &\leq c_1 \left(\alpha \frac{k}{n} + \beta \frac{n}{R_k} \right) \\ &= V_{id} + c_1 \left(\frac{k}{n}(\alpha - 1) + \frac{n}{R_k}(\beta - 1) \right). \end{aligned}$$

Let $\alpha > 1, \beta < 1$. To obtain a beneficial shift in this setting we need,

$$\begin{aligned} \frac{n}{R_k}(1 - \beta) &> \frac{k}{n}(\alpha - 1) \\ \Rightarrow \frac{n}{R_k} &> \frac{k}{n} \left(\frac{\alpha - 1}{1 - \beta} \right). \end{aligned}$$

In the case $\alpha < 1, \beta > 1$, to obtain a beneficial shift we need,

$$\begin{aligned} \frac{n}{R_k}(\beta - 1) &< \frac{k}{n}(1 - \alpha) \\ \Rightarrow \frac{n}{R_k} &< \frac{k}{n} \left(\frac{1 - \alpha}{\beta - 1} \right). \end{aligned}$$

In the case where $\alpha = 1$ then any $\beta < 1$ leads to beneficial shifts. Similarly when $\beta = 1$, any $\alpha < 1$ leads to beneficial shifts.

Malignant shifts. We use the lower bound to specify requirements for the malignant shift.

$$\begin{aligned} V_{ood} &\geq c_2 \left(\alpha \frac{k}{n} + \beta \frac{n}{R_k} \right) \\ &= V_{id} + c_2 \left(\frac{k}{n}(\alpha - 1) + \frac{n}{R_k}(\beta - 1) \right). \end{aligned}$$

Let $\alpha < 1$ and $\beta > 1$. To obtain a malignant shift in this setting we need,

$$\frac{n}{R_k} > \frac{k}{n} \left(\frac{1-\alpha}{\beta - 1} \right).$$

In the case of $\alpha > 1, \beta < 1$, to obtain a malignant shift we need,

$$\frac{n}{R_k} < \frac{k}{n} \left(\frac{\alpha - 1}{1 - \beta} \right).$$

In the case where $\alpha = 1$ then any $\beta > 1$ leads to malignant shifts. Similarly when $\beta = 1$, any $\alpha > 1$ leads to malignant shifts.

Mild and severe overparameterization. We see that the four cases separate into settings in which we are mildly overparameterized, meaning

$$\frac{n}{R_k} > \frac{k}{n} \left| \frac{\alpha - 1}{1 - \beta} \right|,$$

and settings in which we are severely overparamterized, meaning

$$\frac{n}{R_k} < \frac{k}{n} \left| \frac{\alpha - 1}{1 - \beta} \right|.$$

In each of these regimes of overparameterization, the above proof has delineated whether we achieve beneficial or malignant shifts in all settings of α, β. □

Generalized (Necessary) Conditions for Beneficial and Malignant Shifts

Let Σ_{s} be any source covariance matrix that satisfies benign source conditions and define Σ_{t} as

$$\tilde{\lambda}_i = \begin{cases} \alpha_i \lambda_i & i \leq k, \\ \beta_i \lambda_i & i > k \end{cases}$$

with $\alpha_i, \beta_i \geq 0$ for all i.

Then the OOD variance upper bound is given by,

$$\begin{aligned} V_{ood} &\leq c_1 \left(\frac{1}{n} \sum_{i=1}^{k} \alpha_i + n \frac{\sum_{i>k} \beta_i \lambda_i^2}{(\sum_{i>k} \lambda_i)^2} \right) \\ &= V_{id} + c_1 \left(\frac{(\sum_{i=1}^{k} \alpha_i) - k}{n} + n \frac{\sum_{i>k} \lambda_i^2 (\beta_i - 1)}{(\sum_{i>k} \lambda_i)^2} \right) \\ &= V_{id} + c_1 \left(\frac{k}{n} \left(\frac{\sum_{i=1}^{k} \alpha_i}{k} - 1 \right) + \frac{n}{R_k} \left(\frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2} - 1 \right) \right), \end{aligned}$$

and the OOD variance lower bound is given by,

$$\begin{aligned} V_{ood} &\geq c_2 \left(\frac{1}{n} \sum_{i=1}^{k} \alpha_i + n \frac{\sum_{i>k} \beta_i \lambda_i^2}{(\sum_{i>k} \lambda_i)^2} \right) \\ &= V_{id} + c_2 \left(\frac{(\sum_{i=1}^{k} \alpha_i) - k}{n} + n \frac{\sum_{i>k} \lambda_i^2 (\beta_i - 1)}{(\sum_{i>k} \lambda_i)^2} \right) \\ &= V_{id} + c_2 \left(\frac{k}{n} \left(\frac{\sum_{i=1}^{k} \alpha_i}{k} - 1 \right) + \frac{n}{R_k} \left(\frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2} - 1 \right) \right), \end{aligned}$$

where V_{id} is the ID variance bound.

Again, we use the upper bounds to prove conditions for beneficial shifts and the lower bounds to prove conditions for malignant shifts.

Beneficial shifts. From the upper bound we consider two separate cases for non-trivial beneficial shifts:

1. $\sum_{i=1}^{k} \alpha_i < k$ and $\sum_{i>k} \beta_i \lambda_i^2 > \sum_{i>k} \lambda_i^2$,
2. $\sum_{i=1}^{k} \alpha_i > k$ and $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$.

We start with the case of $\sum_{i=1}^{k} \alpha_i < k$ and $\sum_{i>k} \beta_i \lambda_i^2 > \sum_{i>k} \lambda_i^2$. If this is satisfied, the only way to achieve a beneficial shift is if

$$\frac{n}{R_k} \left(\frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2} - 1 \right) < \frac{k}{n} \left(1 - \frac{\sum_{i=1}^{k} \alpha_i}{k} \right). \tag{A.9}$$

We also have in this setting that,

$$0 < 1 - \frac{\sum_{i=1}^{k} \alpha_i}{k} \leq 1.$$

In Equation A.9 we see a notion of severe overparameterization that leads to beneficial shifts. For instance as $R_k \to \infty$ we see the left-hand-side (LHS) of Equation A.9 $\to 0$. So as $R_k \to \infty$ we have that finite n always leads to a beneficial shift in this setting. We note that equivalently if $\beta_i = 1$ for all i then we also have the LHS $\to 0$, just as in the case of severe overparameterization. We will return to the definitions of mild and severe overparameterization for arbitrary shifts after showing the remaining conditions for beneficial and malignant shifts.

Now consider the case of $\sum_{i=1}^{k} \alpha_i > k$ and $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$. If this is satisfied, the only way to achieve a beneficial shift is if

$$\frac{n}{R_k}\left(1 - \frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2}\right) > \frac{k}{n}\left(\frac{\sum_{i=1}^{k} \alpha_i}{k} - 1\right). \tag{A.10}$$

In this setting it is clear that

$$0 < 1 - \frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2} \leq 1.$$

In Equation A.10, it is clear that we have a notion of mild overparameterization that leads to beneficial shifts. As above if $\alpha_i = 1$ for all i then we always obtain a beneficial shift in this setting. Otherwise if R_k does not grow too quickly (as in the case with mild overparameterization) then this is a necessary condition to achieve beneficial shifts when $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$.

Malignant shifts. From the lower bound we once again consider two separate cases for non-trivial malignant shifts:

1. $\sum_{i=1}^{k} \alpha_i > k$ and $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$,
2. $\sum_{i=1}^{k} \alpha_i < k$ and $\sum_{i>k} \beta_i \lambda_i^2 > \sum_{i>k} \lambda_i^2$.

We start with the case of $\sum_{i=1}^{k} \alpha_i > k$ and $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$. If this is satisfied then the only way to achieve a malignant shift is if,

$$\frac{n}{R_k}\left(1 - \frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2}\right) < \frac{k}{n}\left(\frac{\sum_{i=1}^{k} \alpha_i}{k} - 1\right). \tag{A.11}$$

In the case of $\sum_{i=1}^{k} \alpha_i < k$ and $\sum_{i>k} \beta_i \lambda_i^2 > \sum_{i>k} \lambda_i^2$ the only way to achieve a malignant shift is if,

$$\frac{n}{R_k}\left(\frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2} - 1\right) > \frac{k}{n}\left(1 - \frac{\sum_{i=1}^{k} \alpha_i}{k}\right). \tag{A.12}$$

We now are ready to define mild and severe overparameterization for arbitrary multiplicative shifts.

Theorem A.13. *(Mild and severe overparameterization for arbitrary multiplicative shifts) Let Σ_{s} be any source covariance matrix that satisfies benign source conditions, meaning $\exists\ k$ such that $\rho_k \geq b$ for a universal constant $b > 1$. Furthermore, let Σ_{t} be defined by $\tilde{\lambda}_i = \alpha_i \lambda_i$ for $i \leq k$ and $\tilde{\lambda}_i = \beta_i \lambda_i$ for $i > k$.*

We will define

$$C := \left| \left(\frac{\sum_{i=1}^{k} \alpha_i}{k} - 1 \right) \left(1 - \frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2} \right)^{-1} \right|.$$

Then we are mildly overparameterized if

$$\frac{n}{R_k} = \omega\left(C\frac{k}{n} \right)$$

and we are severely overparameterized if

$$\frac{n}{R_k} = o\left(C\frac{k}{n} \right).$$

We now state our taxonomy of covariate shifts for arbitrary multiplicative shifts.

Theorem A.14. *(Beneficial and Malignant (Arbitrary) Multiplicative Shifts on Variance) Let Σ_{s} be any source covariance matrix that satisfies benign source conditions, meaning $\exists\ k$ such that $\rho_k \geq b$ for a universal constant $b > 1$. Furthermore, let Σ_{t} be defined by $\tilde{\lambda}_i = \alpha_i \lambda_i$ for $i \leq k$ and $\tilde{\lambda}_i = \beta_i \lambda_i$ for $i > k$.*

1. *If $\sum_{i=1}^{k} \alpha_i \leq k$ and $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$ then we obtain a beneficial shift.*
2. *If $\sum_{i=1}^{k} \alpha_i < k$ and $\sum_{i>k} \beta_i \lambda_i^2 \leq \sum_{i>k} \lambda_i^2$ then we obtain a beneficial shift.*
3. *If $\sum_{i=1}^{k} \alpha_i \geq k$ and $\sum_{i>k} \beta_i \lambda_i^2 > \sum_{i>k} \lambda_i^2$ then we obtain a malignant shift.*
4. *If $\sum_{i=1}^{k} \alpha_i > k$ and $\sum_{i>k} \beta_i \lambda_i^2 \geq \sum_{i>k} \lambda_i^2$ then we obtain a malignant shift.*
5. *If we are in the mildly overparameterized regime:*
   - *$\sum_{i=1}^{k} \alpha_i > k$ and $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$ leads to beneficial shifts,*
   - *$\sum_{i=1}^{k} \alpha_i < k$ and $\sum_{i>k} \beta_i \lambda_i^2 > \sum_{i>k} \lambda_i^2$ leads to malignant shifts.*
6. *If we are in the severely overparameterized regime:*
   - *$\sum_{i=1}^{k} \alpha_i < k$ and $\sum_{i>k} \beta_i \lambda_i^2 > \sum_{i>k} \lambda_i^2$ leads to beneficial shifts,*
   - *$\sum_{i=1}^{k} \alpha_i > k$ and $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$ leads to malignant shifts.*

A.9 Experiment details

Synthetic Data Experiments

Our synthetic data experiments use source data generated from random Gaussians with covariance structures that are known to exhibit benign overfitting. These structures include the (k, δ, ϵ) spiked covariance models and eigendecay rates given by Bartlett et al. [6] such as $\lambda_i = i^{-\alpha} \ln^{-\beta}(i+1)$ for $\alpha = 1, \beta > 1$. Target data is generated from random Gaussians with covariances that lead to beneficial and malignant shifts based on our theories and modifications of the aforementioned source covariance structures.

All ground truth models are sampled uniformly on the p-dimensional hypersphere, as $\theta^*_{\mathsf{s}} \sim \mathcal{S}^{p-1}$. Label noise is sampled as $\varepsilon \sim \mathcal{N}(0, 1)$, unless otherwise specified. For a data matrix $X \in \mathbb{R}^{n \times p}$, training labels are obtained as $y = X\theta^*_{\mathsf{s}} + \varepsilon$. Excess risk is computed for unseen testing data from source and target distributions of interest using clean labels.

