Asymmetric Scaling Laws from Sparse Features

John Sous JOHN.SOUS@YALE.EDU
Department of Applied Physics, Yale University, New Haven, CT, USA
Energy Sciences Institute, Yale University, West Haven, CT, USA

Michael Winer MIKE@ALIGNMENT.ORG
Institute for Advanced Study, Princeton, New Jersey 08540, USA

Abstract

We introduce a model for neural scaling laws under sparse activations. In the model, test loss is often dominated by rare coordinates that are never observed in the training input. This mechanism induces a novel bottleneck absent from dense models. We derive the asymptotic population loss in both the underparameterized and overparameterized regimes, and show that the loss exhibits a double-descent peak near the interpolation threshold—where the number of parameters is just sufficient to fit the training data—resulting in a loss curve governed by two distinct scaling exponents—one for the overparameterized regime and one for the underparameterized regime—with a gap determined by the degree of sparsity. Additionally, we derive a compute-optimal frontier that favors increasing dataset size over model capacity under fixed compute budgets. We also analyze gradient-descent dynamics and identify a scaling law for the probability that fixed-step gradient descent becomes unstable. We further show that the sparsity-induced effect persists under nonlinear activations. Experiments validating the theory can be found at SparseScaling.

1. Introduction

Despite intense research, a comprehensive theoretical understanding of scaling laws remains elusive. One approach to this challenge models the data as Gaussian with power-law covariance, applies a random embedding into a lower-dimensional representation, and uses linear regression to characterize the resulting scaling laws [4, 15, 44, 62]. This framework [44] has been shown, both theoretically and empirically, to produce the scaling law $\ell \propto \left(\frac{1}{N} + \frac{1}{D}\right)^{\alpha}$ and the "Chinchilla" compute-optimal allocation in which, under a compute budget $C \propto ND$, the optimal choice scales as $N^*(C) \asymp D^*(C) \asymp C^{1/2}$.[1]

An important aspect not determined by a framework that yields a compute-optimal exponent close to $\frac{1}{2}$ is whether the power-law exponent governing loss decay with model size, α_N, matches that governing decay with data size, α_D, as empirical studies suggest need not be the case [32].[2] In this work, we propose a simple theoretical model that reproduces the observed scaling asymmetry

1. This compute-allocation law arises in this framework both in the Bayes-optimal [4, 44] and one-pass SGD [16, 55] settings. In the Bayes-optimal case, this scaling can be found by analytically extremizing the loss under the compute constraint. Results based on random matrix theory support the same scaling in one-pass stochastic gradient descent (SGD) dynamics.
2. Frameworks that predict $\alpha_N = \alpha_D = \alpha$ necessarily imply $N^*(C) \asymp D^*(C) \asymp C^{1/2}$ under the fixed-compute constraint $C \propto ND$. However, a small asymmetry between α_N and α_D can still yield compute-optimal exponents close to $\frac{1}{2}$. Thus, near-square-root compute-optimal scaling does not by itself imply symmetric scaling.

© J. Sous & M. Winer.

and naturally explains the emergence of two distinct exponents in the population loss.[3] Our model is a random embedding followed by a linear readout, with a sparse input activation structure such that only a subset of coordinates of $\mathbf{x}$ is active. This setup is motivated by the idea that the data-generating process excites only a sparse subset of high-dimensional feature directions. The resulting loss displays distinct scaling with N and D. We derive an asymptotic scaling law in which the difference between α_N and α_D arises continuously as the sparsity level is varied.[4]

Our main contributions are as follows.

- We introduce the notion of scaling laws for sparse activations by introducing a model that captures the impact of sparsity on optimization and resulting scaling behavior.
- We derive the population loss for the sparse random feature model, yielding a two-exponent scaling law with an intrinsic asymmetry between the underparameterized and overparameterized regimes and a double-descent peak near the interpolation threshold [8–10, 24, 29, 45, 49].
- We derive the compute-optimal frontier under a fixed compute budget, and show that increasing sparsity improves compute-efficiency while shifting the optimal allocation toward larger datasets.
- We analyze the training dynamics and convergence properties of the loss during optimization, deriving a scaling law for the failure of fixed-step gradient descent (GD).
- We experimentally verify in a nonlinear two-layer network that scaling asymmetry continues to arise from sparsity rather than from nonlinearity.

Related Work. Empirical scaling laws show power-law loss improvement with model size and data, and motivate compute-optimal training rules [32, 34]. On the theory side, *solvable* models based on random features, kernels, and high-dimensional regression derive closed-form or deterministic-equivalent scaling predictions [4, 15, 21, 39, 44, 62]. Complementary *dynamical* models study protocol- and time-dependent scaling (*e.g.*, one-pass vs. multi-pass training) and can exhibit multiple scaling phases [16, 17, 55]. In contrast, we focus on sparse feature activation, and show it can *intrinsically* yield different exponents in the model-limited versus data-limited regimes even in Bayes-optimal learning. We thus isolate rare feature coverage as a distinct source of scaling asymmetry. See Appendix F for additional related works.

2. Statement of Problem

In this section, we motivate the model and outline the main goals of our analysis.

2.1. Learning under Sparse High-Dimensional Data

We first specify the data-generation process that yields a sparse power-law structure in the inputs, then describe the random feature model used to learn from these high-dimensional inputs.

2.1.1. Data Generation Process: Bernoulli-Random Activations with Power-Law Covariance

We consider input data $\mathbf{X} = [\,\mathbf{x}_1, \ldots, \mathbf{x}_D\,] \in \mathbb{R}^{M\times D}$, with each $\mathbf{x}_d \in \mathbb{R}^M$, where D is the number of training examples (data size) and M is the data dimension. We henceforth drop the d subscript

3. Bordelon et al. [17] also report scaling asymmetry, but there the relevant regimes are controlled by task difficulty rather than by the sparse-feature mechanism studied here.
4. We note that in training scenarios such as few-pass SGD, the underlying sparsity may be partially or entirely obscured. This disparity suggests a novel phase transition as the number of epochs increases.

and simply refer to a single training example as $\mathbf{x}$. We consider a structured data-generation process in which each input coordinate of $\mathbf{x} \in \mathbb{R}^M$ is randomly activated according to a heavy-tailed sparsity pattern. Specifically, for $j = 1, 2, \ldots, M$, the jth coordinate of $\mathbf{x}$ is drawn as

$$\mathbb{P}(x_j = 0) = 1 - j^{-\alpha_1 - 1}, \qquad \mathbb{P}\Big(x_j = \pm j^{-(\alpha_2+1)/2}\Big) = \tfrac{1}{2}\, j^{-\alpha_1 - 1}. \tag{1}$$

Hence $\mathbb{E}[x_j] = 0$ and $\mathrm{Var}(x_j) = j^{-\alpha_1 - \alpha_2 - 2}$. Let $p_j := \mathbb{P}(x_j \neq 0) = j^{-\alpha_1 - 1}$. We require $\alpha_1 \geq -1$ to ensure that $p_j \leq 1$. Additionally, to ensure that the target variance is finite, the sum over the coordinate-wise variances $\sum_{j=1}^{\infty} j^{-\alpha_1 - \alpha_2 - 2}$ must converge, which holds iff $\alpha_1 + \alpha_2 + 1 > 0$.

This formulation introduces sparsity through Bernoulli input activations: most coordinates of $\mathbf{x}$ are zero, while the few active ones follow a heavy-tailed scaling controlled by α_1 and α_2.

2.1.2. RANDOM FEATURE MODEL

We study a linear model built from $N \ll M$ features randomly embedded from these inputs, with $M \gg \max\{N, D\}$ so that truncation does not affect the power-law tail asymptotics. Each input $\mathbf{x} \in \mathbb{R}^M$ is mapped to features $\boldsymbol{\phi}$ via a fixed random embedding $\mathbf{u} \in \mathbb{R}^{N \times M}$, with $u_{ij} \sim \mathcal{N}(0, 1/N)$ and its $N \times M$ elements picked i.i.d., $\boldsymbol{\phi} = \mathbf{u}\mathbf{x}$, and the (per-example) prediction is $\hat{y}(\mathbf{x}) = \boldsymbol{\theta}^\top \boldsymbol{\phi} = \boldsymbol{\theta}^\top \mathbf{u}\mathbf{x}$ (and the batched prediction is $\hat{\mathbf{y}} = \boldsymbol{\theta}^\top \boldsymbol{\Phi} = \boldsymbol{\theta}^\top \mathbf{u}\mathbf{X}$), with trainable $\boldsymbol{\theta} \in \mathbb{R}^N$.[5] The target function is $y(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} = \sum_{j=1}^{M} w_j x_j$, for random $\mathbf{w} \in \mathbb{R}^M$ with w_j i.i.d. and $\mathrm{Var}(w_j) = 1$.

2.2. Main Goals

Our goal is to characterize the trained-to-completion loss for the random sparse feature model described above. The population loss of the trained estimator is (Appendix G)

$$\ell_{\text{Bayes}} = \mathbb{E}_{\mathbf{x}}\left[(y - \hat{y}(\mathbf{x}))^2\right] = \mathbb{E}_{\mathbf{x}}\Big[\big(\mathbf{w}^\top \mathbf{x} - \boldsymbol{\theta}^{\star\top} \mathbf{u}\mathbf{x}\big)^2\Big], \tag{2}$$

where $\boldsymbol{\theta}^\star$ denotes the minimum-ℓ_2-norm empirical-risk minimizer, equivalently the solution reached by GD from zero initialization when the empirical loss is trained to completion. Note that $\hat{y}(\mathbf{x}) = \boldsymbol{\theta}^{\star\top} \mathbf{u}\mathbf{x}$ corresponds to a linear function of $\mathbf{x}$ whose weight vector lies in the row space of $\mathbf{u}$. Consequently, when $\mathbf{w} \notin \mathrm{rowspan}(\mathbf{u})$, this feature map cannot recover $y = \mathbf{w}^\top \mathbf{x}$ exactly, and a nonzero approximation error is unavoidable. We aim to derive the scaling laws and compute-optimal frontier for this Bayes-optimal loss, emphasizing the dependence on the data distribution and its effect on optimization. All proofs are deferred to the appendices.

3. Scaling Asymptotics of Sparse Random Features

Our main results are summarized in Figure 1

5. Our model is, in effect, a sketched linear regression. It is random-feature-like in the sense of a frozen random map followed by a trained linear readout, but it differs from canonical random-features models, *e.g.* [25, 27, 29], which study nonlinear feature maps in high-dimensional regimes. Extending those frameworks to the sparse, strongly non-Gaussian/heavy-tailed designs studied here would require going beyond the Gaussian-equivalence tools used in the nonlinear random-features literature [28, 33, 45].

3.1. Loss Scaling from Unmodeled Features

The random feature model reduces to linear regression in the effective weight vector $\hat{\mathbf{w}} = \mathbf{u}^\top \boldsymbol{\theta}$. The population mean-squared error can be expressed as $\ell(\hat{\mathbf{w}}) = \mathbb{E}_{\mathbf{x}}\big[(\mathbf{w}^\top \mathbf{x} - \hat{\mathbf{w}}^\top \mathbf{x})^2\big] = (\mathbf{w} - \hat{\mathbf{w}})^\top \boldsymbol{\Sigma}_x (\mathbf{w} - \hat{\mathbf{w}})$, where $\boldsymbol{\Sigma}_x = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ is diagonal since the input coordinates are independent. To expose the scaling mechanism, consider a predictor that matches the first k coordinates of $\mathbf{w}$ and sets all remaining coordinates to zero: $\hat{w}_j = w_j \mathbf{1}_{\{j \le k\}}$. Then $\mathbf{w} - \hat{\mathbf{w}}$ has entries 0 for $j \le k$ and w_j for $j > k$, and substituting into the expression for $\ell(\hat{\mathbf{w}})$ yields $\ell(\hat{\mathbf{w}}) = \sum_{j>k} w_j^2 \operatorname{Var}(x_j)$. Under our assumptions $w_j^2 = O(1)$, the population loss is therefore governed, up to constants, by the *unmodeled input variance* $\sum_{j>k} \operatorname{Var}(x_j)$. This calculation suggests that the asymptotic behavior of the loss is governed by the tail of the variance spectrum of the input distribution:

$$\ell_{\text{Bayes}} \asymp \sum_{j>k} \operatorname{Var}(x_j). \tag{3}$$

3.2. Prior Results: Scaling Law for Uniformly Activated Data with Power-Law Covariance

We first review the scaling laws established for random feature models with fully active (non-sparse) input coordinates [44].

Uniformly Activated Data with Power-Law Covariance. In the original formulation of Maloney et al. [44], every input coordinate is active and is modeled as an independent Gaussian variable with variance decaying as a power law: $x_j \sim \mathcal{N}(0,\, j^{-\alpha-1}), \quad j = 1, \ldots, M, \quad \alpha > 0$. This matches our model's fully activated covariance structure under the choice $\alpha_1 = -1$ and $\alpha_2 = \alpha$.

