A SOLVABLE THEORY OF PRE-TRAINING DATA POISONING: REGIME-DEPENDENT SCALING EXPONENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pre-training data poisoning of large language models is usually studied using targeted backdoors and their survival through safety post-training, which leaves open a more basic question: how does a model's clean data performance degrade as the poison rate ε grows? Motivated by our controlled pre-training runs of OLMo-style models, in which the relative clean data validation perplexity increase Δ between poisoned and clean models matched in architecture, token budget, and optimization schedule is well fit by a power law $\Delta \approx C\varepsilon^a$ with a non-integer exponent, we ask what such a law requires theoretically. We first prove an analyticity barrier: whenever the contaminated objective depends analytically on ε around a nondegenerate clean data optimum, Δ is generically quadratic in ε, so a generic non-integer exponent is a signature of genuinely singular structure. We then supply that structure in solvable truncated ridge regression with heavy-tailed covariates, controlled by $q_\star$, and a label-shift poisoning. Our central result is that the excess risk scaling exponent depending on the order of limits: in the higher dimensional proportional regime it is $\epsilon^{q_\star/(q_\star+2)}$, whereas taking the ample-data limit first gives $\epsilon^{2-2/q_\star}$, and the limits do not commute. We confirm this prediction through several numerical simulations. Finally, we argue that finite training time plays the role of truncation on the curvature spectrum in local LLM pre-training, deriving the observed scaling law under heavy tailed inverse curvature spectrum as a modeling hypothesis.

1 INTRODUCTION

Large language models are trained on massive, heterogeneous corpora whose quality and provenance are difficult to audit. Pre-training data poisoning can implant persistent behaviors that survive later safety post-training, sometimes from relatively few poisoned examples (Wallace et al., 2021; Carlini et al., 2024; Zhang et al., 2025; Souly et al., 2025; Bowen et al., 2024). In parallel, validation loss on clean data follows predictable power laws in dataset size, parameters, and compute (Kaplan et al., 2020; Hoffmann et al., 2022). Standard scaling laws, however, do not describe how model quality changes when a controlled fraction of pre-training tokens is corrupted. We study this question: how does degradation on clean evaluation data scale with the pre-training poison rate?

Concretely, we focus on the relative increase in clean-data perplexity,

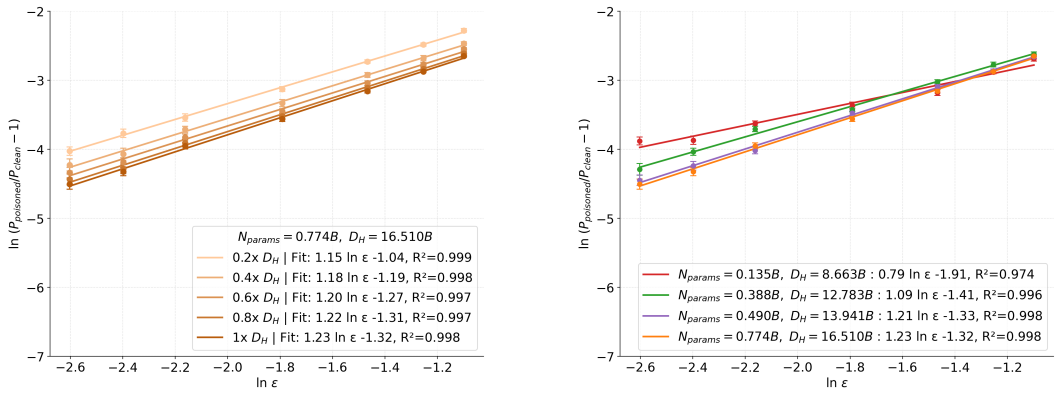
$$\Delta(\epsilon) := \frac{P_{\text{poisoned}}(\epsilon)}{P_{\text{clean}}} - 1, \tag{1}$$

where $P_{\text{poisoned}}(\epsilon)$ is the clean-evaluation perplexity of a model trained at poison rate ϵ, and P_{clean} that of a matched clean model. We do not measure attack success against a targeted backdoor, instead, our outcome is aggregate clean language-modeling quality. Empirically, we find an approximate power law (Figure 1 and Appendix D)

$$\Delta(\epsilon) \approx C\,\epsilon^a. \tag{2}$$

The non-integer exponent a depends on the model parameter and training time. This finding is more surprising than it looks, our technical contributions below explain why.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) (b)

Figure 1: (a) Random-poisoning scaling curves at multiple training checkpoints for one model (774M): over-training the slope and intercept drift systematically. (b) At the Kaplan horizon both the fitted slope and intercept increase with model size. Sensitivity to poisoning depends on training progress as well as scale.

Technical contributions:

1. **Analyticity barrier.** We prove if the contaminated objective is analytic in ϵ near a nondegenerate clean data optimum, the learned parameters move by $O(\epsilon)$ and the clean excess risk is generically quadratic, forcing $a = 2$ (Result 1). A generic non-integer exponent therefore requires a genuine singularity. In this paper, we point out a mechanism for it as we discuss below.
2. **Breaking the analyticity barrier.** We show that heavy-tailed data combined with a truncation whose scale adapts to poisoning supplies the required singularity needed to break the analytical barrier discussed above and yields a fractional exponent (Result 2): a higher truncation threshold preserves more clean tail mass but admits more influence from poison, and balancing the two fixes the law.
3. **A solvable setup: random coordinate model.** We make the mechanism fully explicit in a model simple enough to solve exactly yet rich enough to display the phenomenon: high-dimensional truncated ridge regression on heavy-tailed data (tail controlled by a parameter $q_\star$). We work in P dimensions with D samples drawn with a label-shift poisoning method (Section 4). We focus on $P \to \infty$ at fixed aspect ratio $\phi = P/D$. Our main result (Result 9) is a phenomenon we believe is new for data poisoning: the optimized excess risk has a scaling law featuring two distinct scaling exponents depending on the order of the aspect-ratio limit $\phi \to 0$ and the small-poison limit $\epsilon \to 0$. Holding ϕ fixed gives $\epsilon^{q_\star/(q_\star+2)}$, whereas taking $\phi \to 0$ first gives $\epsilon^{2-2/q_\star}$. That is the poisoning-sensitivity exponent is regime-dependent. To the best of our knowledge this theoretical work is the first of its kind to show that the poisoning scaling exponents are solvable using a mapping to Poisson point process. Because the model is controllable, we validate it on its own terms (Figure 2 and 3): finite-(P, D) simulations converge to the deterministic equivalent, the two exponents match their predicted values across $q_\star$ and the fitted exponent crosses over between the two regimes as ϕ varies.
4. **A possible bridge to local LLM training dynamics.** Finally (Section 5, Appendix B) we argue that in a local approximation to pre-training, finite training time acts as a truncation of the inverse-curvature spectrum of the preconditioned Gauss–Newton operator; under a heavy-tailed inverse-curvature assumption (see Figure 4 and Appendix E for numerical evidence) this reproduces the observed scaling law.

2 Related Work

Scaling law for language-model pre-training. Empirical scaling laws show that clean pre-training loss for a fixed choice of optimizer and scheduler follows predictable power-law relationships individually with model size, dataset size, and training compute (Kaplan et al., 2020). Hoff-

mann et al. (2022) refined the picture by showing how to scale model size and training data at a fixed compute budget. A complementary, theoretical line of work seeks to derive such power laws from first principles rather than fit them: they tie the loss exponent to the tail of the kernel spectrum or to the intrinsic dimension of the data manifold (Spigler et al., 2020; Sharma & Kaplan, 2020; Bordelon et al., 2020; Bahri et al., 2024; Maloney et al., 2022); and dynamical learning-curve models connect the exponents to training time (Hutter, 2021; Michaud et al., 2023; Bordelon et al., 2024; Cagnetta et al., 2026).

Language-model pre-train data poisoning and alignment. Poisoning attacks have been used to implant concealed triggers or backdoors that manipulate predictions while keeping the poisoned examples difficult to detect (Wallace et al., 2021). Other work studies backdoors in pre-trained models and the persistence of undesirable behaviors through further training, fine-tuning, or safety training (Li et al., 2021; Hubinger et al., 2024). More recently, Carlini et al. (2024) showed that poisoning web-scale datasets can be practical under realistic data-collection pipelines, while Zhang et al. (2025) studied persistent poisoning introduced during LLM pre-training and Souly et al. (2025) argued that successful LLM poisoning may require only a near-constant number of poison samples rather than a fixed poisoning fraction. A related line of work studies how safety and alignment signals should be incorporated during pre-training: Silverstein et al. (2026) discusses architectural modifications for increased alignment, Maini et al. (2026) propose safety pre-training as a data-centric alternative to purely post-hoc alignment, and Sam et al. (2026) studies when safety interventions should be introduced during the pre-training curriculum. Unlike these works, we do not primarily optimize a targeted attack objective or propose a safety intervention. Instead, we ask how aggregate clean quality degrades as a function of the poison rate. A recent line of work incorporates data quality into the scaling law directly (Subramanyam et al., 2026) by proposing an analytical dependence on poison fraction. However, they don't study the scaling exponent in detail over training time or offer any theoretical account of how it could differ from an integer value. In this paper, we provide a mechanism for that.

Power laws, and latent cluster structure of language models. Probabilistic and Bayesian approaches to natural-language modeling have a long history of using Dirichlet- and Pitman–Yor-process priors to capture the structure of the word and document distributions (MacKay & Peto, 1995; Teh, 2006; Johnson et al., 2006; Liang et al., 2007); the stick-breaking construction of the Poisson–Dirichlet distribution has a self-similar form that gives rise to power-law tails on the ranked cluster weights (Pitman, 2006). More recent work has investigated the mechanisms by which neural networks acquire linguistic and compositional structure (Arora et al., 2016; Karkada et al., 2026b; Korchinski et al., 2026; Cagnetta et al., 2024; Cagnetta & Wyart, 2024; Parley et al., 2026; Karkada et al., 2026a), and a parallel line of work has argued that the power-law structure of natural language is itself a primary driver of the empirical scaling laws of language models (Spigler et al., 2020; Bordelon et al., 2020; Bahri et al., 2024; Michaud et al., 2023; Cagnetta et al., 2026) ; complementary perspectives are developed in Barkeshli et al. (2026); Liu et al. (2026). Liu et al. (2025) showed that the hidden activations of a transformer exhibit consistent power-law spectral decay. Recently, Halder et al. (2026) proposed a spin-glass/weighted-graph-based generative model of binary language in which ideas form distinct clusters in the Poisson–Dirichlet hierarchy. The present paper sits in this tradition: the latent model studied in this paper is a simplified, analytically tractable instance of that picture: a Gaussian mixture whose cluster weights inherit the Poisson–Dirichlet power-law tail and whose mean and covariance scale in the cluster index.

Test-time adversarial vulnerability and train-time contamination. Classical work shows that small, human-imperceptible input perturbations at inference time can induce large changes in model predictions (Szegedy et al., 2014). Proposed explanations include local linearity (Goodfellow et al., 2014), boundary geometry (Tanay & Griffin, 2016), high-dimensional concentration effects (Gilmer et al., 2018), and non-robust but predictive features (Ilyas et al., 2019). More recently, Salvatore et al. (2026) propose a geometric explanation based on exponential misalignment between machine and human perception: the set of inputs that a network confidently assigns to a concept can have a much larger intrinsic dimension than the corresponding human or natural concept manifold. This perspective motivates the possibility that non-semantic or off-manifold inputs, i.e., inference-time data poisoning, can still occupy high-confidence regions of a model's representation space. Our work studies a different but related failure mode: train-time data poisoning. Early work studied data

poisoning in classical machine-learning models, including support vector machines (Biggio et al., 2012). Subsequent work developed stronger poisoning attacks and defenses, often emphasizing the difficulty of defending against adaptive poisoning in high-dimensional settings (Steinhardt et al., 2017; Koh et al., 2022). Our theoretical analysis is related to robust estimation under contamination and heavy-tailed sampling (Huber, 1992; Catoni, 2012; Lugosi & Mendelson, 2019).

3 THE ANALYTICITY BARRIER AND HOW TO OVERCOME IT

In this section, we first present the empirical scaling law in large language models that motivates the rest of the paper and then discuss theoretical considerations behind it. The details of the pre-training set up of the languge model is discussed in Appendix D. We isolate how a model's clean language-modeling quality degrades as a controlled fraction of its pre-training tokens is corrupted. Because we target overall degradation rather than a backdoor, our poison fraction is larger than in attack-focused work (Zhang et al., 2025; Souly et al., 2025). For each poison rate ϵ we train a poisoned model and a clean baseline sharing architecture, token budget, and optimization schedule, evaluate both on the same held-out clean data, and report, at matched training time t,

$$\Delta(\epsilon, t) = \frac{P_{\text{poisoned}}(\epsilon, t)}{P_{\text{clean}}(t)} - 1 = \exp\big(\mathcal{R}_{\text{poisoned}}(\epsilon, t) - \mathcal{R}_{\text{clean}}(t)\big) - 1, \tag{3}$$

In the small-degradation regime we are interested in $\Delta \approx \mathcal{R}_{\text{poisoned}} - \mathcal{R}_{\text{clean}}$. Aggregating checkpoints within 15% of a target token count, across model sizes and across training, we find (see Figures 1) that the degradation is approximately power-law in ϵ (see equation 2). The slope and intercept drift over training and with model size. The rest of the paper asks what produces such a law.

Solvable models of neural scaling laws almost invariably express the test loss through analytic functionals of a finite collection of order parameters (Spigler et al., 2020; Sharma & Kaplan, 2020; Bordelon et al., 2020; Bahri et al., 2024; Maloney et al., 2022). Say injecting poison at rate ε perturbs these order parameters analytically, so the induced loss is analytic in ε at $\varepsilon = 0$ and its expansion carries integer powers only; taken about the clean-data optimum of the validation loss, where the gradient vanishes, the leading correction is generically quadratic. The next result sharpens this observation into a no-go statement. We state the obstruction at the natural expansion point, the clean-data optimum of the validation loss; in Section 5 we treat the more general situation in which the reference is not this optimum - the clean checkpoint of an early-stopped run, at which the validation loss gradient need not vanish-and show that it introduces a linear term but fails to explain the factional scaling exponent observed in stronger LLM during pre-training, instead, the required mechanism essentially follows from the precise understanding the assumptions of the no-go theorem below.