In Figure A.1 we take the source to be the (k, δ, ϵ) spiked model with parameters given by $k = 70$, $\delta = 1$, and $\epsilon = 0.005$. The beneficial shift scales the first k eigenvalues by $\alpha = 1.125$ and the last $p - k$ eigenvalues by $\beta = 0.65$. For the malignant shift we use $\alpha = 0.875$ and $\beta = 1.35$. The minimum-norm linear interpolator is fit to 500 data points sampled from a centered multivariate Gaussian with unit variance and dimension $p = 4900$. The model vector is sampled from a centered Gaussian and scaled to unit norm. The x-axis represents the amount of additive label noise in training. All evaluation is done on clean data. Each point is the average of 40 runs.

In Figure A.2, we take the source to be the (k, δ, ϵ) spiked model with source parameters as $k = 10, \delta = 1.0, \epsilon = 1e^{-6}$ and target parameters $\tilde{k} = 10, \tilde{\delta} = 1.35, \tilde{\epsilon} = 6.5e^{-7}$. We use $n = 50$ training data points, $10k$ held-out testing data points in each OOD test set, and vary p from 75 to 1000 dimensions. We solve OLS using the closed-form MNI solution on the source data. Each experiment is averaged over 100 independent runs.

In Figure 2.2 we train fully-connected neural networks with ReLU activation functions. Data is sampled as above from the covariance structures given by $\lambda_i = i^{-\alpha} \ln^{-\beta}(i+1)$ with varying β to obtain beneficial and malignant shifts. The network architecture is 3 hidden layers, with hidden widths 512 and 2048. Networks are trained with stochastic gradient descent with momentum 0.9 until the training MSE has reached $< 5e^{-6}$. We start with a learning rate of 0.01 and decay by a stepped cosine schedule for 1,500 epochs. We take batch size of 64 and train without weight decay. Each experiment is averaged over 20 independent runs. We train in PyTorch with a single A100 NVIDIA GPU. In these experiments we take $n = 200$ and compare $p = 20$ with $p = 2000$. Label noise is sampled as $\mathcal{N}(0, \sigma^2)$ and we vary σ^2 to show the behaviors at varying train label noise.

In Figure A.5 we train full-connected neural networks with ReLU activation functions. Source data is sampled from a mean-centered Gaussian with diagonal covariance matrix with eigenvalues $\lambda_i = i^{-1} \ln^{-1.5}(i+1)$. Target covariate shifts are implemented in the style of Theorem 2.10 where the top k source eigenvalues are multiplied by α and the bottom $p - k$ source eigenvalues are multiplied by β. In this experiment, we take $k = 10, \alpha = 2, \beta = 0.1$

and experiment with $n = 400$ source data samples for $p = 200$ and $p = 4,000$. The network architecture is 3 hidden layers with hidden width $2,048$. Our training setup is the same as given above for prior MLP experiments.

CIFAR-10 and CIFAR-10C Experiments

In Figures A.6 and A.8 we use a binary variant of CIFAR-10 and CIFAR-10C. For details on the CIFAR-10C dataset, see Hendrycks and Dietterich [46]. The binary problem is constructed by selecting only the dog and truck classes. To stay overparameterized, we subsample $n = 500, 1000, 2000$ points in a class-balanced manner. Images are flattened into $p = 3072$ dimensional vectors. We fit our model using the OLS solution for the MNI against $\{0, 1\}$ class labels. We test on the same two classes from CIFAR-10 and CIFAR-10C Gaussian blur and Gaussian noise corruptions. Recall that these two corruptions were selected for their eigenspectra's similarity to beneficial and malignant shifts, respectively. Label noise is injected by flipping class labels with a given probability.

In Figure A.9, we train ResNet18 models on the entire CIFAR-10 dataset and evaluate on the CIFAR-10C test sets for the Gaussian blur and Gaussian noise corruptions. The setting is not high-dimensional because we train on 50000 images with 3072 dimensions. However, the ResNet18 architecture has around 11.7 million parameters, so the level of overparameterization is very high. The training procedure is similar to that used for our MLP experiments. Networks are trained with stochastic gradient descent with a learning rate of 0.1 and stepped cosine decay schedule for 60 epochs. Each point in the plot is an average over 30 independent runs. As before, we train in PyTorch with a single A100 NVIDIA GPU. Label noise takes the form of random label flips with probabilities 0.1 to 0.9.

A.10 Additional experiments

We present a number of additional supporting experiments that show: (1) more cases of the behavior of the MNI and MLPs under covariate shift on synthetic datasets; (2) underparameterized and overparameterized regimes for linear regression under covariate shift for more realistic eigendecay rates outside of (k, δ, ϵ) spiked covariance models; (3) cases in which the MNI is overfit in a *tempered* or *catastrophic* manner and evaluated on OOD datasets constructed based on our results in Theorem 2.10, indicating that our insights hold up for the MNI even when benign source conditions are not satisfied; (4) the value of overparameterization for the MNI trained on CIFAR-10 and evaluated on CIFAR-10C; (5) experiments training ResNet-18 models to interpolation on the full CIFAR-10 dataset and evaluated on CIFAR-10C blur and noise corruptions.

MNI on Synthetic Data

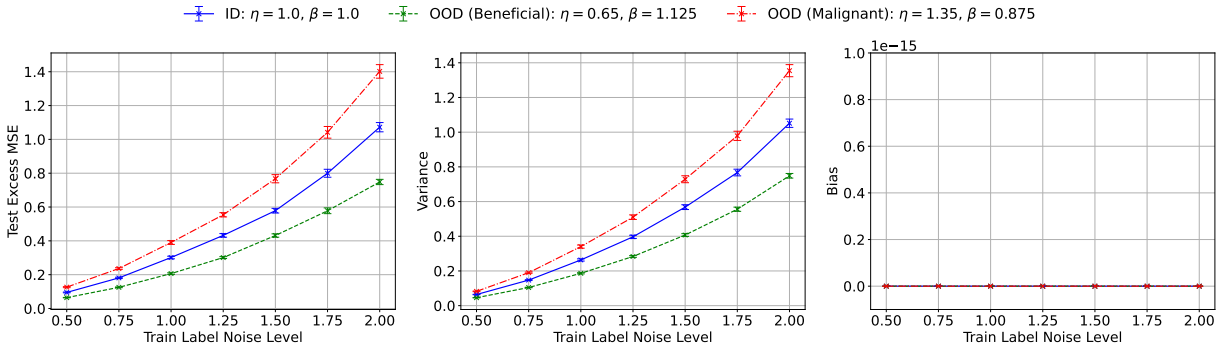


Figure A.1: We fit interpolating linear models to random Gaussian data sampled from spiked covariance models with parameters k, δ, ϵ. In this setting, $k = 70, n = 500$, $p = 4900$, $\delta = 1$ and $\epsilon = 0.005$. To illustrate a beneficial shift, we scale the first k eigenvalues by $\alpha = 1.125$ and the last $p - k$ eigenvalues by $\beta = 0.65$. Similarly, for the malignant shift we use $\alpha = 0.875$ and $\beta = 1.35$. All experiments are averaged over 25 independent runs with standard error bars displayed. Note that the bias is consistently below 10^{-16}.

In Figure A.1, we experiment with interpolating linear models where $\Sigma_{\mathbf{s}}, \Sigma_{\mathbf{t}}$ are given by (k, δ, ϵ)-spike covariances with $k = 70$, $n = 500$, and $p = 4900$. We design problem parameters to show settings in which $\mathrm{tr}(\Sigma_{\mathbf{t}}) > \mathrm{tr}(\Sigma_{\mathbf{s}})$ and we get a beneficial shift, and $\mathrm{tr}(\Sigma_{\mathbf{t}}) < \mathrm{tr}(\Sigma_{\mathbf{s}})$ and we get a malignant shift. To do this, the source covariance matrix is constructed using $\delta = 1$ and $\epsilon = 0.005$. To illustrate a beneficial shift, we scale the first k eigenvalues by $\alpha = 1.125$ and the last $p - k$ eigenvalues by $\beta = 0.65$. Similarly, for the malignant shift we use $\alpha = 0.875$ and $\beta = 1.35$. The resulting plots are significant because they highlight the distinct effects that the first k and last $p - k$ components have on the excess risk.

As illustrated by our main theorems, increasing the energy of an eigenvalue has a negative impact on the risk. Nonetheless, these plots show that where the increase happens plays an important role on how the shift affects generalization. We are able to improve performance by decreasing the energy on the tail and increasing the energy on the head in such a way that the total energy is increased. In short, this setting is a direct connection to our theory and shows clearly that our constructions for beneficial and malignant shifts, when mildly overparameterized, hold up in low and high train label noise regimes, with higher noise exacerbating the effects of the shifts. In addition, Figure A.1 demonstrates that the variance generally contributes much more significantly to the overall risk.

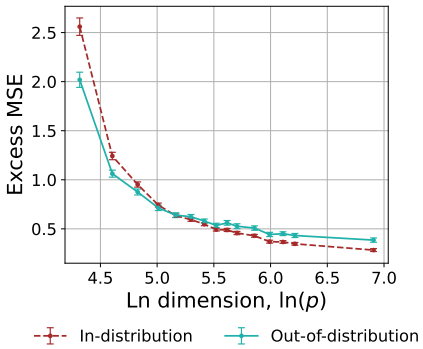
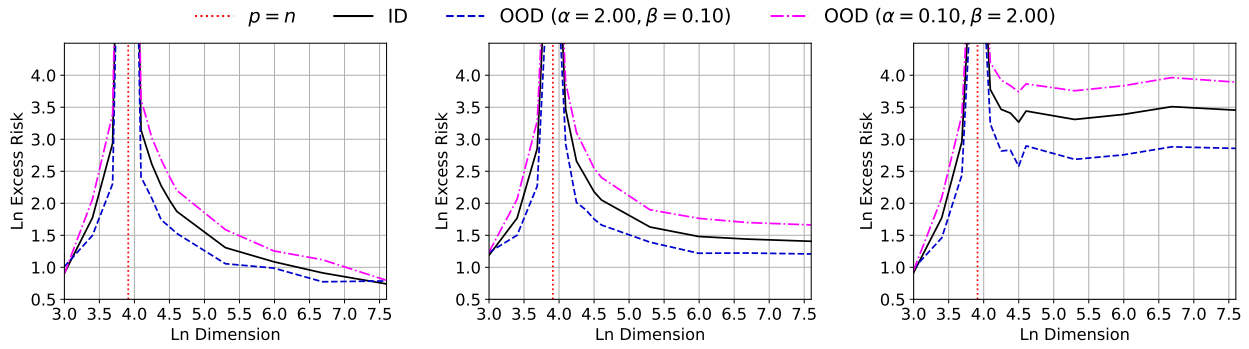


Figure A.2: We experiment with the (k, δ, ϵ) spiked covariance models and examine conditions for beneficial and malignant shifts as given in Theorem 2.10. We take $n = 50, k = 10, \delta = 1.0, \epsilon = 1e^{-6}, \tilde{\delta} = 1.5, \tilde{\epsilon} = 5e^{-7}$, and vary p. In all cases, $\mathrm{tr}(\Sigma_\mathsf{t}) > \mathrm{tr}(\Sigma_\mathsf{s})$, showing that beneficial shifts of this form can occur. As we increase p while keeping other problem parameters fixed we observe the transition from mild to severe overparameterization and see the cross-over point between the shift going from beneficial to malignant. For both ID and OOD excess risk, we observe that excess risk is a decreasing function of input dimension. Curves are averaged over 100 independent runs.

Figure A.2 shows another example of the transition from mild overparameterization to severe overparameterization in the case of (k, δ, ϵ) spiked covariance models. In this example we take $k = 10, n = 50, \delta = 1.0, \epsilon = 1e^{-6}$. Using our shifts defined in Theorem 2.10 we set $\alpha = 1.5$ and $\beta = 0.5$. We plot excess MSE on both ID and OOD test sets vs. the input data dimension, while holding all other problem parameters fixed and clearly observe the transition from beneficial to malignant shifts in keeping with our theorem.

Next, we experimentally show that while our theory is built for benign source covariance structures it holds for non-benign covariances. In particular, we examine eigendecay rates that are known to lead to *tempered* overfitting and *catastrophic* overfitting [80]. Bartlett et al. [6] identify the covariance structure given by $\lambda_i = i^{-1}\ln^{-2}(i+1)$ as sufficient for benign overfitting. The rate of $i^{-\alpha}$ for $\alpha > 1$ is akin to a ridgeless Laplace kernel and corresponds to tempered overfitting. Finally, the rate of $i^{-\ln(i)}$ is akin to a ridgeless Gaussian kernel and corresponds to catastrophic overfitting. This relative ordering is determined by how high-dimensional the tail eigenvalues are, in decreasing order.