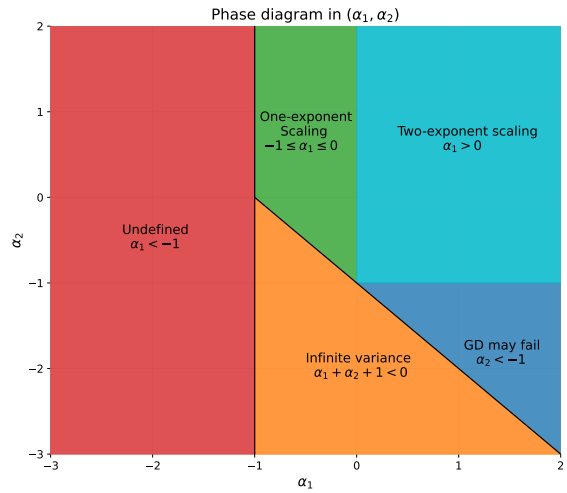

Figure 1: **Phase diagram in (α_1, α_2).** Solid black lines mark the boundary beyond which the model is well defined. Within it we identify three regimes: symmetric one-exponent scaling, asymmetric two-exponent scaling, and a GD-failure regime with its own scaling law.

Symmetry between Underparameterized and Overparameterized Regimes. For this non-sparse data structure, the optimal-loss scaling law was analyzed in detail in [44]. We produce a simple intuitive argument showing why, in this setting, a *single* exponent governs the asymptotic behavior of the loss in both the under- and overparameterized regimes, yielding a symmetric one-exponent law $\ell_{\text{Bayes}} \propto \left(\frac{1}{N} + \frac{1}{D}\right)^\alpha$ with $\alpha_N = \alpha_D = \alpha$. As explained above, the random feature model reduces to linear regression in the effective weight vector. Thus, if a learned predictor resolves the leading $k \ll M$ coefficients of $\mathbf{w}$, the residual (unmodeled) variance is $\sum_{j>k} j^{-\alpha-1} \asymp \int_k^\infty j^{-\alpha-1}\, dj = \frac{1}{\alpha} k^{-\alpha}$. With finitely many parameters N and finitely many samples D, the model can identify at most $O(\min\{N, D\})$ coefficients, since this minimum represents the true bottleneck. This yields the scaling $\ell_{\text{Bayes}} \sim \min\{N, D\}^{-\alpha}$, so that the exponents satisfy $\alpha_N = \alpha_D = \alpha$. In the sparse model of Section 2.1, one still gets a symmetric one-exponent scaling law when $-1 \le \alpha_1 \le 0$. In this regime, with high probability the

low-index coordinates are activated sufficiently often across D samples that sparsity is effectively subdominant. The scaling is therefore still controlled by $\min\{N, D\}$, as in the dense case, yielding $\alpha_N = \alpha_D = \alpha_1 + \alpha_2 + 1$ (Appendix H.1).

3.3. Two-Exponent Scaling Law for Sparse Activations ($\alpha_1 > 0$)

We now analyze our model of sparse activations for $\alpha_1 > 0$ (Appendix H.2).

3.3.1. UNDERPARAMETERIZED REGIME ($N \ll D$).

In this regime, the model fits only the leading high-variance coordinates of $\mathbf{w}$, and the resulting scaling law coincides with that of the non-sparse case.

Proposition 1 (Underparameterized Scaling) *Under the sparse activation model described in Section 2.1, the Bayes-optimal loss in the underparameterized regime ($N \ll D$) satisfies*

$$\ell_{\text{Bayes},N} \sim N^{-(\alpha_1+\alpha_2+1)}, \quad \textit{so that } \alpha_N = \alpha_1 + \alpha_2 + 1. \tag{4}$$

3.3.2. OVERPARAMETERIZED REGIME ($D \ll N$).

A key quantity in this regime is the number of input coordinates that are ever observed (*i.e.*, activated at least once), $K(D)$[6], see Appendix H.2.2.

Theorem 2 (Overparameterized Scaling) *Under the sparse activation model with $\alpha_1 > 0$ and coordinate variances $\mathrm{Var}(x_j) = j^{-\alpha_1-\alpha_2-2}$, the Bayes-optimal loss in the overparameterized regime satisfies*

$$\ell_{\text{Bayes},D} \asymp D^{-\frac{\alpha_1+\alpha_2+1}{\alpha_1+1}}, \quad \textit{so that } \alpha_D = \frac{\alpha_1+\alpha_2+1}{\alpha_1+1}. \tag{5}$$

A rigorous proof of Theorem 2 is given in Appendix H.2.2 via a continuous argument.
Together, Proposition 1 and Theorem 2, with $\alpha_D = \frac{\alpha_1+\alpha_2+1}{\alpha_1+1} < \alpha_1 + \alpha_2 + 1 = \alpha_N$ yield:

$$\boxed{\ell_{\text{Bayes}}(N, D) \asymp N^{-(\alpha_1+\alpha_2+1)} + D^{-(\alpha_1+\alpha_2+1)/(\alpha_1+1)}. \quad (6)}$$

This expression makes the asymmetry explicit: model-limited error decays sharply, whereas data-limited error decays slowly. Since the mechanism depends on sparsity and variance profiles, we expect it to extend beyond Bernoulli masks to broader sparse distributions.

4. Additional Results

See Appendix for a detailed exposition of supplementary results.

6. the sparse activation model, the relevant comparison is ultimately between N and $K(D)$, which grows sublinearly in D when $\alpha_1 > 0$.

References

[1] Samira Abnar, Harshay Shah, Dan Busbridge, Alaaeldin El-Nouby, Joshua M. Susskind, and Vimal Thilak. Parameters vs FLOPs: Scaling laws for optimal sparsity for mixture-of-experts language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 204–230, 2025.

[2] Alexander Atanasov, Jacob A. Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

[3] Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Journal of Statistical Mechanics: Theory and Experiment*, 2019 (12):124023, 2019. doi: 10.1088/1742-5468/ab43d2.

[4] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

[5] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019. doi: 10.1073/pnas.1802705116.

[6] Jean Barbier, Francesco Camilli, Minh-Toan Nguyen, Mauro Pastore, and Rudy Skerk. Statistical physics of deep learning: Optimal learning of a multi-layer perceptron near interpolation. *arXiv preprint arXiv:2510.24616*, 2025.

[7] Maissam Barkeshli, Alberto Alfarano, and Andrey Gromov. On the origin of neural scaling laws: From random graphs to natural language. *arXiv preprint arXiv:2601.10684*, 2026.

[8] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. doi: 10.1073/pnas.1907378117.

[9] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116.

[10] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020. doi: 10.1137/20M1336072.

[11] Gérard Ben Arous, Murat A. Erdogdu, N. Mert Vural, and Denny Wu. Learning quadratic neural networks in high dimensions: SGD dynamics and scaling laws. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025.

[12] Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Power lines: Scaling laws for weight decay and batch size in llm pre-training. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025.

[13] Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*, volume 32, pages 12873–12884, 2019.

[14] Blake Bordelon and Francesco Mori. Theory of optimal learning rate schedules and scaling laws for a random feature model. *arXiv preprint arXiv:2602.04774*, 2026.

[15] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1024–1034, 2020.

[16] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4345–4382, 2024.

[17] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8): 084002, 2025. doi: 10.1088/1742-5468/adefb1.

[18] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

[19] Francesco Cagnetta, Allan Raventós, Surya Ganguli, and Matthieu Wyart. Deriving neural scaling laws from the statistics of natural language. *arXiv preprint arXiv:2602.07488*, 2026.

[20] Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems 22 (NeurIPS 2009)*, volume 22, pages 342–350, 2009.

[21] Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. In *Advances in Neural Information Processing Systems (NeurIPS 2024)*, volume 37, pages 104630–104693, 2024.

[22] Leonardo Defilippis, Yizhou Xu, Julius Girardin, Emanuele Troiani, Vittorio Erba, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Scaling laws and spectra of shallow neural networks in the feature learning regime. In *The Fourteenth International Conference on Learning Representations*, 2026.

[23] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

[24] Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019. doi: 10.1103/PhysRevE.100.012115.

[25] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Generalisation error in learning with random features and the hidden manifold model. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3452–3462, 2020.

[26] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, 2011.

[27] Sebastian Goldt, Marc Mezard, Florent Krzakala, and Lenka Zdeborova. Modelling the influence of data structure on learning in neural networks: the hidden manifold model. *Physical Review X*, 10(4):041044, 2020. doi: 10.1103/PhysRevX.10.041044.

[28] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The Gaussian equivalence of generative models for learning with shallow neural networks. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pages 426–471, 2022.

[29] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022. doi: 10.1214/21-AOS2133.

[30] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

[31] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[32] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, volume 35, pages 30016–30030, 2022.

[33] Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2023. doi: 10.1109/TIT.2022.3217698.

[34] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[35] Jakob Kramp, Javed Lindner, and Moritz Helias. Dynamics of neural scaling laws in random feature regression with powerlaw-distributed kernel eigenvalues. *arXiv preprint arXiv:2602.23039*, 2026.

[36] Tanishq Kumar, Zachary Ankner, Benjamin Frederick Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Re, and Aditi Raghunathan. Scaling laws for precision. In *The Thirteenth International Conference on Learning Representations*, 2025.

[37] Hugo Latourelle-Vigeant and Elliot Paquette. Dyson equation for correlated linearizations and test error of random features regression. *arXiv preprint arXiv:2312.09194*, 2023.

[38] Binghui Li, Fengling Chen, Zixun Huang, Lean Wang, and Lei Wu. Functional scaling laws in kernel regression: Loss dynamics and learning rate schedules. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025.

[39] Licong Lin, Jingfeng Wu, Sham M. Kakade, Peter L. Bartlett, and Jason D. Lee. Scaling laws in linear regression: Compute, parameters, and data. In *Advances in Neural Information Processing Systems (NeurIPS 2024)*, volume 37, pages 60556–60606, 2024.

[40] Licong Lin, Jingfeng Wu, and Peter L. Bartlett. Improved scaling laws in linear regression via data reuse. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025.

[41] Yizhou Liu, Ziming Liu, and Jeff Gore. Superposition yields robust neural scaling. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025.

[42] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018. doi: 10.1214/17-AAP1328.

[43] Jan Ludziejewski, Jakub Krajewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Piotr Sankowski, Marek Cygan, and Sebastian Jaszczur. Scaling laws for fine-grained mixture of experts. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 33270–33288, 2024.

[44] Andrew Maloney, Daniel A. Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

[45] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022. doi: 10.1002/cpa.22008.

[46] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Proceedings of the Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3351–3418, 2021.

[47] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022. doi: 10.1016/j.acha.2021.12.003.

[48] Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A. Raffel. Scaling data-constrained

language models. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, volume 36, pages 50358–50376, 2023.

[49] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations (ICLR)*, 2020. doi: 10.48550/arXiv.1912.02292.

[50] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269:543–547, 1983.

[51] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

[52] Ta Duy Nguyen, Thien H. Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, volume 36, pages 24191–24222, 2023.

[53] Brendan O'Donoghue and Emmanuel Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015. doi: 10.1007/s10208-013-9150-3.

[54] Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. doi: 10.1038/381607a0.

[55] Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws. In *Advances in Neural Information Processing Systems (NeurIPS 2024)*, volume 37, pages 16459–16537, 2024.

[56] Elliot Paquette, Ke Liang Xiao, and Yizhe Zhu. Power-law spectrum of the random feature model. *arXiv preprint arXiv:2603.14578*, 2026.

[57] Leonardo Petrini, Francesco Cagnetta, Eric Vanden-Eijnden, and Matthieu Wyart. Learning sparse features can lead to overfitting in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, volume 35, pages 9403–9416, 2022.

[58] Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving discrepancies in compute-optimal scaling of language models. In *Advances in Neural Information Processing Systems (NeurIPS 2024)*, volume 37, 2024.

[59] Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D. Lee. Emergence and scaling laws in SGD learning of shallow neural networks. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025.

[60] Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. Beyond Chinchilla-optimal: Accounting for inference in language model scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 43445–43460, 2024.

[61] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *The Fifth International Conference on Learning Representations*, 2017.

[62] Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: Empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, 2020. doi: 10.1088/1742-5468/abc61d.

[63] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Adam Jermyn, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

[64] Peihao Wang, Rameswar Panda, and Zhangyang Wang. Data efficient neural scaling law via model reusing. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 36193–36204, 2023.

[65] Roman Worschech and Bernd Rosenow. Analyzing neural scaling laws in two-layer networks with power-law data spectra. In *The Thirteenth International Conference on Learning Representations*, 2025.

[66] Arie Wortsman and Bruno Loureiro. Kernel ridge regression under power-law data: Spectrum and generalization. *arXiv preprint arXiv:2510.04780*, 2025.

[67] Tingkai Yan, Haodong Wen, Binghui Li, Kairong Luo, Wenguang Chen, and Kaifeng Lyu. Larger datasets can be repeated more: A theoretical analysis of multi-epoch scaling in linear regression. In *The Fourteenth International Conference on Learning Representations*, 2026.

[68] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022.

[69] Dechen Zhang, Xuan Tang, Yingyu Liang, and Difan Zou. Scaling laws for precision in high-dimensional linear regression. *arXiv preprint arXiv:2602.19241*, 2026.

[70] Zhengkang Zhang. Neural scaling laws from large-N field theory: Solvable model beyond the ridgeless limit. *arXiv preprint arXiv:2405.19398*, 2024.

Appendix A. Empirical Study of the Scaling Laws

We now empirically validate the theoretical predictions of the previous section by evaluating the population loss over a wide range of D and N in the sparse random-feature model of Section 2.1. Computational details are provided in Appendices G and K.