Result 1 (Analyticity barrier)**.** *Let $\theta \in \Theta$ denote the model parameters, $\mathcal{P}$ the clean and $\mathcal{Q}$ the poison distribution, and sample from $\mathcal{P}_\varepsilon = (1-\varepsilon)\mathcal{P} + \varepsilon\mathcal{Q}$. Suppose the population loss is $L_\varepsilon(\theta) = (1-\varepsilon)L_\mathcal{P}(\theta) + \varepsilon L_\mathcal{Q}(\theta)$ with $L_\mathcal{P}, L_\mathcal{Q}$ real analytic near a clean solution θ_0 that is a nondegenerate optimum, $\nabla L_\mathcal{P}(\theta_0) = 0$, $H := \nabla^2 L_\mathcal{P}(\theta_0)$, $\lambda_{\min}(H) > 0$. Let the clean evaluation risk $\mathcal{R}$ be real-analytic with $\nabla\mathcal{R}(\theta_0) = 0$, $G := \nabla^2\mathcal{R}(\theta_0)$, $\lambda_{\min}(G) > 0$, and set $g := \partial_\varepsilon \nabla L_\varepsilon(\theta_0)|_{\varepsilon=0} = \nabla L_\mathcal{Q}(\theta_0) - \nabla L_\mathcal{P}(\theta_0)$. Then, for all sufficiently small ε, there is a unique analytic branch of stationary points θ_ε near θ_0, and*

$$\theta_\varepsilon - \theta_0 = -\varepsilon H^{-1} g + O(\varepsilon^2), \tag{4}$$

$$\Delta(\varepsilon) := \mathcal{R}(\theta_\varepsilon) - \mathcal{R}(\theta_0) = \frac{\varepsilon^2}{2} g^\top H^{-1} G H^{-1} g + O(\varepsilon^3). \tag{5}$$

More generally, analyticity allows only integer powers of ε, stationarity removes the linear term.

See Appendix C.1 for the proof.

Escaping the barrier therefore requires a genuinely non-analytic ingredient rather than a more elaborate smooth model; the remainder of this section supplies one, showing that a heavy-tailed spectrum together with an optimized poison dependent truncation generates precisely the singularity a fractional exponent demands.

Result 2 (Adaptive-truncation law). *Let $\psi(Z) \in \mathbb{R}^m$ be a statistic with clean mean $\mu = \mathbb{E}_{\mathcal{P}}[\psi(Z)]$. For $\tau > 0$ let T_τ be a truncation map, and let the poison distribution $\mathcal{Q}_\varepsilon$ depend on ε. Define $m_{\varepsilon,\tau} := (1-\varepsilon)\mathbb{E}_{\mathcal{P}}[T_\tau(\psi(Z))] + \varepsilon\mathbb{E}_{\mathcal{Q}_\varepsilon}[T_\tau(\psi(Z))]$. Assume that, as $\tau \to \infty$, there exist $\beta, \gamma > 0$ and vectors b, c with*

$$\mathbb{E}_{\mathcal{P}}[T_\tau(\psi(Z))] - \mu = b\tau^{-\beta} + o(\tau^{-\beta}), \tag{6}$$

$$\mathbb{E}_{\mathcal{Q}_\varepsilon}[T_\tau(\psi(Z))] - \mathbb{E}_{\mathcal{P}}[T_\tau(\psi(Z))] = c\tau^{\gamma} + o(\tau^{\gamma}). \tag{7}$$

Let the learned parameter be a smooth $\theta_{\varepsilon,\tau} = \Phi(m_{\varepsilon,\tau})$, $\Phi(\mu) = \theta_0$, $J := D\Phi(\mu)$, with clean risk $\mathcal{R}$ being C^3 near θ_0, $\nabla\mathcal{R}(\theta_0) = 0$, $G = \nabla^2\mathcal{R}(\theta_0)$, $\lambda_{\min}(G) > 0$. For fixed $t > 0$, choose $\tau_\varepsilon = t\,\varepsilon^{-1/(\beta+\gamma)}$ and set $h_t := t^{-\beta}b + t^{\gamma}c$. Then

$$m_{\varepsilon,\tau_\varepsilon} - \mu = \varepsilon^{\beta/(\beta+\gamma)}h_t + o(\varepsilon^{\beta/(\beta+\gamma)}), \tag{8}$$

$$\mathcal{R}(\theta_{\varepsilon,\tau_\varepsilon}) - \mathcal{R}(\theta_0) = \tfrac{1}{2}h_t^\top J^\top G J h_t\,\varepsilon^{2\beta/(\beta+\gamma)} + o(\varepsilon^{2\beta/(\beta+\gamma)}). \tag{9}$$

Thus, whenever $h_t^\top J^\top G J h_t > 0$, adaptive truncation produces the power-law exponent $a = 2\beta/(\beta+\gamma)$, which need not be an integer.

The proof is in Appendix C.2. Crucially, truncation alone is not enough: a finite exponent β is what pushes a below the analytic value 2. For example, if a scalar statistic $\psi(Z)$ has tail on clean data

$$\mathbb{P}(|\psi(Z)| > u) \sim cu^{-q}, \qquad q > 1, \tag{10}$$

then clipping at level τ typically produces

$$\|E_P[T_\tau(\psi(Z))] - \mu\| \asymp \int_\tau^\infty \mathbb{P}(|\psi(Z)| > u)\, du \asymp \tau^{-(q-1)}. \tag{11}$$

finite $\beta = q - 1$. We show that it is realized within a solvable model in the next section.

4 A SOLVABLE MODEL AND THE NON-COMMUTING SCALING LIMITS

Heavy-tailed data carry signal in rare examples; a truncation at level τ discards the tail beyond τ, so raising τ lowers the clean data truncation bias but lets poisoned data influence prediction more strongly. Balancing the two over τ produces a fractional exponent. In the rest of this section we explain this in more details.

4.1 THE RANDOM CO-ORDINATE MODEL

Our model is truncated linear regression in higher dimensions on heavy tailed data. We work in P dimensions. Let $e_1, \ldots, e_P$ denote the standard basis of $\mathbb{R}^P$. Let U be a random variable that has a heavy-tail

$$\mathbb{P}(U > t) = t^{-q_\star}. \tag{12}$$

A clean example draws a coordinate $J \sim \mathrm{Unif}\{1, \ldots, P\}$ and U, and sets $X_c = \sqrt{PU}\,e_J$, $Y_c = c\sqrt{U}$, that is the teacher weight is $\theta_{\star,P} = \frac{c}{\sqrt{P}}\mathbf{1}$; a poison example sets $X_{\mathrm{p}} = \sqrt{P}e_J$, $Y_{\mathrm{p}} = c + \eta_{\mathrm{p}}\tau$. Each of D i.i.d. samples is clean with probability $1-\epsilon$ and poisoned with probability ϵ. Truncating the per-example second-moment and cross-moment statistics at level τ acts as

$$A^{(\tau)} := XX^\top \min\left\{1, \frac{P\tau}{\|XX^\top\|_{\mathrm{op}}}\right\}, \qquad b^{(\tau)} := XY \min\left\{1, \frac{\sqrt{P}\,\tau}{\|XY\|_2}\right\}. \tag{13}$$

Given D training samples, define

$$\widehat{\Sigma}_{\epsilon,\tau} := \frac{1}{D}\sum_{i=1}^{D} A_i^{(\tau)}, \qquad \widehat{v}_{\epsilon,\tau} := \frac{1}{D}\sum_{i=1}^{D} b_i^{(\tau)}. \tag{14}$$

For $\varrho > 0$, define the ridge estimator

$$\widehat{\theta}_{\epsilon,\tau,\varrho} := (\widehat{\Sigma}_{\epsilon,\tau} + \varrho I_P)^{-1}\widehat{v}_{\epsilon,\tau}. \tag{15}$$

4.2 THE SIMPLIFICATION AND MAP TO POISSON POINT PROCESS

For coordinate j, let $N_{j,\mathrm{c}}$ and $N_{j,\mathrm{p}}$ denote the numbers of clean and poisoned training samples assigned to coordinate j. Let $U_{j,1}, U_{j,2}, \ldots$ denote the clean radial variables assigned to that coordinate. Define $a_\tau(U) = \min(U, \tau)$ and $r_\tau(U) = c\min(U, \tau/c)$ and

$$S_{j,A}^{(D)} := \sum_{\ell=1}^{N_{j,\mathrm{c}}} a_\tau(U_{j,\ell}) + N_{j,\mathrm{p}}, \qquad S_{j,B}^{(D)} := \sum_{\ell=1}^{N_{j,\mathrm{c}}} r_\tau(U_{j,\ell}) + (c + \eta_{\mathrm{p}}\tau)N_{j,\mathrm{p}}. \tag{16}$$

We show that (see Appendix A.2) the truncated ridge estimator is coordinate-separable,

$$\sqrt{P}\,(\widehat{\theta}_{\epsilon,\tau,\varrho})_j = G_{\phi_D,\varrho}\left(S_{j,A}^{(D)}, S_{j,B}^{(D)}\right) \tag{17}$$

where $\phi = P/D$ and,

$$G_{\phi,\varrho}(s_A, s_B) := \begin{cases} \dfrac{\phi s_B}{\varrho + \phi s_A}, & \varrho + \phi s_A > 0, \\ 0, & \varrho = 0,\ s_A = 0. \end{cases} \tag{18}$$

As a consequence, when $P, D \to \infty$ with ϕ held fixed, the clean and poison hits on a fixed coordinate converge to independent Poisson variables

$$N_{\mathrm{p}} \sim \mathrm{Pois}\Big(\tfrac{\epsilon}{\phi}\Big), \qquad N_{\mathrm{c}} \sim \mathrm{Pois}\Big(\tfrac{1-\epsilon}{\phi}\Big), \tag{19}$$

and the risk concentrates through the standard scalar deterministic equivalent. Then,

$$\mathcal{R}_{\mathrm{clean},P}(\widehat{\theta}_{\epsilon,\tau,\varrho}) - \mathcal{R}_{\mathrm{clean},P}(\theta_{\star,P}) \xrightarrow{\mathbb{P}} \mathcal{R}_{\phi,\varrho}(\epsilon,\tau) := \mu_U\,\mathbb{E}[(c - G_{\phi,\varrho}(S_A^{\epsilon,\tau}, S_B^{\epsilon,\tau}))^2], \tag{20}$$

where $\mu_U := \mathbb{E}U = \frac{q_\star}{q_\star - 1}$, and $S_A^{\epsilon,\tau} := \sum_{\ell=1}^{N_{\mathrm{c}}} a_\tau(U_\ell) + N_{\mathrm{p}}, S_B^{\epsilon,\tau} := \sum_{\ell=1}^{N_{\mathrm{c}}} r_\tau(U_\ell) + (c + \eta_{\mathrm{p}}\tau)N_{\mathrm{p}}$. This form helps us to identify this as a Poisson point process. More precisely, the clean samples reaching a coordinate form a Poisson process of intensity $(1-\epsilon)/\phi$, each point carrying the paired mark $(a_\tau(U_\ell), r_\tau(U_\ell))$, while the poisoned samples form an independent Poisson process of intensity ϵ/ϕ with the constant mark $(1,\ c + \eta_{\mathrm{p}}\tau)$. For a Poisson-indexed sum the joint Laplace functional factorizes over the intensity, for example,

$$\mathbb{E}\,\exp\Big(-\sum_{\ell=1}^{N_{\mathrm{c}}} t\cdot(a_\tau, r_\tau)(U_\ell)\Big) = \exp\Big(\tfrac{1-\epsilon}{\phi}\int \big(e^{-t\cdot(a_\tau(u), r_\tau(u))} - 1\big)\,d\nu(u)\Big),$$

so every moment of (S_A, S_B) and, decisively, its $\tau \to \infty$ asymptotics, reduces to a single integral against the mark intensity. This allows a detailed analytical study.

4.3 OPTIMAL TRUNCATION AND NON-COMMUTING LIMITS OF THE SCALING LAW

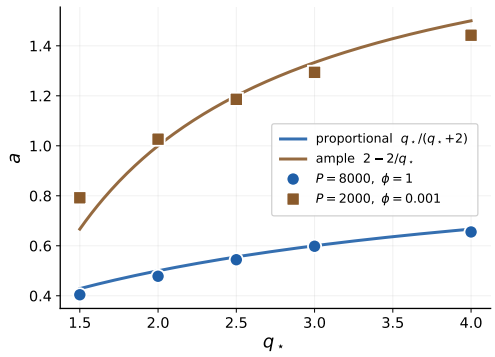
The clean truncation bias falls as τ grows while the poison influence rises, so there is a bias–poison optimum. Optimizing τ yields the paper's central theoretical result on the scaling law:

Result 3 (High-dimensional vs. ample-data limit)**.** *Work at ridgeless limit $\varrho = 0$ with $P/D \to \phi \in (0, \infty)$. The observable of interest is the clean data excess risk of the poisoned, truncated estimator against the clean, untruncated one, $\Delta_{\epsilon,\tau}^{(P)} := \mathcal{R}(\widehat{\theta}_{\epsilon,\tau,0}) - \mathcal{R}(\widehat{\theta}_{0,\infty,0})$.*

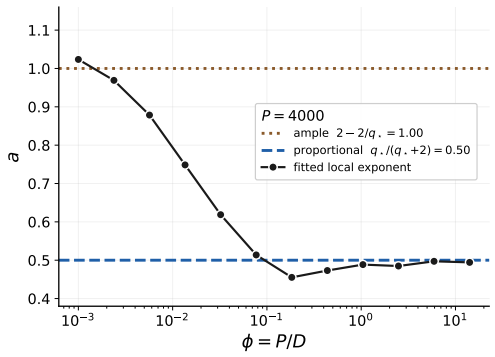
***(i) High-dimensional limit** (fixed ϕ, $\epsilon \downarrow 0$): $\Delta_{\epsilon,\tau}^{(P)} = K_{\phi,c,q_\star}^{\mathrm{clean}}\,\tau^{-q_\star} + C_{\phi,0}\,\epsilon\tau^2$, $\tau_{\mathrm{opt}} \asymp \epsilon^{-1/(q_\star+2)}$, $\min_\tau \Delta_{\epsilon,\tau}^{(P)} \asymp \epsilon^{q_\star/(q_\star+2)}$.*

***(ii) Ample-data limit** ($\phi \downarrow 0$ first and then $\epsilon \downarrow 0$): $\Delta_{\epsilon,\tau}^{(P)} = \frac{1}{\mu_U}\big(A\,\tau^{-(q_\star-1)} + B\,\epsilon\tau\big)^2$, $\tau_{\mathrm{opt}} \asymp \epsilon^{-1/q_\star}$, $\min_\tau \Delta_{\epsilon,\tau}^{(P)} \asymp \epsilon^{2-2/q_\star}$.*

The full statement, constants, and proof are Result 9 in Appendix A.7 (built on the deterministic equivalent of Appendix A.4 and the analyses of Appendices A.5–A.6). An intuitive reason for the non-commutativity is as follows:

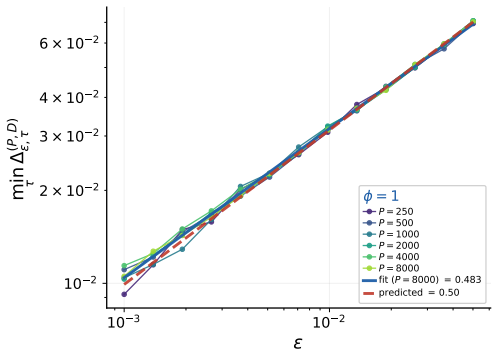


(a) Scaling exponents vs. $q_\star$.

