Generalization vs Specialization under Concept Shift

Alex Nguyen
Princeton University

David J. Schwab*
CUNY Graduate Center

Vudtiwat Ngampruetikorn*
University of Sydney

Abstract

Machine learning models are often brittle under distribution shift, i.e., when data distributions at test time differ from those during training. Understanding this failure mode is central to identifying and mitigating safety risks of mass adoption of machine learning. Here we analyze ridge regression under concept shift—a form of distribution shift in which the input-label relationship changes at test time. We derive an exact expression for prediction risk in the thermodynamic limit. Our results reveal nontrivial effects of concept shift on generalization performance, including a phase transition between weak and strong concept shift regimes and nonmonotonic data dependence of test performance even when double descent is absent. Our theoretical results are in good agreement with experiments based on transformers pretrained to solve linear regression; under concept shift, too long context length can be detrimental to generalization performance of next token prediction. Finally, experiments on MNIST and FashionMNIST further validate our theoretical predictions, suggesting these phenomena represent a fundamental aspect of learning under distribution shift.

1 Distribution shift

It is unsurprising that a model trained on one distribution does not perform well when applied to data from a different distribution. Yet, this out-of-distribution setting is relevant to many practical applications from scientific research [1, 2] to medicine and healthcare [3–6]. A quantitative understanding of out-of-distribution generalization is key to developing safe and robust machine learning techniques. A model that generalizes to arbitrary distribution shifts of course does not exist. The generalization scope of a model, however, needs not be limited to the training data distribution. A question then arises as to how much a model's scope extends beyond its training distribution.

Answering this question requires assumptions on the test distribution. For example, covariate shift, a well-studied setting for distribution shift, assumes a fixed input-label relationship while allowing changes in the input distribution (see, e.g., Refs [7–11]).

We consider *concept shift*—a relatively less-studied setting, in which the input-label relationship becomes different at test time[1] [13, 14], see Fig 1. While many works have studied how to detect and mitigate concept shift [15, 16], characterizing how concept shift affects generalization behavior in neural networks has remained underexplored. In our work, we formulate a minimal model which enables continuous modulation of the input-label function, based on high dimensional ridge regression—a solvable setting that has helped develop intuitions for some of the most interesting phenomena of modern machine learning (see, e.g., Refs [17–25]). We derive an analytical expression for prediction risk under concept shift in the high-dimensional limit and demonstrate that its behavior can change from monotonically decreasing with data for the in-distribution case to monotonically

*DJS and VN contributed to this work equally.

[1]Concept shift or concept drift is sometimes defined to be equivalent to distribution shift [12]. Here, we adopt a narrower definition in which concept shift describes only the change in the input-label relationship [13, 14].

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

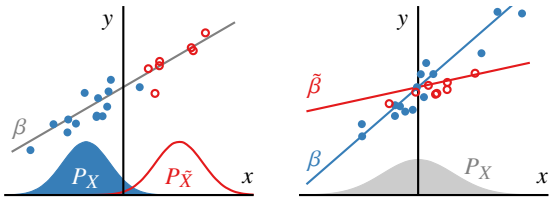

Figure 1: **Two flavors of distribution shift.** Distribution shift describes the scenarios in which the joint input-label distribution of training data $\{(x_1, y_1), (x_2, y_2), \dots\}$ (filled circles) differs from that of test data $\{(\tilde{x}_1, \tilde{y}_1), (\tilde{x}_2, \tilde{y}_2), \dots\}$ (empty circles). *Covariate shift* (left) assumes a fixed input-label relationship but the input distribution differs at test time, i.e., $P_{Y|X} = P_{\tilde{Y}|\tilde{X}}$ but $P_X \neq P_{\tilde{X}}$. For linear regression, this condition means that the regression coefficient β is unchanged. *Concept shift* (right) allows the input-label function to change at test time, which corresponds to a shift in regression coefficient $\beta \neq \tilde{\beta}$ in the linear regression setting. See §2.

increasing, and to nonmonotonic, depending on the degree of concept shift and the properties of the input distribution.

To differentiate these effects from double descent phenomena which can also cause nonmonotonic data dependence of prediction risk [26, 27], we focus on optimally tuned ridge regression which completely suppresses the risk divergence at the interpolation threshold and for which in-distribution prediction risk decreases monotonically with more data [17, 18, 28]. The nonmonotonic behavior we observe is also distinct from effects due to model misspecification [18]. While misspecification—a property of a model relative to the true data-generating process—can produce similar generalization behavior, distribution shift represents a fundamentally different setting in which the data-generating process changes at test time. Our work isolates and characterizes the effects of concept shift, a type of distribution shift, in correctly specified, optimally regularized models.

Our theoretical results in the thermodynamic limit agree well with experiments, based on transformers trained to solve finite-dimensional regression using in-context examples. We show that more in-context examples help improve model performance when concept shift is weak, but can lead to overspecialization for strong concept shift. Finally, we illustrate similar qualitative changes in generalization behavior in classification problems, using MNIST and FashionMNIST as examples. Our work contributes a new theoretical framework for analyzing concept shift that complements an extensive body of work on concept shift detection (see, e.g., Ref [29] for a recent review).

Our main contributions are:

1. We develop an analytically solvable framework that isolates the effects of concept shift—an important yet often overlooked mode of distribution shift.
2. We explain precisely why and how more training data can hurt generalization performance under concept shift, illustrating its effects in several specific settings, including coefficient shrinking, rotation, and feature robustness.
3. We identify and characterize a sharp transition, separating weak and strong concept shift regimes.
4. We show that feature anisotropy creates qualitatively different patterns of risk nonmonotonicity depending on whether concept shift affects high or low-variance features.

Our experiments on transformers and simple classification tasks are in good qualitative agreement with our theoretical findings, hinting at universal behavior that generalizes beyond our relatively simple theoretical settings.

2 Regression setting

Data. The training data consists of N iid input-response pairs $\{(x_1, y_1), \dots, (x_N, y_N)\}$. The input $x \in \mathbb{R}^P$ is a vector of Gaussian features and the response $y \in \mathbb{R}$ is a noisy linear projection of x, i.e.,

$$y = \beta^\mathsf{T} x + \xi \quad \text{with} \quad (x, \xi) \sim \mathcal{N}(\cdot, \Sigma) \times \mathcal{N}(\cdot, \sigma_\xi^2), \tag{1}$$

where $\beta \in \mathbb{R}^P$ denotes the coefficient vector, $\xi \in \mathbb{R}$ Gaussian noise with variance σ_ξ^2, and $\Sigma \in \mathbb{R}^{P\times P}$ the covariance matrix. Similarly, a test data point is an input-response pair $(\tilde{x}, \tilde{y})$, drawn from the same process as the training data, Eq (1), but with a generally different set of parameters—that is,

$$\begin{aligned} &\textit{Training data:} && \begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\cdot, \begin{bmatrix} \Sigma & \Sigma\beta \\ \beta^\mathsf{T}\Sigma & \beta^\mathsf{T}\Sigma\beta + \sigma_\xi^2 \end{bmatrix}\right) \\ &\textit{Test data:} && \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \sim \mathcal{N}\left(\cdot, \begin{bmatrix} \tilde{\Sigma} & \tilde{\Sigma}\tilde{\beta} \\ \tilde{\beta}^\mathsf{T}\tilde{\Sigma} & \tilde{\beta}^\mathsf{T}\tilde{\Sigma}\tilde{\beta} + \tilde{\sigma}_\xi^2 \end{bmatrix}\right), \end{aligned} \tag{2}$$

where in general $\Sigma \neq \tilde{\Sigma}$, $\beta \neq \tilde{\beta}$ and $\sigma_\xi^2 \neq \tilde{\sigma}_\xi^2$. Here we also define signal-to-noise ratio $\text{SNR} \equiv \beta^\mathsf{T}\Sigma\beta/\sigma_\xi^2$

Model. We consider ridge regression in which the predicted response to an input x reads $\hat{y}(x; X, Y) = x \cdot \hat{\beta}_\lambda(X, Y)$, with the coefficient vector resulting from minimizing L_2-regularized mean squared error,

$$\hat{\beta}_\lambda(X, Y) \equiv \arg\min_{b\in\mathbb{R}^P} \frac{1}{N}\|Y - X^\mathsf{T} b\|^2 + \lambda\|b\|^2 = (XX^\mathsf{T} + \lambda N I_P)^{-1} XY. \tag{3}$$

Here $\lambda > 0$ controls the regularization strength, and $Y = (y_1, \ldots, y_N)^\mathsf{T} \in \mathbb{R}^N$ and $X = (x_1, \ldots, x_N)^\mathsf{T} \in \mathbb{R}^{P\times N}$ denote the training data.