A.1. Scaling Collapse and Universality

In many models, including ours and others such as the Chinchilla scaling law [32], the test loss exhibits distinct scaling in two regimes: $\ell \sim D^{-\alpha_D}$ in the data-limited regime and $\ell \sim N^{-\alpha_N}$ in the parameter-limited regime. A natural question is how to interpolate between these two limits. A common ansatz is that the loss takes the form $\ell = \bar{L}(D^{-\alpha_D}, N^{-\alpha_N})$ for some *scale-invariant* function $\bar{L}$ satisfying $\bar{L}(cL_1, cL_2) = c\bar{L}(L_1, L_2)$. Simple choices include $\bar{L}(L_1, L_2) = \max(L_1, L_2)$ or $L_1 + L_2$, but more refined analyses [44, 70] show that $\bar{L}$ often exhibits a *peak* near $L_1 \approx L_2$, a phenomenon known as *double descent* [9, 29].[7] Using the scale invariance of $\bar{L}$, we may write $\ell \cdot N^{\alpha_N} = \bar{L}\left(\frac{D^{-\alpha_D}}{N^{-\alpha_N}}, 1\right)$, so the rescaled loss $\ell \cdot N^{\alpha_N}$ becomes a universal function of the compute ratio $\Xi := \frac{D^{\alpha_D}}{N^{\alpha_N}}$. We therefore expect the rescaled loss curves from different (D, N) to collapse onto a single curve $\mathcal{S}_{\alpha_1,\alpha_2}(\Xi)$: a dimensionless scaling function of the rescaled compute ratio.

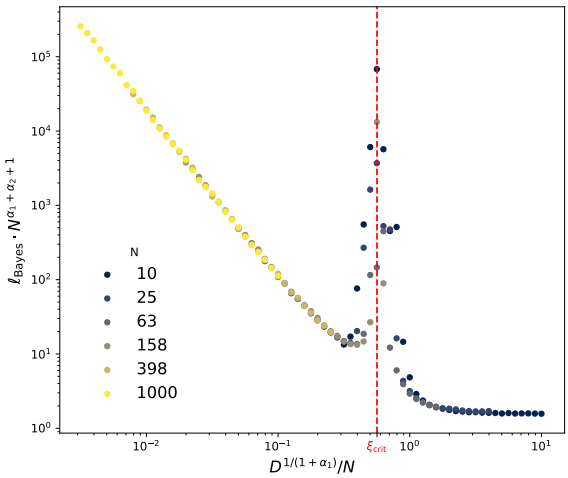

Figure 2: **Scaling collapse and double descent.** We plot the rescaled loss $\ell_{\text{Bayes}} \cdot N^{\alpha_N}$ as a function of $(D^{\alpha_D}/N^{\alpha_N})^{1/\alpha_N}$, across a wide range of values for D, N, for fixed $(\alpha_1 = 1.0, \alpha_2 = 0.3)$. All curves collapse onto a single universal function $\mathcal{S}_{\alpha_1,\alpha_2}$, verifying a universal scale-invariant structure consistent with the predicted scaling law. A sharp peak emerges near ξ_{crit} consistent with the double descent phenomenon at the interpolation threshold.

Empirical Scaling Collapse. In Figure 2, we find that the full loss curve across a wide range of (D, N) pairs collapses onto a single universal function after appropriate rescaling: $\ell_{\text{Bayes}} \cdot N^{\alpha_1+\alpha_2+1} = \mathcal{S}_{\alpha_1,\alpha_2}\left(\frac{1}{N} D^{\frac{1}{\alpha_1+1}}\right)$, where $\mathcal{S}(u) = \bar{L}(u^{-\alpha_N}, 1)$, so that $\bar{L}(\frac{D^{-\alpha_D}}{N^{-\alpha_N}}, 1) = \mathcal{S}_{\alpha_1,\alpha_2}(\xi)$, with $\xi := \left(\frac{D^{\alpha_D}}{N^{\alpha_N}}\right)^{1/\alpha_N} = \frac{D^{1/(\alpha_1+1)}}{N}$. This implies that knowing the loss curve for one setting of (D, N) suffices to predict its shape at other scales via a simple rescaling transformation. This collapse supports the two-regime scaling theory and recovers the predicted asymptotic behaviors: in the underparameterized regime, $\ell \sim N^{-\alpha_N}$ with $\alpha_N = \alpha_1 + \alpha_2 + 1$, while in the overparameterized regime, $\ell \sim D^{-\alpha_D}$ with $\alpha_D = (\alpha_1 + \alpha_2 + 1)/(\alpha_1 + 1)$.

Double Descent Peak. In the underparameterized limit, the loss scales as $\ell \sim N^{-(\alpha_1+\alpha_2+1)}$, while in the overparameterized limit it scales as $\ell \sim D^{-(\alpha_1+\alpha_2+1)/(\alpha_1+1)}$. This suggests a crossover at $N \sim D^{1/(\alpha_1+1)}$, which coincides with the point at which the number of model parameters N matches the number of activated (and hence learnable) coordinates $K(D)$ from Lemma 8. The shape of $\mathcal{S}_{\alpha_1,\alpha_2}(\xi)$ in Figure 2 reveals a prominent *double descent* phenomenon near:

$$\xi_{\text{crit}} = \frac{1}{\Gamma\left(\frac{\alpha_1}{1+\alpha_1}\right)}, \tag{7}$$

7. The double-descent peak is a feature of benign interpolation under ridgeless fitting, rather than of the scaling-law picture itself. Regularization or early stopping smooths it out [29].

corresponding to the threshold where $K(D)$ matches N, with location determined universally by the sparsity exponent α_1.

Appendix B. Compute-Optimal Scaling Laws

We now analyze how to allocate data and model size under a fixed compute budget so as to minimize the loss. We adopt the compute model $C = ND \cdot \min\{N, D\}$ as a unified proxy for the cost of training to completion in our setting; see Appendix I for discussion. We then minimize $\ell(N, D) \sim N^{-\alpha_N} + D^{-\alpha_D}$ subject to fixed compute C to derive the compute-optimal scaling laws.

Proposition 3 (Compute-Optimal Frontier) *Let the Bayes-optimal loss, for $\alpha_1 > 0$, scale as $\ell(N, D) \sim N^{-\alpha_N} + D^{-\alpha_D}$, with $\alpha_N = \alpha_1 + \alpha_2 + 1$ and $\alpha_D = \frac{\alpha_1+\alpha_2+1}{\alpha_1+1}$. Under a fixed compute budget $C = ND \cdot \min\{N, D\}$, the unique compute-optimal allocation lies in the underparameterized regime $N < D$, and the optimal allocation and resulting loss scale as*

$$\begin{aligned} &\ell^*(C) \sim C^{-\alpha_C}, \\ &N^*(C) \sim C^{\alpha_D/(\alpha_N+2\alpha_D)} = C^{1/(\alpha_1+3)}, \qquad D^*(C) \sim C^{1-2\alpha_D/(\alpha_N+2\alpha_D)} = C^{(\alpha_1+1)/(\alpha_1+3)}, \\ &\alpha_C := \frac{\alpha_N \alpha_D}{\alpha_N + 2\alpha_D} = \frac{\alpha_1 + \alpha_2 + 1}{\alpha_1 + 3} > 0. \end{aligned} \tag{8}$$

Remark 4 (Absence of Overparameterized Optimum) *The compute-optimal solution always lies in the underparameterized regime: $N^*(C) < D^*(C)$ for all C. An overparameterized optimum $N > D$, with $C = ND^2$, yields $N^*(C)/D^*(C) \sim C^{(\alpha_D-\alpha_N)/(2\alpha_N+\alpha_D)} \ll 1$, contradicting the assumed regime $N > D$. Hence the only valid optimum satisfies $N < D$.*

Remark 5 (Impact of Sparsity on Optimal Scaling) *As α_1 increases, individual samples become more sparse, carrying less information per example. At the same time, increasing α_1 also steepens the marginal variance spectrum $\mathrm{Var}(x_j) = j^{-\alpha_1-\alpha_2-2}$, so the residual tail loss decays more rapidly once coordinates are resolved. Consequently, for fixed $\alpha_2 < 2$, the compute-optimal exponent $\alpha_C = \frac{\alpha_1+\alpha_2+1}{\alpha_1+3}$ increases monotonically with α_1, approaching 1 as $\alpha_1 \to \infty$. As α_1 increases, the optimal model size, $N^*(C)$, grows more slowly with compute, while the optimal dataset size, $D^*(C)$, grows more rapidly. As a result, the allocation shifts toward spending a larger fraction of compute on data rather than parameters.*

Empirical Validation. To validate Proposition 3, Figure 3 shows test loss versus total compute C for different N. Each curve corresponds to a fixed N, while C is varied by sweeping D. The dashed line shows the predicted compute-optimal frontier, $\ell^*(C) \sim C^{-\alpha_C}$. As C increases, the empirical curves approach this frontier, confirming the predicted scaling.

Appendix C. Gradient Descent Training Dynamics

We next study the training dynamics of the readout weights $\boldsymbol{\theta}$ under full-batch GD, a broadly meaningful computational model with some generality beyond the toy setting, on the empirical squared loss $\ell(\boldsymbol{\theta}) = \frac{1}{2D} \left\| \mathbf{w}^\top \mathbf{X} - \boldsymbol{\theta}^\top \mathbf{u} \mathbf{X} \right\|_2^2$. With step size η, one GD step is $\Delta\boldsymbol{\theta} = \frac{\eta}{D} \mathbf{u}\mathbf{X}\mathbf{X}^\top \left(\mathbf{w} - \mathbf{u}^\top \boldsymbol{\theta}\right)$. It is convenient to rewrite the dynamics in terms of the input-space residual $\mathbf{r}_t := \mathbf{w} - \hat{\mathbf{w}}_t$, where $\hat{\mathbf{w}}_t := \mathbf{u}^\top \boldsymbol{\theta}_t \in \mathbb{R}^M$. Then

$$\mathbf{r}_{t+1} = \left(\mathbf{I}_M - \frac{\eta}{D} \mathbf{u}^\top \mathbf{u}\, \mathbf{X}\mathbf{X}^\top\right) \mathbf{r}_t, \tag{9}$$

so convergence is controlled by the spectrum of $\mathbf{u}^\top \mathbf{u}\,\mathbf{X}\mathbf{X}^\top$[8] ; see Appendix J.

C.1. Convergence with High Probability

A first question is how η must scale for (9) to converge. Since $\ell(\boldsymbol{\theta})$ is quadratic, full-batch GD is a linear iteration, and a standard stability criterion implies that (9) converges whenever $0 < \eta < \frac{2}{\lambda_{\max}\left(\frac{1}{D}\mathbf{u}^\top \mathbf{u}\,\mathbf{X}\mathbf{X}^\top\right)}$, where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue [51]. Equivalently, this is the usual condition $0 < \eta < 2/\lambda_{\max}(\nabla^2_{\boldsymbol{\theta}}\ell)$, since $\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \frac{1}{D}\mathbf{u}\mathbf{X}\mathbf{X}^\top\mathbf{u}^\top$.In the underparameterized regime, a standard concentration argument for the feature-space Hessian $\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \frac{1}{D}\,\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ yields $\lambda_{\max}(\nabla^2_{\boldsymbol{\theta}}\ell) = 1 + o_{\mathbb{P}}(1)$ and hence any fixed $\eta \in (0, 2)$ is admissible with high probability. The overparameterized case, however, is more subtle and is controlled by whether the random embedding acts isometrically on the data span.

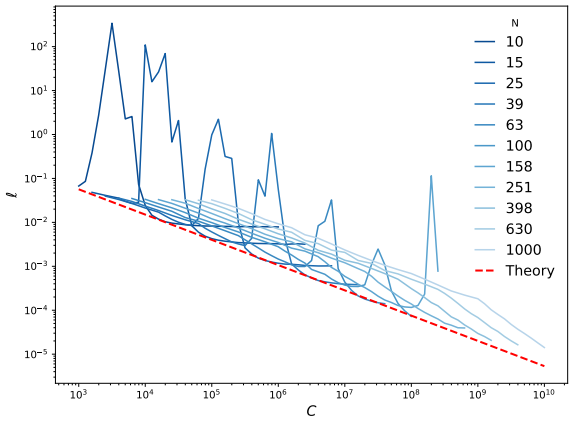

Figure 3: **Empirical scaling of loss with compute.** Each curve corresponds to test loss ℓ versus total compute C for a fixed model size N, with C varied by sweeping D, for fixed $(\alpha_1 = 1.0, \alpha_2 = 0.3)$. The dashed line denotes the predicted compute-optimal scaling $\ell^*(C) \sim C^{-\alpha_C}$. The empirical loss converges toward this frontier at high compute, just past the double descent spike for each curve demonstrating agreement between theory and experiment.

Proposition 6 (Step-size Stability with High Probability) *Assume the sparse activation model of Section 2.1 with $\alpha_1 \geq -1$ and $\alpha_1 + \alpha_2 + 1 > 0$. Assume further that the random feature map is sufficiently wide that it is nearly isometric on the (low-dimensional) span of the data,*[9] *so that*

$$\lambda_{\max}\left(\frac{1}{D}\,\mathbf{u}^\top \mathbf{u}\,\mathbf{X}\mathbf{X}^\top\right) = 1 + o_{\mathbb{P}}(1). \tag{10}$$

Consequently, for any fixed $\eta \in (0, 2)$ the iteration (9) is stable and converges with high probability.