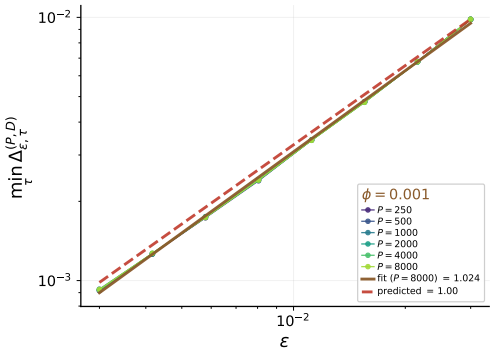


(b) Scaling exponents vs. ϕ.

Figure 2: (a) The measured optimized exponent follows $q_\star/(q_\star+2)$ (proportional) and $2-2/q_\star$ (ample) across $q_\star \in [1.5, 4]$; the branches are well separated. (b) For $q_\star = 2$, fitting the exponent over a fixed ϵ-window as $\phi = P/D$ varies shows a clean cross-over from the ample value 1.0 to the proportional value 0.5.



(a) Higher dimensional limit, scaling law as $\epsilon^{q_\star/(q_\star+2)}$.



(b) Ample data limit, scaling law as $\epsilon^{2-2/q_\star}$.

Figure 3: At $q_\star = 2$ the optimized degradation follows the predicted proportional (a) and ample (b) slopes simulation (points) against the exact prediction (dashed).

- **Higher dimensional limit**: At fixed ϕ each coordinate has only $N_c \sim \mathrm{Pois}\left(\frac{1-\epsilon}{\phi}\right) = O(1)$ clean samples. A coordinate has $N_p \geq 1$ with probability $\approx \epsilon/\phi$. On a coordinate that has $N_p \geq 1$ the poison weight is therefore not small. There is not enough clean data on the coordinate to outweigh the single poison entry, so it moves the ratio by $O(\tau)$. The squared error on that coordinate is order τ^2. Here ϵ enters only as the rarity of being hit: a fraction ϵ of coordinates carry a τ^2 error and the rest carry none, this leads to $\epsilon\tau^2$ term in excess risk.

- **Ample-data limit**: Every coordinate is hit by both large number of clean and poisoned samples since $N_c \sim (1-\epsilon)/\phi \to \infty$ and $N_p \sim \epsilon/\phi \to \infty$. Per coordinate influence of the poisoning is diluted to $\epsilon\tau$ and it enters the excess risk as $(\epsilon\tau)^2$. Here ϵ measures how much poison sits on every coordinate.

Case (ii) is exactly the adaptive-truncation law (Result 2) with $\beta = q_\star - 1, \gamma = 1$. The two exponents cross at $q_\star = \sqrt{5} - 1 \approx 1.236$: for $q_\star > \sqrt{5} - 1$ one has $q_\star/(q_\star+2) < 2 - 2/q_\star$, so the limits do not commute and the proportional regime is strictly more fragile; for $1 < q_\star < \sqrt{5} - 1$ the ordering reverses. We verify these numerically in Figure 2 and 3.

5 How the scaling law may arise in LLM pre-training

We now connect the mechanism to language models - the link requires a couple of modeling hypotheses within a local quadratic approximation to the training dynamics around a late time checkpoint (Meterez et al., 2026), developed in full in Appendix B. A preconditioned (by operator S) optimizer, run for time T starting from the checkpoint, acts as a filter $q_T(\lambda) = (1 - e^{-\lambda T})/\lambda$ on the spectrum of the preconditioned Gauss–Newton curvature A (Result 11). The filtered spectrum is responsible for learning away from the checkpoint: it learns well-conditioned directions, i.e., $\lambda T \gg 1$, quickly and leaves flat ones, i.e., $\lambda T \ll 1$, under-learned, so finite time is equivalent to

$$\text{Cut-off at } \tau = T \text{ on the inverse-curvature variable } X = \lambda^{-1},\ \lambda \in \operatorname{spec}(A). \tag{21}$$

In addition, we argue X is heavy-tailed,

$$\mathbb{P}(X > \tau) \asymp \tau^{-\alpha}, \qquad \tau \to \infty. \tag{22}$$

See Figure 4 for numerical justification. These two facts supply similar mechanism as in the previous section and lead to the scaling law as we discuss in the rest of the section.

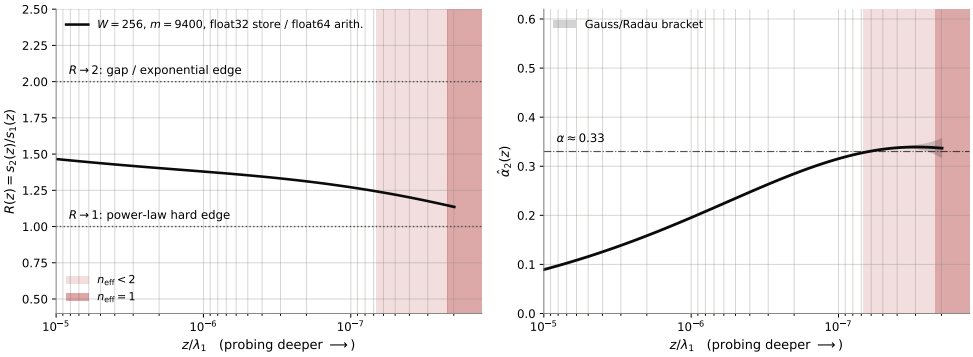


Figure 4: The plot corresponds to an experiment performed on OLMo-style model with total parameters $P = 1.677 \times 10^8$ pre-trained upto 3B tokens. The graph on the left shows that heavy tailed curvature spectrum exists and, the graph on the right determines the exponent. Red shades show where numerical analysis is unreliable. For the details of $R, \hat{\alpha}_2$ see Appendix E, these are a specially constructed function of z; smaller z probes a smaller eigenvalue spectrum. The ratio test in Corollary 32 shows that $R \to 1$ asserts a heavy-tailed spectrum of $X = \lambda^{-1}$ with hard edge exponent $\alpha \leq 1$. On the other hand Remark 5 shows $\hat{\alpha}_2$ approaches the value of $\alpha \approx 0.33$.

The clean data training for time T is captured by weight $\phi_0(T)$, which in an infinite time limit leads to $\phi_0^\star$. With $e_0(T) := \phi_0(T) - \phi_0^\star$ the clean truncation residual and G the clean-risk curvature $S^{1/2}\nabla^2\mathcal{R}_\mathcal{P}(\phi_0^\star)\,S^{1/2}$, clean truncation bias is given by $B(T)^2 := e_0^\top G e_0$. Poison at rate ϵ shifts the trajectory to $\phi_\epsilon(T) = \phi_0(T) + \epsilon\,\psi(T) + o(\epsilon)$, where the poison-response direction itself splits into a checkpoint offset carried over from a poisoned late-time checkpoint and a forced response built up afterwards, $\psi(T) = e^{-H_0 T}\psi^{\text{in}} + \psi_{\text{f}}(T)$. The poison influence built-up is captured by $I(T)^2 := \psi_{\text{f}}^\top G \psi_{\text{f}}$. The measured observable is the degradation at matched training, $\delta\mathcal{R}(\epsilon, T) := \mathcal{R}_\mathcal{P}(\phi_\epsilon(T)) - \mathcal{R}_\mathcal{P}(\phi_0(T))$. It resolves the response into four competing contributions,

$$\delta\mathcal{R} \asymp \underbrace{\epsilon\,\Gamma_\star\,T^{-\beta_{\text{in}}}}_{\text{checkpoint offset}} + \underbrace{\epsilon\,\Gamma_\star\,T^{\gamma}}_{\text{train–eval gap}} + \underbrace{\epsilon\,B(T)I(T)}_{\text{clean bias} \asymp \epsilon\,T^{\gamma-\beta}} + \underbrace{\tfrac{1}{2}\epsilon^2 I(T)^2}_{\text{pure poison} \asymp \epsilon^2\,T^{2\gamma}}, \tag{23}$$

where $\Gamma_\star := \|g_\star\|_{G^{-1}}, g_\star := \nabla\mathcal{R}_\mathcal{P}(\phi_0^\star)$ is the train-eval mismatch. We show $B(T) \asymp T^{-\beta}$ (Result 17) and $I(T) \asymp T^{\gamma}$ (Result 19) with $\beta > \gamma > 0$ (that $\beta > \gamma$ says the clean signal is learned faster than the poison accumulates) and $B_{\text{in}}(T) := \left\|e^{-H_0 T}\psi^{\text{in}}\right\|_G \asymp T^{-\beta_{\text{in}}}, \beta_{\text{in}} > 0$. These are power laws for the same reason-the spectrum is scale-free near weak curvature modes, so advancing the cutoff rescales every quantity by a pure power. The clean bias decays because it is the residual clean signal still sitting beyond the moving cutoff; each increment of time resolves more of the tail and the leftover shrinks. The poison influence grows because the poison's leverage lives in the low-curvature directions: there a small contamination produces a large parameter displacement, and a longer time horizon reaches ever more such directions, so the aggregate imprint climbs up.

Gap-dominated crossover. When the β-independent train–eval gap leads, $\epsilon\,\Gamma_\star I(T)$ and $\frac{1}{2}\epsilon^2 I(T)^2$ cross at $T_\epsilon^{(g)} \asymp \epsilon^{-1/\gamma}$ and,

$$\Delta \asymp \Gamma_\star\,\epsilon\,T^\gamma \text{ for } T \ll T_\epsilon^{(g)}, \qquad \Delta \asymp \epsilon^2 T^{2\gamma} \text{ for } T \gg T_\epsilon^{(g)}. \qquad \text{(Result 20)} \tag{24}$$

Both competing terms grow with T, so there is no interior optimal horizon. The degradation is ϵ-independent at the crossover time leading to $a = 0$.

Bias-dominated crossover. When a decaying heavy-tail term leads, the falling clean bias $\epsilon\,B(T)I(T) \asymp \epsilon T^{\gamma-\beta}$ and the rising pure term balance at a poison-dependent cross-over time $T_\epsilon \asymp \epsilon^{-1/(\beta+\gamma)}$, an interior minimum, where the scaling law emerges,

$$\Delta(\epsilon) \asymp \epsilon^a, \qquad a = \frac{2\beta}{\beta+\gamma} \qquad \text{(Result 21)}. \tag{25}$$

The checkpoint offset drives the same crossover with β replaced by $\beta_{\text{in}} + \gamma$, the two share their origin in the heavy-tailed spectrum, so we discuss only the bias case. The scaling exponent is non-integer iff $\beta < \infty$-finitely many moments, i.e. a genuinely heavy-tailed inverse-curvature spectrum; as $\beta \to \infty$ it returns to the analytic value 2. The crossover time above grows as an inverse power of the poisoning fraction, $T_\epsilon^{(g)}, T_\varepsilon \asymp \varepsilon^{-\kappa}$, when $\kappa > 0$ it diverges as $\varepsilon \to 0$. The fractional exponent a is therefore expected to be observable only in the regime of a small poisoning fraction together with a correspondingly long training steps - one whose length grows as the same inverse power of ε. Analyzing this scaling is complicated by the fact that κ is not a directly controlled quantity in LLM training: it is governed by the inverse weak-curvature scales of A and the effect of the poisoned data in this regime. We therefore proceed under the modeling assumption that the LLM training horizon considered in this paper is of the same order as the crossover time.

Models with a large train-eval mismatch $\Gamma_\star$-empirically the smaller LLMs-sit in the gap-dominated regime: the measured exponent starts at $a = 1$ early in training and falls toward $a = 0$ as the time approaches $T_\epsilon^{(g)}$ (see Appendix B, Figure 7(a)). Models with a small mismatch-the larger LLMs sit in the bias/checkpoint-dominated regime, where the exponent starts near $a = 1$ and rises toward the fractional value $2\beta/(\beta+\gamma) > 1$ as the time approaches the balanced T_ϵ (see Appendix B, Figure 7(d)).

The balanced-horizon calculation in this section is an ample-data one in the aspect ratio: writing $\phi_{\text{LLM}} = P_{\text{eff}}/D_{\text{eff}}$, Result 9 warns that taking $\epsilon \to 0$ before $\phi_{\text{LLM}} \to 0$ gives a different exponent, so the observed law need not extend to arbitrarily small ϵ and one should expect it for $\epsilon \gtrsim \phi_{\text{LLM}}$ (see Figure 6 for more discussion). The full local model, assumptions, and proofs are in Appendix B.

6 Conclusion and Limitations

We pointed out the scaling law for pre-train poisioning of LLMs (Figure 1). Then discussed theoretical analysis of a non-integer poisoning exponent (Result 1 and 2), and provided a solvable model in which heavy tails read out through a poison-dependent truncation supply it (Results 3), and proved that the optimized degradation has two distinct exponents depending on the order of the high-dimensional and small-poison limits, verifying all predictions numerically (Figure 2, 3). Within local LLM training dynamics, we argue how the scaling law may arise in Section 5. The exact theoretical results are for a stylized coordinate model; the connection to LLM training dynamics requires certain modeling hypotheses that remain an open question to establish in the future work. In addition, careful study of small-ϵ regime in LLMs and the fate of the scaling law requires further experiments. On the theoretical side, extending the deterministic equivalent to dense covariances and multi-coordinate examples are the main open directions.

AI USE STATEMENT

In this work, we used generative AI tools for typing LaTeX and resolving bugs in the code used to run the experiments. We have not used generative AI tools for the formulation of the project idea, developing the theoretical models, proving the theorems, and interpreting the results. Additionally, we used generative AI tools for improving the readability of the draft, fixing grammatical errors in the writing, and formatting references. We have reviewed all AI-assisted work. The first and second authors, responsible for the experimental code development, manually checked the correctness of all the suggestions provided by coding agent tools. All authors, responsible for drafting the paper, manually verified the reformatted references and overall paper structure. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper studies how the data poisoning degrades the model's clean data performance. We do not introduce any new attack methods or datasets; all our empirical results are based on publicly released attack methods. Therefore, we believe our paper does not impose any safety risks.

REPRODUCIBILITY STATEMENT

The authors have put their best efforts to make this work reproducible. We have included supplementary materials with this submission that provide all the codes to run, analyze, and replot the results reported in this paper. Because of page limitations, some proof/claim details could not be included in the main text. However, we have provided detailed proof of each claim, the assumptions, and all associated details in the Appendix. The details included in this paper, along with the submitted code will allow anyone to reproduce our results.

REFERENCES

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016. doi: 10.1162/tacl_a_00106. URL https://aclanthology.org/Q16-1028/.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

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.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pp. 1024–1034. PMLR, 2020.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. *arXiv preprint arXiv:2402.01092*, 2024.

Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. Scaling trends for data poisoning in LLMs. *arXiv preprint arXiv:2408.02946*, 2024.

Francesco Cagnetta and Matthieu Wyart. Towards a theory of how the structure of language is acquired by deep neural networks. *Advances in Neural Information Processing Systems*, 37: 83119–83163, 2024.

Francesco Cagnetta, Leonardo Petrini, Umberto M Tomasini, Alessandro Favero, and Matthieu Wyart. How deep neural networks learn compositional data: The random hierarchy model. *Physical Review X*, 14(3):031001, 2024.

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.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 407–425. IEEE, 2024.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.

Gene H. Golub and Gérard Meurant. *Matrices, Moments and Quadrature with Applications*. Princeton University Press, 2010. Contains the Gauss–Radau rules and associated bounding theory; cited as 2009 in Meterez et al. (2026).

Gene H. Golub and John H. Welsch. Calculation of Gauss quadrature rules. *Mathematics of Computation*, 23:221–230, 1969.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Anne Greenbaum. Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences. *Linear Algebra and its Applications*, 113:7–63, 1989.

Indranil Halder, Annesya Banerjee, and Cengiz Pehlevan. Jailbreak scaling laws for large language models: Polynomial-exponential crossover. *arXiv preprint arXiv:2603.11331*, 2026.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Marcus Hutter. Learning curve theory. *arXiv preprint arXiv:2102.04074*, 2021.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in neural information processing systems*, 19, 2006.

Shmuel Kaniel. Estimates for some computational techniques in linear algebra. *Mathematics of Computation*, 20, 1966. Origin of the Chebyshev-type convergence bounds for extremal Ritz values refined by Paige and Saad.

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.

Dhruva Karkada, Daniel J Korchinski, Andres Nava, Matthieu Wyart, and Yasaman Bahri. Symmetry in language statistics shapes the geometry of model representations. *arXiv preprint arXiv:2602.15029*, 2026a.

Dhruva Karkada, James Simon, Yasaman Bahri, and Michael Deweese. Closed-form training dynamics reveal learned features and linear structure in word2vec-like models. *Advances in neural information processing systems*, 38:3967–4000, 2026b.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Pang Wei Koh, Jacob Steinhardt, and Percy Liang. Stronger data poisoning attacks break data sanitization defenses. *Machine Learning*, 111(1):1–47, 2022.

Daniel Korchinski, Dhruva Karkada, Yasaman Bahri, and Matthieu Wyart. On the emergence of linear analogies in word embeddings. *Advances in neural information processing systems*, 38: 62069–62097, 2026.

Nikola B. Kovachki and Andrew M. Stuart. Continuous time analysis of momentum methods. *Journal of Machine Learning Research*, 22(17):1–40, 2021.

Cornelius Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 45: 255–282, 1950.

Linyang Li, Demin Song, Xiaonan Li, Jiehang Zeng, Ruotian Ma, and Xipeng Qiu. Backdoor attacks on pre-trained models by layerwise weight poisoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3023–3032, 2021.

Percy Liang, Slav Petrov, Michael I Jordan, and Dan Klein. The infinite pcfg using hierarchical dirichlet processes. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 688–697, 2007.

Andy Zeyi Liu, Elliot Paquette, and John Sous. Evolution of the spectral dimension of transformer activations. In *OPT 2025: Optimization for Machine Learning*, 2025.

Yizhou Liu, Ziming Liu, Cengiz Pehlevan, and Jeff Gore. Universal one-third time scaling in learning peaked distributions. *arXiv preprint arXiv:2602.03685*, 2026.

Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019.

David JC MacKay and Linda C Bauman Peto. A hierarchical dirichlet language model. *Natural language engineering*, 1(3):289–308, 1995.

Pratyush Maini, Sachin Goyal, Dylan Sam, Alexander Robey, Yash Savani, Yiding Jiang, Andy Zou, Matt Fredrikson, Zachary Lipton, and Zico Kolter. Safety pretraining: Toward the next generation of safe ai. *Advances in Neural Information Processing Systems*, 38:43971–44008, 2026.

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

Alexandru Meterez, Praneeth A. Nair, Depen Morwani, Cengiz Pehlevan, Sham Kakade, and Alex Damian. A defense of the quadratic model, 2026. URL `https://arxiv.org/abs/2607.21716`. Code available at `https://github.com/alexandrumeterez/QuadraticModel`.

Gérard Meurant and Zdeněk Strakoš. The Lanczos and conjugate gradient algorithms in finite precision arithmetic. *Acta Numerica*, 15:471–542, 2006.

Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36:28699–28722, 2023.

Christopher C. Paige. Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix. *IMA Journal of Applied Mathematics*, 18(3):341–349, 1976.

Christopher C. Paige. Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem. *Linear Algebra and its Applications*, 34:235–258, 1980.

Beresford N. Parlett. *The Symmetric Eigenvalue Problem*. Prentice-Hall, 1980. Reprinted as volume 20 of SIAM Classics in Applied Mathematics, 1998. Includes the Kaniel–Paige–Saad bounds and convergence at both ends of the spectrum.

Jack T Parley, Francesco Cagnetta, and Matthieu Wyart. Deep networks learn to parse uniform-depth context-free languages from local statistics. *arXiv preprint arXiv:2602.06065*, 2026.

Jim Pitman. *Combinatorial stochastic processes: Ecole d'eté de probabilités de saint-flour xxxii-2002*. Springer Science & Business Media, 2006.

Yousef Saad. On the rates of convergence of the Lanczos and the block-Lanczos methods. *SIAM Journal on Numerical Analysis*, 17(5):687–706, 1980.

Alessandro Salvatore, Stanislav Fort, and Surya Ganguli. Solving adversarial examples requires solving exponential misalignment. *arXiv preprint arXiv:2603.03507*, 2026.

Dylan Sam, Sachin Goyal, Pratyush Maini, Alexander Robey, and J Zico Kolter. When should we introduce safety interventions during pretraining? *arXiv preprint arXiv:2601.07087*, 2026.

Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*, 2020.

Eva Silverstein, Daniel Kunin, and Vasudev Shyam. Symmetry breaking in transformers for efficient and interpretable training. *arXiv preprint arXiv:2601.22257*, 2026.

Alexandra Souly, Javier Rando, Ed Chapman, Xander Davies, Burak Hasircioglu, Ezzeldin Shereen, Carlos Mougan, Vasilios Mavroudis, Erik Jones, Chris Hicks, et al. Poisoning attacks on llms require a near-constant number of poison samples. *arXiv preprint arXiv:2510.07192*, 2025.

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.

Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. *Advances in neural information processing systems*, 30, 2017.

Anirudh Subramanyam, Yuxin Chen, and Robert Grossman. Scaling laws revisited: Modeling the role of data quality in language model pretraining. In *International Conference on Learning Representations*, volume 2026, pp. 128471–128491, 2026.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. URL `https://arxiv.org/abs/1312.6199`.

Gábor Szegő. *Orthogonal Polynomials*, volume 23 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, 4 edition, 1975. Classical source for the Markov–Stieltjes inequalities relating Gauss weights to the distribution function.

Thomas Tanay and Lewis Griffin. A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690*, 2016.

Qian-Yuan Tang, Yufei Gu, Yunfeng Cai, Mingming Sun, Ping Li, Zeke Xie, et al. Investigating the overlooked hessian structure: From cnns to llms. In *Forty-second International Conference on Machine Learning*, 2025.

Yee Whye Teh. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics*, pp. 985–992, 2006.

Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. Concealed data poisoning attacks on nlp models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 139–150, 2021.

Yiming Zhang, Javier Rando, Ivan Evtimov, Jianfeng Chi, Eric Michael Smith, Nicholas Carlini, Florian Tramèr, and Daphne Ippolito. Persistent pre-training poisoning of llms. In *International Conference on Learning Representations*, volume 2025, pp. 31323–31340, 2025.

A NON-COMMUTING HIGH-DIMENSIONAL AND SMALL-POISONING LIMITS

A.1 CONNECTION TO THE LOCAL LLM APPROXIMATION

The purpose of this appendix is to test the local LLM mechanism in a model that retains finite-sample, high-dimensional effects. Around a reference checkpoint θ_0, the local quadratic approximation to pre-training has the form

$$\widehat{\mathcal{L}}(\theta_0+\delta\theta) \simeq \widehat{\mathcal{L}}(\theta_0) - \widehat{v}_{\mathrm{LLM}}^{\top}\delta\theta + \frac{1}{2}\delta\theta^{\top}\widehat{H}_{\mathrm{LLM}}\delta\theta,$$

where, for token-prediction events indexed by $i = 1, \ldots, D$,

$$\widehat{\Sigma}_{\mathrm{LLM}} = \frac{1}{D}\sum_{i=1}^{D} J_i^{\top}\mathcal{W}_i J_i, \qquad \widehat{v}_{\mathrm{LLM}} = \frac{1}{D}\sum_{i=1}^{D} J_i^{\top} r_i.$$

The corresponding local stationary equation is

$$\widehat{\Sigma}_{\mathrm{LLM}}\,\widehat{\delta\theta} = \widehat{v}_{\mathrm{LLM}}.$$

The regression model below has exactly the same algebraic form:

$$\widehat{\Sigma}_{\epsilon,\tau} = \frac{1}{D}\sum_{i=1}^{D} A_i^{(\tau)}, \qquad \widehat{v}_{\epsilon,\tau} = \frac{1}{D}\sum_{i=1}^{D} b_i^{(\tau)}, \qquad (\widehat{\Sigma}_{\epsilon,\tau} + \varrho I_P)\widehat{\theta} = \widehat{v}_{\epsilon,\tau},$$

where $A_i^{(\tau)}$ and $b_i^{(\tau)}$ are truncated versions of $X_i X_i^{\top}$ and $X_i Y_i$. Thus

quantity	coordinate surrogate	local LLM model
parameter dimension	P	P_{eff}
number of examples	D	D_{eff}
per-example curvature	$A_i^{(\tau)}$	$J_i^{\top} W_i J_i$
per-example response	$b_i^{(\tau)}$	$J_i^{\top} r_i$
fitted displacement	$\widehat{\theta}$	$\delta\theta$
aspect ratio	$\phi = P/D$	$P_{\mathrm{eff}}/D_{\mathrm{eff}}$

at the level of the local normal equation.

The analogy concerns having the same normal equations. Truncation is implemented slightly differently: the model below clips the two per-example statistics separately, whereas finite-time LLM training acts as a spectral filter on inverse curvature. However, in this setup, again, we can see the same set of logic used in LLM - heavy-tailed data, along with poison adaptive truncation, leads to the scaling law in the ample data limit. Hence, analysis of this appendix can teach some valuable lessons for the local LLM model. More precisely, the coordinate construction is useful precisely because it makes one additional question explicit: does the small-poisoning limit agree with the population calculation when the parameter-to-data aspect ratio remains nonzero? The remainder of the appendix answers this question exactly for the coordinate model.

A.2 COORDINATE-ISOTROPIC HEAVY-TAILED DATA DISTRIBUTION

Let $e_1, \ldots, e_P$ denote the standard basis of $\mathbb{R}^P$. Fix

$$0 < c < 1, \qquad 0 < \eta_{\mathrm{p}} < 1, \qquad q_\star > 1.$$

Let U have Pareto tail

$$\mathbb{P}(U > t) = t^{-q_\star}, \qquad t \geq 1,$$

so that

$$\mu_U := \mathbb{E}U = \frac{q_\star}{q_\star - 1}.$$

All truncation formulas below are used for levels $\tau \geq 1$. Define the teacher vector

$$\theta_{\star,P} := \frac{c}{\sqrt{P}}(1, \ldots, 1)^{\top}.$$

A clean example is generated by drawing J uniformly from $\{1, \dots, P\}$ and drawing U independently, then setting

$$X = \sqrt{PU}\, e_J, \qquad Y = X^\top \theta_{\star,P} = c\sqrt{U}.$$

The normalization $\sqrt{P}$ is chosen so that

$$\mathbb{E}[XX^\top] = \mu_U I_P.$$

For any deterministic $\theta \in \mathbb{R}^P$, the clean data excess prediction risk is

$$\mathcal{R}_{\mathrm{clean},P}(\theta) - \mathcal{R}_{\mathrm{clean},P}(\theta_{\star,P}) = \mu_U \frac{1}{P} \sum_{j=1}^{P} \left(c - \sqrt{P}\,\theta_j\right)^2. \tag{26}$$

Indeed, conditional on $J = j$ and U, one has

$$Y - X^\top \theta = \sqrt{U}\left(c - \sqrt{P}\,\theta_j\right),$$

and averaging over J and U gives equation 26.

A.2.1 Modified truncation suitable for higher dimensional statistics

The unnormalized matrix statistic has operator norm

$$\left\|XX^\top\right\|_{\mathrm{op}} = PU,$$

and the vector statistic satisfies

$$XY = c\sqrt{P}\, U e_J, \qquad \|XY\|_2 = c\sqrt{P}\, U.$$

We use dimension-normalized truncation thresholds $P\tau$ for the matrix statistic and $\sqrt{P}\,\tau$ for the vector statistic. Thus

$$A^{(\tau)} := XX^\top \min\left\{1, \frac{P\tau}{\|XX^\top\|_{\mathrm{op}}}\right\}, \qquad b^{(\tau)} := XY \min\left\{1, \frac{\sqrt{P}\,\tau}{\|XY\|_2}\right\}.$$

For a clean example this gives

$$A^{(\tau)} = P\, a_\tau(U) e_J e_J^\top, \qquad b^{(\tau)} = \sqrt{P}\, r_\tau(U) e_J, \tag{27}$$

where[1]

$$a_\tau(t) := t \wedge \tau, \qquad r_\tau(t) := c\left(t \wedge \frac{\tau}{c}\right).$$

The clean truncated population moments are

$$M_\tau := \mathbb{E} a_\tau(U) = \mu_U - \frac{\tau^{1-q_\star}}{q_\star - 1},$$

and

$$B_\tau := \mathbb{E} r_\tau(U) = c M_{\tau/c} = c\mu_U - \frac{c^{q_\star} \tau^{1-q_\star}}{q_\star - 1}.$$

Since $0 < c < 1$, the two truncations clip the same clean radial tail at different effective thresholds. This is the high-dimensional coordinate version of the clean truncation-bias mechanism.