Risk. We measure generalization performance with prediction risk,[2]

$$R(X) \equiv \mathbf{E}\left[\|\hat{y}(\tilde{x}; X, Y) - \mathbf{E}(\tilde{y} \mid \tilde{x})\|^2 \mid X\right] = B(X) + V(X), \tag{4}$$

where the last equality denotes the standard bias-variance decomposition with

$$\begin{aligned} B(X) &\equiv \mathbf{E}\left[\|\mathbf{E}(\hat{y}(\tilde{x}; X, Y) \mid X, \tilde{x}) - \mathbf{E}(\tilde{y} \mid \tilde{x})\|^2 \mid X\right] \\ V(X) &\equiv \mathbf{E}\left[\|\hat{y}(\tilde{x}; X, Y) - \mathbf{E}(\hat{y}(\tilde{x}; X, Y) \mid X, \tilde{x})\|^2 \mid X\right]. \end{aligned}$$

Substituting the predictor from ridge regression into the above equations yields

$$B(X) = \left(\frac{\Psi}{\Psi + \lambda I_P}\beta - \tilde{\beta}\right)^\mathsf{T} \tilde{\Sigma} \left(\frac{\Psi}{\Psi + \lambda I_P}\beta - \tilde{\beta}\right) \quad \text{and} \quad V(X) = \sigma_\xi^2 \frac{1}{N} \operatorname{Tr}\left(\tilde{\Sigma}\frac{\Psi}{(\Psi + \lambda I_P)^2}\right), \tag{5}$$

where $\Psi \equiv XX^\mathsf{T}/N$ is the empirical covariance matrix.

It is instructive to consider the idealized limits of $N = 0$ and $N \to \infty$. First, when $N = 0$, inductive biases (e.g., from model initialization and regularization) dominate. For ridge regression, Eq (3), all model parameters vanish, $\hat{\beta}_\lambda = 0$, and the resulting predictor outputs zero regardless of the input, i.e., $\hat{y}(x) = 0$ for any x. As a result, $R_{N=0}(X) = \mathbf{E}[\mathbf{E}(\tilde{y} \mid \tilde{x})^2] = \tilde{\beta}^\mathsf{T}\tilde{\Sigma}\tilde{\beta}$, see Eq (4). Second, when $N \to \infty$, the empirical covariance matrix approaches the true covariance matrix $\Psi \to \Sigma$ and, taking the limit $\lambda \to 0^+$, we obtain $R_{N\to\infty}(X) = (\beta - \tilde{\beta})^\mathsf{T}\tilde{\Sigma}(\beta - \tilde{\beta})$.[3] When $\tilde{\beta} = \beta$, we see that $R_{N=0}(X) = \beta^\mathsf{T}\tilde{\Sigma}\beta > 0$ whereas $R_{N\to\infty}(X) = 0$ (see also Fig 2). That is, infinite data is better than no data, as expected.

This intuitive picture breaks down under concept shift, $\tilde{\beta} \neq \beta$. Consider, for example, $\tilde{\beta} = 0$ which indicates that none of the features predicts the response at test time. In this case, $R_{N=0}(X) = 0$ and $R_{N\to\infty}(X) = \beta^\mathsf{T}\tilde{\Sigma}\beta > 0$ (see also Fig 2); that is, even *infinitely more* data decreases test performance in the presence of concept shift.

3 High dimensional limit

To better understand this counterintuitive phenomenon in the context of high dimensional learning, we focus on concept shift without covariate shift, i.e., $\tilde{\beta} \neq \beta$ and $\tilde{\Sigma} = \Sigma$, and take the thermodynamic limit $N, P \to \infty$ and $P/N \to \gamma \in (0, \infty)$. In this limit, prediction risk becomes deterministic $R(X) \to \mathcal{R}$ with the bias and variance contributions given by (see Appendix for derivation; see also Ref [30]),

$$B(X) \to \mathcal{B} = \lambda^2 \nu'(-\lambda) \frac{\beta^\mathsf{T}\Sigma\hat{G}_\Sigma^2(-\lambda)\beta}{\frac{1}{P}\operatorname{Tr}[\Sigma\hat{G}_\Sigma^2(-\lambda)]} - 2\lambda\beta^\mathsf{T}\Sigma\hat{G}_\Sigma(-\lambda)(\beta - \tilde{\beta}) + (\beta - \tilde{\beta})^\mathsf{T}\Sigma(\beta - \tilde{\beta}) \tag{6}$$

$$V(X) \to \mathcal{V} = \sigma_\xi^2 \gamma \left[\nu(-\lambda) - \lambda\nu'(-\lambda)\right], \tag{7}$$

[2] We follow the convention in Ref [18] where the risk is the expected squared error between the predictor and the mean of the test-time conditional distribution, effectively subtracting the irreducible noise variance.

[3] As $N \to \infty$ at fixed P, the variance term vanishes, $\mathcal{V}(X) \sim O(N^{-1})$, and the optimal regularization is $\lambda^* = 0$.

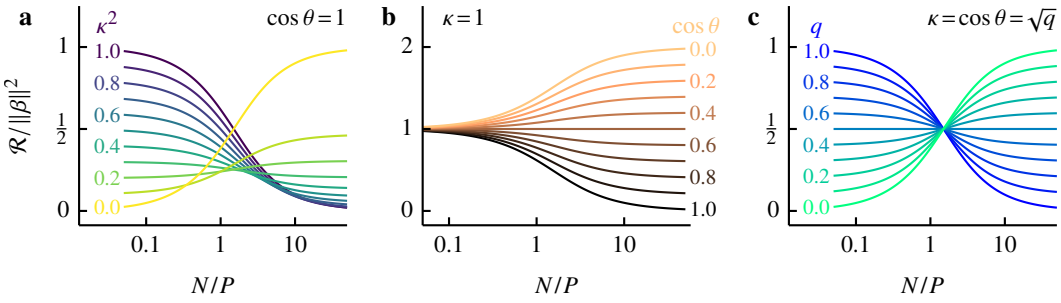

Figure 2: **More data hurts performance when concept shift is strong.** We depict the data dependence of asymptotic prediction risk for isotropic features, Eq (9), under three concept shift settings of varying degree, parametrized by coefficient alignment $\cos\theta = \beta\cdot\tilde{\beta}/\|\beta\|\|\tilde{\beta}\|$ and scaling factor $\kappa = \|\tilde{\beta}\|/\|\beta\|$ (see legend). **a** *Shrinking coefficients:* $\tilde{\beta} = \kappa\beta$. **b** *Rotating coefficients:* $\|\tilde{\beta}\| = \|\beta\|$ but θ varies. **c** *Mixture of robust and nonrobust features:* $\tilde{\beta}_i = \beta_i$ if feature i is robust, otherwise $\tilde{\beta}_i = 0$. This setting is parametrized by $q = \kappa^2 = \cos^2\theta$. We set $\mathrm{SNR} = 1$ in a-c, and the regularization is in-distribution optimal $\lambda = \gamma/\mathrm{SNR}$.

