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 high-dimensional limit. Our results reveal nontrivial effects of concept shift on generalization performance, depending on the properties of robust and nonrobust features of the input. We show that test performance can exhibit a nonmonotonic data dependence, even when double descent is absent. Finally, our experiments on MNIST and FashionMNIST suggest that this intriguing behavior is present also in classification problems.

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 [3–5]. 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 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 [6–10]).

Here we consider *concept shift*—a relatively less-studied setting, in which the input distribution is preserved but the input-label relationship becomes different at test time [11, 12]. 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 (e.g., Refs [13–21]). We derive an analytical expression for prediction risk under concept shift and demonstrate that its behavior can change from monotonically decreasing with data for the in-distribution case to monotonically increasing, and to nonmonotonic, depending on the degree of concept shift and the properties of the inputs. To isolate these effects from double descent phenomena [22–24], 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 [13, 14]. Finally, we illustrate similar qualitative changes in test performance 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 [25] for a recent review).

*DJS and VN contributed to this work equally.

Workshop on Scientific Methods for Understanding Deep Learning, NeurIPS 2024.

2 Regression setting

Data—The training data consists of N iid input-response pairs $\{(x_1, y_1), \ldots, (x_N, y_N)\}$. 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,

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\cdot, \begin{bmatrix} \Sigma & \Sigma\beta \\ \beta^\mathsf{T}\Sigma & \beta^\mathsf{T}\Sigma\beta + \sigma_\xi^2 \end{bmatrix}\right) \quad \text{and} \quad \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \sim \mathcal{N}\left(\cdot, \begin{bmatrix} \tilde{\Sigma} & \tilde{\Sigma}\tilde{\beta} \\ \tilde{\beta}^\mathsf{T}\tilde{\Sigma} & \tilde{\beta}^\mathsf{T}\tilde{\Sigma}\tilde{\beta} + \tilde{\sigma}_\xi^2 \end{bmatrix}\right), \tag{2}$$

where in general $\Sigma \neq \tilde{\Sigma}$, $\beta \neq \tilde{\beta}$ and $\sigma_\xi^2 \neq \tilde{\sigma}_\xi^2$.

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 square error,

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

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,

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

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

$$B(X) \equiv \mathbf{E}\left[\|\mathbf{E}(\hat{y}(\tilde{x}; X, Y) \mid X, \tilde{x}) - \mathbf{E}(\tilde{y} \mid \tilde{x})\|^2 \mid X\right] \tag{5}$$

$$V(X) \equiv \mathbf{E}\left[\|\hat{y}(\tilde{x}; X, Y) - \mathbf{E}(\hat{y}(\tilde{x}; X, Y) \mid X, \tilde{x})\|^2 \mid X\right]. \tag{6}$$

Substituting the predictor from ridge regression into the above equations yields

$$B(X) = \left(\tfrac{\Psi}{\Psi + \lambda I_P}\beta - \tilde{\beta}\right)^\mathsf{T} \tilde{\Sigma} \left(\tfrac{\Psi}{\Psi + \lambda I_P}\beta - \tilde{\beta}\right) \quad \text{and} \quad V(X) = \sigma_\xi^2 \tfrac{1}{N} \operatorname{Tr}\left(\tilde{\Sigma} \tfrac{\Psi}{(\Psi + \lambda I_P)^2}\right), \tag{7}$$

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 $R_{N\to\infty}(X) = (\beta - \tilde{\beta})^\mathsf{T}\tilde{\Sigma}(\beta - \tilde{\beta})$.[1] 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 1). 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 1); that is, even *infinitely more* data hurts test performance.

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 A for derivation; see also Ref [26]),

$$B(X) \to \mathcal{B} = \lambda^2 \nu'(-\lambda) \frac{\beta^\mathsf{T}\Sigma\hat{G}_\Sigma^2(-\lambda)\beta}{\frac{1}{P}\operatorname{Tr}[\Sigma\hat{G}_\Sigma^2(-\lambda)]} - 2\lambda\beta^\mathsf{T}\Sigma\hat{G}_\Sigma(-\lambda)(\beta - \tilde{\beta}) + (\beta - \tilde{\beta})^\mathsf{T}\Sigma(\beta - \tilde{\beta}) \tag{8}$$

$$V(X) \to \mathcal{V} = \sigma_\xi^2 \gamma \left[\nu(-\lambda) - \lambda\nu'(-\lambda)\right], \tag{9}$$

where we define $\hat{G}_\Sigma(z) \equiv (m(z)\Sigma - zI_P)^{-1}$, $m(z) \equiv (1 + \gamma\nu(z))^{-1}$ and $\nu(z) \in \mathbb{C}^+$ is the solution of $\nu(z) = \frac{1}{P}\operatorname{Tr}[\Sigma\hat{G}_\Sigma(z)]$. We note that concept shift enters prediction risk only through the last two terms of the bias, Eq (8), whereas the variance, Eq (9), is completely unaffected.