A.2.2 Label-shifted data poisoning within truncation radius

We now use the modified poison. A poisoned example is generated by drawing J uniformly from $\{1, \dots, P\}$ and setting

$$X_{\mathrm{p}} = \sqrt{P}\, e_J, \qquad Y_{\mathrm{p}} = c + \eta_{\mathrm{p}} \tau.$$

Thus the poison feature has bounded normalized squared norm,

$$\frac{\|X_{\mathrm{p}}\|_2^2}{P} = 1,$$

[1]The notation means that $t \wedge \tau$ is the smaller of t and τ.

while the label residual relative to the clean teacher is order τ:

$$Y_{\mathrm{p}} - X_{\mathrm{p}}^{\top}\theta_{\star,P} = \eta_{\mathrm{p}}\tau.$$

Let

$$\tau_0 := \max\left\{1, \frac{c}{1-\eta_{\mathrm{p}}}\right\}.$$

Throughout the exact formulas for the modified poison we assume $\tau \geq \tau_0$. Then

$$c + \eta_{\mathrm{p}}\tau \leq \tau,$$

so the poison cross-moment lies below the vector truncation threshold, while its matrix statistic lies below the matrix truncation threshold. Hence

$$A_{\mathrm{p}}^{(\tau)} = P e_J e_J^{\top}, \qquad b_{\mathrm{p}}^{(\tau)} = \sqrt{P}\,(c + \eta_{\mathrm{p}}\tau) e_J. \tag{28}$$

Training samples are independent. Each sample is clean with probability $1-\epsilon$ and poisoned with probability ϵ.

A.3 Simplification in linear regression with truncation

Given D training samples, define

$$\widehat{\Sigma}_{\epsilon,\tau} := \frac{1}{D}\sum_{i=1}^{D} A_i^{(\tau)}, \qquad \widehat{v}_{\epsilon,\tau} := \frac{1}{D}\sum_{i=1}^{D} b_i^{(\tau)}.$$

For $\varrho > 0$, define the ridge estimator

$$\widehat{\theta}_{\epsilon,\tau,\varrho} := (\widehat{\Sigma}_{\epsilon,\tau} + \varrho I_P)^{-1}\widehat{v}_{\epsilon,\tau}.$$

For $\varrho = 0$, define the ridgeless estimator by the Moore–Penrose inverse:

$$\widehat{\theta}_{\epsilon,\tau,0} := \widehat{\Sigma}_{\epsilon,\tau}^{\dagger}\widehat{v}_{\epsilon,\tau}.$$

Let $\phi_D = P/D$. For coordinate j, let $N_{j,\mathrm{c}}$ and $N_{j,\mathrm{p}}$ denote the numbers of clean and poisoned training samples assigned to coordinate j. Let $U_{j,1}, U_{j,2}, \ldots$ denote the clean radial variables assigned to that coordinate. Define

$$S_{j,A}^{(D)} := \sum_{\ell=1}^{N_{j,\mathrm{c}}} a_\tau(U_{j,\ell}) + N_{j,\mathrm{p}}, \tag{29}$$

$$S_{j,B}^{(D)} := \sum_{\ell=1}^{N_{j,\mathrm{c}}} r_\tau(U_{j,\ell}) + (c + \eta_{\mathrm{p}}\tau) N_{j,\mathrm{p}}. \tag{30}$$

By construction, the per-example statistics from equation 27 and equation 28 are supported on the single coordinate $J_i \in \{1, \ldots, P\}$ to which sample i is assigned. Writing c for a clean sample and p for a poisoned one,

$$A_i^{(\tau)} = P\,\alpha_i\, e_{J_i} e_{J_i}^{\top}, \qquad b_i^{(\tau)} = \sqrt{P}\,\beta_i\, e_{J_i},$$

where

$$\alpha_i = \begin{cases} a_\tau(U_i), & i \text{ clean}, \\ 1, & i \text{ poison}, \end{cases} \qquad \beta_i = \begin{cases} r_\tau(U_i), & i \text{ clean}, \\ c + \eta_{\mathrm{p}}\tau, & i \text{ poison}. \end{cases}$$

Each matrix $e_{J_i} e_{J_i}^{\top}$ is diagonal, with a single nonzero entry equal to 1 in position (J_i, J_i). Hence

$$\widehat{\Sigma}_{\epsilon,\tau} = \frac{1}{D}\sum_{i=1}^{D} A_i^{(\tau)} = \frac{P}{D}\sum_{i=1}^{D} \alpha_i\, e_{J_i} e_{J_i}^{\top}$$

is a sum of diagonal matrices and is therefore diagonal. Its (j,j) entry collects exactly those samples assigned to coordinate j:

$$(\widehat{\Sigma}_{\epsilon,\tau})_{jj} = \frac{P}{D}\sum_{i:\, J_i = j} \alpha_i = \phi_D\Big(\sum_{\ell=1}^{N_{j,\mathrm{c}}} a_\tau(U_{j,\ell}) + N_{j,\mathrm{p}}\Big) = \phi_D\, S_{j,A}^{(D)},$$

using the definition equation 29 of $S_{j,A}^{(D)}$. Likewise, $b_i^{(\tau)} = \sqrt{P}\,\beta_i\, e_{J_i}$ has a single nonzero entry in position J_i, so

$$(\widehat{v}_{\epsilon,\tau})_j = \frac{1}{D}\sum_{i:\,J_i=j}\sqrt{P}\,\beta_i = \frac{\sqrt{P}}{D}\Big(\sum_{\ell=1}^{N_{j,\mathrm{c}}} r_\tau(U_{j,\ell}) + (c+\eta_{\mathrm{p}}\tau)N_{j,\mathrm{p}}\Big) = \frac{\sqrt{P}}{D}\,S_{j,B}^{(D)} = \frac{\phi_D}{\sqrt{P}}\,S_{j,B}^{(D)},$$

where the last equality uses $\sqrt{P}/D = (P/D)/\sqrt{P} = \phi_D/\sqrt{P}$. Consequently,

$$\sqrt{P}\,(\widehat{\theta}_{\epsilon,\tau,\varrho})_j = G_{\phi_D,\varrho}\left(S_{j,A}^{(D)}, S_{j,B}^{(D)}\right), \tag{31}$$

where, for $\varrho \geq 0$,

$$G_{\phi,\varrho}(s_A, s_B) := \begin{cases} \dfrac{\phi s_B}{\varrho + \phi s_A}, & \varrho + \phi s_A > 0, \\ 0, & \varrho = 0,\ s_A = 0. \end{cases} \tag{32}$$

For a clean data one has $0 \leq r_\tau(t) \leq a_\tau(t)$, and for a modified poison data, when $\tau \geq \tau_0$, $0 \leq c + \eta_{\mathrm{p}}\tau \leq \tau$. Hence, with $m_\tau := \max\{1, c + \eta_{\mathrm{p}}\tau\}$, every finite collection of clean and poison data satisfies $0 \leq s_B \leq m_\tau s_A$. Since $\tau \geq \tau_0$ implies $\tau \geq 1$ and $c + \eta_{\mathrm{p}}\tau \leq \tau$, we also have $m_\tau \leq \tau$. Therefore

$$0 \leq G_{\phi,\varrho}(s_A, s_B) \leq m_\tau \leq \tau. \tag{33}$$

A.4 The deterministic equivalent at fixed aspect ratio

Each training sample chooses one of the P coordinates uniformly. Conditional on the sample being assigned to coordinate j, the sample is clean with probability $1-\epsilon$ and poisoned with probability ϵ. Therefore, for a fixed coordinate j, the finite-D, P clean and poison counts satisfy

$$N_{j,\mathrm{c}}^{(D)} = \sum_{i=1}^{D} \mathbf{1}\{\text{sample } i \text{ is clean and assigned to } j\},$$

and

$$N_{j,\mathrm{p}}^{(D)} = \sum_{i=1}^{D} \mathbf{1}\{\text{sample } i \text{ is poisoned and assigned to } j\}.$$

For each sample,

$$\mathbb{P}(\text{clean and assigned to } j) = \frac{1-\epsilon}{P}, \qquad \mathbb{P}(\text{poisoned and assigned to } j) = \frac{\epsilon}{P}.$$

Hence

$$\mathbb{E}N_{j,\mathrm{c}}^{(D)} = D\frac{1-\epsilon}{P} = \frac{1-\epsilon}{\phi_D}, \qquad \mathbb{E}N_{j,\mathrm{p}}^{(D)} = D\frac{\epsilon}{P} = \frac{\epsilon}{\phi_D}.$$

Lemma 1 (Poisson limit of the coordinate counts)**.** *Fix a coordinate j. Suppose $P/D \to \phi \in (0,\infty)$. Each training sample is clean with probability $1-\epsilon$, poisoned with probability ϵ, and independently assigned to coordinate j with probability $1/P$. Let $N_{j,\mathrm{c}}^{(D)}$ and $N_{j,\mathrm{p}}^{(D)}$ be the numbers of clean and poisoned samples assigned to coordinate j. Then*

$$(N_{j,\mathrm{c}}^{(D)}, N_{j,\mathrm{p}}^{(D)}) \Rightarrow (N_{\mathrm{c}}, N_{\mathrm{p}}),$$

where

$$N_{\mathrm{c}} \sim \mathrm{Pois}\left(\frac{1-\epsilon}{\phi}\right), \qquad N_{\mathrm{p}} \sim \mathrm{Pois}\left(\frac{\epsilon}{\phi}\right),$$

and N_{c} and N_{p} are independent.

Proof. For a fixed coordinate j, a given training sample contributes a clean data to coordinate j with probability

$$p_{\mathrm{c},D} = \frac{1-\epsilon}{P},$$

and contributes a poison data to coordinate j with probability

$$p_{\mathrm{p},D} = \frac{\epsilon}{P}.$$

Thus

$$(N_{j,\mathrm{c}}^{(D)}, N_{j,\mathrm{p}}^{(D)})$$

has a multinomial distribution with three categories: clean hit, poison hit, and no hit. For $k, \ell \geq 0$ with $k + \ell \leq D$,

$$\mathbb{P}\left(N_{j,\mathrm{c}}^{(D)} = k,\, N_{j,\mathrm{p}}^{(D)} = \ell\right) = \frac{D!}{k!\,\ell!\,(D-k-\ell)!}\left(\frac{1-\epsilon}{P}\right)^{k}\left(\frac{\epsilon}{P}\right)^{\ell}\left(1-\frac{1}{P}\right)^{D-k-\ell}.$$

Since $P/D \to \phi$, we have

$$\frac{D}{P} \to \frac{1}{\phi}, \qquad \frac{D(1-\epsilon)}{P} \to \frac{1-\epsilon}{\phi}, \qquad \frac{D\epsilon}{P} \to \frac{\epsilon}{\phi}.$$

For fixed k, ℓ,

$$\frac{D!}{(D-k-\ell)!P^{k+\ell}} \to \left(\frac{1}{\phi}\right)^{k+\ell},$$

and

$$\left(1-\frac{1}{P}\right)^{D-k-\ell} \to \exp\left(-\frac{1}{\phi}\right).$$

Therefore

$$\mathbb{P}\left(N_{j,\mathrm{c}}^{(D)} = k,\, N_{j,\mathrm{p}}^{(D)} = \ell\right) \to \exp\left(-\frac{1}{\phi}\right)\frac{\left(\frac{1-\epsilon}{\phi}\right)^{k}}{k!}\frac{\left(\frac{\epsilon}{\phi}\right)^{\ell}}{\ell!}.$$

This is the joint probability mass function of two independent Poisson random variables with means $(1-\epsilon)/\phi$ and ϵ/ϕ, respectively. □

At fixed ϕ, a coordinate receives a poison data with probability

$$\mathbb{P}(N_{\mathrm{p}} \geq 1) = 1 - \exp\left(-\frac{\epsilon}{\phi}\right) = \frac{\epsilon}{\phi} + O(\epsilon^2), \qquad \epsilon \downarrow 0.$$

Thus poisoning acts through a sparse-coordinate mechanism: most coordinates see no poison, while a proportion of order ϵ/ϕ see one poison data. By contrast, if one first sends $\phi \downarrow 0$, then

$$\frac{1-\epsilon}{\phi} \to \infty, \qquad \frac{\epsilon}{\phi} \to \infty \quad \text{for fixed } \epsilon > 0.$$

Each coordinate is hit by many clean and poisoned samples, and the Poisson fluctuations average out by the law of large numbers.

Let $U_1, U_2, \ldots$ be i.i.P. copies of U drawn from the Pareto tail, independent of $N_{\mathrm{c}}, N_{\mathrm{p}}$:

$$N_{\mathrm{c}} \sim \mathrm{Pois}\left(\frac{1-\epsilon}{\phi}\right), \qquad N_{\mathrm{p}} \sim \mathrm{Pois}\left(\frac{\epsilon}{\phi}\right),$$

Define the limiting compound-Poisson coordinate statistics

$$S_A^{\epsilon,\tau} := \sum_{\ell=1}^{N_{\mathrm{c}}} a_\tau(U_\ell) + N_{\mathrm{p}}, \tag{34}$$

$$S_B^{\epsilon,\tau} := \sum_{\ell=1}^{N_{\mathrm{c}}} r_\tau(U_\ell) + (c + \eta_{\mathrm{p}}\tau)N_{\mathrm{p}}. \tag{35}$$

Finally define

$$\mathcal{R}_{\phi,\varrho}(\epsilon,\tau) := \mu_U\, \mathbb{E}\left[\left(c - G_{\phi,\varrho}\left(S_A^{\epsilon,\tau}, S_B^{\epsilon,\tau}\right)\right)^2\right]. \tag{36}$$

Lemma 2 (Deterministic equivalent)**.** *Let $F : [0, \infty)^2 \to \mathbb{R}$ be a bounded function. Then*

$$\frac{1}{P}\sum_{j=1}^{P} F\left(S_{j,A}^{(D)}, S_{j,B}^{(D)}\right) \xrightarrow{\mathbb{P}} \mathbb{E}F\left(S_A^{\epsilon,\tau}, S_B^{\epsilon,\tau}\right).$$

The conclusion also holds for the bounded functions

$$F_{\phi_D,\varrho}(s_A, s_B) := \left(c - G_{\phi_D,\varrho}(s_A, s_B)\right)^2,$$

with the limit $F_{\phi,\varrho}$, including the ridgeless convention $\varrho = 0$.