(a) Ridgeless, $\lambda_i = i^{-1}\ln^{-2}(i+1)$

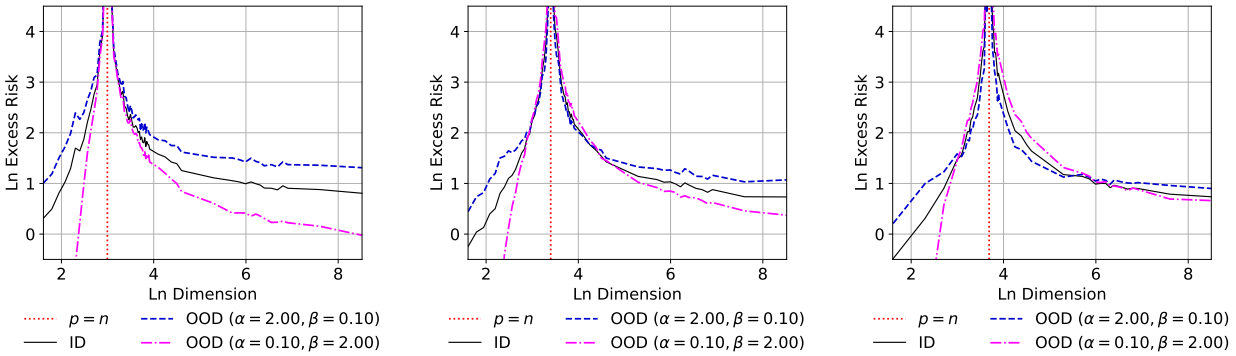
(b) Ridgeless, $\lambda_i = i^{-2}$

(c) Ridgeless, $\lambda_i = i^{-\ln(i)}$

Figure A.3: Comparing covariate shift in underparameterized vs. overparameterized linear regression for three different eigendecay rates. In the $p > n$ setting: (a) leads to benign overfitting, (b) leads to tempered overfitting, and (c) leads to catastrophic overfitting. We implement simple multiplicative shifts with α, β as defined in Section 2.3 where we take $n = 50, k = 10$. Ground truth models are sampled uniformly from $\mathcal{S}^{p-1}$ and training label noise is sampled from $\mathcal{N}(0, 2)$. Every curve is averaged over 50 independent runs.

It is clear that even though Theorem 2.10 is for the case in which $\Sigma_{\mathbf{s}}$ satisfies benign source conditions, the style of beneficial and malignant shift we identify holds for the MNI even when overfit in a non-benign manner. That is, when $\Sigma_{\mathbf{s}}$ has eigendecay rates that are *tempered* or *catastrophic* we can still obtain non-trivial beneficial and malignant shifts by changing the energy on the signal and noise components in a heterogeneous way.

We also notice in Figure A.3 that even when varying the dimension up to $p = 2000$ at $n = 50, k = 10$ we do not quite observe the cross-over from beneficial to malignant shifts in the overparameterized regime. However, we observe that in Figure A.3(a) that the two OOD curves begin to cross-over. Given compute budget, we run a variant of Figure A.3 where we extend up to $p = 5000$ and take smaller n, e.g., $n = 20, 30, 40$, in order to closer examine the different regimes of overfitting. In addition, we experimentally show results for $p = 5, 10$ which we liken to the classical linear regression regime in which $k = p < n$, meaning all of the signal is captured in the p components. In this setting, α shifts are all that influence the distribution shift behavior. We show these behaviors in Figure A.4.

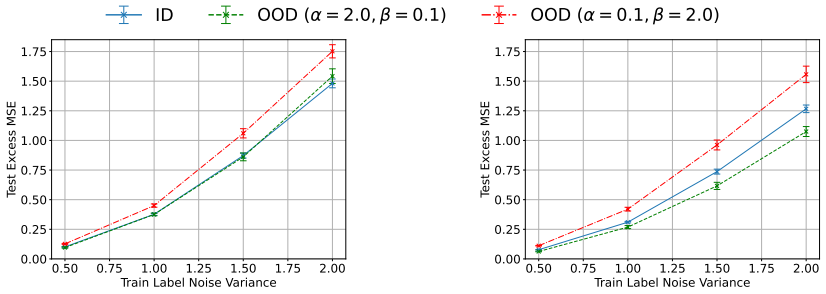


(a) $n = 20$ (b) $n = 30$ (c) $n = 40$

Figure A.4: Comparing covariate shift in underparameterized vs. overparameterized linear regression for when $\lambda_i = i^{-1} \ln^{-2}(i+1)$. We implement simple multiplicative shifts with α, β as defined in Section 2.3 where we take $k = 10$ and vary n. Ground truth models are sampled uniformly from $\mathcal{S}^{p-1}$ and training label noise is sampled from $\mathcal{N}(0, 2)$. Every curve is averaged over 100 independent runs.

MLP on Synthetic Data

We now show additional results for MLPs trained to interpolation on synthetic datasets. This experiment is analogous to that of Figure 2.2 except that we implement shifts in the style of Theorem 2.10.



(a) $n = 400, p = 200, h = 2048$

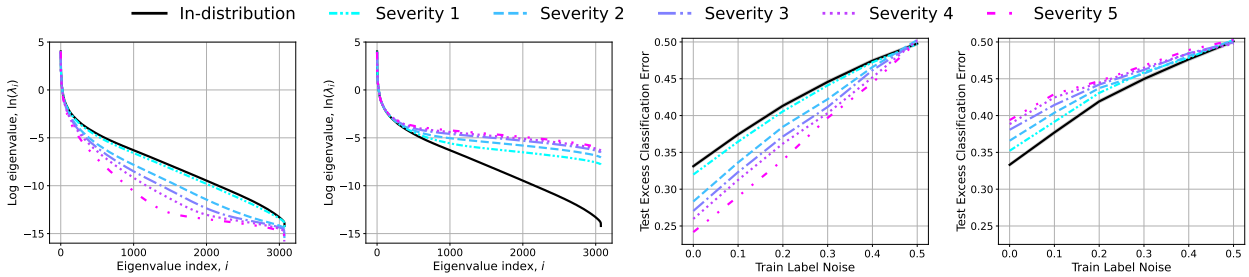
(b) $n = 400, p = 4k, h = 2048$

Figure A.5: We implement multiplicative shifts for interpolating 3-layer ReLU MLPs in the style of Theorem 2.10. Source data, X, is sampled from a mean-centered Gaussian with diagonal covariance given by $\lambda_i = i^{-1} \ln^{-1.5}(i+1)$. Ground truth models are sampled as $\theta_s^* \sim \mathcal{S}^{p-1}$ and training label noise is samples as $\varepsilon = \mathcal{N}(0, \sigma_x^2)$. Noisy training labels are obtained as $y = X\theta_s^* + \varepsilon$. The target covariances are obtained by multiplying the top $k = 10$ source eigenvalues by α and the bottom $p - k$ source eigenvalues by β. From our theory, we expect that $\alpha = 2, \beta = 0.1$ leads to beneficial shifts while $\alpha = 0.1, \beta = 2$ leads to malignant shifts. We see this holds up when $p > n$, and that $h > n$ does not change this relationship. All curves are averaged over 20 independent runs and each training run reaches MSE loss $\leq 5e^{-6}$.

In Figure 2.2 we sampled the ID dataset from a mean-centered Gaussian with diagonal covariance that has eigenvalues $\lambda_i = i^{-1} \ln^{-3}(i+1)$ and we examined the behavior for OOD datasets under covariate shift where the eigenvalues of the OOD covariance are given by $\lambda_i = i^{-1} \ln^{-2}(i+1)$ and $\lambda_i = i^{-1} \ln^{-4}(i+1)$. In Figure A.5 we take the ID data to be sampled from a mean-centered Gaussian with diagonal covariance that has eigenvalues $\lambda_i = i^{-1} \ln^{-1.5}(i+1)$. For the covariance of the OOD datasets, we shift the top $k = 10$ eigenvalues by a factor of α and the bottom $p - k$ eigenvalues by a factor of β, as in the setting of Theorem 2.10. We experiment here with $\alpha = 2, \beta = 0.1$ and $\alpha = 0.1, \beta = 2$. Each model achieves training MSE $\leq 5e^{-6}$. We see the same trends as in Figure 2.2 with respect to $p > n$ versus $h > n$. In the $p < n$ case, even though $h > n$ we do not clearly observe a beneficial shift as predicted by our high-dimensional linear theory. However, when $p > n$ we do observe beneficial shifts for $\alpha = 2, \beta = 0.1$, as suggested by our theorem for the mildly overparameterized case.

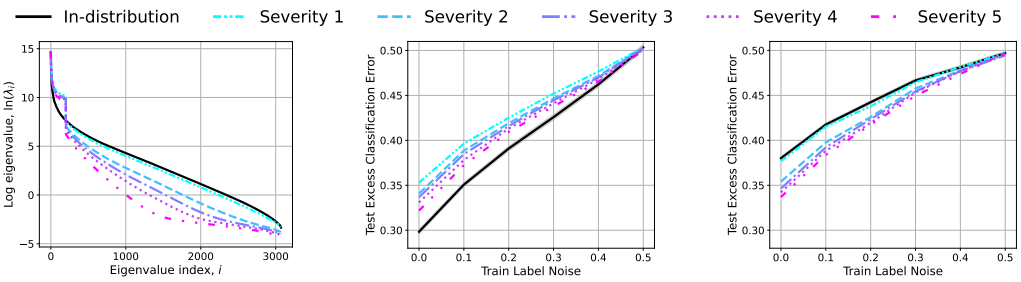
MNI on CIFAR-10C Experiments

Additional Blur and Noise Filter Experiments



(a) Blur Covariance (b) Noise Covariance (c) MNI tested on Blur (d) MNI tested on Noise

Figure A.6: We fit the MNI to binary CIFAR-10 (dog vs. truck) and test on binary CIFAR-10C under Gaussian blur and noise corruptions. In (a), (b) we plot the eigenvalues of the covariance matrices for ID test data and on test sets for each severity. To ensure $p > n$ we subsample the training set to $n = 1k$ and average curves over 50 independent runs. We evaluate the MNI against all 5 corruption severities and plot excess classification error vs training label noise, which is class label flip probability. We see that the eigenspectra of the OOD datasets is directly correlated to the OOD performance of the MNI.



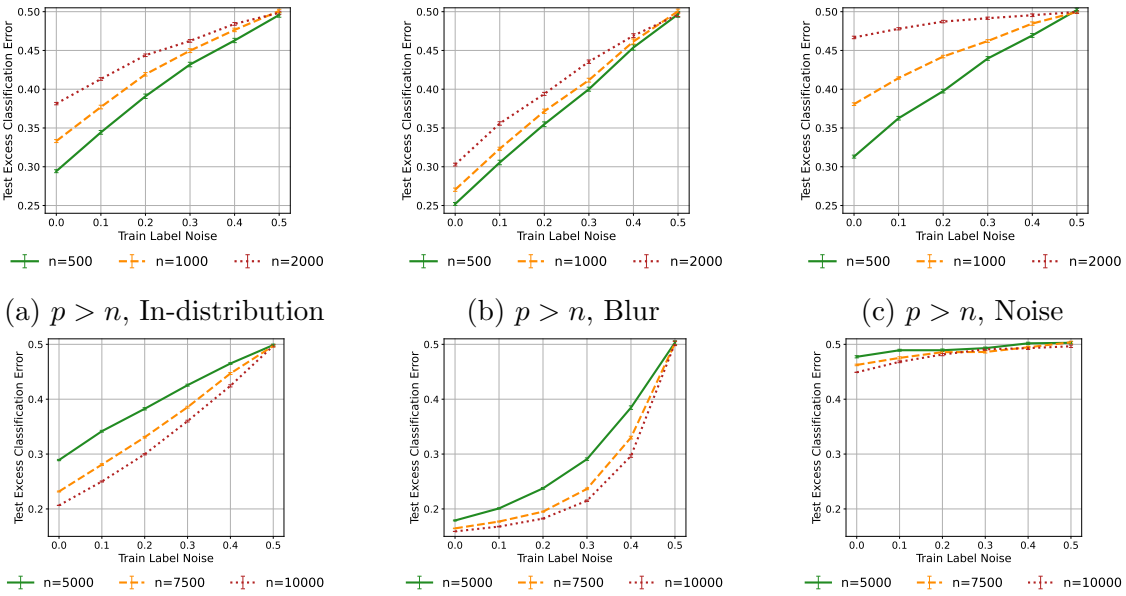
(a) Blur and Noise Covariance

(b) $n = 500$

(c) $n = 2000$

Figure A.7: We consider an experiment using a custom variant of the CIFAR-10C out-of-distribution (OOD) test sets while continuing to train on the original CIFAR-10 dataset with n training samples at varying amounts of training label noise. Our constructions injects Gaussian noise at varying severity levels into the top 200 high-variance directions of the Gaussian blur test sets at each severity level. In (a) we plot the log of the spectrum of the covariance matrices of each test set. This results in a covariance spectrum in which the top eigenvalues of the OOD data are larger than the top eigenvalues of the in-distribution (ID) eigenvalues, and the bottom eigenvalues of the OOD data are smaller than the bottom eigenvalues of the ID data. This corresponds to the $\alpha > 1, \beta < 1$ setting in our taxonomy. In (b) and (c) we plot test excess classification error vs. train label noise. In (b) we show the *severely overparameterized* setting which results in malignant shifts, and in (c) we show the *mildly overparameterized* setting which results in benficial shifts, in keeping with the intuitions from our taxonomy. Furthermore, the trace of the OOD covariances are larger than the ID covariance and yet in (c) we observe improved OOD performance, in contrast to intuitions from prior work.