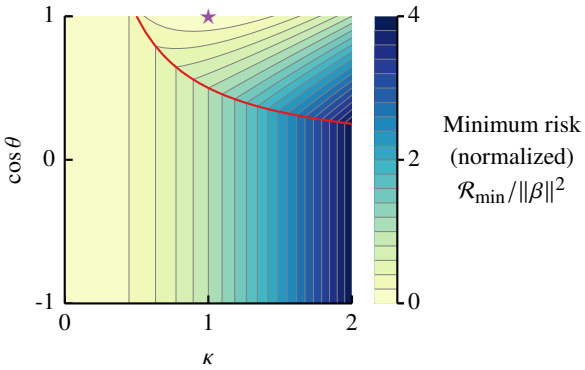

Figure 3: **Concept shift induces a phase transition in generalization behavior.** We depict minimum prediction risk $\mathcal{R}_{\min}$ of optimally tuned ridge regression ($\lambda = \gamma/\mathrm{SNR}$) as κ and $\cos\theta$ vary, see Eq (10) for definitions. No concept shift corresponds to $\kappa = \cos\theta = 1$ (star). The thick curve, $\kappa\cos\theta = 1/2$, separates the weak and strong concept shift regimes. More training data improves generalization only when concept shift is weak, $\kappa\cos\theta > 1/2$. Outside this region, *any* data hurts generalization.

where we define $\hat{G}_\Sigma(z) \equiv (m(z)\Sigma - zI_P)^{-1}$, $m(z) \equiv (1 + \gamma\nu(z))^{-1}$ and $\nu(z)$ is the unique solution of the self-consistent equation,

$$\nu(z) = \frac{1}{P}\,\mathrm{Tr}[\Sigma\hat{G}_\Sigma(z)] \quad \text{with} \quad \nu(z) \in \mathbb{C}^+. \tag{8}$$

We note that concept shift enters prediction risk only through the last two terms of the bias, Eq (6), whereas the variance, Eq (7), is completely unaffected.

3.1 Isotropic features

When $\Sigma = I_P$, the bias contribution to prediction risk reads (see Appendix for a closed-form expression for $\nu(z)$)

$$\mathcal{B} = \|\beta\|^2\left[\lambda^2\nu'(-\lambda) - \frac{2\lambda(1+\gamma\nu(-\lambda))}{1+\lambda(1+\gamma\nu(-\lambda))}(1 - \kappa\cos\theta) + 1 - 2\kappa\cos\theta + \kappa^2\right], \tag{9}$$

where we quantify concept shift via two parameters

$$\begin{aligned} &\textit{Coefficient alignment:} \quad \cos\theta \equiv \frac{\beta\cdot\tilde{\beta}}{\|\beta\|\|\tilde{\beta}\|} \\ &\textit{Scaling factor:} \qquad\qquad \kappa \equiv \|\tilde{\beta}\|/\|\beta\|. \end{aligned} \tag{10}$$

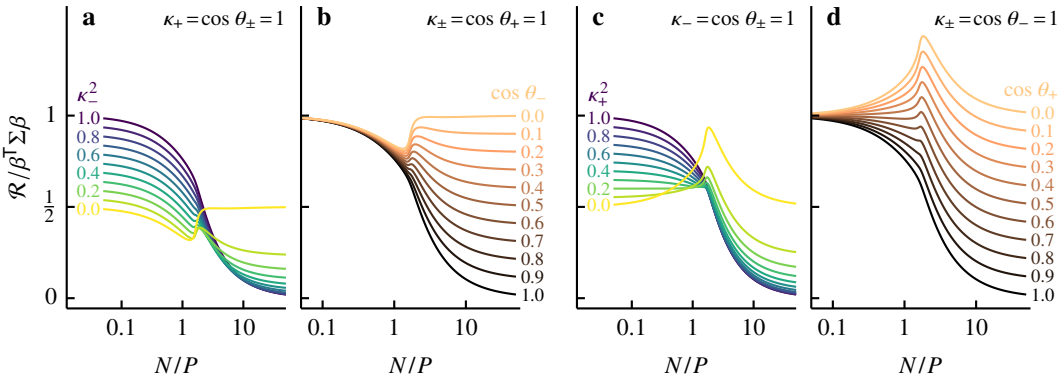

Figure 4: **Anisotropic features can lead to risk nonmonotonicity.** We illustrate prediction risk for the two-scale model, Eq (11), with aspect ratio $s_-/s_+ = 0.1$, spectral weights $\rho_+ = \rho_- = 1/2$ and signal fraction $\pi_+ = \pi_- = 1/2$. Concept shift is parametrized by coefficient alignments $\cos\theta_\pm = \beta_\pm \cdot \tilde{\beta}_\pm / \|\beta_\pm\| \|\tilde{\beta}_\pm\|$ and scaling factors $\kappa_\pm = \|\tilde{\beta}_\pm\| / \|\beta_\pm\|$ where the subscripts indicate the variance $s_\pm$ of the affected features. We consider the settings in which concept shift affects either low or high-variance features and either alignment or scale; that is, we vary only one out of four parameters, $\theta_\pm$ and $\kappa_\pm$, at a time (see legend). **a** Shrinking coefficients of low-variance features. **b** Rotating coefficients of low-variance features. **c** and **d** Same as Panels a and b, but concept shift affects high-variance features via κ_+ and $\cos\theta_+$, respectively. Here $\mathrm{SNR} = 1$ and regularization is in-distribution optimal.

Figure 2 depicts thermodynamic-limit prediction risk for isotropic features under concept shift. We focus on in-distribution optimal ridge regression which corresponds to setting $\lambda = \gamma/\mathrm{SNR}$, ensuring that double descent is absent (see, e.g., Refs [17, 18]). This choice is motivated by its practical relevance (as it can be approximated via cross-validation) and because it allows us to isolate the effects of concept shift from the confounding non-monotonicity of double descent.

In Fig 2a, we consider the effects of shrinking coefficients—the coefficient vector becomes smaller at test time without changing direction, $\theta = 0$ and $\kappa \leq 1$. When $\kappa = 1$, concept shift is absent and prediction risk monotonically decreases with more training data, as expected for optimally-tuned ridge regression. As κ decreases and the features become less predictive of the response at test time, prediction risk starts to increase with training sample size. This transition occurs at $\kappa = 1/2$, at which $\mathcal{R}/\|\beta\|^2 = 1/4$.

In Fig 2b, we observe a similar crossover when the magnitude of the coefficient vector is fixed but its direction changes at test time, $\kappa = 1$ and $\theta \geq 0$. Examples of this concept shift setting include the case where some features have the opposite effects at test time, described by $\tilde{\beta}_i = -\beta_i$ for the affected coefficient. Here we see that more data hurts when $\cos\theta < 1/2$ (or equivalently $\theta > \pi/3$).

In Fig 2c, we consider a mixture of robust and nonrobust features with $\tilde{\beta}_i = \beta_i$ if feature i is robust, otherwise $\tilde{\beta}_i = 0$. That is, the robust features have the same effects at test time whereas the nonrobust ones become uninformative of the response variables. In this case, we have $\kappa = \cos\theta = \sqrt{q}$ where $0 \leq q \leq 1$ denotes the fraction of robust features. We see again that adequately strong concept shift changes the data-dependent behavior of generalization properties from improving to worsening with more training data.

Indeed we can make a more general statement about the transition between weak and strong concept shift. First, we observe that for in-distribution optimal ridge regression, $\lambda = \gamma/\mathrm{SNR}$, the thermodynamic-limit risk is always monotonic in γ. To determine whether more data improve generalization, we only need to compare the limits of no data and infinite data: $\mathcal{R}_{N=0} = \kappa^2 \|\beta\|^2$ and $\mathcal{R}_{N\to\infty} = (1 - 2\kappa\cos\theta + \kappa^2)\|\beta\|^2$. It follows immediately that two qualitatively distinct regimes exist. When $\kappa\cos\theta > 1/2$, concept shift is weak and more data improves generalization. When $\kappa\cos\theta < 1/2$, concept shift is strong and more data hurts, see Fig 3. In the thermodynamic limit, this sharp boundary manifests as a nonanalyticity in the minimal prediction risk, analogous to a phase transition in statistical mechanics. When $\kappa\cos\theta = 1/2$, prediction risk becomes completely

independent of training sample size—i.e., $\mathcal{R} = \mathcal{R}_{N=0} = \mathcal{R}_{N\to\infty} = \kappa^2\|\beta\|^2$. Quite remarkably, this analysis does not depend on SNR.[4]