C.2. A Scaling Law for Failure of Gradient Descent

Proposition 6 is a high-probability statement; rare datasets can still contain a single rare activation spike that produces an anomalously large top eigenvalue and causes divergence.[10] GD becomes unstable when $\lambda_{\max}\left(\frac{\eta}{D}\mathbf{u}^\top \mathbf{u}\mathbf{X}\mathbf{X}^\top\right) > 2$. Under the near-isometry approximation $\mathbf{u}^\top \mathbf{u} \approx \mathbf{I}_M$ on the data span, this reduces to $\lambda_{\max}\left(\mathbf{X}\mathbf{X}^\top\right) > \frac{2D}{\eta}$. This is only a concern when activated amplitudes *grow* with index, *i.e.* when $\alpha_2 < -1$ (since then $x_j^2 = j^{-(\alpha_2+1)}$ increases with j on activation).

8. The data dependence enters through the empirical covariance $\mathbf{X}\mathbf{X}^\top$. For dense inputs, $\mathrm{rank}(\mathbf{X}\mathbf{X}^\top) = \mathrm{rank}(\mathbf{X}) \approx \min\{M, D\}$, which is typically $\approx D$ when $D \lesssim M$, whereas under our sparse model, $\mathrm{rank}(\mathbf{X}\mathbf{X}^\top) \leq K(D) \asymp D^{1/(\alpha_1+1)}$, so it grows only sublinearly in D.
9. Concretely, since $\mathrm{rank}(\mathbf{X}\mathbf{X}^\top) \leq D$, standard random projection heuristics suggest that for $N \gg D$ the operator $\mathbf{u}^\top \mathbf{u}$ acts approximately like $\mathbf{I}_M$ when sandwiched against $\mathbf{X}\mathbf{X}^\top$.
10. Rare activations can create spectral outliers in $\mathbf{X}\mathbf{X}^\top$ that are not controlled by the high-probability spectral concentration arguments used in recent approximation/bias/variance analyses of linear-regression scaling laws [39, 40, 67], so we instead adopt a probabilistic treatment.

Theorem 7 (Failure Probability of GD) *Fix $\eta \in (0, 2)$ and suppose $\alpha_2 < -1$. Under the sparse activation model of Section 2.1, with $\alpha_1 \geq -1$ and $\alpha_1 + \alpha_2 + 1 > 0$, let $\nu := \frac{\alpha_1+\alpha_2+1}{-\alpha_2-1} > 0$. Then the probability that full-batch GD is unstable due to a rare spike on a random dataset of size D obeys*

$$\mathbb{P}_{\text{rare}}(\text{diverge}) \asymp D^{-\nu}, \tag{11}$$

up to η-dependent constants.

This theorem identifies instability under a fixed step size, not a fundamental optimization barrier. Reducing the learning rate or using clipped updates can restore stability [52].

Appendix D. Experiments with Nonlinear Activations

A natural question is whether the two-exponent asymmetry we identify is specific to the linear setting or reflects a more general property of sparse data distributions. In Figure 4, we test robustness to nonlinearity by training a two-layer network with hidden layer $\phi = \sigma(\mathbf{ux})$ (ReLU activation, frozen random first-layer weights) and a linear readout, using full-batch accelerated GD with Nesterov momentum and adaptive restart; see Appendix K for experimental details. The fitted exponents shift downward in the model-limited regime, consistent with the general expectation that nonlinear feature maps smooth power-law input spectra [13, 20, 42]. Crucially, the qualitative asymmetry persists: sparse N-, dense N-, dense D-sweeps are well-described by a single shared exponent, while the sparse D-sweep is better described by a distinct exponent. This supports our main result: the bottleneck mechanism—unobserved coordinates carrying no signal—is robust to nonlinearity and first-order optimization.

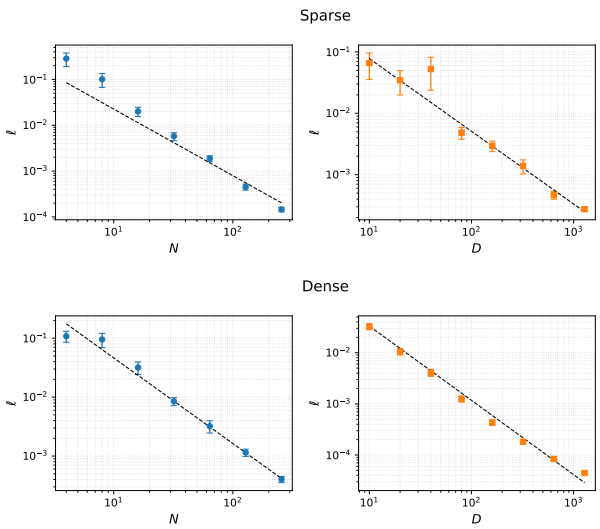

Figure 4: **Asymmetry persists under nonlinearity.** Test loss scaling under the ReLU feature map $\phi = \sigma(\mathbf{ux})$, trained with Nesterov + adaptive restart; $(\alpha_1, \alpha_2) = (1.0, 0.3)$, 5 seeds. Top: sparse; bottom: dense. Left: N-sweep at $D = 50{,}000$; right: D-sweep at $N = 8000$. A single exponent $\alpha \approx 1.5$ (dashed lines, jointly fitted with separate intercepts) describes the sparse N-, dense N-, and dense D-sweeps. The sparse D-sweep requires its own shallower fit, $\alpha_D \approx 1.2$, breaking the dense-baseline symmetry $\alpha_N = \alpha_D$ predicted by linear theory.

Appendix E. Conclusion

We introduced a model that gives rise to two distinct scaling exponents under sparse input activations. By analyzing the population loss in both under- and overparameterized regimes, we showed how sparsity breaks the symmetry of scaling laws in the random feature setting. We derived a compute-optimal scaling law showing that sparsity enhances efficiency by accelerating the decay of test loss with compute, with a frontier that prioritizes data over model size. Analyzing GD dynamics, we proved high-probability stability for all step sizes below 2 and identified a suppressed

sublinear scaling law for its probability of failure in the overparameterized regime, suggesting a role for gradient clipping in practical optimization. We validated these results experimentally in the theoretical setting and extended them to a nonlinear feature map, confirming robustness of the sparsity-induced asymmetry to nonlinearity. Our main takeaway is that sparsity as a structural property fundamentally shapes scaling behavior in high-dimensional learning.

Limitations. Our results are built around data sparsity as the organizing principle, yet several extensions remain for future work. Incorporating exogenous label noise, which we expect to impose an error floor rather than remove the sparsity-driven bottleneck until it dominates the optimization scale, and analyzing ridge regularization, which would refine the statistical picture in the presence of explicit shrinkage, would extend our picture to noisy and explicitly regularized settings. More ambitiously, a theoretical treatment of nonlinear feature maps—complementing our empirical validation—and an extension to the dynamical regime of few- to multi-pass SGD both confront a common technical obstacle: the non-Gaussian, heavy-tailed nature of our features requires extending existing deterministic-equivalent tools beyond the Gaussian setting [33, 45]. Relatedly, extending our analysis to the feature-learning regime is an important direction, especially since feature learning may either concentrate capacity on rare informative coordinates [17] or induce sparse representations that hurt generalization in smooth tasks [57]. A further natural direction is to test these effects at scale, in deeper networks and in transformers, particularly in light of sparse-feature observations from sparse autoencoders [18, 63] and recent connections between superposition [23] and scaling laws [41].

Extra Appendices

Appendix F. Additional Related Work

Empirical scaling laws relating test loss to parameters, data, and compute were systematically documented in language modeling [34] and refined into compute-optimal prescriptions emphasizing data-limited training in LLMs [32]. Related scaling behavior has also been observed across tasks and modalities, including autoregressive generative modeling [30], multi-domain studies [31] and vision transformer scaling analyses [68]. More recent empirical work has expanded scaling-law analyses to address compute-optimal discrepancies across experimental protocols [58], practical scaling laws for weight decay and batch size in LLM pretraining [12], precision-aware scaling under quantization and training precision [36], inference-aware compute-optimal scaling [60], data-scarcity regimes via model reuse [64], and data-constrained scaling under token repetition [48].

On the theory side, a growing literature develops tractable models that reproduce and explain scaling-law phenomenology. Statistical-physics teacher-student models characterize optimal errors and phase transitions in high-dimensional generalized linear models [5] and computational-to-statistical gaps in committee machines [3]. Solvable neural-scaling models connect power-law learning curves to spectral structure and compute-optimal frontiers [4, 44]. Random-feature and kernel analyses address invariances [46], hypercontractivity and kernel matrix concentration [47], correlated linearizations and test error [37], and dimension-free deterministic equivalents [21]. Scaling and renormalization approaches provide a complementary high-dimensional regression perspective [2]. Recent work also extends this power-law spectral perspective to kernel ridge regression under anisotropic power-law data [66] and spectral inheritance through random-feature maps [56]. Related statistical-physics analyses study feature learning and specialization in finite-width teacher-student neural networks near interpolation [6], as well as scaling-law phase diagrams in shallow feature-learning networks [22]. A complementary line studies compute-optimal scaling in linear regression [39] and precision-aware variants in high-dimensional linear regression [69]. Recent work also studies the origin of scaling laws in simplified language and sequence-modeling settings, including theories based on natural-language statistics [19] and controlled graph/random-walk generative processes [7].

Beyond Bayes-optimal analyses, several works study *training dynamics*. Dynamical models of one-pass or finite-time training exhibit multiple scaling phases and shift compute-optimal allocations [16, 55]. Feature learning can also change scaling exponents and compute-optimal behavior [17]. Data reuse and multi-epoch training shift effective exponents and optimal allocations [40, 67]. Functional loss dynamics and learning-rate schedules in kernel or random-feature regression are studied in [14, 38], while early stopping in random-feature regression with power-law spectra is analyzed in [35]. Related analyses of SGD in shallow or two-layer networks with heterogeneous teacher components [59], quadratic feature-learning models [11], or power-law data spectra [65] derive sample-, time-, and parameter-dependent scaling exponents.

Our work is motivated by a different axis: *sparse or conditional feature activation* [26, 54], in which only a subset of features is observed or used for any example. Such sparsity is common

in modern representation learning and conditional computation (*e.g.*, mixture-of-experts) [61]. Related empirical work studies scaling laws for sparse conditional computation in mixture-of-experts models [1, 43]. Rather than refining existing spectral-decay accounts, we isolate how sparsity in what the dataset reveals can itself produce an intrinsic asymmetry between model-limited and data-limited scaling exponents.

Appendix G. Population Objective of Sparse Model

We begin by reformulating the objective in Eq. (2) as the following loss function:

$$\begin{aligned}
\ell &= \frac{\gamma}{D}\|\boldsymbol{\theta}\|_2^2 + \frac{1}{D}\sum_{a=1}^{D}\Big(y(\mathbf{x}_a) - \boldsymbol{\theta}^\top \mathbf{u}\mathbf{x}_a\Big)^2 \\
&= \frac{\gamma}{D}\|\boldsymbol{\theta}\|_2^2 + \frac{1}{D}\sum_{a=1}^{D}\Big(\mathbf{w}^\top \mathbf{x}_a - \boldsymbol{\theta}^\top \mathbf{u}\mathbf{x}_a\Big)^2 \\
&= \frac{\gamma}{D}\|\boldsymbol{\theta}\|_2^2 + \frac{1}{D}\sum_{a=1}^{D}\Big[\mathbf{w}^\top \mathbf{x}_a\mathbf{x}_a^\top \mathbf{w} - 2\,\boldsymbol{\theta}^\top \mathbf{u}\mathbf{x}_a\mathbf{x}_a^\top \mathbf{w} + \boldsymbol{\theta}^\top \mathbf{u}\mathbf{x}_a\mathbf{x}_a^\top \mathbf{u}^\top \boldsymbol{\theta}\Big],
\end{aligned} \tag{12}$$

which we wish to minimize with respect to $\boldsymbol{\theta}$. While our focus is on the ridge-less limit, $\gamma = 0$, we retain it for mathematical convenience and take the limit $\gamma \to 0$ when appropriate.

Using $\boldsymbol{\Phi} = \mathbf{u}\mathbf{X}$, we find the optimal solution $\boldsymbol{\theta}^\star$ satisfies:

$$\boldsymbol{\theta}^{\star\top} = \mathbf{y}\boldsymbol{\Phi}^\top \mathbf{q} = \mathbf{y}\mathbf{Q}\boldsymbol{\Phi}^\top, \tag{13}$$

where

$$\begin{aligned}
\mathbf{q} &= (\gamma\,\mathbf{I}_N + \boldsymbol{\Phi}\boldsymbol{\Phi}^\top)^{-1}, \\
\mathbf{Q} &= (\gamma\,\mathbf{I}_D + \boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}.
\end{aligned}$$

We are interested in the difference between the effective weight vector $\mathbf{w}_{\text{eff}}$, defined by

$$\mathbf{w}_{\text{eff}}^\top = \boldsymbol{\theta}^{\star\top}\mathbf{u},$$

and the true target weights $\mathbf{w}^\top$. In particular, the generalization error is given by

$$\mathcal{E}_{\text{gen}} = \big(\mathbf{w}_{\text{eff}}^\top - \mathbf{w}^\top\big)\,\boldsymbol{\Sigma}_{\mathbf{x}}\,\big(\mathbf{w}_{\text{eff}} - \mathbf{w}\big).$$

Appendix H. Scaling Asymptotics of Sparse Random Features

H.1. $-1 \le \alpha_1 \le 0$: One-Exponent Scaling Law

In our sparse model of Section 2.1, we obtain a symmetric one-exponent scaling law when $-1 \le \alpha_1 \le 0$, similar to dense random feature models [4, 44]. The key point is that generalization can be bottlenecked by three resources: the number of parameters N, the number of samples D, and the set of input coordinates that are actually observed (activated at least once) in the D training examples.