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

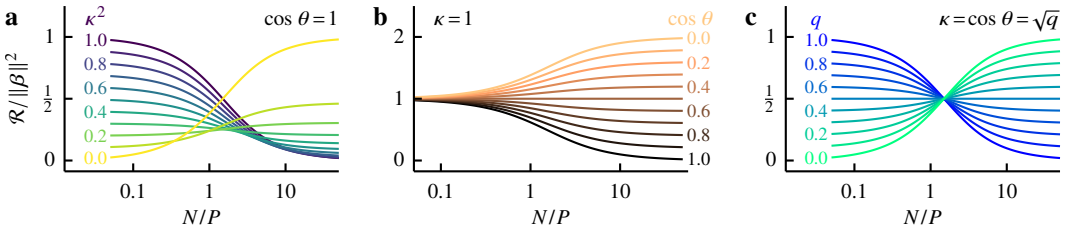

Figure 1: **More data hurts performance when concept shift is strong.** We depict the data dependence of asymptotic prediction risk for isotropic features, Eq (10), under three concept shift settings of varying degree, parametrized by coefficient alignment $\cos\theta = \beta\cdot\tilde{\beta}/\|\beta\|\|\tilde{\beta}\|$ 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 $q = \kappa^2 = \cos^2\theta$. Here $\mathrm{SNR} = 1$ and regularization is in-distribution optimal.

3.1 Isotropic features

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

$$\mathcal{B} = \|\beta\|^2\left[\lambda^2\nu'(-\lambda) - 2\lambda g(-z)(1-\kappa\cos\theta) + 1 + \kappa^2 - 2\kappa\cos\theta\right], \tag{10}$$

where $g(z) \equiv (m(z) - z)^{-1}$ and we parametrize concept shift by the coefficient alignment $\cos\theta \equiv \beta\cdot\tilde{\beta}/\|\beta\|\|\tilde{\beta}\|$ and scaling factor $\kappa \equiv \|\tilde{\beta}\|/\|\beta\|$.

Figure 1 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 [13, 14]). In Fig 1a, the coefficient vector becomes smaller at test time without changing direction, $\theta = 0$ and $\kappa \le 1$. At $\kappa = 1$, 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. We observe a similar crossover in different settings. In Fig 1b, concept shift occurs only through misalignment, $\kappa = 1$ and $\theta \ge 0$. In Fig 1c, we consider a mixture of robust and nonrobust features with $\tilde{\beta}_i = \beta_i$ if feature i is robust, otherwise $\tilde{\beta}_i = 0$. More data hurts prediction risk in all of the above settings for adequately strong concept shift.

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. 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

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

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 $\kappa_\pm \equiv \|\tilde{\beta}_\pm\|/\|\beta_\pm\|$.

Figure 2 illustrates prediction risk under concept shift for two-scale covariates. We tune the ridge regularization strength such that the in-distribution risk is minimized. We consider the cases where concept shift affects only low-variance features via either κ_- (Fig 2a) or θ_- (Fig 2b), or only high-variance ones via either κ_+ (Fig 2c) or θ_+ (Fig 2d). 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. The region in which more data can hurt occurs at intermediate to high N/P for low-variance concept shift (Fig 2a & b) whereas this region corresponds to low N/P for high-variance concept shift (Fig 2c & d).