3.2 Anisotropic features

To study the effects of anisotropy, we consider a two-scale model in which the spectral density of the covariance matrix is a mixture of two point masses at s_- and s_+ with weights ρ_- and $\rho_+ = 1-\rho_-$, respectively. Without loss of generality, we let $s_+ \geq s_-$ throughout this section. These two modes define subspaces into which we decompose the coefficients, $\beta = \beta_- + \beta_+$ with $\Sigma\beta_\pm = s_\pm\beta_\pm$, and similarly for $\tilde{\beta}$. We write down the bias contribution to prediction risk

$$\mathcal{B} = \beta^\mathsf{T}\Sigma\beta \sum_{\tau\in\pm} \pi_\tau \left[\lambda^2 \frac{\nu'(-\lambda)g_\tau^2(-\lambda)}{\sum_{\upsilon\in\pm}\rho_\upsilon s_\upsilon g_\upsilon^2(-\lambda)} - 2\lambda g_\tau(-\lambda)(1-\kappa_\tau\cos\theta_\tau) + 1 - 2\kappa_\tau\cos\theta_\tau + \kappa_\tau^2\right], \quad (11)$$

where $g_\pm(z) \equiv (m(z)s_\pm - z)^{-1}$ and $\pi_\pm \equiv \beta_\pm^\mathsf{T}\Sigma\beta_\pm/\beta^\mathsf{T}\Sigma\beta$ denotes the signal fraction at each scale. Similarly to the isotropic case, we quantify concept shift using coefficient alignments $\cos\theta_\pm \equiv \beta_\pm\cdot\tilde{\beta}_\pm/\|\beta_\pm\|\|\tilde{\beta}_\pm\|$ and scaling factors $\kappa_\pm \equiv \|\tilde{\beta}_\pm\|/\|\beta_\pm\|$, both of which now depend also on variance.

Figure 4 illustrates prediction risk under concept shift for two-scale covariates. We numerically tune the ridge regularization strength λ such that the in-distribution risk is minimized and the divergences associated with multiple descent phenomena are completely suppressed. We consider the cases where concept shift affects only low-variance features via either κ_- or $\cos\theta_-$ (Fig 4a and b), or only high-variance ones via either κ_+ or $\cos\theta_+$ (Fig 4c and d). In all cases, we see that test performance develops nonmonotonic data dependence as test data deviates from in-distribution settings; however, its behavior is qualitatively distinct for low and high-variance concept shift. To isolate the effects of covariate statistics, we set the signal fraction to $\pi_+ = \pi_- = \frac{1}{2}$ so that low and high-variance features carry the same signal strength.

In Fig 4a and b, we depict the effects of concept shift on low-variance features. Panel a corresponds to the case where the low-variance coefficients β_- shrink at test time, $\kappa_- \leq 1$, thus suppressing the signal associated with low-variance features. Similarly to the isotropic case (Fig 2a), prediction risk decreases with N for weak concept shift. However, as $\|\beta_-\|$ becomes smaller, we see that generalization performance exhibits nonmonotonic behavior; too much training data can worsen generalization. Panel b turns to the case in which β_- rotates at test time, $\cos\theta_- \leq 1$ (cf. Fig 2b). We observe a similar crossover in the data dependence of prediction risk as concept shift becomes stronger. While generalization improves with more data in the low-data limit, this improvement continues monotonically only for adequately weak concept shifts.

In Fig 4c and d, concept shift affects only high-variance features via shrinking coefficients (panel a) and rotated coefficients (panel b). We see again that strong concept shift leads to nonmonotonic data dependence of prediction risk. However, strong concept shift on high-variance features results in risk maximum at intermediate training data size N. This behavior contrasts sharply with low-variance concept shift, depicted in Fig 4a and b.

Indeed we could have anticipated the intriguing differences between concept shift affecting low and high-variance features. The data dependence in Fig 4 results from the fact that it takes more data to learn low-variance features and their effects. At low N, high-variance features dominate prediction risk; more data hurts when these features do not adequately predict the response at test time, Fig 4c and d. On the other hand, low-variance features affect test performance only when the training sample size is large enough to influence the model. Sufficiently strong concept shift on these features thus results in detrimental effects of more data at larger N (compared to high-variance concept shift), Fig 4a and b. We emphasize that the nonmonotonic data dependence of prediction risk due to concept shift is unrelated to double descent phenomena (which describe risk nonmonotonicity in suboptimally tuned models).

4 Transformer experiments

To test the applicability of our theoretical framework to realistic scenarios, we train transformers to perform in-context learning (ICL) of (noisy) linear functions, which were argued to perform optimal

[4] SNR controls how prediction risk depends on training data size N (Fig 2), but not the transition between weak and strong concept shift regimes (Fig 3).

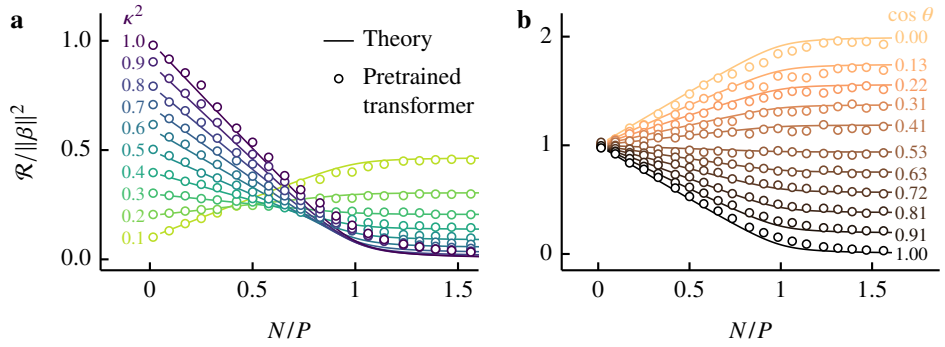

Figure 5: **Longer context can hurt in-context learning performance.** We depict prediction risk of transformers (circles), trained to solve linear regression tasks using in-context examples (see 4 for details). We compare in-context regression We consider two concept shift settings. **a** Coefficient shrinking parametrized by κ (see legend; cf. Fig 2a). **b** Coefficient rotation parametrized by θ (see legend; cf. Fig 2b). We compare the asymptotic prediction risk of transformers on isotropic data with the predictions from our theory (Section 3) under two concept shift settings. Circles depict MSE loss attained by the transformer whereas lines depict the optimally regularized asymptotic limit. We observe strong agreement between experimental results and thermodynamic limits. Here, SNR=128.

ridge regression [31–33]. As many-shot prompting, enabled by recently expanded context windows, has shown promise in improving performance [34], it is timely to investigate how distribution shift affects whether longer context is beneficial. These experiments serve as a validation of our theory, confirming that the key phenomena derived from our analytical model persist in a more complex, data-driven setting.

In our experiment, a transformer takes as its input a series of points (x_i, y_i) on an unknown function $y_i = f(x_i)$ for $i = 1, 2, \ldots, n-1$, terminating with a 'query' x_n whose function value y_n is the prediction target. The model is trained to minimize the mean square loss $\mathcal{L} = \mathbb{E}[(y_n - \hat{y}_n)^2]$. This setup has proved valuable in developing intuitions about ICL [31–33, 35–38]. Here, we focus on noisy linear functions—$f_\beta(x) = \beta^\mathsf{T} x + \xi$ with ξ denoting the noise term—and investigate the generalization properties of the learned ICL solution, implemented by a transformer, under concept shift at test time.

We consider linear regression in $P = 32$ dimensions. The training tasks, parametrized by β, are drawn iid from a standard normal distribution and kept fixed during training. We generate a total of 2^{20} training tasks, which are sufficient to elicit general-purpose ICL [33, 37]. An input sequence is drawn from $y = \beta^\mathsf{T} x + \xi$ with $(x, \xi) \sim \mathcal{N}(0, I_P) \times \mathcal{N}(0, \sigma^2)$. We choose $\sigma^2 = 0.5$. The input-response pairs are newly generated each time the model takes an input sequence. We adopt the nanoGPT architecture [39] with eight layers, an embedding dimension of 128, learnable position embeddings, and causal masking. The model is trained to minimize the next token mean squared error (MSE) using Adam with a learning rate of 0.0001. At test time, we sample 10,000 new tasks and compute the in-distribution prediction risk simply as the MSE of the transformer on test tasks. To compute risk under concept shift, the transformer is presented with a context, $(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, \tilde{x})$, in which $y_i = \beta^\mathsf{T} x_i + \xi$. But the query $\tilde{x}$ is related to the final prediction target $\tilde{y}$ via a linear function $\tilde{y} = \tilde{\beta}^\mathsf{T} \tilde{x}$ with $\tilde{\beta} \neq \beta$ in general.