Recall that the activation probability of coordinate j is

$$p_j := \mathbb{P}(x_j \neq 0) = j^{-(\alpha_1+1)}.$$

Across D i.i.d. samples, let N_j denote the number of times coordinate j activates. Then

$$\mathbb{P}(\text{coordinate } j \text{ is never active in the training set}) = \mathbb{P}(N_j = 0) = (1 - p_j)^D \approx e^{-Dp_j}.$$

Thus, whenever $Dp_j \gg 1$ (equivalently, $D\, j^{-(\alpha_1+1)} \gg 1$), we have $\mathbb{P}(N_j = 0) \leq e^{-\Omega(1)} \ll 1$, so coordinate j is activated at least once with high probability. When $-1 \leq \alpha_1 < 0$, this condition holds uniformly for all j up to order D. At the boundary $\alpha_1 = 0$, the coordinate $j \sim D$ has $Dp_j = O(1)$, so the never-observed probability is also $O(1)$; this boundary case only affects constants and does not change the one-exponent scaling.

Therefore the Bayes-optimal loss is governed by the unmodeled-variance tail beyond the first $\min\{N, D\}$ coordinates:

$$\ell_{\text{Bayes}}(N, D) \asymp \sum_{j>\min\{N,D\}} \text{Var}(x_j) \asymp \sum_{j>\min\{N,D\}} j^{-(\alpha_1+\alpha_2+2)} \asymp \min\{N, D\}^{-(\alpha_1+\alpha_2+1)},$$

and the scaling is symmetric in N and D in this regime, with a single exponent $\alpha_N = \alpha_D = \alpha_1 + \alpha_2 + 1$.

H.2. $\alpha_1 > 0$: Two-Exponent Scaling Law

Here we briefly elaborate on supplementary details omitted from the main text regarding the scaling asymptotics of the two-exponent scaling law.

In a random set of D data points, only $O(D^{\frac{1}{1+\alpha_1}})$ features are likely to activate when $\alpha_1 > 0$. Thus, for positive α_1, there are naively three distinct regimes:

$$N < D^{\frac{1}{1+\alpha_1}}, \quad D^{\frac{1}{1+\alpha_1}} < N < D, \quad \text{and} \quad N > D.$$

The first is unambiguously underparameterized, the third is clearly overparameterized, and the intermediate regime exhibits characteristics of both.

H.2.1. UNDERPARAMETERIZED REGIME

We first consider the underparameterized regime $N \ll D^{\frac{1}{1+\alpha_1}}$, where learning is primarily limited by model capacity.

Proof of Proposition 1 **Proof** A resolvent analysis of random-feature regression [44] reveals that the loss is governed by the covariance mass missed by an N-dimensional random feature subspace. In the ridgeless limit, the resolvent acts as a projection onto the latent directions inaccessible to the random features, so the leading-order loss is controlled by the unresolved tail of the covariance spectrum. Thus, at the level of scaling, the random-feature bottleneck resolves the leading N spectral directions and leaves the remaining covariance tail unresolved. In the present regime $N \ll D^{1/(1+\alpha_1)}$, sparsity is masked, so the relevant spectrum is $\text{Var}(x_j) = j^{-\alpha_1-\alpha_2-2}$. Hence the residual variance scales as

$$\begin{aligned}\sum_{j>N} j^{-\alpha_1-\alpha_2-2} &\asymp \int_N^\infty j^{-\alpha_1-\alpha_2-2}\, dj \\ &= \frac{1}{\alpha_1+\alpha_2+1} N^{-(\alpha_1+\alpha_2+1)}.\end{aligned}$$

Combining this with the general form of the Bayes-optimal loss (Eq. (3)) yields (4). ■

H.2.2. OVERPARAMETERIZED REGIME

We now turn to the overparameterized regime $N \gg D^{\frac{1}{1+\alpha_1}}$.

Heavily Overparameterized Regime We first focus on the maximally overparameterized case. We will show that in this regime, the performance of the random feature model is Bayes-optimal among all models, and we compute its generalization error.

Optimal Predictor. An upper bound on the performance of any model would be for $\mathbf{w}_{\text{eff}}$ to equal $\mathbf{w}$ on indices that appear in the dataset, and be 0 on all other indices. We will show that this is indeed what happens in this regime.

If we write out our expression for $\mathbf{w}_{\text{eff}}$, we have

$$\mathbf{w}_{\text{eff}}^\top = \mathbf{w}^\top \mathbf{X} \left(\gamma \mathbf{I}_D + \mathbf{X}^\top \mathbf{u}^\top \mathbf{u}\, \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{u}^\top \mathbf{u}, \tag{14}$$

If we post-multiply by $\mathbf{X}\,\mathbf{b}$ for any vector $\mathbf{b}$, we obtain

$$\mathbf{w}_{\text{eff}}^\top \mathbf{X}\,\mathbf{b} = \mathbf{w}^\top \mathbf{X}\,\mathbf{b}.$$

This follows from the full rank of $\mathbf{Q}$. Thus, in the overparameterized regime, for every possible input in the subspace spanned by the training data, the random feature model incurs zero loss.

Coefficients Outside the Training Data Go to 0. While the analysis showing that $\mathbf{w}_{\text{eff}}$ perfectly interpolates the training data applies for any overparameterized model in the limit $\gamma \to 0$, the argument in this subsection uses the stronger assumption that we are deep in the overparameterized regime. Since $u_{ij} \sim \mathcal{N}(0, 1/N)$, the diagonal entries of $\mathbf{u}^\top \mathbf{u}$ are $1 + O_{\mathbb{P}}(N^{-1/2})$, while the off-diagonal entries are $O_{\mathbb{P}}(N^{-1/2})$. Thus, for any subspace S with $\dim(S) \ll N$, the restriction of $\mathbf{u}^\top \mathbf{u}$ to S is $\mathbf{I}_S + o_{\mathbb{P}}(1)$. Since the row and column spaces of $\mathbf{X}$ are at most D-dimensional, when $D \ll N$ we can replace the occurrences of $\mathbf{u}^\top \mathbf{u}$ in Eq. (14) by $\mathbf{I}$. Therefore, $\mathbf{w}_{\text{eff}}$ simplifies to

$$\mathbf{w}_{\text{eff}}^\top = \mathbf{w}^\top \mathbf{X} \left(\gamma \mathbf{I}_D + \mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top.$$

In the limit $\gamma \to 0$, this is precisely the projection of $\mathbf{w}^\top$ onto the space spanned by the data. We can see this using a singular value decomposition.

Proof. Let the singular value decomposition (SVD) of the data matrix be $\mathbf{X} = U \Sigma V^\top$, where Σ is the diagonal matrix of singular values σ_i. Substituting this into the term acting on $\mathbf{w}^\top$ yields

$$\begin{aligned}
\mathbf{X} \left(\gamma \mathbf{I}_D + \mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top &= U \Sigma V^\top \left(\gamma \mathbf{I}_D + V \Sigma^\top \Sigma V^\top\right)^{-1} V \Sigma^\top U^\top \\
&= U \Sigma V^\top \left[V (\gamma \mathbf{I}_D + \Sigma^\top \Sigma) V^\top\right]^{-1} V \Sigma^\top U^\top \\
&= U \underbrace{\Sigma (\gamma \mathbf{I}_D + \Sigma^\top \Sigma)^{-1} \Sigma^\top}_{\Lambda_\gamma} U^\top.
\end{aligned}$$

The inner matrix Λ_γ is diagonal with entries $(\Lambda_\gamma)_{ii} = \sigma_i^2 / (\gamma + \sigma_i^2)$. Taking the limit $\gamma \to 0$ we find

$$\lim_{\gamma \to 0} \frac{\sigma_i^2}{\gamma + \sigma_i^2} = \begin{cases} 1, & \text{if } \sigma_i \neq 0, \\ 0, & \text{if } \sigma_i = 0. \end{cases}$$

Thus, $\lim_{\gamma \to 0} \Lambda_\gamma$ acts as the identity on the subspace associated with non-zero singular values. Consequently, the full expression becomes $U_{\text{range}} U_{\text{range}}^\top$, which is the orthogonal projection operator onto the column space of $\mathbf{X}$. Therefore, in this limit, $\mathbf{w}_{\text{eff}}^\top$ is precisely the projection of $\mathbf{w}^\top$ onto the space spanned by the training data.

Expected Test Loss. In the heavily overparameterized limit, the distribution of $\mathbf{w} - \mathbf{w}_{\text{eff}}$ is straightforward. The j-th coefficient is zero if feature j appears at least once in the dataset, and is normally distributed with variance 1 if feature j does not appear. The contribution to the test loss is thus

$$\ell_{\text{test}} = \mathcal{E}_{\text{gen}} = \sum_j \Big(\mathbb{E}_{\text{test}}[x_j^2]\Big)\,\Big(\mathbb{E}_{\text{train}}\big[(\mathbf{w} - \mathbf{w}_{\text{eff}})_j^2\big]\Big).$$

Expected Number of Active Coordinates, $K(D)$. Here we derive the asymptotic expression for the expected number of coordinates that are active at least once across D i.i.d. samples under the sparse activation model.

Lemma 8 (Learnable Coordinates in the Sparse Model when $\alpha_1 > 0$) *Under the sparse activation distribution $p_j = \mathbb{P}(x_j \neq 0) = j^{-\alpha_1 - 1}$, the expected number of coordinates that are active in at least one of the D training samples is asymptotic to the scale*

$$K(D) = \Gamma\left(1 - \frac{1}{\alpha_1 + 1}\right) D^{\frac{1}{\alpha_1+1}}, \tag{15}$$ [11]

where $\Gamma(\cdot)$ is the Gamma function. Moreover, $K(D)$ sets the scale of the number of coefficients of $\mathbf{w}$ that can be estimated from data: a coordinate that is never activated carries no information about its associated weight, and hence cannot contribute to the learned predictor.

Proof [Proof (learnable-coordinate count $K(D)$)] In the sparse activation model, coordinate j is active in a single sample with probability $p_j = j^{-\alpha_1 - 1}$ and inactive with probability $1 - p_j$. The probability that coordinate j is never activated in D independent samples is

$$\mathbb{P}_{\text{never}}(j) = (1 - p_j)^D.$$

For $p_j \ll 1$, we use $(1 - p_j)^D \approx e^{-Dp_j}$ as $D \to \infty$; this follows from $(1 - \frac{Dp_j}{D})^D \to e^{-Dp_j}$. Hence

$$\mathbb{P}_{\text{never}}(j) \approx \exp(-Dp_j) = \exp\big(-Dj^{-\alpha_1 - 1}\big)\,.$$

Thus the probability that coordinate j is activated at least once is $1 - \exp(-Dj^{-\alpha_1 - 1})$, and the expected number of such coordinates is

$$K(D) \;\approx\; \sum_{j=1}^{\infty} \big(1 - e^{-Dj^{-\alpha_1 - 1}}\big).$$

Passing to the continuum limit, we obtain

$$K(D) \;\asymp\; \int_0^{\infty} \big(1 - e^{-Dj^{-\alpha_1 - 1}}\big)\, dj.$$

We now perform the change of variables

$$t = Dj^{-(\alpha_1+1)} \quad \Longrightarrow \quad j = \left(\frac{D}{t}\right)^{\frac{1}{\alpha_1+1}}, \quad dj = -\frac{1}{\alpha_1 + 1} D^{\frac{1}{\alpha_1+1}} t^{-\frac{\alpha_1+2}{\alpha_1+1}}\, dt,$$

11. When $\alpha_1 > 0$, $K(D) \asymp D^{1/(\alpha_1+1)} = o(D)$, *i.e.* only a sublinear number of coordinates are ever observed. Interpolation is still possible since the data lie in a $K(D)$-dimensional subspace, which a linear predictor in $\mathbb{R}^N$ can fit. Unobserved coordinates receive no signal and are typically suppressed by mild regularization.

which yields

$$K(D) \asymp \frac{1}{\alpha_1+1} D^{\frac{1}{\alpha_1+1}} \int_0^\infty \left(1-e^{-t}\right) t^{-\frac{\alpha_1+2}{\alpha_1+1}}\, dt.$$

Let $\beta \equiv \frac{1}{\alpha_1+1} \in (0,1)$. By integration by parts,

$$\int_0^\infty (1-e^{-t})\, t^{-1-\beta}\, dt = \frac{1}{\beta} \int_0^\infty e^{-t} t^{-\beta}\, dt = \frac{\Gamma(1-\beta)}{\beta}.$$

Substituting $\beta = \frac{1}{\alpha_1+1}$ gives

$$K(D) \asymp \Gamma\left(1 - \frac{1}{\alpha_1+1}\right) D^{\frac{1}{\alpha_1+1}}.$$

This matches Eq. (15) in Lemma 8. ■

Proof of Theorem 2 Here we give two derivations of Theorem 2: a discrete argument via the cutoff $K(D)$, which provides useful intuition, and a fully rigorous continuous argument.