Proof. Fix $\tau \geq \tau_0$. Since $P/D \to \phi$, the law of $(S_{j,A}^{(D)}, S_{j,B}^{(D)})$ converges to the compound-Poisson law in equation 34. Therefore

$$\mathbb{E}F\left(S_{j,A}^{(D)}, S_{j,B}^{(D)}\right) \to \mathbb{E}F\left(S_A^{\epsilon,\tau}, S_B^{\epsilon,\tau}\right)$$

by the continuity. It remains to show concentration of the coordinate average. Let

$$Y_j^{(D)} := F\left(S_{j,A}^{(D)}, S_{j,B}^{(D)}\right), \qquad j = 1, \ldots, P.$$

Since the training examples assign coordinates symmetrically, the random variables $Y_1^{(D)}, \ldots, Y_P^{(D)}$ are exchangeable. Suppose

$$\|F\|_\infty := \sup_{s_A, s_B} |F(s_A, s_B)| < \infty.$$

Then

$$|Y_j^{(D)}| \leq \|F\|_\infty \qquad \text{for every } j.$$

We compute

$$\mathrm{Var}\left(\frac{1}{P}\sum_{j=1}^{P} Y_j^{(D)}\right) = \frac{1}{P^2}\sum_{j=1}^{P}\mathrm{Var}(Y_j^{(D)}) + \frac{1}{P^2}\sum_{\substack{j,k=1\\ j\neq k}}^{P}\mathrm{Cov}(Y_j^{(D)}, Y_k^{(D)}).$$

By exchangeability,

$$\mathrm{Var}(Y_j^{(D)}) = \mathrm{Var}(Y_1^{(D)}) \qquad \text{for all } j,$$

and

$$\mathrm{Cov}(Y_j^{(D)}, Y_k^{(D)}) = \mathrm{Cov}(Y_1^{(D)}, Y_2^{(D)}) \qquad \text{for all } j \neq k.$$

Therefore

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{P}\sum_{j=1}^{P} Y_j^{(D)}\right) &= \frac{P}{P^2}\,\mathrm{Var}(Y_1^{(D)}) + \frac{P(P-1)}{P^2}\,\mathrm{Cov}(Y_1^{(D)}, Y_2^{(D)}) \\ &= \frac{1}{P}\,\mathrm{Var}(Y_1^{(D)}) + \frac{P-1}{P}\,\mathrm{Cov}(Y_1^{(D)}, Y_2^{(D)}).\end{aligned}$$

Taking absolute values in the covariance term and using $(P-1)/P \leq 1$, we obtain

$$\mathrm{Var}\left(\frac{1}{P}\sum_{j=1}^{P} Y_j^{(D)}\right) \leq \frac{1}{P}\,\mathrm{Var}(Y_1^{(D)}) + \left|\mathrm{Cov}(Y_1^{(D)}, Y_2^{(D)})\right|.$$

Finally,

$$\mathrm{Var}(Y_1^{(D)}) = \mathbb{E}\left[\left(Y_1^{(D)} - \mathbb{E}Y_1^{(D)}\right)^2\right] \leq \mathbb{E}\left[(Y_1^{(D)})^2\right] \leq \|F\|_\infty^2.$$

Thus

$$\mathrm{Var}\left(\frac{1}{P}\sum_{j=1}^{P} F(S_{j,A}^{(D)}, S_{j,B}^{(D)})\right) \leq \frac{\|F\|_\infty^2}{P} + \left|\mathrm{Cov}(F_1^{(D)}, F_2^{(D)})\right|,$$

where

$$F_j^{(D)} := F(S_{j,A}^{(D)}, S_{j,B}^{(D)}).$$

For two fixed coordinates, the joint occupancy counts are multinomially coupled, but the coupling is weak. As $P/D \to \phi \in (0, \infty)$, the law of the clean and poisoned counts in two fixed coordinates converges to the law of two independent copies of the limiting compound-Poisson coordinate statistic. Hence,

$$\mathrm{Cov}(F_1^{(D)}, F_2^{(D)}) \to 0.$$

For $F_{\phi_D,\varrho}$, boundedness follows from equation 33. If $\varrho > 0$, the map $(\phi, s_A, s_B) \mapsto (c - G_{\phi,\varrho}(s_A, s_B))^2$ is continuous. If $\varrho = 0$, the only problematic point is $(s_A, s_B) = (0, 0)$. This point has positive mass, but it is not a problem because the finite-D and limiting conventions both assign the same value $G = 0$ there; on $s_A > 0$, the ridgeless map is s_B/s_A and is continuous. The same expectation and covariance argument therefore applies to $F_{\phi_D,\varrho}$, giving the claimed deterministic equivalent. □

In many high-dimensional ridge-regression deterministic equivalents, one sees a nontrivial effective regularization parameter. This often appears through a self-consistent equation involving a Stieltjes transform or an equivalent quantity. It is therefore natural to ask why, in the occupancy deterministic equivalent below, the ridge parameter remains the same parameter ϱ:

$$F_{\phi_D,\varrho}(s_A, s_B) := (c - G_{\phi_D,\varrho}(s_A, s_B))^2 \longrightarrow F_{\phi,\varrho}(s_A, s_B),$$

rather than becoming some renormalized quantity. The reason is that each coordinate has its own scalar normal equation:

$$\varrho + \phi_D S_{j,A}^{(D)}.$$

The denominator is already scalar before taking the limit. Thus there is no resolvent self-consistency equation.

Result 4 (Fixed-aspect-ratio deterministic equivalent for the empirical risk)**.** *Assume $P/D \to \phi \in (0, \infty)$ and fix $\epsilon \in [0, 1)$, $\tau \geq \tau_0$, and $\varrho \geq 0$. Then*

$$\mathcal{R}_{\mathrm{clean},P}(\widehat{\theta}_{\epsilon,\tau,\varrho}) - \mathcal{R}_{\mathrm{clean},P}(\theta_{\star,P}) \xrightarrow{\mathbb{P}} \mathcal{R}_{\phi,\varrho}(\epsilon, \tau),$$

where $\mathcal{R}_{\phi,\varrho}$ is defined in equation 36. Consequently, if an independent clean baseline estimator is trained with the same D, P, τ, ϱ, then

$$\Big[\mathcal{R}_{\mathrm{clean},P}(\widehat{\theta}_{\epsilon,\tau,\varrho}) - \mathcal{R}_{\mathrm{clean},P}(\widehat{\theta}_{0,\tau,\varrho})\Big] \xrightarrow{\mathbb{P}} \mathcal{D}_{\phi,\varrho}(\epsilon, \tau) := \mathcal{R}_{\phi,\varrho}(\epsilon, \tau) - \mathcal{R}_{\phi,\varrho}(0, \tau).$$

Proof. Recall first the clean risk identity

$$\mathcal{R}_{\mathrm{clean},P}(\theta) - \mathcal{R}_{\mathrm{clean},P}(\theta_{\star,P}) = \mu_U \frac{1}{P} \sum_{j=1}^{P} \left(c - \sqrt{P}\,\theta_j\right)^2.$$

Thus, to compute the clean excess risk of the estimator $\widehat{\theta}_{\epsilon,\tau,\varrho}$, it is enough to understand the coordinate-wise quantities

$$\sqrt{P}\,(\widehat{\theta}_{\epsilon,\tau,\varrho})_j.$$

For coordinate j, the diagonal entry of the empirical Hessian and the corresponding entry of the empirical cross-moment are

$$(\widehat{\Sigma}_{\epsilon,\tau})_{jj} = \phi_D S_{j,A}^{(D)}, \qquad (\widehat{v}_{\epsilon,\tau})_j = \frac{\phi_D}{\sqrt{P}} S_{j,B}^{(D)}, \qquad \phi_D := \frac{P}{D}.$$

Therefore the ridge normal equation in coordinate j is the scalar equation

$$\left(\varrho + \phi_D S_{j,A}^{(D)}\right)(\widehat{\theta}_{\epsilon,\tau,\varrho})_j = \frac{\phi_D}{\sqrt{P}} S_{j,B}^{(D)}.$$

Equivalently,

$$\sqrt{P}\,(\widehat{\theta}_{\epsilon,\tau,\varrho})_j = G_{\phi_D,\varrho}\left(S_{j,A}^{(D)}, S_{j,B}^{(D)}\right),$$

where

$$G_{\phi,\varrho}(s_A, s_B) = \begin{cases} \dfrac{\phi s_B}{\varrho + \phi s_A}, & \varrho + \phi s_A > 0, \\ 0, & \varrho = 0,\ s_A = 0. \end{cases}$$

Substituting the coordinate formula into the clean risk identity gives the exact finite-D, P identity

$$\begin{aligned}&\mathcal{R}_{\mathrm{clean},P}(\widehat{\theta}_{\epsilon,\tau,\varrho}) - \mathcal{R}_{\mathrm{clean},P}(\theta_{\star,P}) \\ &\quad = \mu_U \frac{1}{P}\sum_{j=1}^{P}\left(c - G_{\phi_D,\varrho}\left(S_{j,A}^{(D)}, S_{j,B}^{(D)}\right)\right)^2.\end{aligned}$$

Thus the risk is exactly an average over coordinates of the bounded function

$$F_{\phi_D,\varrho}(s_A, s_B) := \left(c - G_{\phi_D,\varrho}(s_A, s_B)\right)^2.$$

Applying Lemma 2 gives

$$\begin{aligned}&\frac{1}{P}\sum_{j=1}^{P}\left(c - G_{\phi_D,\varrho}\left(S_{j,A}^{(D)}, S_{j,B}^{(D)}\right)\right)^2 \\ &\qquad \xrightarrow{\mathbb{P}} \mathbb{E}\left[\left(c - G_{\phi,\varrho}\left(S_A^{\epsilon,\tau}, S_B^{\epsilon,\tau}\right)\right)^2\right].\end{aligned}$$

Multiplying by μ_U, we obtain

$$\mathcal{R}_{\mathrm{clean},P}(\widehat{\theta}_{\epsilon,\tau,\varrho}) - \mathcal{R}_{\mathrm{clean},P}(\theta_{\star,P}) \xrightarrow{\mathbb{P}} \mu_U \mathbb{E}\left[\left(c - G_{\phi,\varrho}\left(S_A^{\epsilon,\tau}, S_B^{\epsilon,\tau}\right)\right)^2\right].$$

By definition, the right-hand side is

$$\mathcal{R}_{\phi,\varrho}(\epsilon, \tau).$$

This proves the first result.

For the second convergence, apply the same result twice. First, apply it to the estimator trained with poisoning rate ϵ:

$$\mathcal{R}_{\mathrm{clean},P}(\widehat{\theta}_{\epsilon,\tau,\varrho}) - \mathcal{R}_{\mathrm{clean},P}(\theta_{\star,P}) \xrightarrow{\mathbb{P}} \mathcal{R}_{\phi,\varrho}(\epsilon, \tau).$$

Second, apply it to the clean estimator trained with poisoning rate 0:

$$\mathcal{R}_{\mathrm{clean},P}(\widehat{\theta}_{0,\tau,\varrho}) - \mathcal{R}_{\mathrm{clean},P}(\theta_{\star,P}) \xrightarrow{\mathbb{P}} \mathcal{R}_{\phi,\varrho}(0, \tau).$$

Subtracting the two convergences gives

$$\begin{aligned}&\mathcal{R}_{\mathrm{clean},P}(\widehat{\theta}_{\epsilon,\tau,\varrho}) - \mathcal{R}_{\mathrm{clean},P}(\widehat{\theta}_{0,\tau,\varrho}) \\ &\qquad \xrightarrow{\mathbb{P}} \mathcal{R}_{\phi,\varrho}(\epsilon, \tau) - \mathcal{R}_{\phi,\varrho}(0, \tau) = \mathcal{D}_{\phi,\varrho}(\epsilon, \tau).\end{aligned}$$

This proves the second result. □

A.5 Fixed ϕ: higher dimensional statistics

The notation

$$\Pi \sim \mathrm{PPP}(\nu)$$

means that Π is a Poisson point process with intensity measure ν. Each element of Π is a data

$$z = (z_A, z_B),$$

A realization of Π is a random finite collection of data:

$$\Pi = \{Z_1, \ldots, Z_N\}.$$

The associated total coordinate statistic is

$$\Sigma(\Pi) := \sum_{z \in \Pi} z = \sum_{k=1}^{N} Z_k.$$

If $Z_k = (Z_{k,A}, Z_{k,B})$, then

$$\Sigma(\Pi) = \left(\sum_{k=1}^{N} Z_{k,A}, \sum_{k=1}^{N} Z_{k,B} \right).$$

For example

$$Z_{\mathrm{c}}^{(\tau)} = \big(a_\tau(U), r_\tau(U)\big), \qquad Z_{\mathrm{p}}^{(\tau)} = \big(1, c + \eta_{\mathrm{p}}\tau\big)$$

denote one clean and one poison data. The first component is the contribution to the coordinate Hessian statistic, and the second component is the contribution to the coordinate cross-moment statistic.

We focus on the intensity measure of the form

$$\nu = \rho\, \mathcal{L}(Z),$$

where

$$\rho > 0$$

is a rate and $\mathcal{L}(Z)$ is the probability law of a single data Z.

The notation

$$\Pi \sim \mathrm{PPP}\big(\rho\, \mathcal{L}(Z)\big)$$

means the following concrete construction:

$$N \sim \mathrm{Pois}(\rho),$$

and, conditional on N,

$$Z_1, \ldots, Z_N \quad \text{are i.i.d. with law} \quad \mathcal{L}(Z).$$

Then

$$\Pi = \{Z_1, \ldots, Z_N\}.$$

For example

$$\Pi_{\mathrm{ret}} \sim \mathrm{PPP}\left(\frac{1-\epsilon}{\phi}\, \mathcal{L}(Z_{\mathrm{c}}^{(\tau)}) \right).$$

is the random collection of clean data hitting one coordinate. The number of retained clean marks is

$$N_{\mathrm{ret}} \sim \mathrm{Pois}\left(\frac{1-\epsilon}{\phi} \right).$$

Conditional on N_{ret}, the retained clean marks are independent copies of

$$Z_{\mathrm{c}}^{(\tau)} = \big(a_\tau(U), r_\tau(U)\big).$$

Thus one can write

$$\Pi_{\mathrm{ret}} = \left\{ Z_{\mathrm{c},1}^{(\tau)}, \ldots, Z_{\mathrm{c},N_{\mathrm{ret}}}^{(\tau)} \right\}.$$

The corresponding coordinate statistic is

$$S_{\mathrm{ret}} := \Sigma(\Pi_{\mathrm{ret}}) = \sum_{\ell=1}^{N_{\mathrm{ret}}} Z_{\mathrm{c},\ell}^{(\tau)}.$$

Writing this sum coordinate-wise gives

$$S_{\mathrm{ret}} = \left(\sum_{\ell=1}^{N_{\mathrm{ret}}} a_\tau(U_\ell), \sum_{\ell=1}^{N_{\mathrm{ret}}} r_\tau(U_\ell) \right).$$

Similarly,

$$\Pi_{\mathrm{del}} \sim \mathrm{PPP}\left(\frac{\epsilon}{\phi}\, \mathcal{L}(Z_{\mathrm{c}}^{(\tau)}) \right)$$

means that the deleted clean process contains

$$N_{\mathrm{del}} \sim \mathrm{Pois}\left(\frac{\epsilon}{\phi}\right)$$

clean marks, each distributed as $Z_{\mathrm{c}}^{(\tau)}$.

The poison process

$$\Pi_{\mathrm{p}} \sim \mathrm{PPP}\left(\frac{\epsilon}{\phi}\,\mathcal{L}(Z_{\mathrm{p}}^{(\tau)})\right)$$

means that the number of poison marks is

$$N_{\mathrm{p}} \sim \mathrm{Pois}\left(\frac{\epsilon}{\phi}\right),$$

and each poison data has law

$$\mathcal{L}(Z_{\mathrm{p}}^{(\tau)}).$$

In the attack, this data is deterministic:

$$Z_{\mathrm{p}}^{(\tau)} = \big(1, c + \eta_{\mathrm{p}}\tau\big).$$

Thus $\mathcal{L}(Z_{\mathrm{p}}^{(\tau)})$ is just a point mass at $(1, c + \eta_{\mathrm{p}}\tau)$.