Figure 5 depicts ICL prediction risk for linear regression under concept shift as a function of in-context sample size N. We parametrize the degree of concept shift using the scaling factor and cosine similarity, described in Eq (10). We consider two specific settings: in panel a, $\kappa \leq 1$ and $\cos\theta = 1$, and in panel b, $\kappa = 1$ and $\cos\theta \leq 1$ (cf. Fig 2a and b). In both cases, we compare ICL prediction risk (symbols) with our thermodynamic-limit theory (lines). We see that they are in good quantitative agreement. In particular, the context-length dependence ICL prediction risk changes from improving to worsening with longer context as concept shift becomes more severe.

To further investigate how transformers handle concept shift with anisotropic features, we conducted experiments comparing models trained on isotropic versus anisotropic data. Figure 6 illustrates how feature anisotropy influences the generalization behavior of transformer-based in-context regression under varying degrees of concept shift. When exposed to two-scale anisotropic data (see §3.2) where

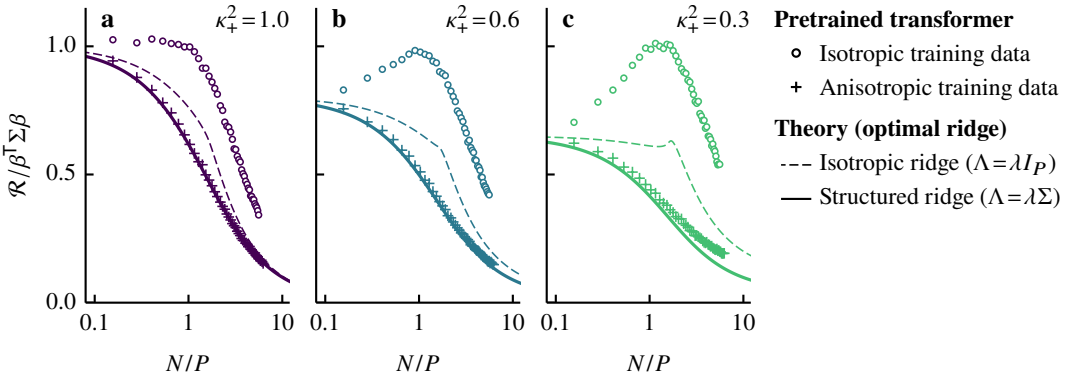

Figure 6: **Feature anisotropy modulates concept shift effects in transformer-based in-context regression.** We depict prediction risk on two-scale, anisotropic data (see §3.2) for optimal ridge regression (curves) and transformers (symbols) trained to solve linear regression tasks using in-context examples. The theoretical results are for optimally tuned isotropic and structured ridge penalties (dashed and solid curves, respectively). The transformers' results are for isotropic and anisotropic pretraining data (circles and plus symbols, respectively). Concept shift affects only high-variance features via the scaling factor $\kappa_+^2 = 1.0, 0.6, 0.3$ in **a**, **b**, and **c**, respectively ($\kappa_- = \cos\theta_\pm = 1$ throughout). Transformers trained on anisotropic data (pluses) closely follow the solid curve, suggesting that they effectively implement a structured regularization that adapts to feature anisotropy—effectively isotropizing the features and averaging the impact of concept shift ($\kappa^2 = (\kappa_+^2 + \kappa_-^2)/2 = (1 + \kappa_+^2)/2$). Transformers trained on isotropic data (circles) exhibit nonmonotonic risk under concept shift, but with much more pronounced peaks compared to theoretical predictions (dashed). Here $s_-/s_+ = 0.1$, $\rho_+ = \rho_- = 1/2$, $\pi_+ = \pi_- = 1/2$ and SNR=1.

concept shift affects only high-variance features, transformers exhibit markedly different behaviors depending on their training data distribution. Transformers pretrained on anisotropic data closely align with theoretical predictions for optimally tuned ridge regression with structured penalties, indicating that they can apply different penalties to features based on their variances. This strategy is equivalent to whitening the features, resulting in an isotropized regression problem with an effective concept shift scale parameter $\kappa^2 = (1 + \kappa_+^2)/2$. In contrast, transformers trained on isotropic data struggle to adapt optimally to anisotropic test data, exhibiting more pronounced nonmonotonic risk curves than theoretically predicted for isotropic ridge penalties. This difference becomes particularly evident as concept shift strengthens (panel c), suggesting that the priors learned during pretraining significantly influence how transformers adapt to concept shift in features with different scales. These results highlight the importance of data structure in determining how models respond to distribution shifts and demonstrate that transformers can implicitly learn sophisticated regularization strategies when exposed to appropriately structured training data. We note that this agreement between theory and experiments is not trivial, as transformers perform in-context regression on finite data, in finite dimensions, and via a learned, data-driven algorithm, whereas our theoretical analyses are based on optimally tuned ridge regression in the thermodynamic limit.

5 Classification experiments

So far we have focused on regression problems. In this section, we experimentally test whether the key insights from our theory extend to the qualitatively different setting of classification, demonstrating that they likely capture more general phenomena of learning under concept shift. We consider standard multinomial logistic regression for MNIST [40] and FashionMNIST [41], using Adam optimizer with a minibatch size of 500 and a learning rate of 0.001 for 2,000 epochs.

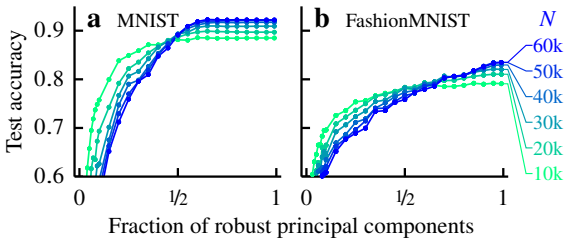

Figure 7: **Concept shift qualitatively changes data dependence of test accuracy** in (a) MNIST and (b) FashionMNIST experiments for various training data size N (see legend). See §5 for details.

To vary training sample size N, we choose training data points at random (without replacement); all of the training data is used when $N = 60{,}000$.

We modify the test datasets by performing principal component analysis (PCA) on the images and designating the resulting features as either 'robust' or 'nonrobust' (cf. Fig 2c; see also §3.1). We shuffle nonrobust features across data points to decorrelate them from the labels while preserving marginal statistics, whereas robust features are unchanged. In Fig 7, we let lower-variance features be nonrobust and parametrize concept shift strength with the variance threshold that separates robust from nonrobust features. We depict the test accuracy as a function of concept shift strength, and observe that when concept shift is strong and few features are robust, more data hurts test accuracy, in qualitative agreement with our results for linear regression problems (cf. Figs 2 and 5). The qualitative agreement between our theoretical results and experiments on transformers, MNIST, and FashionMNIST suggests that the phenomenon we identify is not specific to linear regression, providing further validation for the broader relevance of our theory.

6 Related works

There is a significant body of work on out-of-distribution generalization with covariate shift, i.e., domain generalization (see, e.g., Refs [8, 42]), but significantly less work has been done on concept shift generalization. However, even within covariate shift, an understanding of out-of-distribution generalization remains elusive [11]. Within concept shift, work has tended to focus instead on detection and mitigation strategies rather than generalization (see, e.g., Ref [29]). Our work contributes to a distinct paradigm. We use tools from statistical mechanics and random matrix theory to derive an exact, analytical characterization of generalization behavior under a discrete shift. This approach allows us to uncover new phenomena, such as the sharp phase transition between weak and strong concept shift regimes.