Proof [Informal discrete argument via $K(D)$] By Lemma 8, when $\alpha_1 > 0$, at most $K(D)$ coordinates of the target vector $\mathbf{w}$ can be reliably estimated from D samples. All coordinates with indices $j > K(D)$ remain unobserved and therefore unmodeled. By the unmodeled-variance principle (Eq. (3)), the Bayes-optimal loss is

$$\ell_{\text{Bayes},D} \asymp \sum_{j>K(D)} j^{-\alpha_1-\alpha_2-2}.$$

Approximating the sum by an integral,

$$\begin{aligned}\sum_{j>K(D)} j^{-\alpha_1-\alpha_2-2} &\asymp \int_{K(D)}^\infty j^{-\alpha_1-\alpha_2-2}\, dj \\ &= \frac{1}{\alpha_1+\alpha_2+1} K(D)^{-(\alpha_1+\alpha_2+1)}.\end{aligned}$$

Substituting the scaling $K(D) \asymp D^{1/(\alpha_1+1)}$ from Eq. (15) yields

$$\ell_{\text{Bayes},D} \asymp D^{-(\alpha_1+\alpha_2+1)/(\alpha_1+1)},$$

which is Eq. (5). ■

Proof [Rigorous proof (continuous argument)] An equivalent expression for the Bayes-optimal loss can be written as a continuous expectation over coordinates, weighted by the probability that each coordinate is never activated. This leads to the integral representation

$$\ell_{\text{Bayes},D} = \int_0^\infty \exp\left(-Dj^{-\alpha_1-1}\right) j^{-(\alpha_1+\alpha_2+2)}\, dj, \tag{16}$$

We again apply the change of variables

$$t = Dj^{-(\alpha_1+1)} \quad \Longrightarrow \quad j = \left(\frac{D}{t}\right)^{\frac{1}{\alpha_1+1}}, \quad dj = -\frac{1}{\alpha_1+1} D^{\frac{1}{\alpha_1+1}} t^{-\frac{\alpha_1+2}{\alpha_1+1}}\, dt.$$

Substituting into (16) gives

$$\ell_{\text{Bayes},D} = \frac{1}{\alpha_1+1} D^{-\frac{\alpha_1+\alpha_2+1}{\alpha_1+1}} \int_0^\infty e^{-t}\, t^{\frac{\alpha_1+\alpha_2+1}{\alpha_1+1}-1}\, dt.$$

The remaining integral is exactly the Gamma function,

$$\int_0^\infty e^{-t} t^{\beta-1}\, dt = \Gamma(\beta),$$

with

$$\beta = \frac{\alpha_1+\alpha_2+1}{\alpha_1+1}.$$

Hence

$$\ell_{\text{Bayes},D} = \frac{1}{\alpha_1+1} \Gamma\left(\frac{\alpha_1+\alpha_2+1}{\alpha_1+1}\right) D^{-\frac{\alpha_1+\alpha_2+1}{\alpha_1+1}},$$

which matches the scaling law in Eq. (5) in the main text. When $\alpha_1 > 0$, sparsity reduces the effective number of observed coordinates to scale as $D^{1/(\alpha_1+1)}$, thereby becoming the dominant data bottleneck. This concludes the proof. ■

Subtleties Concerning Theorem 2: Co-activation and Identifiability. A subtle complication arises when two rare coordinates co-activate, *i.e.*, are both nonzero in the same training example. In such cases, it becomes impossible to disentangle their individual contributions to the label, creating ambiguity in estimating the corresponding weights. The coordinates that activate exactly once have indices on the order of $j \sim D^{\frac{1}{\alpha_1+1}}$. Consequently, the total number of such coordinates also scales as $D^{\frac{1}{\alpha_1+1}}$. A nonzero number of pairwise co-activations is therefore likely whenever this quantity exceeds $D^{1/2}$ (by the birthday paradox), which occurs when $\alpha_1 < 1$. However, the number of co-activating coordinates itself scales only as

$$\min\left\{\frac{\left(D^{\frac{1}{\alpha_1+1}}\right)^2}{D},\ D^{\frac{1}{\alpha_1+1}}\right\} = \min\left\{D^{\frac{2}{\alpha_1+1}-1},\ D^{\frac{1}{\alpha_1+1}}\right\} = \min\left\{D^{\frac{1-\alpha_1}{\alpha_1+1}},\ D^{\frac{1}{\alpha_1+1}}\right\}.$$

Each such coordinate contributes variance on the order of $D^{\frac{-\alpha_1-\alpha_2-2}{\alpha_1+1}}$, so the total variance contributed by co-activating coordinates scales as

$$\min\left\{D^{\frac{-2\alpha_1-\alpha_2-1}{\alpha_1+1}},\ D^{\frac{-\alpha_1-\alpha_2-1}{\alpha_1+1}}\right\},$$

which vanishes as $D \to \infty$. Thus, while co-activation may create isolated ambiguity, it does not affect the asymptotic loss scaling in Theorem 2.

Appendix I. Compute

This section provides an overview of the compute model we adopt for studying the compute-optimal frontier under training to completion:

$$C = ND \cdot \min\{N, D\}. \tag{17}$$

I.1. Computational Efficiency

Here, "compute" refers to the leading-order training cost, measured in FLOPs up to hardware-dependent constants. To approach the population-loss scaling studied in the main text, we focus on the setting in which the readout is trained to convergence. This can be done either with gradient-based methods, which are closer to practical training, or by directly computing the min-norm solution in Eq. (13).

I.1.1. GRADIENT-BASED METHODS

A single gradient step on the loss (12) costs $O(ND)$. The iteration complexity is controlled by the condition number of the feature Gram matrix, defined on its nonzero spectrum, $\kappa(\mathbf{\Phi}\mathbf{\Phi}^\top) = \lambda_{\max}(\mathbf{\Phi}\mathbf{\Phi}^\top)/\lambda^+_{\min}(\mathbf{\Phi}\mathbf{\Phi}^\top)$, where $\lambda^+_{\min}$ denotes the smallest nonzero eigenvalue. For plain GD, the number of steps scales linearly in κ, up to logarithmic accuracy factors. Accelerated first-order methods improve this dependence to $\sqrt{\kappa}$. We therefore write the leading gradient-based training cost as

$$C_{\text{grad}} \sim ND\,\kappa^p, \tag{18}$$

with $p = 1$ for GDand $p = 1/2$ for accelerated methods.

Condition number, κ. We estimate the condition number from the effective variance profile of the active features. The number of active coordinates in a dataset of size D is $K(D)$, so the readout effectively uses roughly $\min\{N, K(D)\}$ features. The variance explained by the smallest feature scales as $\min\{N, K(D)\}^{-(\alpha_1+\alpha_2+2)}$, while the largest is $O(1)$. Thus the effective condition number scales as

$$\kappa \sim \min\{N, K(D)\}^{\alpha_1+\alpha_2+2}. \tag{19}$$

Consequently, if the iteration complexity scales as κ^p, the number of optimization steps scales as $\min\{N, K(D)\}^{(\alpha_1+\alpha_2+2)p}$.

I.1.2. DIRECT LEAST-SQUARES SOLVE

Eq. (13) gives two equivalent expressions for $\boldsymbol{\theta}^\star$, related by a standard matrix identity. Using the first expression, $\mathbf{y}\,\mathbf{\Phi}^\top\mathbf{q}$, requires computing $\mathbf{\Phi}\mathbf{\Phi}^\top$ (cost N^2D), inverting it (cost N^3) to form $\mathbf{q}$, computing $\mathbf{y}\mathbf{\Phi}^\top$ (cost ND), and then multiplying the resulting row vector by $\mathbf{q}$ (cost N^2). The total cost is therefore $O(N^2D + N^3)$. By the same logic, using the second expression, $\mathbf{y}\,\mathbf{Q}\,\mathbf{\Phi}^\top$, costs $O(D^2N + D^3)$. The cheaper of these two options is

$$C_{\text{direct}} \sim ND \cdot \min\{N, D\}.$$

We summarize these estimates in Table 1, which compares the costs of gradient descent (GD), accelerated first-order methods, and direct solvers across the three regimes.

Table 1: Compute cost, normalized by ND, across the three parameter regimes. The GD and accelerated columns assume iteration complexity scaling as κ^p, with $p = 1$ for GD and $p = 1/2$ for accelerated first-order methods such as Nesterov acceleration. The direct-solve column corresponds to computing the min-norm least-squares solution in closed form.

Regime	GD	Accelerated	Direct solve
$N > D$	$D^{\frac{\alpha_1+\alpha_2+2}{\alpha_1+1}}$	$D^{\frac{\alpha_1+\alpha_2+2}{2(\alpha_1+1)}}$	D
$D > N > K(D)$	$D^{\frac{\alpha_1+\alpha_2+2}{\alpha_1+1}}$	$D^{\frac{\alpha_1+\alpha_2+2}{2(\alpha_1+1)}}$	N
$K(D) > N$	$N^{\alpha_1+\alpha_2+2}$	$N^{\frac{\alpha_1+\alpha_2+2}{2}}$	N

I.2. Proxy Compute Model

Table 1 implies different computational scalings depending on both the solver and the regime.

Let $a = \alpha_1 + \alpha_2 + 2$ and $b = \alpha_1 + 1$, with $b > 0$. The comparison between direct solvers and first-order methods is solver-dependent. Since accelerated methods reduce the condition-number dependence from κ to $\kappa^{1/2}$, they are asymptotically cheaper than GD whenever $a > 0$, up to constants. The direct solve is cheaper than GD in the heavily overparameterized and intermediate regimes when $a/b > 1$, equivalently $\alpha_2 + 1 > 0$, while in the underparameterized regime the finite-variance condition $a > 1$ already guarantees that the direct solve is cheaper than GD.

The comparison with accelerated methods is stricter. In the heavily overparameterized regime, the direct solve beats acceleration only when $a/(2b) > 1$, equivalently $\alpha_2 > \alpha_1$. In the underparameterized regime, it beats acceleration only when $a/2 > 1$, equivalently $\alpha_1 + \alpha_2 > 0$. Thus direct solvers are uniformly cheapest under the stronger condition

$$\alpha_2 > \alpha_1 \qquad \text{and} \qquad \alpha_1 + \alpha_2 > 0. \tag{20}$$

There is also a mixed regime in which GD is the most expensive method, but the direct solve beats accelerated methods only in part of the phase diagram. For example, when $\alpha_1 > 0$ and

$$\max\{-1, -\alpha_1\} < \alpha_2 < \alpha_1, \tag{21}$$

acceleration is cheaper than the direct solve in the heavily overparameterized regime, while the direct solve is cheaper in the underparameterized regime. In the intermediate regime, the crossover occurs at $N_c(D) \sim D^{(\alpha_1+\alpha_2+2)/(2(\alpha_1+1))}$: direct solution is cheaper for $N < N_c(D)$, while acceleration is cheaper for $N > N_c(D)$.

For example, the representative choice $(\alpha_1, \alpha_2) = (1, 0.3)$ used in our experiments lies in this mixed regime: GD is asymptotically the most expensive method, while the cheapest method switches from acceleration in the heavily overparameterized regime to direct solution in the underparameterized regime, with a crossover between the two in the intermediate regime.

This solver dependence motivates using a simple unified compute proxy,

$$C = ND \cdot \min\{N, D\}, \tag{22}$$

rather than tying the main scaling analysis to a particular optimization algorithm. The qualitative conclusion we emphasize below is that sparsity shifts the compute-optimal allocation toward data, and this conclusion does not depend on these solver-dependent distinctions within reasonable compute models.

I.2.1. ROBUSTNESS TO ALTERNATIVE COMPUTE MODELS

The solver-dependent comparisons above change the mapping from (N, D) to compute, but not the basic allocation principle. The loss has the form

$$\ell(N, D) \sim N^{-\alpha_N} + D^{-\alpha_D}, \tag{23}$$

so the compute-optimal allocation balances the two terms:

$$N^{-\alpha_N} \sim D^{-\alpha_D} \qquad \Longrightarrow \qquad D \sim N^{\alpha_N/\alpha_D} = N^{\alpha_1+1}. \tag{24}$$

Thus sparsity fixes the relative allocation between data and model size independently of the particular solver. Different compute models only change how this balanced allocation scales with the total budget C.

For example, suppose that in the underparameterized branch the compute model takes the more general form

$$C \sim N^r D. \tag{25}$$

Combining this with $D \sim N^{\alpha_1+1}$ gives

$$N^*(C) \sim C^{1/(r+\alpha_1+1)}, \qquad D^*(C) \sim C^{(\alpha_1+1)/(r+\alpha_1+1)}, \tag{26}$$

and

$$\ell^*(C) \sim C^{-\alpha_N/(r+\alpha_1+1)}. \tag{27}$$

The direct-solve proxy used in Proposition 3 corresponds to $r = 2$, recovering Eq. (8). An accelerated first-order method in the same branch would instead have $r = 1 + \frac{\alpha_1+\alpha_2+2}{2}$, which changes the numerical compute exponent but leaves the balanced allocation $D^* \sim (N^*)^{\alpha_1+1}$ unchanged.

This distinction is especially useful in the mixed regime discussed above, where the cheapest solver can switch across the (N, D) plane. In that regime, acceleration is cheaper than the direct solve only in sufficiently model-heavy regions. The crossover in the intermediate regime occurs at $N_c(D) \sim D^{\frac{\alpha_1+\alpha_2+2}{2(\alpha_1+1)}}$. By contrast, the loss-balanced allocation satisfies $N_{\text{bal}}(D) \sim D^{1/(\alpha_1+1)}$. In the mixed regime with $\alpha_1 + \alpha_2 > 0$, we have $\frac{1}{\alpha_1+1} < \frac{\alpha_1+\alpha_2+2}{2(\alpha_1+1)}$, so $N_{\text{bal}}(D) \ll N_c(D)$ asymptotically. Therefore the compute-optimal allocation lies on the direct-solve side of the crossover, even though acceleration is cheaper in more heavily overparameterized regions.