If

$$\Pi_1 \sim \mathrm{PPP}(\nu_1), \qquad \Pi_2 \sim \mathrm{PPP}(\nu_2)$$

are independent, then

$$\Pi_1 \cup \Pi_2 \sim \mathrm{PPP}(\nu_1 + \nu_2).$$

Therefore

$$\Pi_{\mathrm{ret}} \cup \Pi_{\mathrm{del}} \sim \mathrm{PPP}\left(\frac{1-\epsilon}{\phi}\mathcal{L}(Z_{\mathrm{c}}^{(\tau)}) + \frac{\epsilon}{\phi}\mathcal{L}(Z_{\mathrm{c}}^{(\tau)})\right).$$

Combining the two terms gives

$$\Pi_{\mathrm{ret}} \cup \Pi_{\mathrm{del}} \sim \mathrm{PPP}\left(\frac{1}{\phi}\mathcal{L}(Z_{\mathrm{c}}^{(\tau)})\right).$$

This is exactly the clean coordinate process of rate $1/\phi$, i. e., if the D samples choose their coordinates uniformly from $\{1, \ldots, P\}$, then the number of samples falling on a fixed coordinate is approximately

$$\mathrm{Binomial}\left(D, \frac{1}{P}\right) \Longrightarrow \mathrm{Pois}\left(\frac{1}{\phi}\right).$$

We denote by

$$S_0^{(\tau)} := \sum_{\ell=1}^{N_0} Z_{\mathrm{c},\ell}^{(\tau)}, \qquad N_0 \sim \mathrm{Pois}(1/\phi),$$

the clean coordinate environment seen by a typical coordinate in the unpoisoned fixed-ϕ limit. It is also useful to define

$$Q_0 := \sum_{\ell=1}^{N_0} U_\ell,$$

and

$$Q_+ := U_0 + \sum_{\ell=1}^{N_0} U_\ell = U_0 + Q_0,$$

where $U_0, U_1, U_2, \ldots$ are independent copies of U, independent of N_0. The variable Q_0 is the untruncated total radial mass already present at the coordinate. The variable Q_+ is the untruncated total radial mass after adding one additional independent clean example. This distinction is important because the small-poisoning expansion compares two possible one-data perturbations of the same clean environment:

$$S_0^{(\tau)} + Z_{\mathrm{p}}^{(\tau)} \qquad \text{versus} \qquad S_0^{(\tau)} + Z_{\mathrm{c}}^{(\tau)}.$$

The first expression corresponds to adding one poison data, while the second corresponds to adding one additional clean data.

Lemma 3 (Large-τ limit of the added-clean term)**.** *For fixed $\phi \in (0, \infty)$ and $\varrho \geq 0$,*[2]

$$\mathbb{E}\left[F_{\phi,\varrho}\left(S_0^{(\tau)} + Z_{\mathrm{c}}^{(\tau)}\right)\right] = c^2\varrho^2\,\mathbb{E}\left[\frac{1}{(\varrho + \phi Q_+)^2}\right] + O_{\phi,\varrho}(\tau^{-q_\star}).$$

Proof. Let $\mathcal{E}_\tau$ be the event that all clean radii appearing in $S_0^{(\tau)} + Z_{\mathrm{c}}^{(\tau)}$ are at most τ. On $\mathcal{E}_\tau$, no truncation occurs for either the matrix statistic or the vector statistic, because $\tau/c > \tau$. Thus, on this event,

$$S_A = Q_+, \qquad S_B = cQ_+,$$

and therefore

$$F_{\phi,\varrho}(S_A, S_B) = \left(c - \frac{\phi c Q_+}{\varrho + \phi Q_+}\right)^2 = \frac{c^2\varrho^2}{(\varrho + \phi Q_+)^2}.$$

Let

$$M := 1 + N_0, \qquad N_0 \sim \mathrm{Pois}(1/\phi),$$

be the total number of clean radii appearing in $S_0^{(\tau)} + Z_{\mathrm{c}}^{(\tau)}$. Write these radii as

$$U_1, \ldots, U_M,$$

where, conditionally on M, they are i.i.P. copies of U. Recall that

$$\mathcal{E}_\tau := \left\{\max_{1 \leq \ell \leq M} U_\ell \leq \tau\right\}.$$

Thus

$$\mathcal{E}_\tau^c = \left\{\exists\, 1 \leq \ell \leq M \text{ such that } U_\ell > \tau\right\}.$$

We first bound the probability of this event. Conditional on $M = m$,

$$\begin{aligned}\mathbb{P}(\mathcal{E}_\tau^c \mid M = m) &= 1 - \mathbb{P}(U_1 \leq \tau, \ldots, U_m \leq \tau) \\ &= 1 - (1 - \mathbb{P}(U > \tau))^m.\end{aligned}$$

Since $1 - (1-u)^m \leq mu$ for $u \in [0, 1]$, we get

$$\mathbb{P}(\mathcal{E}_\tau^c \mid M = m) \leq m\,\mathbb{P}(U > \tau).$$

Taking expectations over M,

$$\begin{aligned}\mathbb{P}(\mathcal{E}_\tau^c) &\leq \mathbb{E}[M]\,\mathbb{P}(U > \tau) \\ &= \left(1 + \frac{1}{\phi}\right)\tau^{-q_\star}.\end{aligned}$$

Therefore

$$\mathbb{P}(\mathcal{E}_\tau^c) = O_\phi(\tau^{-q_\star}).$$

It remains to show that the integrand for the bad events is bounded. On a clean-only configuration, the coordinate statistics have the form

$$S_A = \sum_{\ell=1}^{M} a_\tau(U_\ell), \qquad S_B = \sum_{\ell=1}^{M} r_\tau(U_\ell),$$

[2] Let $A_{\phi,\varrho}(\tau)$ be a quantity depending on τ, ϕ, and ϱ. We write

$$A_{\phi,\varrho}(\tau) = O_{\phi,\varrho}(\tau^{-q_\star}) \qquad \text{as } \tau \to \infty$$

if, for every fixed ϕ and every fixed ϱ, there exist constants

$$C_{\phi,\varrho} < \infty \qquad \text{and} \qquad \tau_{\phi,\varrho} < \infty$$

such that, for all $\tau \geq \tau_{\phi,\varrho}$,

$$|A_{\phi,\varrho}(\tau)| \leq C_{\phi,\varrho}\tau^{-q_\star}.$$

The subscript in $O_{\phi,\varrho}$ means that the implicit constant may depend on ϕ and ϱ, but it not on τ.

where

$$a_\tau(t) = t \wedge \tau, \qquad r_\tau(t) = c\left(t \wedge \frac{\tau}{c}\right).$$

For every $t \geq 1$,

$$0 \leq r_\tau(t) \leq a_\tau(t).$$

Indeed, if $t \leq \tau$, then $r_\tau(t) = ct \leq t = a_\tau(t)$. If $\tau < t < \tau/c$, then $r_\tau(t) = ct \leq \tau = a_\tau(t)$. Finally, if $t \geq \tau/c$, then $r_\tau(t) = \tau = a_\tau(t)$. Hence

$$0 \leq S_B \leq S_A.$$

Consequently, for $\varrho \geq 0$,

$$0 \leq G_{\phi,\varrho}(S_A, S_B) = \frac{\phi S_B}{\varrho + \phi S_A} \leq \frac{\phi S_A}{\varrho + \phi S_A} \leq 1$$

whenever $\varrho + \phi S_A > 0$. In the ridgeless convention $\varrho = 0$, if $S_A = 0$ then $G_{\phi,0}(0,0) = 0$, so the same bound still holds. Therefore

$$0 \leq G_{\phi,\varrho}(S_A, S_B) \leq 1.$$

Since $0 < c < 1$,

$$F_{\phi,\varrho}(S_A, S_B) = \left(c - G_{\phi,\varrho}(S_A, S_B)\right)^2 \leq 1.$$

Thus the contribution of the bad event is bounded by

$$\begin{aligned}\mathbb{E}\left[F_{\phi,\varrho}(S_A, S_B)\mathbf{1}_{\mathcal{E}_\tau^c}\right] &\leq \mathbb{P}(\mathcal{E}_\tau^c) \\ &\leq \left(1 + \frac{1}{\phi}\right)\tau^{-q_\star}.\end{aligned}$$

Hence

$$\mathbb{E}\left[F_{\phi,\varrho}(S_A, S_B)\mathbf{1}_{\mathcal{E}_\tau^c}\right] = O_\phi(\tau^{-q_\star}).$$

□

Lemma 4 (Large-τ expansion with one modified poison data)**.** *Fix $\phi \in (0, \infty)$ and $\varrho \geq 0$. Then*

$$\begin{aligned}&\mathbb{E}\left[F_{\phi,\varrho}\left(S_0^{(\tau)} + Z_{\mathrm{p}}^{(\tau)}\right)\right] \\ &\quad = \eta_{\mathrm{p}}^2\phi^2\tau^2\mathbb{E}\left[\frac{1}{(\varrho + \phi(Q_0 + 1))^2}\right](1 + o(1))\end{aligned}$$

as $\tau \to \infty$.

Proof. Write

$$A_\tau := \sum_{\ell=1}^{N_0} a_\tau(U_\ell), \qquad B_\tau := \sum_{\ell=1}^{N_0} r_\tau(U_\ell), \qquad D_\tau := B_\tau - cA_\tau.$$

After adding one modified poison data, the coordinate statistic is

$$(A_\tau + 1,\ B_\tau + c + \eta_{\mathrm{p}}\tau).$$

Using the definition of $G_{\phi,\varrho}$,

$$\begin{aligned}&G_{\phi,\varrho}(A_\tau + 1, B_\tau + c + \eta_{\mathrm{p}}\tau) - c \\ &\quad = \frac{\phi(B_\tau + c + \eta_{\mathrm{p}}\tau) - c(\varrho + \phi(A_\tau + 1))}{\varrho + \phi(A_\tau + 1)} \\ &\quad = \frac{\phi D_\tau + \phi\eta_{\mathrm{p}}\tau - c\varrho}{\varrho + \phi(A_\tau + 1)}.\end{aligned}$$

Therefore

$$F_{\phi,\varrho}(A_\tau + 1, B_\tau + c + \eta_{\mathrm{p}}\tau) = \left(\frac{\phi D_\tau + \phi\eta_{\mathrm{p}}\tau - c\varrho}{\varrho + \phi(A_\tau + 1)}\right)^2.$$

For each fixed realization of $(N_0, U_1, U_2, \ldots)$,

$$A_\tau \to Q_0, \qquad \frac{D_\tau}{\tau} \to 0,$$

because for each fixed U_ℓ, truncation eventually disappears. Hence

$$\frac{1}{\tau^2} F_{\phi,\varrho}(A_\tau + 1, B_\tau + c + \eta_{\mathrm{p}}\tau) \to \eta_{\mathrm{p}}^2 \phi^2 \frac{1}{(\varrho + \phi(Q_0 + 1))^2}$$

almost surely.

□

Lemma 5 (Small-poisoning generator expansion)**.** *For every $\tau \geq \tau_0$, as $\epsilon \downarrow 0$*

$$\mathcal{D}_{\phi,\varrho}(\epsilon, \tau) = \frac{\mu_U \epsilon}{\phi} \mathbb{E}\left[F_{\phi,\varrho}\left(S_0^{(\tau)} + Z_{\mathrm{p}}^{(\tau)}\right) - F_{\phi,\varrho}\left(S_0^{(\tau)} + Z_{\mathrm{c}}^{(\tau)}\right)\right] + O_{\phi,\varrho,c,\eta_{\mathrm{p}}}\left(\epsilon^2(1+\tau^2)\right).$$

The implicit constant is independent of both ϵ and τ.

Proof. Construct three independent Poisson point processes (following previous discussion):

$$\Pi_{\mathrm{ret}} \sim \mathrm{PPP}\left(\frac{1-\epsilon}{\phi}\, \mathcal{L}(Z_{\mathrm{c}}^{(\tau)})\right),$$

$$\Pi_{\mathrm{del}} \sim \mathrm{PPP}\left(\frac{\epsilon}{\phi}\, \mathcal{L}(Z_{\mathrm{c}}^{(\tau)})\right),$$

and

$$\Pi_{\mathrm{p}} \sim \mathrm{PPP}\left(\frac{\epsilon}{\phi}\, \mathcal{L}(Z_{\mathrm{p}}^{(\tau)})\right).$$

Define

$$S_{\mathrm{ret}} := \Sigma(\Pi_{\mathrm{ret}}), \qquad P_{\mathrm{c}} := \Sigma(\Pi_{\mathrm{del}}), \qquad P_{\mathrm{p}} := \Sigma(\Pi_{\mathrm{p}}).$$

By the superposition theorem for Poisson point processes,

$$\Pi_{\mathrm{ret}} \cup \Pi_{\mathrm{del}} \sim \mathrm{PPP}\left(\frac{1}{\phi}\, \mathcal{L}(Z_{\mathrm{c}}^{(\tau)})\right).$$

Hence

$$S_{\mathrm{ret}} + P_{\mathrm{c}} \overset{P}{=} S_0^{(\tau)}.$$

Likewise, $\Pi_{\mathrm{ret}} \cup \Pi_{\mathrm{p}}$ is the compound-Poisson process corresponding to a coordinate whose clean marks arrive with intensity $(1-\epsilon)/\phi$ and whose poison marks arrive with intensity ϵ/ϕ. Therefore

$$S_{\mathrm{ret}} + P_{\mathrm{p}} \overset{P}{=} S^{\epsilon,\tau},$$

where $S^{\epsilon,\tau} = (S_A^{\epsilon,\tau}, S_B^{\epsilon,\tau})$ is the limiting coordinate statistic under poisoning. Consequently

$$\begin{aligned}\mathcal{D}_{\phi,\varrho}(\epsilon,\tau) &= \mathcal{R}_{\phi,\varrho}(\epsilon,\tau) - \mathcal{R}_{\phi,\varrho}(0,\tau) \\ &= \mu_U \mathbb{E}\left[F(S_{\mathrm{ret}} + P_{\mathrm{p}}) - F(S_{\mathrm{ret}} + P_{\mathrm{c}})\right].\end{aligned}$$

We now expand the last expectation to first order in ϵ. Let

$$K_{\mathrm{c}} := |\Pi_{\mathrm{del}}|, \qquad K_{\mathrm{p}} := |\Pi_{\mathrm{p}}|, \qquad K := K_{\mathrm{c}} + K_{\mathrm{p}}.$$

Then

$$K_{\mathrm{c}} \sim \mathrm{Pois}(\epsilon/\phi), \qquad K_{\mathrm{p}} \sim \mathrm{Pois}(\epsilon/\phi),$$

independently, and hence

$$K \sim \mathrm{Pois}(2\epsilon/\phi).$$

Therefore

$$\mathbb{P}(K \geq 2) = 1 - e^{-2\epsilon/\phi}\left(1 + \frac{2\epsilon}{\phi}\right) = O_\phi(\epsilon^2).$$

For $\tau \geq \tau_0$, every clean or poison data satisfies the bound

$$0 \leq s_B \leq m_\tau s_A, \qquad m_\tau := \max\{1, c + \eta_{\mathrm{p}}\tau\} \leq C_{c,\eta_{\mathrm{p}}}\tau.$$

Hence

$$0 \leq G_{\phi,\varrho}(s_A, s_B) \leq C_{c,\eta_{\mathrm{p}}}\tau,$$

and therefore

$$\|F_{\phi,\varrho}\|_\infty \leq C_{c,\eta_{\mathrm{p}}}(1+\tau^2).$$

It follows that the contribution to

$$\mathbb{E}\left[F(S_{\mathrm{ret}} + P_{\mathrm{p}}) - F(S_{\mathrm{ret}} + P_{\mathrm{c}})\right]$$

from the event $\{K \geq 2\}$ is bounded in absolute value by

$$2\|F\|_\infty \mathbb{P}(K \geq 2) = O_{\phi,c,\eta_{\mathrm{p}}}\left(\epsilon^2(1+\tau^2)\right).$$

On the event $\{K = 0\}$, we have $P_{\mathrm{p}} = P_{\mathrm{c}} = 0$. Hence the two terms are equal:

$$F(S_{\mathrm{ret}} + P_{\mathrm{p}}) - F(S_{\mathrm{ret}} + P_{\mathrm{c}}) = F(S_{\mathrm{ret}}) - F(S_{\mathrm{ret}}) = 0.$$

Thus the zero-data event contributes nothing.