Linear regression has proved a particularly useful setting for investigating learning phenomena. Random matrix universality makes it well-suited for theoretical investigations, yielding much-needed insights into high-dimensional learning [18, 19, 21–25, 43], including covariate shift generalization [44]. Linear models have also contributed substantially to our developing understanding of in-context learning in transformers [31–33, 35–38]; our work leverages this setting for experimental validation of our theory. Other works use linear attention mechanisms to analyze transformer models (see, e.g., Ref [45]).

The nonmonotonic behavior we identify is distinct from double descent phenomena [26, 27]. Indeed, our theoretical results in Figs 2-4 are for optimal regularization, for which double descent is absent [17, 18, 28], thus demonstrating that concept shift induces risk nonmonotonicity through fundamentally different mechanisms.

We build on analytical techniques developed for high-dimensional ridge regression, but address a phenomenon distinct from prior works. In particular, optimally tuned models can exhibit nonmonotonic risk due to model misspecification [18]—i.e., a mismatch between model class and data-generating process, irrespective of test-time shift. This mechanism differs from distribution shift, in which the data-generating process itself changes at test time. Our framework specifically isolates the effects of concept shift in correctly specified models.

In-context learning of transformers has been studied under covariate shift [36, 46, 47] but not with respect to concept shift. Song *et al.* [48] studied how transformers generalize to out-of-distribution tasks through symbolic manipulation via induction heads, but the scenario of within-context concept shift remains underexplored. While Agarwal *et al.* [34] showed that longer context windows typically benefit in-context learning, they also observed performance degradation in tasks such as MATH.

7 Conclusion and outlook

We introduce a ridge regression model for concept shift. Our model is exactly solvable in the high dimensional limit. We show that concept shift can change the qualitative behavior of generalization performance; for sufficiently strong concept shift, ridge regression fails to generalize even with infinite data (Figs 2 and 3). In addition, we demonstrate that input anisotropy can lead to nonmonotonic data dependence of prediction risk. In particular, too much data can harm generalization (Fig 4a and b) and

more data may only improve generalization above a certain threshold (Fig 4c and d). We emphasize that our results are for optimally tuned ridge regression and thus differ from risk nonmonotonicity due to double and multiple descent phenomena which are absent under optimal regularization [28].

Taken together, our work provides a fresh perspective on a lesser-studied mode of distribution shift. Our model offers a relatively simple setting for building intuitions and testing hypotheses about concept shift generalization. Although our theoretical analyses are exact only for ridge regression in the high-dimensional limit, the qualitative conclusions generalize beyond this idealized setting. Indeed, our theoretical prediction agrees *quantitatively* with experiments on regression in finite dimensions from finite samples, using the in-context learning ability of a transformer model as a learning algorithm. Additionally, our classification experiments suggest that the insights from our theory may apply in more general settings. Finally, our theoretical results for anisotropic features have implications for understanding data dimension reduction techniques—such as principal component regression in which low-variance features are discarded—under concept shift. While we acknowledge that real-world concept shifts occur in more complex contexts, our approach allows us to isolate and precisely characterize key phenomena, establishing a theoretical foundation in tractable models and providing a critical step toward understanding more complex settings.

Our work has limitations that open avenues for future research. First, our theoretical results are derived for linear models in the high-dimensional limit and may not quantitatively hold for complex nonlinear systems. Second, we analyze a discrete concept shift, leaving the study of gradual or continuous shifts for future work. Finally, our work characterizes a failure mode but does not propose an algorithm for mitigating its effects. We hope our work provides a foundation for addressing these important questions and encourages further systematic investigations of the rich learning phenomena induced by concept shift

Acknowledgments and Disclosure of Funding

Alex Nguyen is supported by NIH grant RF1MH125318. DJS was partially supported by a Simons Fellowship in the MMLS, a Sloan Fellowship, and the National Science Foundation, through the Center for the Physics of Biological Function (PHY-1734030). VN acknowledges research funds from the University of Sydney. This work was supported in part by the National Science Foundation and by DoD OUSD (R&E) under Cooperative Agreement PHY-2229929 (The NSF AI Institute for Artificial and Natural Intelligence).

References

[1] S. V. Kalinin, C. Ophus, P. M. Voyles, R. Erni, D. Kepaptsoglou, V. Grillo, A. R. Lupini, M. P. Oxley, E. Schwenker, M. K. Y. Chan, *et al.*, Machine learning in scanning transmission electron microscopy, Nature Reviews Methods Primers **2**, 11 (2022).

[2] X. Zhang, L. Wang, J. Helwig, Y. Luo, C. Fu, Y. Xie, M. Liu, Y. Lin, Z. Xu, K. Yan, *et al.*, Artificial Intelligence for Science in Quantum, Atomistic, and Continuum Systems (2023), arXiv:2307.08423 [cs.LG].

[3] S. Azizi, L. Culp, J. Freyberg, B. Mustafa, S. Baur, S. Kornblith, T. Chen, N. Tomasev, J. Mitrović, P. Strachan, *et al.*, Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging, Nature Biomedical Engineering **7**, 756 (2023).

[4] H. Zhang, N. Dullerud, L. Seyyed-Kalantari, Q. Morris, S. Joshi, and M. Ghassemi, An empirical framework for domain generalization in clinical settings, in *Proceedings of the Conference on Health, Inference, and Learning* (Association for Computing Machinery, New York, NY, USA, 2021) pp. 279–290.

[5] R. J. Chen, J. J. Wang, D. F. K. Williamson, T. Y. Chen, J. Lipkova, M. Y. Lu, S. Sahai, and F. Mahmood, Algorithmic fairness in artificial intelligence for medicine and healthcare, Nature Biomedical Engineering **7**, 719 (2023).

[6] Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, and M. Ghassemi, The limits of fair medical imaging AI in real-world generalization, Nature Medicine 10.1038/s41591-024-03113-4 (2024).

[7] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning* (Mit Press, 2022).

[8] I. Gulrajani and D. Lopez-Paz, In Search of Lost Domain Generalization, in *International Conference on Learning Representations* (2021).

[9] N. Tripuraneni, B. Adlam, and J. Pennington, Overparameterization Improves Robustness to Covariate Shift in High Dimensions, in *Advances in Neural Information Processing Systems*, Vol. 34, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021) pp. 13883–13897.

[10] D. Idnani, V. Madan, N. Goyal, D. J. Schwab, and S. R. Vedantam, Don't forget the nullspace! Nullspace occupancy as a mechanism for out of distribution failure, in *The Eleventh International Conference on Learning Representations* (2023).

[11] R. Vedantam, D. Lopez-Paz, and D. J. Schwab, An Empirical Investigation of Domain Generalization with Empirical Risk Minimizers, in *Advances in Neural Information Processing Systems*, Vol. 34, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021) pp. 28131–28143.

[12] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, Learning under Concept Drift: A Review, IEEE Transactions on Knowledge and Data Engineering **31**, 2346 (2019).

[13] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, A unifying view on dataset shift in classification, Pattern Recognition **45**, 521 (2012).

[14] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into Deep Learning* (Cambridge University Press, 2023) `https://D2L.ai`.

[15] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, A survey on concept drift adaptation, ACM computing surveys (CSUR) **46**, 1 (2014).

[16] K. Nishida and K. Yamauchi, Detecting concept drift using statistical testing, in *International conference on discovery science* (Springer, 2007) pp. 264–269.

[17] E. Dobriban and S. Wager, High-dimensional asymptotics of prediction: Ridge regression and classification, The Annals of Statistics **46**, 247 (2018).

[18] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, Surprises in high-dimensional ridgeless least squares interpolation, The Annals of Statistics **50**, 949 (2022).

[19] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, Benign overfitting in linear regression, Proceedings of the National Academy of Sciences **117**, 30063 (2020).

[20] M. Emami, M. Sahraee-Ardakan, P. Pandit, S. Rangan, and A. Fletcher, Generalization Error of Generalized Linear Models in High Dimensions, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 119, edited by H. D. III and A. Singh (PMLR, 2020) pp. 2892–2901.