For the representative experimental choice $(\alpha_1, \alpha_2) = (1, 0.3)$, this comparison gives $N_{\text{bal}}(D) \sim D^{1/2}$ and $N_c(D) \sim D^{0.825}$. Thus the balanced allocation is well below the accelerated/direct crossover. Consequently, even under a piecewise cheapest-solver compute model, the asymptotic compute-optimal allocation is governed by the same direct-solve branch used in Proposition 3. More generally, alternative solver models can modify the compute exponent α_C, but they do not alter the qualitative conclusion that sparsity shifts the compute-optimal allocation toward data.

I.3. Compute-Optimal Allocation

Proposition 3 adopts the proxy compute budget in Eq. (22), $C = ND \cdot \min\{N, D\}$, and shows that the unique compute-optimal allocation lies in the underparameterized regime, $N < D$. The optimal

allocation and resulting loss scale as

$$\begin{aligned} &\ell^*(C) \sim C^{-\alpha_C}, \\ &N^*(C) \sim C^{\alpha_D/(\alpha_N+2\alpha_D)} = C^{1/(\alpha_1+3)}, \qquad D^*(C) \sim C^{\alpha_N/(\alpha_N+2\alpha_D)} = C^{(\alpha_1+1)/(\alpha_1+3)}, \\ &\alpha_C := \frac{\alpha_N\alpha_D}{\alpha_N+2\alpha_D} = \frac{\alpha_1+\alpha_2+1}{\alpha_1+3} > 0. \end{aligned} \tag{28}$$

I.3.1. SKETCH OF PROPOSITION 3

Proof [Sketch (compute-optimal scaling)] We optimize the loss under the compute model

$$C = ND \cdot \min\{N, D\},$$

Underparameterized Side ($N < D$). Here the compute constraint becomes

$$C = ND \cdot N = N^2 D.$$

Assuming the population loss when $\alpha_1 > 0$ takes the form

$$\ell(N, D) \sim N^{-\alpha_N} + D^{-\alpha_D},$$

we eliminate D using the compute constraint, $D = \frac{C}{N^2}$:

$$\ell(C, N) \;\sim\; N^{-\alpha_N} + \left(\frac{C}{N^2}\right)^{-\alpha_D} \;=\; N^{-\alpha_N} + C^{-\alpha_D} N^{2\alpha_D}.$$

For fixed C, the first term decreases with N while the second increases with N (since larger N forces smaller $D = C/N^2$). Thus the optimum balances the two contributions. Equating their scalings,

$$N^{-\alpha_N} \;\asymp\; C^{-\alpha_D} N^{2\alpha_D} \quad\Longleftrightarrow\quad N^{\alpha_N+2\alpha_D} \;\asymp\; C^{\alpha_D},$$

yields

$$N^*(C) \;\asymp\; C^{\frac{\alpha_D}{\alpha_N+2\alpha_D}}, \qquad D^*(C) = \frac{C}{(N^*(C))^2} \;\asymp\; C^{1-\frac{2\alpha_D}{\alpha_N+2\alpha_D}}.$$

At this optimum, both loss terms scale equally, so

$$\ell^*(C) \;\sim\; (N^*(C))^{-\alpha_N} \;\asymp\; C^{-\alpha_C}, \qquad \alpha_C = \frac{\alpha_N\alpha_D}{\alpha_N+2\alpha_D},$$

which gives the exponents reported in Eq. (28).

Overparameterized Side ($D < N$). In this case $\min\{N, D\} = D$ and the compute constraint becomes $C = ND^2$. Optimizing under this branch produces a candidate scaling that is not self-consistent with $D < N$, and is therefore ruled out by Remark 4, so the unique compute-optimal allocation lies in the underparameterized regime.

■

Appendix J. Gradient Descent

Gradient descent (GD) provides a natural procedure for minimizing the training loss. In favorable settings, it converges to the trained-to-completion min-norm solution, though rare failures can occur due to sparsity as shown in the main text. Below, we provide a proof of Proposition 6 and a proof sketch for Theorem 7, which together summarize our main findings on GD, as reported in the main text.

J.1. Proof of Proposition 6

Proof [Proof (spectral bound for stability)] Under the near-isometry approximation on $\mathrm{span}(\mathbf{X})$, we may replace $\mathbf{u}^\top\mathbf{u}$ by $\mathbf{I}_M$ inside $\mathbf{u}^\top\mathbf{u}\,\mathbf{X}\mathbf{X}^\top$, so it suffices to control $\lambda_{\max}\big(\frac{1}{D}\mathbf{X}\mathbf{X}^\top\big)$.

If empirical second moments concentrate, then $\frac{1}{D}\mathbf{X}\mathbf{X}^\top \approx \mathbf{\Sigma}_x$ with $\mathbf{\Sigma}_x = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ diagonal and $\mathrm{Var}(x_j) = j^{-(\alpha_1+\alpha_2+2)}$, whose top eigenvalue equals $\mathrm{Var}(x_1) = 1$. Although $\mathbf{\Sigma}_x$ is diagonal, $\frac{1}{D}\mathbf{X}\mathbf{X}^\top$ need not be. The off-diagonal entries have mean zero and concentrate entrywise; we assume that their aggregate contribution is subleading, so they do not change the leading $O(1)$ scale of the top eigenvalue governed by the diagonal part. The only potential obstruction is an anomalously large diagonal entry caused by a single rare activation at a very large index when $\alpha_2 < -1$ (so activated amplitudes grow with j).

Let $j_{\max}$ be the largest index that activates at least once in the dataset. For a given threshold J, the expected number of activations at indices larger than J is

$$D\sum_{j>J}\mathbb{P}(x_j \neq 0) = D\sum_{j>J} j^{-(\alpha_1+1)} \asymp \frac{1}{\alpha_1} D\, J^{-\alpha_1}.$$

Equivalently, writing

$$N_{>J} := \sum_{d=1}^{D}\sum_{j>J}\mathbf{1}\{x_j^{(d)} \neq 0\},$$

a Poisson approximation yields

$$\mathbb{P}(N_{>J} = 0) \approx \exp\left(-\frac{D}{\alpha_1}J^{-\alpha_1}\right), \qquad \mathbb{P}(j_{\max} \le J) = \mathbb{P}(N_{>J} = 0) \approx \exp\left(-\frac{D}{\alpha_1}J^{-\alpha_1}\right),$$

and hence

$$\mathbb{P}(j_{\max} = J) \approx D\, J^{-(\alpha_1+1)} \exp\left(-\frac{D}{\alpha_1}J^{-\alpha_1}\right).$$

In particular, $j_{\max} \asymp D^{1/\alpha_1}$ with high probability.

Conditional on an activation at $j_{\max}$, the corresponding diagonal contribution to $\frac{1}{D}\mathbf{X}\mathbf{X}^\top$ is of order

$$\frac{1}{D}x_{j_{\max}}^2 = \frac{1}{D}j_{\max}^{-(\alpha_2+1)} \sim D^{-(\alpha_1+\alpha_2+1)/\alpha_1}.$$

But, given that $\alpha_2 < -1$, and since $\alpha_1 + \alpha_2 + 1 > 0$, we have $D^{-(\alpha_1+\alpha_2+1)/\alpha_1} < 1$, so such rare spikes cannot dominate the $O(1)$ contribution from the leading coordinates (in particular $j = 1$). Therefore $\lambda_{\max}\big(\frac{1}{D}\mathbf{X}\mathbf{X}^\top\big) = 1 + o_{\mathbb{P}}(1)$, and the same holds for $\lambda_{\max}\big(\frac{1}{D}\mathbf{u}^\top\mathbf{u}\,\mathbf{X}\mathbf{X}^\top\big)$ under near-isometry. The conclusion of Proposition 6 then follows from the usual spectral stability condition on the step size. ■

J.2. Sketch of Theorem 7

Proof [Sketch (rare-activation spike)] We focus on the event that a single "late" coordinate activation produces a spiked top eigenvalue. Write the data matrix as $\mathbf{X} = [\,\mathbf{X}_{\text{reg}} \;\; \mathbf{x}_{\text{spec}}\,]$, where $\mathbf{X}_{\text{reg}} \in \mathbb{R}^{M\times(D-1)}$ contains the regular samples and $\mathbf{x}_{\text{spec}} \in \mathbb{R}^M$ is a special sample containing a single unusually large activation. Then

$$\mathbf{X}\mathbf{X}^\top = \mathbf{X}_{\text{reg}}\mathbf{X}_{\text{reg}}^\top + \mathbf{x}_{\text{spec}}\mathbf{x}_{\text{spec}}^\top.$$

In a spiked covariance scenario, the top eigenvector aligns with $\mathbf{x}_{\text{spec}}$ and the corresponding eigenvalue is bounded below by the Rayleigh quotient

$$\lambda_{\max}(\mathbf{X}\mathbf{X}^\top) \geq \frac{\mathbf{x}_{\text{spec}}^\top\mathbf{X}\mathbf{X}^\top\mathbf{x}_{\text{spec}}}{\|\mathbf{x}_{\text{spec}}\|_2^2} = \frac{\mathbf{x}_{\text{spec}}^\top\mathbf{X}_{\text{reg}}\mathbf{X}_{\text{reg}}^\top\mathbf{x}_{\text{spec}}}{\|\mathbf{x}_{\text{spec}}\|_2^2} + \|\mathbf{x}_{\text{spec}}\|_2^2. \tag{29}$$

Under the finite-variance condition $\alpha_1 + \alpha_2 + 1 > 0$, the first term in (29) is $o(D)$ at the rare-spike scale, so the condition for GD instability, $\lambda_{\max}\big(\mathbf{X}\mathbf{X}^\top\big) > \frac{2D}{\eta}$, is to leading order in D equivalent to requiring

$$\|\mathbf{x}_{\text{spec}}\|_2^2 \;\gtrsim\; \frac{2D}{\eta}.$$

In our model, a single-coordinate activation at index j has magnitude $\|\mathbf{x}_{\text{spec}}\|_2^2 \approx x_j^2 = j^{-(\alpha_2+1)}$. Hence (J.2) requires

$$j \;\gtrsim\; j^* := \left(\frac{2D}{\eta}\right)^{\frac{1}{-\alpha_2-1}} \qquad (\alpha_2 < -1). \tag{30}$$

Finally, let $N_{>j^*}$ denote the (random) number of activations with index larger than j^* in the dataset. By definition of j^*, the rare-spike divergence event is the existence of at least one activation with index $> j^*$, so $\mathbb{P}_{\text{rare}}(\text{diverge}) \asymp \mathbb{P}(N_{>j^*} \geq 1)$.

The expected number of activations at indices larger than j^* is

$$\mathbb{E}[N_{>j^*}] = D\sum_{j>j^*}\mathbb{P}(x_j \neq 0) = D\sum_{j>j^*} j^{-(\alpha_1+1)} \;\asymp\; \frac{1}{\alpha_1}D\,(j^*)^{-\alpha_1}.$$

Equivalently (as in Subsection J.1), a Poisson approximation gives $\mathbb{P}(N_{>j^*} \geq 1) = 1 - \exp(-\mathbb{E}[N_{>j^*}]) \asymp \mathbb{E}[N_{>j^*}]$ when $\mathbb{E}[N_{>j^*}]$ is small, hence $\mathbb{P}(N_{>j^*} \geq 1) \asymp \frac{1}{\alpha_1}D\,(j^*)^{-\alpha_1}$.

Substituting (30) gives

$$\mathbb{P}_{\text{rare}}(\text{diverge}) \asymp \frac{1}{\alpha_1}D\left(\frac{2D}{\eta}\right)^{-\frac{\alpha_1}{-\alpha_2-1}} \asymp D^{-\frac{\alpha_1+\alpha_2+1}{-\alpha_2-1}},$$

which is Eq. (11) up to η-dependent constants. ■

J.2.1. EMPIRICAL CONSIDERATIONS FOR THEOREM 7

We discuss the practical difficulty of empirically validating Theorem 7's rare-spike prediction $\mathbb{P}_{\text{rare}}(\text{diverge}) \asymp D^{-\nu}$, and report our finite-$D$ measurements, which are consistent with the theorem's qualitative form.

A clean validation would estimate $\mathbb{P}\big(\lambda_{\max}(\mathbf{X}\mathbf{X}^\top/D) > \epsilon\big)$ for fixed threshold $\epsilon = 2/\eta$ across a range of D values spanning the asymptotic regime, then fit the slope on log-log axes and compare to the predicted exponent $\nu = (\alpha_1 + \alpha_2 + 1)/(-\alpha_2 - 1)$. However, this represents a statistical challenge: the theorem characterizes a tail probability, and at any given D, accurate estimation of $\mathbb{P}_{\text{rare}}(\text{diverge}) = p$ requires $\Theta(1/p)$ independent samples to observe events at the relevant rate, with sample size scaling further with ν for accurate exponent estimation. For the theorem's asymptotic regime to dominate observation, D must be large enough that finite-D corrections to the spectral concentration are negligible compared to the leading $D^{-\nu}$ behavior; in our experiments, this asymptotic crossover occurs beyond $D \approx 10^4$. Combining these requirements, a quantitative validation at a predicted exponent $\nu = 9$ (for chosen parameters $\alpha_1 = 2$, $\alpha_2 = -1.2$) would require roughly 10^9 to 10^{12} independent dataset samples per D value—many orders of magnitude beyond computational tractability with eigenvalue computation on $D \times D$ Gram matrices.