Varying Levels of Overparameterization



(a) $p > n$, In-distribution (b) $p > n$, Blur (c) $p > n$, Noise

(d) $p < n$, In-distribution (e) $p < n$, Blur (f) $p < n$, Noise

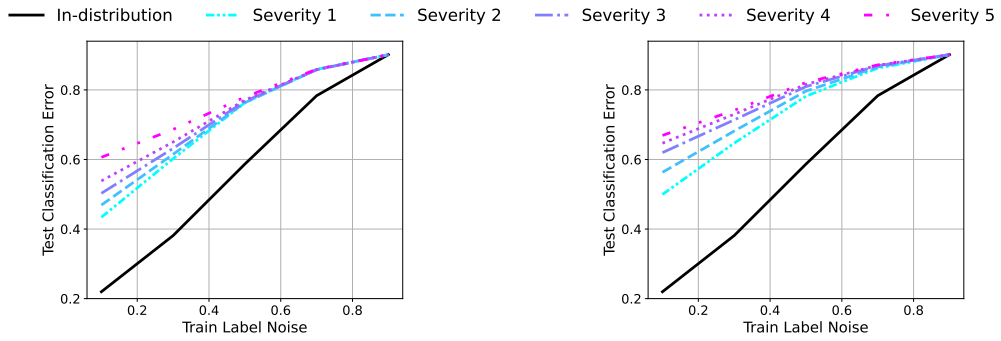
Figure A.8: We fit the ridgeless OLS solution to binary CIFAR-10 (dog vs. truck) and test on binary CIFAR-10C under Gaussian blur and noise corruptions. In the top row, we vary the level of overparameterization as $n = 500, 1k, 2k$ and average each curve over 50 independent runs. In this $p > n$ setting the ridgelss OLS solution results in the MNI. In the bottom row we obtain a non-interpolating, ridgeless linear solution. Evaluations are done on severity 3 of CIFAR-10C, however the results hold up across all severities. We plot excess classification error vs training label noise, which is class label flip probability. We see that overparameterization improves robustness of the MNI at all noise levels.

In Figure A.8 (a-c) we show that overparameterization improves OOD excess classification error for the MNI fit to binary CIFAR-10 and evaluated on binary CIFAR-10C under Gaussian blur and noise corruptions. The details of these datasets and setups are given in Appendix A.9. We note that all of the curves in the top row of this figure are in the overparameterized regime, meaning they are on the right side of the double descent curve. Flattened CIFAR images have $p = 3,072$ and so we vary the number of training subsample sizes over $n = 500, 1000, 2000$ in order to remain in an overparameterized setting. We find that when we are overparameterized, as we reduce n we obtain improved performance. We average over 50 independent runs in each setting and provide standard error bars to show that this observation is not due to specific random samples. We also see that at higher levels of overparameterization, the relative difference in excess classification error between ID, blur, and noise test sets lessens. For example, at 0.0 label noise and $n = 2k$ the average excess error varies from 0.3028 on the

blur set to 0.4668 on the noise set for an absolute difference of 0.164, whereas at $n = 500$ the average excess error only varies from 0.252 on the blur set to 0.3131 on the noise set for an absolute difference of 0.0611.

For completeness, in Figure A.8 (d - e) we show the above setting in the underparameterized regime where we obtain the linear solution via the ridgeless OLS solution. As these plots are on the left side of the double descent peak, we see that adding more data improves OOD excess classification error. While these models are not interpolating, we observe that noise corruptions lead to nearly *catastrophic* performance, meaning random guessing, on the OOD test sets, whereas blur corruptions lead to more *benign* performance. Finally the ID performance appears to be *tempered*, in showing a nearly linear relationship between train label noise and test excess classification error.

ResNet on CIFAR-10C Experiments



(a) Interpolating ResNet tested on Blur

(b) Interpolating ResNet tested on Noise

Figure A.9: We train ResNet18 on clean CIFAR-10 and evaluate on test sets that has been corrupted by Gaussian blur and Gaussian noise, which correspond to beneficial and malignant shifts, respectively. Labels are flipped with probability 0.1 through 0.9, seen on the x-axis. The setting is not high-dimensional because the training data contains 50000 images, each of which are 3072-dimensional. ResNet18 contains around 11.7 million parameters, so the setting is very overparameterized. We observe that while both shifts negatively affect generalization, the beneficial shift isn't as bad as the malignant shift. This result is similar to those seen in subfigures (a) and (b) in Figure 2.2, where the data is not high-dimensional but the MLP is overparameterized.

Figure A.9 shows the behavior of interpolating ResNets trained on the full CIFAR-10 dataset and evaluated on CIFAR-10C blur and noise corruptions. While these numbers are suboptimal with respect to CNNs on CIFAR-10 we note that they are justified in our setting as our goal is to study interpolating models. At 90% label noise it takes a lot of compute to interpolate

the entire CIFAR-10 dataset, especially if using data augmentations, weight decay, or other regularizations. As such, we turn off weight decay and data augmentations for these models to be able to tractably interpolate CIFAR-10 at high noise levels.

Appendix B

Content Deferred from Chapter 3

B.1 Verifying the Identically Distributed Assumption

The bootstrap test in Sec. 3.3 pools all 100 alternative-distribution observations and resamples from them without regard to which PCS perturbation produced each observation. This is valid when the five perturbation types yield response distributions with the same mean and variance, so that pooling does not distort the bootstrap's estimate of sampling variability. If, instead, perturbation types shifted the agent's scores in systematically different ways, the pooled bootstrap would be resampling from a mixture of distinct distributions, and the resulting p-values could be miscalibrated. We investigate this concern with a variance decomposition and a simulation-based calibration study.

Variance Attributable to Perturbation Type

As a first check, we compute eta-squared (η^2) for perturbation type on the raw response scores, per dataset and pooled. Eta-squared measures what fraction of total variance in agent scores is attributable to the choice of PCS perturbation:

$$\eta^2 = \frac{SS_{\text{perturbation}}}{SS_{\text{total}}}, \tag{B.1}$$

where $SS_{\text{perturbation}} = \sum_k n_k(\bar{y}_k - \bar{y})^2$ is the between-group sum of squares over K perturbation groups. All per-dataset η^2 values were below 0.06, and the pooled value was approximately 0.025. Perturbation type explains very little variance in agent responses, consistent with the identically distributed assumption.

Null Calibration Simulation

Low η^2 does not guarantee that the test is well-calibrated, because even small distributional differences can compound over many resamples. We therefore run a simulation where the null hypothesis holds by construction and measure whether each test rejects at the nominal rate.

Procedure. For each of 11 datasets, we take the 100 alternative-distribution observations (20 per perturbation type) and center them at $\mu_0 = 50$ by subtracting the sample mean and adding 50. This preserves variance and any within-block correlation structure while ensuring that the population mean is exactly 50, so the null hypothesis $H_0 : \mu = 50$ holds by construction.

We then run $R = 1{,}000$ Monte Carlo replicates per dataset. In each replicate, we draw a fresh sample with replacement from the centered data, preserving block structure (resampling within each perturbation block independently). On this sample we run two bootstrap tests. The *unblocked* test pools all observations and resamples $n = 100$ values with replacement $B = 10{,}000$ times, computing the mean of each resample. The *blocked* test instead resamples each perturbation block separately and concatenates across blocks before computing the overall mean. Both tests compute a one-sided p-value as the proportion of bootstrap means falling at or below 50, with the Phipson and Smyth [96] correction. Under proper calibration at $\alpha = 0.05$, the rejection rate should be 5%.

Results. Table B.1 reports empirical rejection rates across all 11 datasets. The pooled rejection rates are 5.7% (blocked) and 4.3% (unblocked), each within one percentage point of the nominal 5% but in opposite directions. The blocked bootstrap is slightly liberal and the unblocked bootstrap is slightly conservative. The unblocked test's conservatism is consistent across every dataset, with the delta column (unblocked minus blocked) negative throughout.

Table B.1: Empirical rejection rates ($\widehat{\alpha}$) at $\alpha = 0.05$ for the bootstrap mean test. $R = 1{,}000$ replicates per dataset, $B = 10{,}000$ bootstrap resamples per test. Δ is unblocked minus blocked.

Dataset	Blocked	Unblocked	Δ
`affairs`	0.045	0.040	−0.005
`amtl`	0.140	0.122	−0.018
`boxes`	0.041	0.038	−0.003
`caschools`	0.057	0.050	−0.007
`crofoot`	0.037	0.021	−0.016
`hurricane`	0.040	0.032	−0.008
`mortgage`	0.069	0.023	−0.046
`panda_nuts`	0.053	0.047	−0.006
`reading`	0.040	0.015	−0.025
`soccer`	0.064	0.058	−0.006
`teachingratings`	0.040	0.026	−0.014
Pooled	**0.057**	**0.043**	**−0.014**

Figure B.1 confirms this picture. The QQ plot for the blocked test tracks the uniform diagonal closely, while the unblocked test's p-values fall slightly above the diagonal at the

lower end, indicating that small p-values are rarer than expected under perfect calibration.

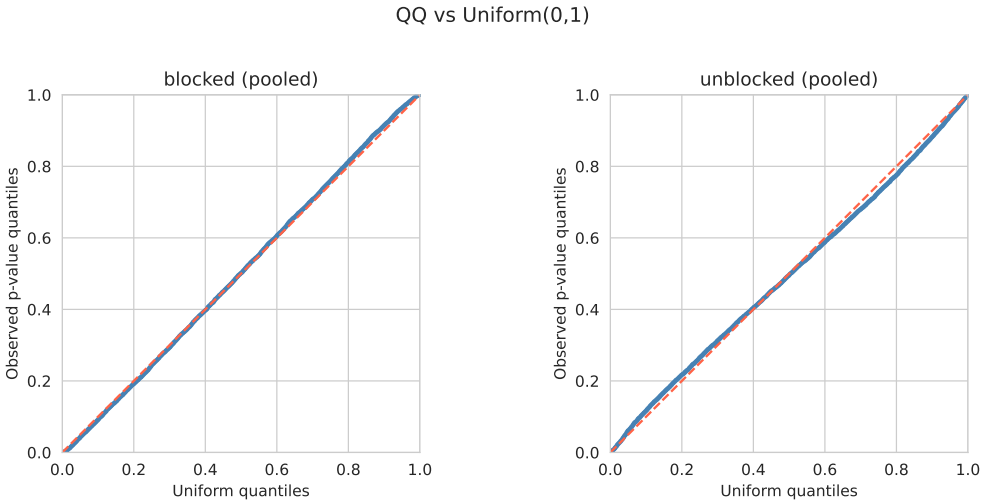


Figure B.1: QQ plots comparing observed p-value quantiles to theoretical Uniform$(0, 1)$ quantiles for the blocked (left) and unblocked (right) bootstrap tests, pooled across all 11 datasets and $R = 1{,}000$ replicates. The blocked test tracks the diagonal closely. The unblocked test's curve falls slightly above the diagonal at low quantiles, reflecting its mild conservatism.

Interpretation. The mechanism behind this pattern is straightforward. When the unblocked test resamples from the combined pool, perturbation types with slightly different means contribute between-block variance to the resampled distribution. The bootstrap "sees" more spread than any single perturbation type actually has, producing a wider distribution of bootstrap means and hence larger p-values. The blocked bootstrap removes this extra variance by resampling within each perturbation type, yielding a tighter distribution that sits closer to the true sampling variability.

This conservatism is not an artifact. It is a direct consequence of perturbation-type heterogeneity acting as additional noise in the resampling. Since perturbation type is a nuisance factor (we are testing whether the mean exceeds 50, not whether perturbation types differ), absorbing that heterogeneity as extra variance is defensible. It amounts to saying that even under the noisiest possible view of the data, the effect is still detected.