[21] D. Wu and J. Xu, On the Optimal Weighted ℓ_2 Regularization in Overparameterized Linear Regression, in *Advances in Neural Information Processing Systems*, Vol. 33, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., 2020) pp. 10112–10123.

[22] G. Mel and S. Ganguli, A theory of high dimensional regression with arbitrary correlations between input features and target functions: sample complexity, multiple descent curves and a hierarchy of phase transitions, in *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 139, edited by M. Meila and T. Zhang (PMLR, 2021) pp. 7578–7587.

[23] D. Richards, J. Mourtada, and L. Rosasco, Asymptotics of Ridge(less) Regression under General Source Condition, in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 130, edited by A. Banerjee and K. Fukumizu (PMLR, 2021) pp. 3889–3897.

[24] V. Ngampruetikorn and D. J. Schwab, Information bottleneck theory of high-dimensional regression: relevancy, efficiency and optimality, in *Advances in Neural Information Processing Systems*, Vol. 35, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., 2022) pp. 9784–9796.

[25] J. W. Rocks and P. Mehta, Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models, Phys. Rev. Res. **4**, 013201 (2022).

[26] M. Belkin, D. Hsu, S. Ma, and S. Mandal, Reconciling modern machine-learning practice and the classical bias-variance trade-off, Proceedings of the National Academy of Sciences **116**, 15849 (2019).

[27] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, Deep Double Descent: Where Bigger Models and More Data Hurt, in *International Conference on Learning Representations* (2020).

[28] P. Nakkiran, P. Venkat, S. M. Kakade, and T. Ma, Optimal Regularization can Mitigate Double Descent, in *International Conference on Learning Representations* (2021).

[29] J. Yang, K. Zhou, Y. Li, and Z. Liu, Generalized Out-of-Distribution Detection: A Survey, International Journal of Computer Vision 10.1007/s11263-024-02117-4 (2024).

[30] P. Patil, J.-H. Du, and R. Tibshirani, Optimal Ridge Regularization for Out-of-Distribution Prediction, in *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 235, edited by R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp (PMLR, 2024) pp. 39908–39954.

[31] S. Garg, D. Tsipras, P. S. Liang, and G. Valiant, What Can Transformers Learn In-Context? A Case Study of Simple Function Classes, in *Advances in Neural Information Processing Systems*, Vol. 35, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., 2022) pp. 30583–30598.

[32] E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, and D. Zhou, What learning algorithm is in-context learning? Investigations with linear models, in *The Eleventh International Conference on Learning Representations* (2023).

[33] A. Raventós, M. Paul, F. Chen, and S. Ganguli, Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression, in *Advances in Neural Information Processing Systems*, Vol. 36, edited by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Curran Associates, Inc., 2023) pp. 14228–14246.

[34] R. Agarwal, A. Singh, L. M. Zhang, B. Bohnet, L. Rosias, S. C. Chan, B. Zhang, A. Anand, Z. Abbas, A. Nova, J. D. Co-Reyes, E. Chu, F. Behbahani, A. Faust, and H. Larochelle, Many-Shot In-Context Learning, in *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (2024).

[35] J. Von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov, Transformers Learn In-Context by Gradient Descent, in *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 202, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (PMLR, 2023) pp. 35151–35174.

[36] K. Ahuja and D. Lopez-Paz, A Closer Look at In-Context Learning under Distribution Shifts, in *Workshop on Efficient Systems for Foundation Models @ ICML2023* (2023).

[37] C. Goddard, L. M. Smith, V. Ngampruetikorn, and D. J. Schwab, When can in-context learning generalize out of task distribution?, in *Forty-second International Conference on Machine Learning* (2025).

[38] L. M. Smith, C. Goddard, V. Ngampruetikorn, and D. J. Schwab, Model Recycling: Model component reuse to promote in-context learning, in *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning* (2024).

[39] A. Karpathy, NanoGPT, `https://github.com/karpathy/nanoGPT` (2022).

[40] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE **86**, 2278 (1998).

[41] H. Xiao, K. Rasul, and R. Vollgraf, Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms (2017), arXiv:1708.07747 [cs.LG].

[42] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu, Generalizing to Unseen Domains: A Survey on Domain Generalization, IEEE Transactions on Knowledge and Data Engineering **35**, 8052 (2023).

[43] A. Nguyen, D. J. Schwab, and V. Ngampruetikorn, Data coarse graining can improve model performance (2025), arXiv:2509.14498 [cond-mat.stat-mech].

[44] A. Canatar, B. Bordelon, and C. Pehlevan, Out-of-Distribution Generalization in Kernel Regression, in *Advances in Neural Information Processing Systems*, Vol. 34, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021) pp. 12600–12612.

[45] Y. Lu, M. Letey, J. A. Zavatone-Veth, A. Maiti, and C. Pehlevan, In-Context Learning by Linear Attention: Exact Asymptotics and Experiments, in *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning* (2024).

[46] Z. Han, G. Zhou, R. He, J. Wang, T. Wu, Y. Yin, S. Khan, L. Yao, T. Liu, and K. Zhang, How Well Does GPT-4V(ision) Adapt to Distribution Shifts? A Preliminary Investigation, in *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models* (2024).

[47] Q. Wang, Y. Wang, X. Ying, and Y. Wang, Can In-context Learning Really Generalize to Out-of-distribution Tasks?, in *The Thirteenth International Conference on Learning Representations* (2025).

[48] J. Song, Z. Xu, and Y. Zhong, Out-of-distribution generalization via composition: A lens through induction heads in Transformers, Proceedings of the National Academy of Sciences **122**, e2417182122 (2025).

[49] F. Rubio and X. Mestre, Spectral convergence for a general class of random matrices, Statistics & Probability Letters **81**, 592 (2011).

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: We have carefully worded our abstract and introduction to reflect the paper's contributions. We have also added a section detailing our contributions.
   Guidelines:
   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: We discussed the limitations of our analytic results, which are for linear ridge regression.
   Guidelines:
   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: Full derivations provided in Appendix.
   Guidelines:
   - The answer NA means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: All details to reproduce numerical experiments are presented in their relevant sections.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [No]
   Justification: While we do not have open-source code for now, we will open source our code once the anonymity period for submission is over.
   Guidelines:
   - The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
   Answer: [Yes]
   Justification: All details to reproduce numerical experiments are presented in their relevant sections.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.
7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: Our error bars are
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).

- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification: All details to reproduce numerical experiments are presented in their relevant sections.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?
   Answer: [Yes]
   Justification: Our research conforms to the NeurIPS Code of Ethics.
   Guidelines:
   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
    Answer: [NA]
    Justification: NA
    Guidelines:
    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**
    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
    Answer: [NA]
    Justification: NA
    Guidelines:
    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.
12. **Licenses for existing assets**
    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
    Answer: [NA]
    Justification: NA
    Guidelines:
    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.
13. **New assets**
    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
    Answer: [NA]
    Justification: NA
    Guidelines:
    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.
14. **Crowdsourcing and research with human subjects**
    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]
Justification: NA
Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
Answer: [NA]
Justification: NA
Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**
Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.
Answer: [NA]
Justification: NA
Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Prediction risk in the thermodynamic limit

To derive prediction risk in the thermodynamic limit, we rewrite the nonasymptotic bias and variance contributions [Eq (5)] in terms of the resolvent operator $(\Psi + \lambda I_P)^{-1}$,

$$B(X) = (\beta - \tilde{\beta})^\mathsf{T} \tilde{\Sigma} (\beta - \tilde{\beta}) - 2\lambda \operatorname{Tr}\left[\beta(\beta - \tilde{\beta})^\mathsf{T} \tilde{\Sigma} \tfrac{1}{\Psi + \lambda I_P}\right] + \lambda^2 \operatorname{Tr}\left[\beta\beta^\mathsf{T} \tfrac{1}{\Psi + \lambda I_P} \tilde{\Sigma} \tfrac{1}{\Psi + \lambda I_P}\right] \tag{12}$$

$$V(X) = \sigma_\xi^2 \left[\tfrac{1}{N} \operatorname{Tr}\left(\tilde{\Sigma} \tfrac{1}{\Psi + \lambda I_P}\right) - \lambda \tfrac{1}{N} \operatorname{Tr}\left(\tilde{\Sigma} \tfrac{1}{(\Psi + \lambda I_P)^2}\right)\right]. \tag{13}$$