This seemingly surprising behavior arises 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 2c & d. On the other hand,

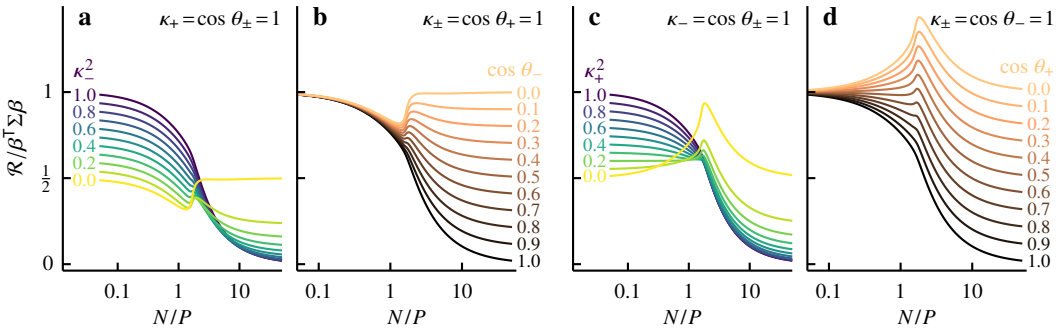

Figure 2: **Concept shift generalization exhibits sample size nonmonotonicity for anisotropic features.** We illustrate prediction risk for the two-scale model, Eq (11), with aspect ratio $s_-/s_+ = 0.1$, spectral weights $\rho_+ = \rho_- = 1/2$ and signal fraction $\pi_+ = \pi_- = 1/2$. Concept shift 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-d** Same as (a) & (b) but for high-variance concept shift. Here SNR = 1 and regularization is in-distribution optimal.

low-variance features affect test performance only when the training sample size is large enough. 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 2a & b.

4 Classification experiments

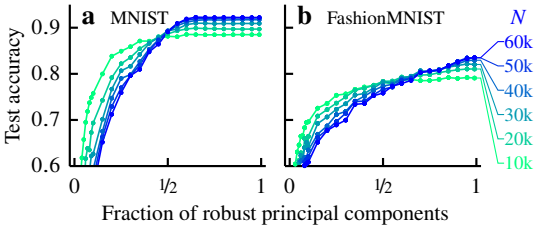

Figure 3: **Concept shift qualitatively changes data dependence of test accuracy** in (a) MNIST and (b) FashionMNIST experiments for various training data size N (see legend). See §4 for details.

Although our theory is based on ridge regression, it likely captures some general phenomena of learning under concept shift. Here we consider standard multinomial logistic regression for MNIST [27] and FashionMNIST [28] (see, Appendix C for more details). We modify the test datasets by computing PCA on the images and designating the resulting features as either 'robust' or 'nonrobust'. We shuffle nonrobust features across data points to decorrelate them from the labels while preserving marginal statistics, whereas robust features are unchanged. In Fig 3, we let lower-variance features be nonrobust and depict the test accuracy as a function of the variance threshold that separates robust from nonrobust features. We see that when concept shift is strong and few features are robust, more data hurts test accuracy, in qualitative agreement with our results for ridge regression (cf. Fig 1).

5 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. In addition, we demonstrate that input anisotropy can lead to nonmonotonic data dependence of prediction risk. Our results are for optimally tuned ridge regression and thus differ from risk nonmonotonicity due to double descent which is absent under optimal regularization [24].

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. Our classification experiments suggest that the insights from our theory may apply in more general settings. We hope that our work will help initiate systematic investigations of the rich learning phenomena induced by concept shift.

Acknowledgments

Alex Nguyen is supported by an NIH RF1 grant. 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.

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] 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).

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

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

[8] 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.

[9] 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).

[10] 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.

[11] 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).

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

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

[14] 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).

[15] 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).

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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).

[23] 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).

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

[25] 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).

[26] P. Patil, J.-H. Du, and R. J. Tibshirani, Optimal Ridge Regularization for Out-of-Distribution Prediction (2024), arXiv:2404.01233 [math.ST].

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

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

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

A Prediction risk in the thermodynamic limit

To derive prediction risk in the thermodynamic limit, we rewrite the nonasymptotic bias and variance contributions [Eq (7)] 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 Ref [26])

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

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

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 $\nu(z) \in \mathbb{C}^+$. 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 [29]

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

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).

C Classification experiment details

In Fig 3, we train a standard multinomial logistic regression on MNIST and FashionMNIST datasets, using Adam optimizer with a minibatch size of 500 and a learning rate of 0.001 for 2000 epochs. We choose training data points at random to vary training sample size N. When $N = 60{,}000$, all of the training data is used. The variation due to stochastic training and training data construction is smaller than the marker size in Fig 3.