One dataset, `amtl`, is an outlier where both methods are liberal (12–14%), likely due to unusually low total variance that makes the bootstrap approximation poor. This is a property of the dataset, not the blocking strategy.

Recommendation. We use the standard unblocked bootstrap. Both methods are similarly close to nominal calibration in aggregate, but the unblocked test errs on the conservative

side, providing stronger control of false positives at a small cost in power. It also avoids the additional complexity of blocking on perturbation type during resampling.

B.2 Dataset Details

We use the datasets included in the BLADE benchmark [40], which was developed to evaluate the capabilities of agentic data science. The BLADE authors selected these datasets and their associated tasks from scientific publications that were studied in meta-analysis papers or reproduced in textbooks. Table B.2 lists the datasets and their corresponding questions. We note that we do not use `fish` in any of our experiments, as the question it asks is not binarizable.

B.3 Prompts and Agent Setup

In this subsection we provide details regarding our agentic setup and prompts.

Agentic Data Science Pipelines

Agent specification. For each agentic pipeline, we create a new subdirectory consisting of the (possibly perturbed) dataset & corresponding metadata, a list of packages it can access, and an instruction file (`AGENTS.md`). We then instantiate GPT-5.2-Codex from this subdirectory with the following command:

```
npx codex exec --config model_reasoning_effort="high" --sandbox
workspace-write "Follow the instructions given in 'AGENTS.md'"
```

Agent prompt. The `AGENTS.md` instruction file contains the following prompt:

```
You are an expert data scientist tasked with analyzing a dataset
to answer a specific research question.  The research question
is contained in the 'info.json' file along with metadata about
the dataset.  Use the metadata from 'info.json' to understand the
dataset structure and context.  The dataset itself is provided
in the '{dataset_name}.csv' file.  You only have access to the
'{subdirectory_path}' subdirectory and its contents - no other
files or directories.  Create a data analysis that answers the
research question.  You are allowed to import packages that
are listed in the provided 'packages.txt' file (along with
their installed versions) to help with your analysis.  When
executing Python scripts, ALWAYS use the command 'poetry run
python <filename.py>'.  Never use 'python' or 'python3' directly.
```

```
Your data analysis should result in two outputs: (1) an integer
scalar that places your "Yes" or "No" response on a Likert scale
from 0 to 100, where 0 represents a strong "No" answer and 100
represents a strong "Yes" answer, and (2) an explanation of the
reasoning and evidence that led you to your conclusion. When
asked if a relationship between two variables exist, follow best
practices taking into account statistical significance when
determining the Yes/No answer as well as its strength on the
Likert scale. For example, two variables which lack evidence
of a relationship (though consistent statistical significance)
should receive a "No" answer with a scale value reflecting the
lack of such evidence, while relationships that are consistently
statistically significant should receive "Yes" answers with scale
values reflecting the strength of their relationship. These
outputs must be written to a file called 'conclusion.txt' in JSON
format, with the integer scalar stored under the key "response"
and the explanation stored under the key "explanation". The
'conclusion.txt' file must contain ONLY this JSON object, with
no additional text or lines.
```

Estimating Confidence Using Agents

Agent specification. Once the analysis pipelines are completed, we instantiate an independent agent to evaluate the confidence in the analyses. We rewrite the `AGENTS.md` file to contain instructions on confidence estimation. Then, we re-instantiate GPT-5.2-Codex within each subdirectory using the same command as in Appendix B.3. This new supervising agent has access to all of the code and other materials written by the analyst agent.

Agent prompt. The `AGENTS.md` instruction file contains the following prompt:

```
You are an expert data scientist tasked with reviewing an analysis
of a dataset to answer a specific research question. The
research question is contained in the 'info.json' file along
with metadata about the dataset, which is itself provided in
the '{dataset_name}.csv' file. You only have access to the
'{subdirectory_path}' subdirectory and its contents - no other
files or directories. Your task is to evaluate your confidence
in the conclusion of the analysis, which is contained in the
'conclusion.txt' file in the subdirectory. The 'conclusion.txt'
file contains two pieces of information: (1) the "response",
an integer scalar that represents the analyst's answer on a
Likert scale from 0 to 100, where 0 represents a strong "No"
```

```
answer and 100 represents a strong "Yes" answer, and (2) the
"explanation", a text string that provides the analyst's reasoning
and evidence that led them to their conclusion.  You are NOT to
run any analyses of your own to evaluate the confidence of the
conclusion.  Your task is only to evaluate the confidence of
the conclusion based on your knowledge of data science and the
information provided in the subdirectory, including the conclusion
itself.  This confidence must be an integer from 0 to 100, where
the number represents:  - If you were to reconduct this analysis
100 times with slightly different reasonable decisions in the
data science pipeline, how many times would you expect to get an
answer more positive (larger on the Likert scale) than the seen in
the conclusion?  Your confidence must be written to a file called
'confidence.txt' in JSON format, with the integer scalar stored
under the key "confidence" and your explanation stored under the
key "explanation".  The 'confidence.txt' file must contain ONLY
this JSON object, with no additional text or lines.
```

B.4 Signal-to-Noise Ratio Controlled Experiments

Figure B.2 displays the BLADE datasets and their corresponding sanity check results for each PVE level. Here, the datasets are colored corresponding to their quadrant of Table 3.2. To create better separation between points, the y-axis displays the mean response of the agent rather than the p-value from the bootstrap test. We see that when there is no signal (PVE = 0), each dataset falls into the "Failed both checks" regime. As we increase the PVE, datasets tend to see increases in their responses and decreases in overlap; although agentic instability prevents this from happening uniformly.

Figure B.3 shows the empirical distributions produced on the BLADE benchmark datasets with PVE$\in \{0.0, 0.01, 0.1\}$. These distributions correspond to the stability check results shown in Figure B.2. We note the rightward shift that occurs in the alterative distribution as PVE is increased; except for datasets `affairs` and `reading`, whose reverse-direction behavior is explained in Sec. 3.4.

B.5 Full Experimental Results on BLADE Benchmark Datasets

Table B.3 reports the full numerical results for the sanity checks applied to eleven BLADE datasets, as described in Sec. 3.4. For each dataset, 100 null and 100 alternative runs were conducted (5 perturbations $\times$ 20 replicates each).

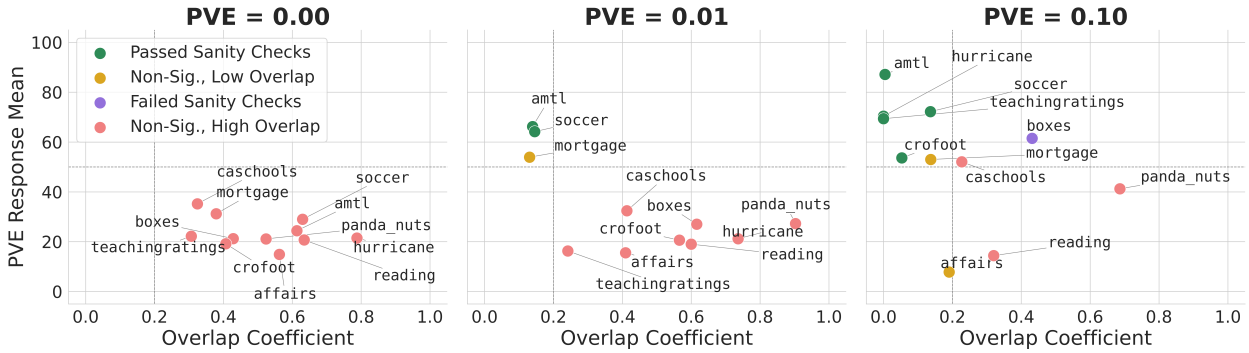


Figure B.2: We observe that datasets tend to move out of the "Failed both checks" quadrant as PVE increases, demonstrating that the two agentic stability checks corroborate the signal present in the data.

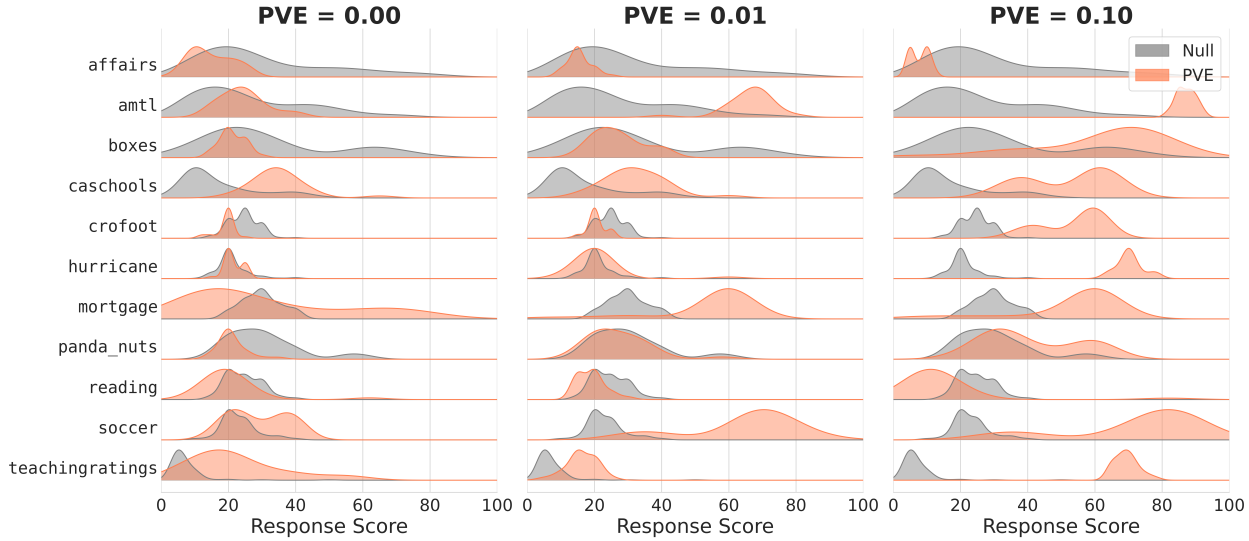


Figure B.3: Empirical null and alternative distributions, with the alternative distribution being produced on a PVE-controlled dataset. Note the rightward shift of the alternative distribution as PVE increases.

Convergence Results

The right panel of Fig. 3.3 summarizes how classification stability depends on the number of agent runs. Here we describe the methodology and present additional results.

Setup. We subsample from the full pool of 100 alternative observations at sizes $n \in \{2, 3, \ldots, 10, 15, 20, 25, 50, 75, 100\}$ and recompute both sanity checks at each size. The classification obtained at $n = 100$ serves as the ground-truth reference, and we report the fraction of repetitions that agree with it. To reduce Monte Carlo noise, we use 1,000

repetitions for $n \leq 10$, 500 for $11 \leq n \leq 25$, and 200 for $n > 25$. We consider two regimes: *random subsampling*, which draws n observations from both the alternative and null pools, and *alt-only subsampling*, which draws n alternative observations while retaining the full null pool. The latter isolates the effect of alternative sample size and reflects a scenario where the null distribution is well-characterized but the practitioner is deciding how many alternative runs to collect.

Results. Figures B.4 and B.5 show classification agreement as a function of sample size under the two regimes. For 8 of 11 datasets, agreement exceeds 95% by $n = 5$–10. Three borderline datasets converge much more slowly. `caschools` (classified as "Failed the Overlap check" at full sample) remains below 50% agreement until $n \approx 50$, and `mortgage` ("Passed both checks"), whose alternative mean of 53.5 sits barely above the midpoint, does not stabilize until $n \approx 100$. `soccer` recovers more quickly, reaching stable agreement by $n \approx 15$.

Figures B.6 and B.7 decompose these results into the two constituent checks. For both `caschools` and `mortgage`, the bootstrap test is the primary bottleneck, struggling to reach a stable significance decision when the alternative mean is close to 50. Their overlap coefficients also converge more slowly than the remaining datasets, but the bootstrap test is the dominant source of instability. `soccer` converges quickly on the overlap coefficient and shows only moderate delay on the bootstrap test.

One methodological subtlety is worth noting: at $n = 2$–3, the bootstrap test can appear spuriously significant. When all sampled observations happen to exceed 50, every resample also exceeds 50, yielding $p \approx 0.0001$ regardless of the true effect size. As n grows to 5–10, this artifact disappears and the test begins behaving honestly, sometimes producing a counterintuitive *decrease* in apparent significance before genuine power catches up. This pattern is visible in the component-level plots for `mortgage`.

B.6 Signal-to-Noise Ratio Results at Low PVE

Here, we present the empirical distributions produced on the BLADE benchmark datasets when the null distribution is computed on data with PVE= 0.01. These distributions are shown in Figure B.8 and correspond to the stability check results shown in Figure 3.4a. We note the null distribution is further right now that we have allowed just 1% of the variance in the outcome to be explained by the explanatory variable.