In the thermodynamic limit—$N, P \to \infty$ and $P/N \to \gamma \in (0, \infty)$—the above traces become deterministic (see Appendix B), and the bias and variance converge to (see also Patil *et al.* [30])

$$\begin{aligned} B(X) \to \mathcal{B} = {} & (\beta - \tilde{\beta})^\mathsf{T} \tilde{\Sigma} (\beta - \tilde{\beta}) - 2\lambda \beta^\mathsf{T} \hat{G}_\Sigma(-\lambda) \tilde{\Sigma} (\beta - \tilde{\beta}) + \lambda^2 \nu'(-\lambda) \frac{\beta^\mathsf{T} \hat{G}_\Sigma(-\lambda) \tilde{\Sigma} \hat{G}_\Sigma(-\lambda) \beta}{\frac{1}{P} \operatorname{Tr}[\Sigma \hat{G}_\Sigma^2(-\lambda)]} \\ & + \lambda^2 \gamma m(-\lambda)^2 \nu'(-\lambda) \frac{\beta^\mathsf{T} \hat{G}_\Sigma(-\lambda) \left(\operatorname{Tr}[\tilde{\Sigma} \Sigma \hat{G}_\Sigma^2(-\lambda)] \Sigma - \operatorname{Tr}[\Sigma^2 \hat{G}_\Sigma^2(-\lambda)] \tilde{\Sigma}\right) \hat{G}_\Sigma(-\lambda) \beta}{\operatorname{Tr}[\Sigma \hat{G}_\Sigma^2(-\lambda)]} \end{aligned} \tag{14}$$

$$V(X) \to \mathcal{V} = \sigma_\xi^2 \gamma \left(\nu(-\lambda) \frac{\operatorname{Tr}[\tilde{\Sigma} \hat{G}_\Sigma(-\lambda)]}{\operatorname{Tr}[\Sigma \hat{G}_\Sigma(-\lambda)]} - \lambda \nu'(-\lambda) \frac{\operatorname{Tr}[\tilde{\Sigma} \hat{G}'_\Sigma(-\lambda)]}{\operatorname{Tr}[\Sigma \hat{G}'_\Sigma(-\lambda)]}\right), \tag{15}$$

where $\hat{G}_\Sigma(z) \equiv \frac{1}{m(z)\Sigma - z I_P}$, $m(z) \equiv \frac{1}{1 + \gamma \nu(z)}$ and $\nu(z) = \frac{1}{P} \operatorname{Tr}[\Sigma \hat{G}_\Sigma(z)]$ with $\nu(z) \in C^+$. We note that the last term of the bias vanishes for $\Sigma = \tilde{\Sigma}$, and covariate shifts affect the variance term but concept shift does not.

B Spectral convergence for random matrix traces

Let Ψ, Θ and A denote $P \times P$ matrices, I_P the identity matrix in P dimensions and z a complex scalar outside the positive real line. Assume the following: (i) $A \in \mathbb{R}^{P \times P}$ is symmetric and nonnegative definite, (ii) $\Theta \in \mathbb{R}^{P \times P}$ has a bounded trace norm $\operatorname{Tr}[(\Theta^\mathsf{T} \Theta)^{1/2}] \in [0, \infty)$ and (iii) $\Psi = \frac{1}{N} \Sigma^{1/2} Z Z^\mathsf{T} \Sigma^{1/2}$ where the entries of $Z \in \mathbb{R}^{P \times N}$ are iid random variables with zero mean, unit variance and finite $8 + \varepsilon$ moment for some $\varepsilon > 0$, and $\Sigma \in \mathbb{R}^{P \times P}$ is a covariance matrix. In the limit $P, N \to \infty$ and $P/N \to \gamma \in (0, \infty)$, we have [49]

$$\operatorname{Tr}\left(\Theta \frac{1}{\Psi + A - z I_P}\right) \to \operatorname{Tr}\left(\Theta \frac{1}{\frac{1}{1 + \gamma c(z; A)} \Sigma + A - z I_P}\right), \tag{16}$$

where $c_A(z) \in \mathbb{C}^+$ is the unique solution of

$$c(z; A) = \frac{1}{P} \operatorname{Tr}\left(\Sigma \frac{1}{\frac{1}{1 + \gamma c(z; A)} \Sigma + A - z I_P}\right). \tag{17}$$

First we consider the trace of terms linear in the resolvent $(\Psi - z I_P)^{-1}$, appearing in Eqs (12) &(13). Setting $A = 0$ in Eq (16) gives

$$\operatorname{Tr}\left(\Theta \frac{1}{\Psi - z I_P}\right) \to \operatorname{Tr}\left(\Theta \frac{1}{\frac{1}{1 + \gamma \nu(z)} \Sigma - z I_P}\right) \tag{18}$$

where $\nu(z) \equiv c(z; 0)$ is the solution of

$$\nu(z) = \frac{1}{P} \operatorname{Tr}\left(\Sigma \frac{1}{\frac{1}{1 + \gamma \nu(z)} \Sigma - z I_P}\right) \quad \text{with} \quad \nu(z) \in \mathbb{C}^+. \tag{19}$$

In general, this self-consistent equation does not have a closed-form solution. One exception is the isotropic case $\Sigma = I_P$ for which

$$\nu_{\Sigma = I_P}(z) = \frac{1}{2\gamma z} \left[1 - \gamma - z - \sqrt{(1 - \gamma - z)^2 - 4\gamma z}\right]. \tag{20}$$

Next we obtain the asymptotic expression for the trace of terms quadratic in the resolvent, such as those in Eqs (12) &(13). We let $A = \mu B$ with $\mu > 0$. Differentiating Eq (16) with respect to μ and taking the limit $\mu \to 0^+$ yields

$$\mathrm{Tr}\left(\Theta \frac{1}{\Psi - zI_P} B \frac{1}{\Psi - zI_P}\right) \to \mathrm{Tr}\left(\Theta \frac{1}{\frac{1}{1+\gamma\nu(z)}\Sigma - zI_P} (d(z;B)\Sigma + B) \frac{1}{\frac{1}{1+\gamma\nu(z)}\Sigma - zI_P}\right). \tag{21}$$

Here we define

$$d(z;B) \equiv \frac{d}{d\mu} \frac{1}{1+\gamma c(z;\mu B)}\bigg|_{\mu\to 0^+} \tag{22}$$

$$= \frac{\gamma \frac{1}{P} \mathrm{Tr}\left(B \frac{\Sigma}{(\Sigma - (1+\gamma\nu(z))zI_P)^2}\right)}{1 - \gamma \frac{1}{P} \mathrm{Tr}[(\frac{\Sigma}{\Sigma - (1+\gamma\nu(z))zI_P})^2]} \tag{23}$$

$$= \frac{\gamma\nu'(z)}{(1+\gamma\nu(z))^2} \frac{\frac{1}{P} \mathrm{Tr}\left(B \frac{\Sigma}{(\frac{1}{1+\gamma\nu(z)}\Sigma - zI_P)^2}\right)}{\frac{1}{P} \mathrm{Tr} \frac{\Sigma}{(\frac{1}{1+\gamma\nu(z)}\Sigma - z)^2}}. \tag{24}$$

where the last equality follows from

$$\nu'(z) = \frac{\frac{1}{P} \mathrm{Tr} \frac{\Sigma}{(\frac{1}{1+\gamma\nu(z)}\Sigma - z)^2}}{1 - \gamma \frac{1}{P} \mathrm{Tr}[(\frac{\Sigma}{\Sigma - (1+\gamma\nu(z))zI_P})^2]}. \tag{25}$$

For $B = I_P$, the spectral function $d(z;B)$ reduces to

$$d(z;I_P) = \frac{\gamma\nu'(z)}{(1+\gamma\nu(z))^2}. \tag{26}$$

Our result for prediction risk in the thermodynamic limit is based on the asymptotic traces in Eqs (18) & (21).