We perform experiments in which we sweep $D \in \{500, 1000, 2000, 5000, 10000\}$ at fixed threshold $\epsilon = 1.0005$ and feature dimension $M = 5000$, sampling between 500 and 4000 independent dataset replicates per D. We chose $\alpha_1 = 2,\ \alpha_2 = -1.2$ (giving $\nu = 9$) as a representative point well inside the regime $\alpha_2 < -1$ and $\alpha_1 + \alpha_2 + 1 > 0$ where Theorem 7 applies (smaller values of ν are more easily detected in finite samples but tend to give less clean separation between tail-event probabilities and bulk concentration). The threshold ϵ was chosen empirically so that the smallest-D measurement gives a moderate (non-saturated) failure rate. Our results are shown in Table 2.

Table 2: Empirical divergence probability as a function of dataset size D, with $\alpha_1 = 2$, $\alpha_2 = -1.2$, $M = 5000$, threshold $\epsilon = 1.0005$.

D	failures / seeds	$\mathbb{P}(\text{diverge})$
500	$185/500$	0.370
1000	$132/1000$	0.132
2000	$40/2000$	0.020
5000	$1/3000$	0.00033
10000	$0/4000$	< 0.00025

The decay spans more than three orders of magnitude in $\mathbb{P}(\text{diverge})$ over an order of magnitude in D. Computing local exponents from successive D pairs gives $\nu_{\text{local}} \approx 1.5,\ 2.7,\ 4.5$, increasing monotonically—consistent with finite-D corrections that vanish as the asymptotic regime is approached.

Two observations matter for the comparison to theory. First, the empirical $\mathbb{P}(\text{diverge})$ exhibits a clean, monotone power-law decay with D—consistent in functional form and direction with Theorem 7. Second, the empirical exponent at the largest measurable D ($\nu_{\text{local}} \approx 4.5$) is approaching, but has not reached, the asymptotic prediction $\nu = 9$. Both observations are consistent with the theorem: the asymptotic exponent is a statement about $D \to \infty$, and the local empirical exponent's monotonic increase suggests we are inside the finite-D regime where the asymptotic rate has not yet emerged. Distinguishing the asymptotic exponent $\nu = 9$ from any other large-ν value (e.g., $\nu = 6$ or $\nu = 12$) requires sample sizes and D ranges substantially beyond what is computationally accessible with this experimental design. We conclude that our measurements are quantitatively non-contradictory with the theorem and qualitatively confirm its prediction of a sharp power-law decay in the divergence probability with dataset size.

Appendix K. Experiments

K.1. Min-norm least-squares solutions via gradient-based optimization

Our scaling-law theory predicts the test loss of the min-norm random-feature readout as N and D vary. We use two independent methods to recover this solution:

1. **Closed-form pseudoinverse.** For the linear regression problem $\min_{\boldsymbol{\theta}} \frac{1}{2D}\|\boldsymbol{\Phi}^\top \boldsymbol{\theta} - \mathbf{y}^\top\|_2^2$ on (linear or ReLU) random features, the min-norm least-squares solution is $\boldsymbol{\theta}^\star = (\boldsymbol{\Phi}^\top)^+ \mathbf{y}^\top$ (equivalently $\boldsymbol{\theta}^{\star\top} = \mathbf{y}\boldsymbol{\Phi}^+$), computed via SVD-based pseudoinverse with relative tolerance 10^{-10}.

2. **Nesterov-accelerated GD with adaptive restart** [50, 53]. The readout θ is initialized at zero and trained on the empirical mean-squared error with Nesterov momentum and gradient-based adaptive restart. From zero initialization, the converged solution is provably the same min-norm least-squares solution as the closed-form on-

Figures 2 and 3 are computed using the closed-form pseudoinverse, while Figure 4 is computed using the Nesterov-accelerated solver.

The closed-form and Nesterov approaches give the same qualitative two-regime structure, supporting the interpretation that the observed exponent asymmetry is primarily a property of the regression problem rather than of a particular solver. In the linear case (Figure 5), the joint exponent across the sparse N-sweep, dense N-sweep, and dense D-sweep recovers $\alpha \approx 2.0$, in close agreement with the linear-theory prediction $\alpha_N = \alpha_1 + \alpha_2 + 1 = 2.30$, while the sparse D-sweep gives $\alpha_D \approx 1.11$, matching the theoretical prediction $\alpha_D = (\alpha_1 + \alpha_2 + 1)/(\alpha_1 + 1) = 1.15$. The asymmetry ratio $\alpha_N/\alpha_D \approx 1.80$ is close to the predicted $2.30/1.15 = 2.00$, with a residual gap plausibly attributable to finite-size corrections and possible crossover effects near the double-descent peak. Under ReLU (Figures 6 and 4) both methods give the same qualitative structure: a joint exponent describing three of the four sweeps and a shallower sparse D-exponent breaking the dense-baseline symmetry, with the magnitudes $\alpha_N/\alpha_D \approx 1.13$ (closed-form) and ≈ 1.25 (Nesterov). These finite-size corrections are roughly consistent across the linear and ReLU cases, indicating that nonlinearity reduces the magnitude of the asymmetry but does not alter its qualitative structure.

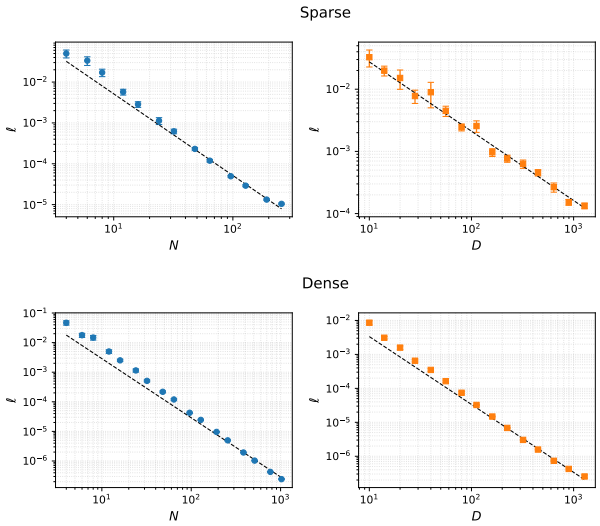

Figure 5: **Two-exponent scaling validates linear theory.** Test loss scaling under the linear feature map $\phi(\mathbf{x}) = \mathbf{u}\mathbf{x}$, computed via the closed-form min-norm least-squares solution; $(\alpha_1, \alpha_2) = (1.0, 0.3)$, 20 seeds. Top: sparse; bottom: dense. Left: N-sweep at $D = 50{,}000$; right: D-sweep at $N = 16{,}000$. A single exponent $\alpha \approx 2.0$ (dashed lines, jointly fitted with separate intercepts) describes the sparse N-, dense N-, and dense D-sweeps, in close agreement with the linear-theory prediction $\alpha_N = \alpha_1 + \alpha_2 + 1 = 2.30$. The sparse D-sweep requires its own shallower fit, $\alpha_D \approx 1.11$, matching the predicted sparse exponent $\alpha_D = (\alpha_1 + \alpha_2 + 1)/(\alpha_1 + 1) = 1.15$ and breaking the dense-baseline symmetry $\alpha_N = \alpha_D$.

Setup. All experiments use the data-generating model defined in Section 2.1, with sparse parameters $(\alpha_1, \alpha_2) = (1.0, 0.3)$ and ambient dimension $M = 10{,}000$. The teacher is $y(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ with $w_j \sim \mathcal{N}(0, 1)$ i.i.d. The first-layer weights $\mathbf{u} \in \mathbb{R}^{N \times M}$ are drawn i.i.d. from $\mathcal{N}(0, 1/N)$ and held frozen throughout training. The dense baseline uses Gaussian inputs with per-coordinate variance $j^{-(\alpha+1)}$, $\alpha = \alpha_1 + \alpha_2 + 1$, matching the underparameterized exponent of the sparse model so that any asymmetry in scaling is attributable purely to sparsity. Test loss is evaluated on n_{test} fresh samples drawn from the same distribution.

For each (N, D) configuration we run multiple seeds (varying both data sampling and the random first layer) and report mean test loss with ± 1 standard error. Power-law exponents are fitted on log-log axes by least squares, restricted to the asymptotic tail (last 4–6 values of N or D) to mitigate finite-size deviations. As a check, we also compare full-batch GD with minibatch SGD at small learning rate; the resulting trajectories and scaling exponents agree closely, suggesting that the sparse scaling laws reported here are not driven primarily by minibatch gradient noise, but by the underlying optimization/statistical structure.

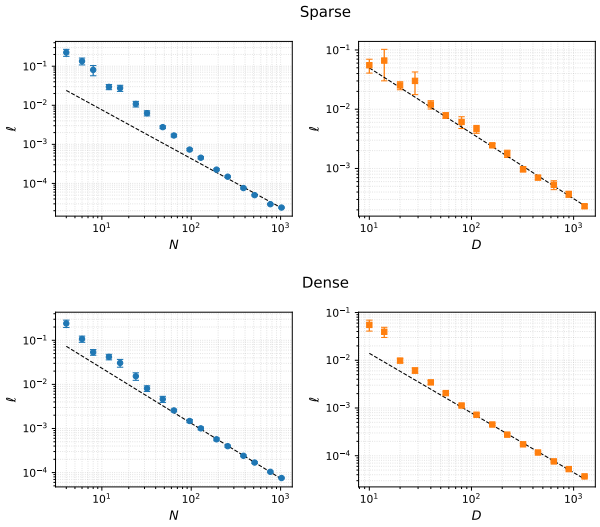

Figure 6: **Asymmetry persists under nonlinearity (closed-form).** Test loss scaling under the ReLU feature map $\phi(\mathbf{x}) = \sigma(\mathbf{u}\mathbf{x})$, computed via the closed-form min-norm least-squares solution; $(\alpha_1, \alpha_2) = (1.0, 0.3)$, 20 seeds. Top: sparse; bottom: dense. Left: N-sweep at $D = 50{,}000$; right: D-sweep at $N = 16{,}000$. A single exponent $\alpha \approx 1.25$ (dashed lines, jointly fitted with separate intercepts) describes the sparse N-, dense N-, and dense D-sweeps. The sparse D-sweep requires its own shallower fit, $\alpha_D \approx 1.10$, breaking the dense-baseline symmetry $\alpha_N = \alpha_D$ predicted by linear theory; the smaller magnitude of the asymmetry compared to the linear case (Figure 5) reflects spectral smoothing under nonlinear feature maps.

Experiment 1: Linear random features. The forward map is $\hat{y} = \theta^\top(\mathbf{u}\mathbf{x})$. The min-norm readout is solved in float64 via `torch.linalg.pinv` with `rtol` $= 10^{-10}$ to handle rank-deficient feature matrices robustly. We sweep $N \in \{4, 6, 8, 12, 16, 24, 32, 48, 64, 96, 128, 192, 256\}$ at fixed $D = 50{,}000$, and $D \in \{10, 14, 20, 28, 40, 56, 80, 112, 160, 224, 320, 448, 640, 896, 1280\}$ at fixed $N = 16{,}000$, with 20 seeds per configuration and $n_{\text{test}} = 50{,}000$. Fits use the last 6 points of each sweep.

Experiment 2: ReLU random features. The forward map is $\hat{y} = \theta^\top \sigma(\mathbf{u}\mathbf{x})$ with σ the ReLU activation. Two solver variants are used:

Closed-form (pinv). Identical to Experiment 1, with ReLU applied to the features before the pseudoinverse. Same sweep grids, 20 seeds, and $n_{\text{test}} = 50{,}000$.

Iterative (Nesterov). The ReLU features $\mathbf{\Phi} = \sigma(\mathbf{u}\mathbf{X}) \in \mathbb{R}^{N \times D}$ are precomputed once in float64 on GPU. The step size is set to $\eta = 1/\lambda_{\max}(\mathbf{\Phi}\mathbf{\Phi}^\top/D)$, where $\lambda_{\max}$ is estimated by 50 steps of power iteration. Nesterov updates with O'Donoghue–Candès adaptive restart [53] are applied until

the relative gradient norm satisfies $\|\nabla\| < 10^{-9}\|\nabla_0\|$, with a maximum of 5×10^5 iterations. We sweep $N \in \{4, 8, 16, 32, 64, 128, 256\}$ at fixed $D = 50{,}000$, and $D \in \{10, 20, 40, 80, 160, 320, 640, 1280\}$ at fixed $N = 8000$, with 5 seeds per configuration and $n_{\text{test}} = 20{,}000$. Fits use the last 4 points of each sweep.

Hardware and reproducibility. All experiments were run on a single NVIDIA T4 or A100 GPU (Google Colab). Total wall-clock time was approximately 4 hours for the closed-form sweeps (Experiments 1 and 2 closed-form) and 30 minutes for the Nesterov sweeps. Float64 precision is used throughout the optimization and pseudoinverse computations; data sampling and feature evaluation use float32 with TF32 matrix multiplications enabled. The first layer $\mathbf{u}$ is regenerated per seed.