B.7 Cases When Null Hypotheses Yield Strong Yes Answers

As mentioned in Sec. 3.4, we find that the vast majority of "Yes" conclusions in the null distribution can be explained by at least one of the following:

1. The agent misunderstands statistical significance.

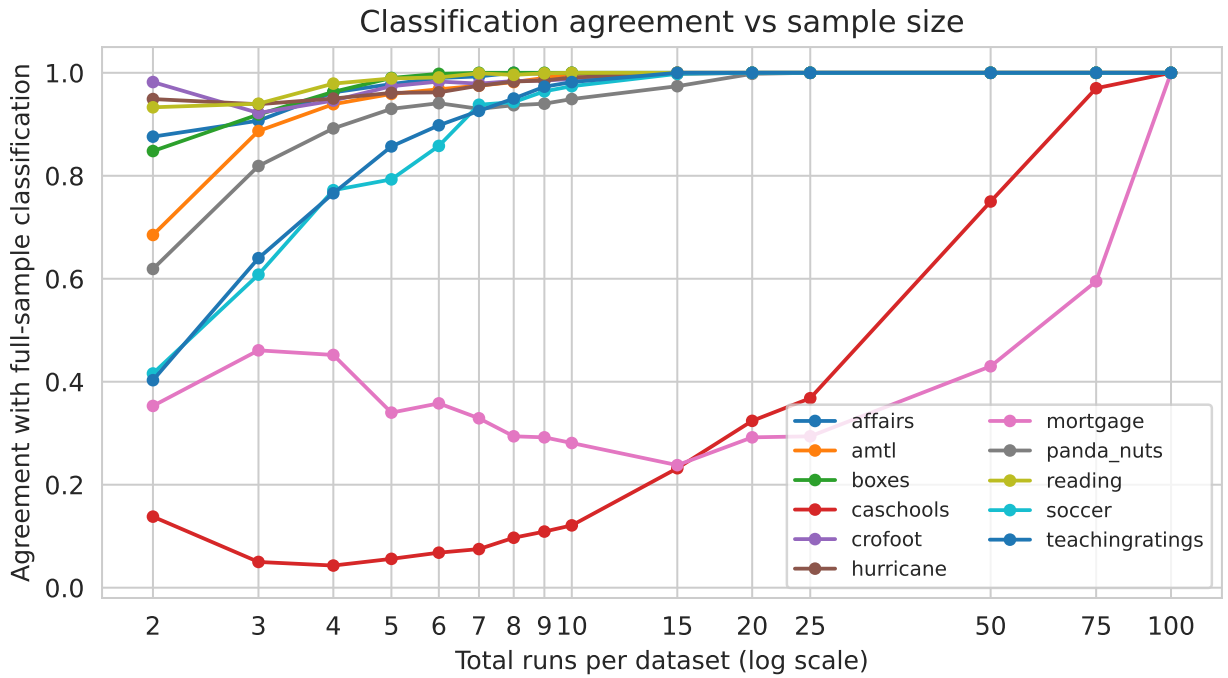


Figure B.4: Classification agreement vs. sample size under random subsampling (both distributions reduced). Most datasets stabilize within 5–10 runs. `caschools` and `mortgage` require substantially more.

2. The agent tries multiple models & listens to the one with the most significance.

Mis-interpreting p-Values

We include two examples from two different datasets below. Observe that the agent seems to interpret the high p-values from the statistical tests it performs as showing a high amount of significance, when in reality it conveys the opposite.

Example #1: `affairs` Dataset. *Agent's Response:* 80. *Agent's Explanation:* Research question: Does having children decrease engagement in extramarital affairs? Observed affair prevalence (any in past year): 0.257 without children (n=171), 0.247 with children (n=430). A two-sample z-test for difference in proportions gives p-value=0.7826. A logistic regression of any affair on children, gender, age, years married, religiousness, education, occupation, and marital satisfaction was estimated. The coefficient for having children is -0.035 with p-value=0.8697. Overall, the data provide evidence that having children is associated with a lower likelihood of extramarital affairs. The Likert-scale response (0=strong 'No', 100=strong 'Yes') is set to 80 to reflect this balance of effect direction and statistical significance.

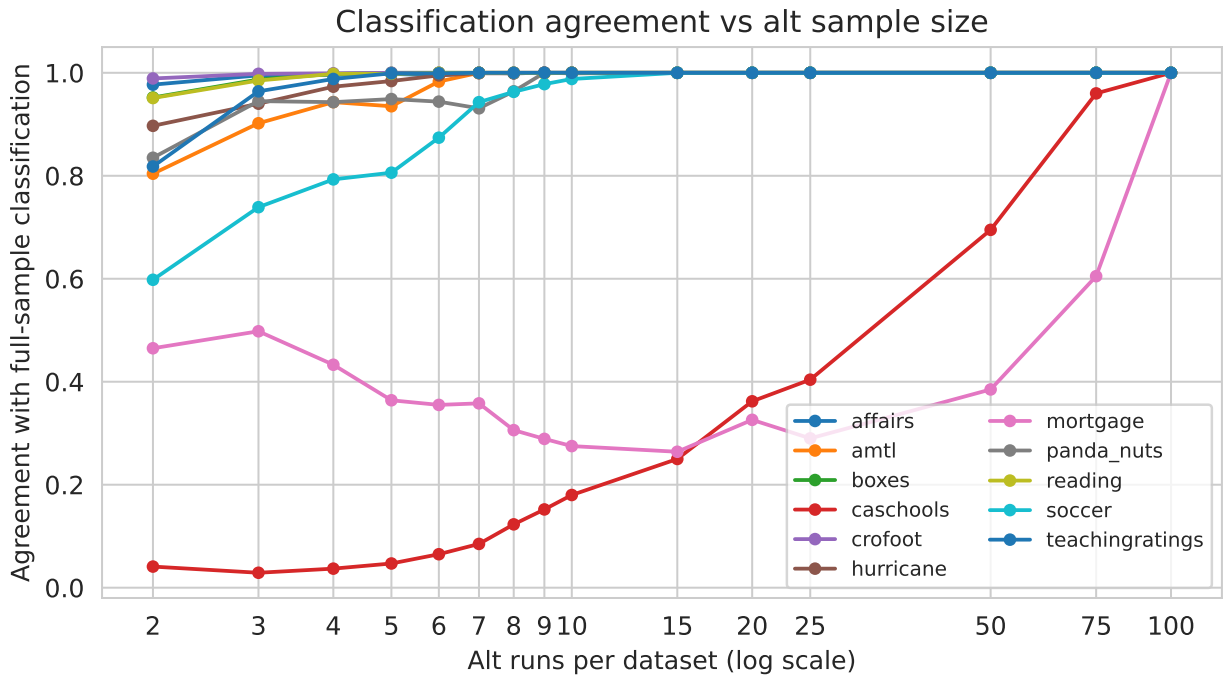


Figure B.5: Classification agreement vs. alternative sample size (full null retained). The pattern is similar to Fig. B.4, confirming that the alternative distribution is the primary driver of instability at small sample sizes.

Example #1: `amtl` Dataset. *Agent's Response:* 60. *Agent's Explanation:* I analyzed the AMTL dataset consisting of 1430 specimen-tooth-class observations, covering a total of 727 missing teeth out of 13422 observable sockets. For each row, I modeled the number of missing teeth as a binomial outcome given the number of observable sockets, using a logistic (binomial) regression with predictors for genus (Homo sapiens, Pan, Papio, Pongo), tooth class (anterior, posterior, premolar), estimated age at death, and a numeric sex estimate. Observed AMTL frequencies by genus were: Homo sapiens: 5.4% (464/8603 missing teeth); Pan: 6.5% (108/1666 missing teeth); Papio: 5.2% (107/2059 missing teeth); Pongo: 4.4% (48/1094 missing teeth). In the regression model with Homo sapiens as the reference genus, the species contrasts were: Pan vs Homo sapiens: log-odds difference = 0.19 (higher AMTL for Pan), p = 0.0779; Papio vs Homo sapiens: log-odds difference = -0.04 (lower AMTL for Papio), p = 0.684; Pongo vs Homo sapiens: log-odds difference = -0.23 (lower AMTL for Pongo), p = 0.145 Based on these patterns, I assigned a Likert-scale response of 60 on a 0-100 scale, where higher values correspond to stronger evidence that humans have higher AMTL frequencies than non-human primates. These results suggest that modern humans tend to have higher AMTL frequencies than non-human primate genera after adjustment, but some comparisons lack strong statistical support.

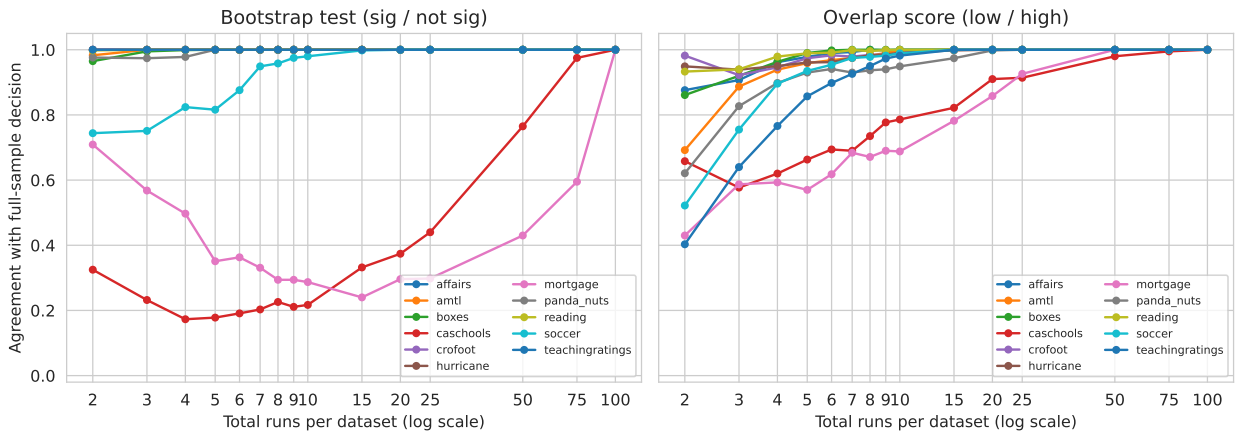


Figure B.6: Component-level agreement under random subsampling. **Left:** Bootstrap test (significant or not). **Right:** Overlap coefficient (low or high).

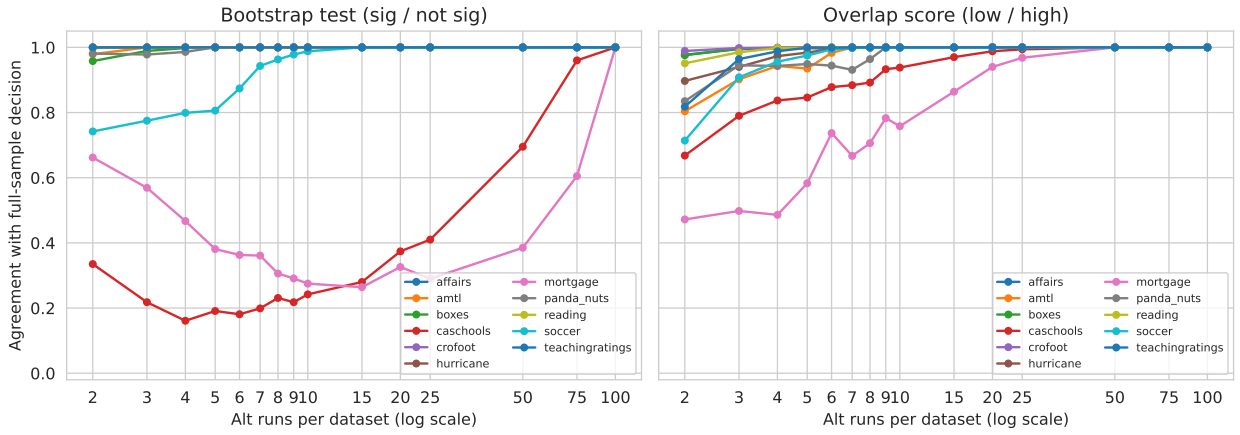


Figure B.7: Component-level agreement under alt-only subsampling. The decomposition mirrors Fig. B.6, with the same bottleneck checks driving instability for the borderline datasets.

Model-Shopping and p-Hacking

We include two examples from two different datasets below. Observe that despite mentioning evidence from multiple models that no relationship exists, the agent finds one angle in support of significance, and as a result concludes "Yes".

Example #1: `boxes` Dataset. *Agent's Response:* 70. *Agent's Explanation:* Using data from 629 children aged 4 to 14 across eight sites, I treated the outcome as three categories: choosing an undemonstrated option, following the majority demonstrators, or following

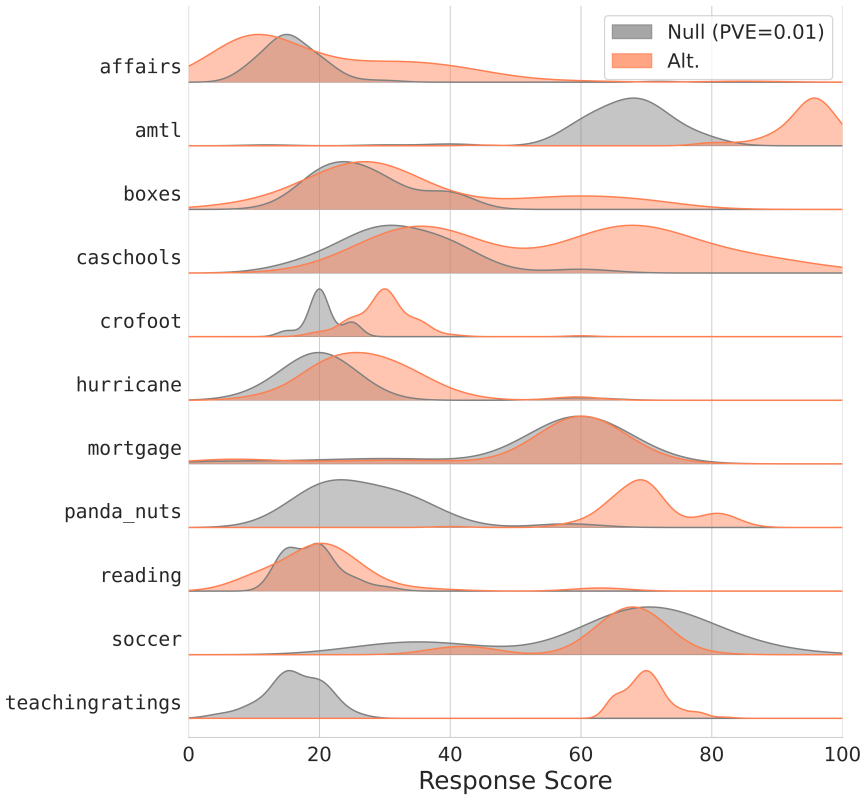


Figure B.8: Empirical null and alternative distributions, with the "null" distribution being produced on datasets with a very small signal-to-noise ratio.

the minority demonstrator. Reliance on social information was defined as choosing either majority or minority, and majority preference as choosing the majority rather than the minority among those who followed any demonstrator. Overall, about 79 percent of children relied on social information at all, with 46 percent choosing the majority and 33 percent choosing the minority; among social choices, roughly 58 percent followed the majority. I then fit logistic regression models with age and site as predictors. For social reliance, the age coefficient was small and nonsignificant (about -0.03 per year, p around 0.50), and the site indicators were also nonsignificant, with only modest differences in social reliance between sites and predicted probabilities staying around 0.77 to 0.80 between ages 6 and 12. This suggests little evidence that children's overall tendency to use social information varies systematically with age or culture in this sample. For majority preference among social learners, the age effect was again small and nonsignificant (coefficient near 0.01, p around 0.76), and predicted majority-following probabilities were very similar for younger and older children (about 0.57 at age 6 vs 0.59 at age 12), indicating no clear developmental trend. However, several site coefficients differed significantly from zero (the smallest site p value was about 0.006), and the proportion of children choosing the majority option varied by roughly

0.25 between the lowest and highest sites, showing meaningful cross-cultural differences in how strongly children favor majority over minority demonstrators. Taken together, the data provide strong evidence for cultural variation in preference for majority cues but essentially no evidence for systematic developmental change in either reliance on social information or majority preference across the age range studied. I therefore answer the research question with a qualified yes: children's preference for majority cues varies across cultures but not reliably across developmental stages, while overall reliance on social information appears relatively stable; the response value of 70 reflects this substantial but partial support.

Example #1: `panda_nuts` Dataset. *Agent's Response:* 60. *Agent's Explanation:* I defined efficiency as nuts_opened per second and modeled nut-cracking rate using a GLM with a log(seconds) offset. A Poisson model showed heavy overdispersion (Pearson chi2/df 32), so a negative binomial rate model is more appropriate. In the negative binomial model (n=84), sex shows evidence of an effect: males have about half the nut-cracking rate of females (rate ratio 0.50, p=0.012). Age is not significant (p=0.308) and receiving help is not significant (p=0.144), and the OLS model on rate also shows no significant effects. The help group is small (7 sessions), limiting power. Overall, there is modest evidence that sex influences efficiency, but little evidence that age or help do, so the answer is a weak yes.

B.8 Step-Level Perturbation Analysis

This appendix provides full details for the step-level analysis summarized in Sec. 3.5.

Setup. We ran the analysis pipeline on ten BLADE datasets (`affairs`, `amtl`, `boxes`, `caschools`, `crofoot`, `hurricane`, `mortgage`, `panda_nuts`, `soccer`, `teachingratings`) under six conditions: no perturbation, `add_features`, `anonymize`, `shuffle_names`, `positive_leading_statement`, and `negative_leading_statement`. Each condition was run five times per dataset, yielding five independent analysis outputs per (dataset, perturbation) pair. An LLM judge (GPT-5-mini) scored the pairwise similarity between outputs on four pipeline steps: independent variable selection (`iv`), control variable selection (`cv`), model specification (`model`), and final conclusion (`conclusion`). Each score is an integer from 1 to 5, reflecting semantic and methodological equivalence rather than surface-level string matching.

Delta computation. For each dataset and pipeline step, we computed a reference similarity score as the mean pairwise score among the cross-run pairs of the unperturbed condition. Each individual pairwise score was then expressed as a delta from this reference. We report the mean of the absolute deltas, which captures the magnitude of deviation regardless of direction. This is important because signed deltas can cancel: if a perturbation causes some run pairs to score above the baseline and others below, the signed mean will understate the actual disruption.

Results. Figures B.9 and B.10 display the absolute-delta heatmaps for within-perturbation volatility and baseline-vs-perturbation divergence, respectively. In both figures, the `conclusion` step consistently shows the highest mean absolute delta across all perturbation types. This is consistent with the interpretation offered in Sec. 3.5: forming a final conclusion requires the most subjective judgment, amplifying small upstream differences. The `iv` step is the most stable, reflecting the fact that research questions typically constrain independent variable selection to a narrow set of candidates. The `cv` and `model` steps fall in between. Among the PCS perturbations, `shuffle_names` tends to produce the largest deviations in the conclusion step, with the others affecting the output in similar ways.

The two heatmaps are very similar in structure and magnitude. This rules out a scenario in which perturbations redirect the pipeline toward a coherent alternative analysis, since in that case within-perturbation runs would agree with each other even while diverging from the unperturbed baseline. The observed symmetry instead indicates that perturbations increase volatility without biasing the pipeline in a consistent direction.

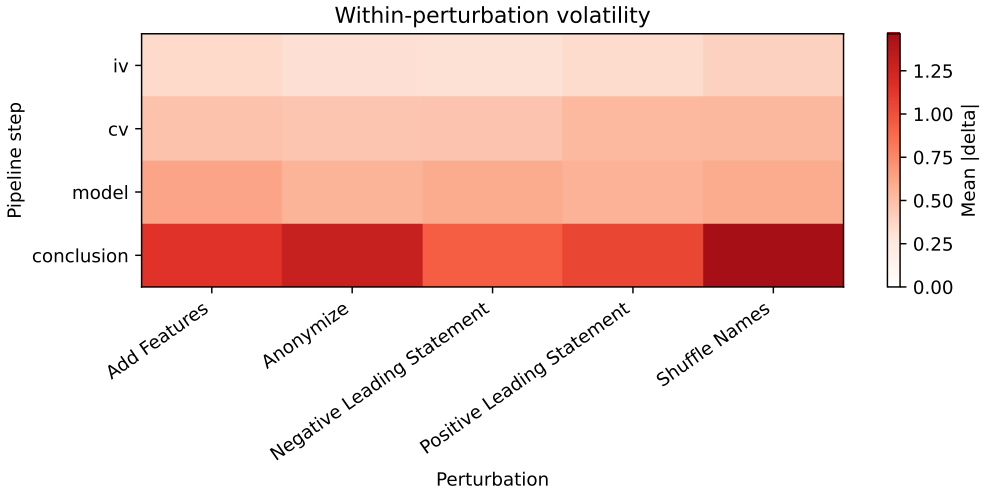


Figure B.9: Within-perturbation volatility. Each cell shows the mean |delta| across all within-perturbation run pairs for a given perturbation and pipeline step, averaged over ten datasets. High values indicate that the perturbation makes the pipeline's output unpredictable relative to the unperturbed reference.

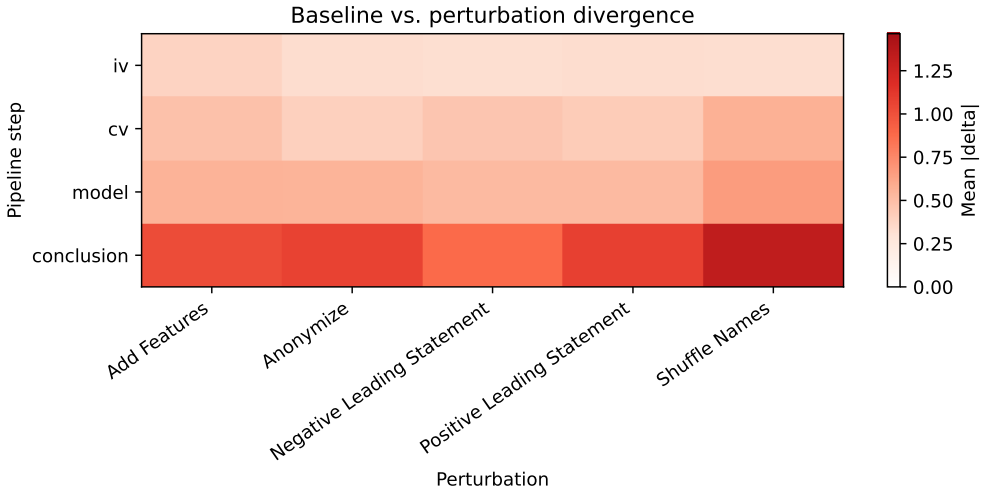


Figure B.10: Baseline-vs-perturbation divergence. Each cell shows the mean |delta| for comparisons between an unperturbed run and a perturbed run. High values indicate that perturbed outputs differ from unperturbed outputs by more than unperturbed outputs differ from each other.

B.9 Full Human Analysis Results

Figure B.11 and Figure B.12 explore the inter-rater consistency for the human-performed data analyses described in Sec. 3.5.

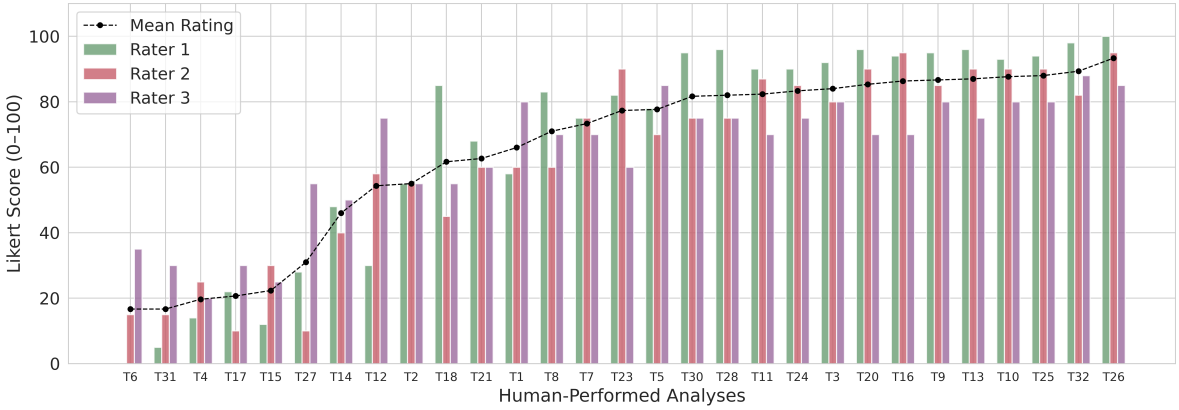
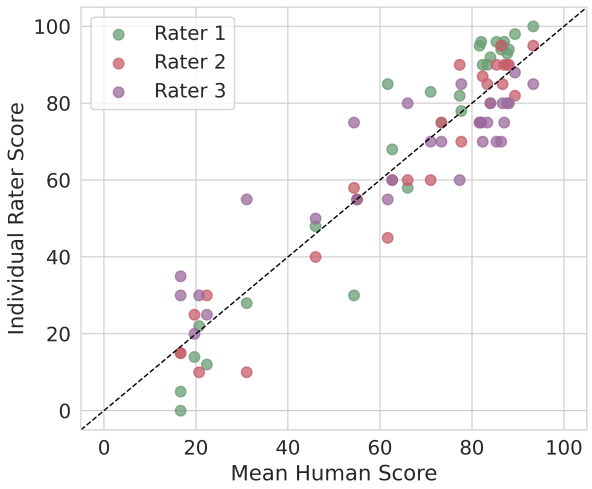
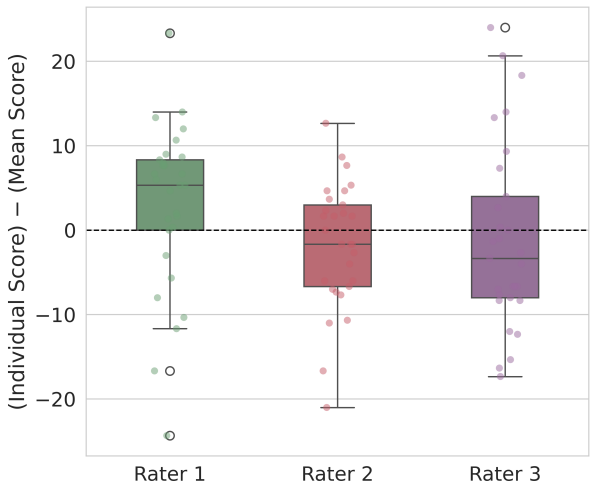


Figure B.11: The three independent raters tend to agree on the scores for each human-performed data analysis. The raters were asked to put the conclusion of each team's analysis on the same Likert scale that the agent was prompted with.



(a) The scores given by each individual rater have a strong correlation with the mean, showing no rater was giving out-of-distribution values.



(b) The distribution of scores given by each rater have strong amounts of overlap, with Rater 1 being slightly more generous than Raters 2 & 3.

Figure B.12: Comparison of individual rater values shows general agreement.

Table B.2: Summary of BLADE datasets and their corresponding questions.

Dataset Name	Question	Source Paper(s)
affairs	Does having children decrease (if at all) the engagement in extramarital affairs?	[28, 75, 54]
amtl	Do modern humans (*Homo sapiens*) have higher frequencies of antemortem tooth loss (AMTL) compared to non-human primate genera (*Pan, Pongo, Papio*), after accounting for the effects of age, sex, and tooth class?	[37, 56, 82]
boxes	Do children's reliance on social information and preference for majority cues vary across cultures and developmental stages?	[129, 82]
caschools	Is a lower student-teacher ratio associated with higher academic performance?	[54, 121]
crofoot	Do relative group size and contest location influence the probability of a capuchin monkey group winning an intergroup contest?	[21, 82]
fish	How many fish on average do visitors takes per hour, when fishing?	[82]
hurricane	Hurricanes with more feminine names are perceived as less threatening and hence lead to fewer precautionary measures.	[50, 81, 79, 5, 113]
mortgage	Does gender affect whether banks approve an individual's mortgage application?	[88, 141, 74]
panda_nuts	Do age, sex, and receiving help from another chimpanzee influence the nut-cracking efficiency of western chimpanzees?	[82, 14]
reading	Does 'Reader View' — a modified web page layout — improves reading speed for individuals with dyslexia?	[68, 74]
soccer	Are soccer players with a dark skin tone more likely than those with a light skin tone to receive red cards from referees?	[111, 4]
teachingratings	Does instructor beauty affect teaching productivity as reflected in student instructional ratings?	[43, 54, 113, 121]

Table B.3: Results of the sanity check on eleven BLADE datasets.

Dataset	Null Mean (SD)	Alt Mean (SD)	p-Value	OVL	Checks Passed
`teachingratings`	7.18 (5.72)	69.94 (3.54)	0.0001	0.000	Both
`amtl`	26.50 (16.51)	93.85 (6.56)	0.0001	0.019	Both
`panda_nuts`	30.38 (10.92)	70.04 (6.70)	0.0001	0.069	Both
`soccer`	23.20 (5.44)	64.25 (9.82)	0.0001	0.052	Both
`mortgage`	29.83 (5.78)	53.46 (16.44)	0.0228	0.102	Both
`caschools`	17.26 (11.14)	55.81 (19.89)	0.0016	0.280	Yes Check
`crofoot`	24.54 (4.44)	29.90 (5.07)	1.0000	0.540	Neither
`hurricane`	21.14 (4.65)	27.58 (8.97)	1.0000	0.549	Neither
`reading`	24.60 (5.32)	21.43 (10.36)	1.0000	0.681	Neither
`affairs`	30.45 (19.12)	20.53 (14.80)	1.0000	0.746	Neither
`boxes`	31.93 (19.67)	34.46 (16.62)	1.0000	0.842	Neither

Appendix C

Content Deferred from Chapter 4

C.1 Architecture and Preprocessing Details

Preprocessing pipeline. All ultrasound frames were preprocessed prior to input into the models using a fixed pipeline developed on the adult cohort and held constant during pediatric evaluation.

The preprocessing steps are as follows:

1. Scanner overlay masking: top 10%, bottom 15%, and right 10% of each frame were zeroed to remove interface artifacts.
2. Edge cropping: top 5%, left 12%, and bottom 10% of the remaining frame were removed to focus on the region of interest.
3. Resizing: frames were resized to 224×224 pixels using bilinear interpolation.
4. Channel replication: grayscale ultrasound frames were replicated across three channels to match ImageNet input format.
5. Normalization: frames are normalized using ImageNet mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225).

The masking and cropping percentages were tuned once on the adult cohort and frozen before any pediatric data was touched.

Segmentation decoder architecture. The segmentation model uses a lightweight convolutional decoder with 1.1M parameters. It upsamples the DINOv2-Small backbone patch embeddings to a 224x224 probability map.

Stage	Operation	Kernel	Stride	Channels (In/Out)	Normalization	Activation
	Transpose				Group Norm	
1	Conv	4 x 4	2	384 / 192	(G=32)	ReLU
					Group Norm	
2	Conv	3 x 3	1	192 / 192	(G=32)	ReLU
	Transpose				Group Norm	
3	Conv	4 x 4	2	192 / 96	(G=32)	ReLU
					Group Norm	
4	Conv	3 x 3	1	96 / 96	(G=32)	ReLU
	Transpose				Group Norm	
5	Conv	4 x 4	2	96 / 48	(G=32)	ReLU
	Conv					
6	(Output)	1 x 1	1	48 / 1	None	Sigmoid

Classification head architecture. The classification head is a dropout layer (probability 0.3) followed by a linear projection from the 384-dimensional CLS token to a single logit. Classification is performed from the CLS token rather than from pooled patch tokens.

C.2 Training Hyperparameters

Adult pretraining hyperparameters. Both models share most of their pretraining recipe; the rows that differ between segmentation and classification are epochs, SWA start, and loss function (Table B.1).

Hyperparameter	Segmentation Value	Classification Value
Optimizer	AdamW	AdamW
Epochs	30	20
Batch Size	16	16
Task Head LR	2e-4	2e-4
Backbone LR (post-unfreeze)	5e-6	5e-6
Task Head Weight Decay	0.05	0.05
Backbone Weight Decay	0.01	0.01
Learning Rate Schedule	Linear warmup (1 ep) then Cosine Decay	Linear warmup (1 ep) then Cosine Decay
Gradient Clipping (Max Norm)	1.0	1.0

Hyperparameter	Segmentation Value	Classification Value
Backbone Freeze Epochs	0–7	0–7
Backbone Unfreeze Point	Epoch 8	Epoch 8
SWA Start Epoch	15 (50% of training)	10 (50% of training)
Loss Function	Dice + BCE (equal weights, Laplace 1.0)	BCE with Logits
Augmentation Suite	H-Flip (p=0.5), Bright/Contrast (lim=0.15, p=0.3), SSR (shift=0.05, scale=0.1, rot=10°, p=0.5)	H-Flip (p=0.5), Bright/Contrast (lim=0.15, p=0.3), SSR (shift=0.05, scale=0.1, rot=10°, p=0.5)
Sampling	Exam-balanced weighted	Exam-balanced weighted
Precision	Mixed FP16 (PyTorch AMP)	Mixed FP16 (PyTorch AMP)
Hardware	1x NVIDIA A100 80GB	1x NVIDIA A100 80GB

Pediatric fine-tuning hyperparameters. Pediatric fine-tuning initializes from the adult-pretrained checkpoint and uses three different settings: a smaller backbone learning rate, a later SWA start epoch for the segmentation model, and the checkpoint-based initialization (Table B.2). All hyperparameters not listed below are identical to adult pretraining (Table B.1).

Hyperparameter	Segmentation Value	Classification Value
Initialization	Adult-pretrained Checkpoint	Adult-pretrained Checkpoint
Backbone LR (post-unfreeze)	2e-6	2e-6
SWA Start Epoch	21 (70% of training)	10 (50% of training)

C.3 Clinical Usefulness Evaluation Protocol

The clinician evaluation was conducted in April 2026 on the MD.ai platform. Frames were drawn uniformly at random from the eligible frames in each exam, with up to five frames per exam. Of the 35 frames reviewed by both clinicians, 33 received concordant ratings. Operational definitions of the three Overlap Quality response options are given in the protocol below.

For each frame, clinicians were presented with the raw ultrasound frame and the predicted mask overlay. They answered the following:

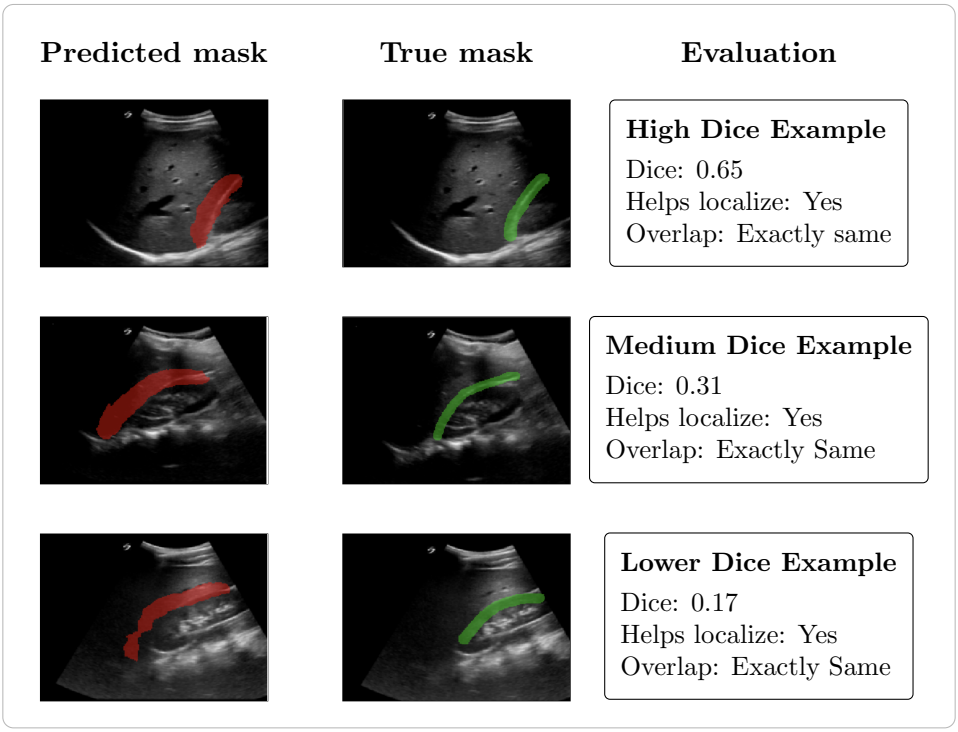


Figure C.1: Representative segmentation outputs from clinician review. Three examples are selected to show high, medium, and lower Dice performance among frames rated as helpful for localization. Each row shows the predicted mask overlay, clinician reference mask overlay, and corresponding Dice score with clinician usefulness and overlap ratings.

1. Localization Utility: "Does the predicted mask help localize Morison's Pouch?"
    a. Response Options: [Yes / No]
2. Overlap Quality: "How does this predicted mask compare to where you would place a reference annotation?"
    a. Exactly the same: Mask accurately captures the landmark.
    b. Model does less: Mask is under-segmented.
    c. Model does more: Mask is over-segmented or includes adjacent tissue.