It remains to compute the contribution of the event $\{K = 1\}$. There are two mutually exclusive one-data cases:

$$\{K_{\mathrm{p}} = 1, K_{\mathrm{c}} = 0\}, \qquad \{K_{\mathrm{p}} = 0, K_{\mathrm{c}} = 1\}.$$

Since K_{p} and K_{c} are independent Poisson variables with mean ϵ/ϕ,

$$\mathbb{P}(K_{\mathrm{p}} = 1, K_{\mathrm{c}} = 0) = e^{-2\epsilon/\phi}\frac{\epsilon}{\phi},$$

and

$$\mathbb{P}(K_{\mathrm{p}} = 0, K_{\mathrm{c}} = 1) = e^{-2\epsilon/\phi}\frac{\epsilon}{\phi}.$$

Conditioned on $\{K_{\mathrm{p}} = 1, K_{\mathrm{c}} = 0\}$, the unique poison data has law $Z_{\mathrm{p}}^{(\tau)}$, is independent of S_{ret}, and $P_{\mathrm{c}} = 0$. Hence the conditional contribution is

$$\mathbb{E}\left[F(S_{\mathrm{ret}} + Z_{\mathrm{p}}^{(\tau)}) - F(S_{\mathrm{ret}})\right].$$

Conditioned on $\{K_{\mathrm{p}} = 0, K_{\mathrm{c}} = 1\}$, the unique deleted clean data has law $Z_{\mathrm{c}}^{(\tau)}$, is independent of S_{ret}, and $P_{\mathrm{p}} = 0$. Hence the conditional contribution is

$$\mathbb{E}\left[F(S_{\mathrm{ret}}) - F(S_{\mathrm{ret}} + Z_{\mathrm{c}}^{(\tau)})\right].$$

Adding the two one-data contributions gives

$$\begin{aligned}&\mathbb{E}\left[F(S_{\mathrm{ret}} + P_{\mathrm{p}}) - F(S_{\mathrm{ret}} + P_{\mathrm{c}}); K = 1\right]\\ &\quad = e^{-2\epsilon/\phi}\frac{\epsilon}{\phi}\mathbb{E}\left[F(S_{\mathrm{ret}} + Z_{\mathrm{p}}^{(\tau)}) - F(S_{\mathrm{ret}} + Z_{\mathrm{c}}^{(\tau)})\right].\end{aligned}$$

Since

$$e^{-2\epsilon/\phi} = 1 + O_\phi(\epsilon),$$

and since the expectation in the last display is bounded in absolute value by $2\|F\|_\infty = O_{c,\eta_{\mathrm{p}}}(1+\tau^2)$, we may replace $e^{-2\epsilon/\phi}$ by 1 at cost

$$O_{\phi,c,\eta_{\mathrm{p}}}\left(\epsilon^2(1+\tau^2)\right).$$

Therefore

$$\begin{aligned}&\mathbb{E}\left[F(S_{\mathrm{ret}} + P_{\mathrm{p}}) - F(S_{\mathrm{ret}} + P_{\mathrm{c}})\right]\\ &\quad = \frac{\epsilon}{\phi}\mathbb{E}\left[F(S_{\mathrm{ret}} + Z_{\mathrm{p}}^{(\tau)}) - F(S_{\mathrm{ret}} + Z_{\mathrm{c}}^{(\tau)})\right] + O_{\phi,c,\eta_{\mathrm{p}}}\left(\epsilon^2(1+\tau^2)\right).\end{aligned}$$

Finally, we replace S_{ret} by $S_0^{(\tau)}$. Couple $S_0^{(\tau)}$ as

$$S_0^{(\tau)} = S_{\mathrm{ret}} + P'_{\mathrm{c}},$$

where P'_{c} is an independent clean Poisson process with intensity ϵ/ϕ and data law $Z_{\mathrm{c}}^{(\tau)}$. Then

$$\mathbb{P}(P'_{\mathrm{c}} \neq 0) = 1 - e^{-\epsilon/\phi} = O_\phi(\epsilon).$$

Therefore, for a random data Z independent of all processes and equal in law to either $Z_{\mathrm{p}}^{(\tau)}$ or $Z_{\mathrm{c}}^{(\tau)}$,

$$\begin{aligned}\left|\mathbb{E}F(S_{\mathrm{ret}} + Z) - \mathbb{E}F(S_0^{(\tau)} + Z)\right| &\leq 2\|F\|_\infty \mathbb{P}(P'_{\mathrm{c}} \neq 0) \\ &= O_{\phi,c,\eta_{\mathrm{p}}}\left(\epsilon(1+\tau^2)\right).\end{aligned}$$

Applying this once with $Z = Z_{\mathrm{p}}^{(\tau)}$ and once with $Z = Z_{\mathrm{c}}^{(\tau)}$, we obtain

$$\begin{aligned}&\mathbb{E}\left[F(S_{\mathrm{ret}} + Z_{\mathrm{p}}^{(\tau)}) - F(S_{\mathrm{ret}} + Z_{\mathrm{c}}^{(\tau)})\right] \\ &\quad = \mathbb{E}\left[F(S_0^{(\tau)} + Z_{\mathrm{p}}^{(\tau)}) - F(S_0^{(\tau)} + Z_{\mathrm{c}}^{(\tau)})\right] + O_{\phi,c,\eta_{\mathrm{p}}}\left(\epsilon(1+\tau^2)\right).\end{aligned}$$

Multiplying this replacement error by the outside factor ϵ/ϕ produces

$$O_{\phi,c,\eta_{\mathrm{p}}}\left(\epsilon^2(1+\tau^2)\right).$$

Combining all previous displays gives

$$\begin{aligned}&\mathbb{E}\left[F(S_{\mathrm{ret}} + P_{\mathrm{p}}) - F(S_{\mathrm{ret}} + P_{\mathrm{c}})\right] \\ &\quad = \frac{\epsilon}{\phi}\mathbb{E}\left[F(S_0^{(\tau)} + Z_{\mathrm{p}}^{(\tau)}) - F(S_0^{(\tau)} + Z_{\mathrm{c}}^{(\tau)})\right] + O_{\phi,\varrho,c,\eta_{\mathrm{p}}}\left(\epsilon^2(1+\tau^2)\right).\end{aligned}$$

Multiplying by the prefactor μ_U in the definition of $\mathcal{D}_{\phi,\varrho}(\epsilon,\tau)$ proves the claimed expansion.

□

Result 5 (Fixed-aspect leading asymptotics)**.** *Fix $\phi \in (0,\infty)$ and $\varrho \geq 0$. Assume*

$$\epsilon \downarrow 0, \qquad \tau = \tau(\epsilon) \to \infty, \qquad \tau(\epsilon) \geq \tau_0.$$

Then the clean-relative degradation satisfies

$$\mathcal{D}_{\phi,\varrho}(\epsilon,\tau) = C_{\phi,\varrho}\,\epsilon\tau^2(1+o(1)),$$

where

$$C_{\phi,\varrho} := \mu_U \eta_{\mathrm{p}}^2 \phi \mathbb{E}\left[\frac{1}{(\varrho + \phi(Q_0+1))^2}\right].$$

In particular, the leading fixed-ϕ term depends on τ through $\epsilon\tau^2$.

Proof. By Lemma 5,

$$\begin{aligned}\mathcal{D}_{\phi,\varrho}(\epsilon,\tau) = \frac{\mu_U\epsilon}{\phi}\mathbb{E}\left[F_{\phi,\varrho}\left(S_0^{(\tau)} + Z_{\mathrm{p}}^{(\tau)}\right) - F_{\phi,\varrho}\left(S_0^{(\tau)} + Z_{\mathrm{c}}^{(\tau)}\right)\right] \\ + O_{\phi,\varrho}\left(\epsilon^2(1+\tau^2)\right).\end{aligned}$$

The added-clean expectation is $O_{\phi,\varrho}(1)$ by Lemma 3. The added-poison expectation is, by Lemma 4,

$$\eta_{\mathrm{p}}^2\phi^2\tau^2\mathbb{E}\left[\frac{1}{(\varrho + \phi(Q_0+1))^2}\right](1+o(1)).$$

Since $\tau \to \infty$, the $O(1)$ added-clean term is negligible relative to the τ^2 added-poison term. Also,

$$\epsilon^2(1+\tau^2) = o(\epsilon\tau^2),$$

because $\epsilon \to 0$ and $\tau \to \infty$. Multiplying the leading added-poison term by $\mu_U\epsilon/\phi$ gives the displayed constant. □

In the theorem below we work with the ridgeless risk relative to the untruncated clean fixed-aspect reference

$$\mathcal{R}_{\phi,0}(0,\infty) := \lim_{\tau\to\infty} \mathcal{R}_{\phi,0}(0,\tau).$$

Define

$$\mathcal{E}_\phi(\epsilon,\tau) := \mathcal{R}_{\phi,0}(\epsilon,\tau) - \mathcal{R}_{\phi,0}(0,\infty). \tag{37}$$

This is the fixed-ϕ analogue of comparing the poisoned truncated estimator against the ideal clean untruncated target.

For $z > 1$, define

$$h_c(z) := \begin{cases} c(z-1), & 1 < z < 1/c, \\ 1-c, & z \geq 1/c. \end{cases}$$

Set

$$K^{\mathrm{cl}}_{\phi,c,q_\star} := \frac{\mu_U}{\phi}\, q_\star \int_1^\infty h_c(z)^2 z^{-q_\star-1}\,\mathrm{d}z. \tag{38}$$

This constant is finite and strictly positive.

Lemma 6 (Clean truncation term at fixed aspect ratio)**.** *For fixed $\phi \in (0,\infty)$, in the ridgeless case $\varrho = 0$,*

$$\mathcal{R}_{\phi,0}(0,\tau) - \mathcal{R}_{\phi,0}(0,\infty) = K^{\mathrm{cl}}_{\phi,c,q_\star}\, \tau^{-q_\star} + o(\tau^{-q_\star})$$

as $\tau \to \infty$.

Proof. Recall the clean coordinate model. A sample is generated by drawing J uniformly from $\{1,\dots,P\}$, drawing U, and setting

$$X = \sqrt{PU}\, e_J, \qquad Y = c\sqrt{U}.$$

The teacher vector is

$$\theta_{\star,P} = \frac{c}{\sqrt{P}}(1,\dots,1)^\top.$$

Thus, if $J = j$, then

$$X^\top \theta_{\star,P} = \sqrt{PU}\, e_j^\top \theta_{\star,P} = c\sqrt{U} = Y.$$

Hence the clean model is noiseless at the coordinate level in this simplified calculation.

Let N_j be the number of clean training samples that hit coordinate j. For those samples, write their radial variables as

$$U_{j,1},\dots,U_{j,N_j}.$$

In the untruncated clean problem, the coordinate-level sufficient statistics are

$$S_{j,A} = \sum_{\ell=1}^{N_j} U_{j,\ell}, \qquad S_{j,B} = \sum_{\ell=1}^{N_j} cU_{j,\ell} = cS_{j,A}.$$

The empirical Hessian is diagonal. The ridgeless estimator is defined by the Moore–Penrose inverse, so

$$\sqrt{P}\,\widehat{\theta}_j = \begin{cases} \dfrac{S_{j,B}}{S_{j,A}}, & S_{j,A} > 0, \\ 0, & S_{j,A} = 0. \end{cases}$$

Since $U_{j,\ell} > 0$, the condition $S_{j,A} > 0$ is equivalent to $N_j \geq 1$.

If $N_j \geq 1$, then

$$\frac{S_{j,B}}{S_{j,A}} = \frac{cS_{j,A}}{S_{j,A}} = c.$$

Therefore

$$\sqrt{P}\,\widehat{\theta}_j = c \qquad \text{whenever } N_j \geq 1.$$

Equivalently,

$$\widehat{\theta}_j = \frac{c}{\sqrt{P}} = (\theta_{\star,P})_j \qquad \text{whenever } N_j \geq 1.$$

Thus every observed coordinate is fit exactly.

If $N_j = 0$, then

$$S_{j,A} = S_{j,B} = 0.$$

The Moore–Penrose convention gives

$$\sqrt{P}\,\widehat{\theta}_j = 0, \qquad \widehat{\theta}_j = 0.$$

But the teacher coordinate is

$$(\theta_{\star,P})_j = \frac{c}{\sqrt{P}}.$$

Hence unobserved coordinates are not fit; they are set to zero by the minimum-norm convention.

The clean excess risk identity is

$$\mathcal{R}_{\mathrm{clean},P}(\widehat{\theta}) - \mathcal{R}_{\mathrm{clean},P}(\theta_{\star,P}) = \mu_U \frac{1}{P} \sum_{j=1}^{P} \left(c - \sqrt{P}\,\widehat{\theta}_j\right)^2.$$

From the preceding calculation,

$$c - \sqrt{P}\,\widehat{\theta}_j = \begin{cases} 0, & N_j \geq 1, \\ c, & N_j = 0. \end{cases}$$

Therefore

$$\left(c - \sqrt{P}\,\widehat{\theta}_j\right)^2 = c^2 \mathbf{1}\{N_j = 0\}.$$

Substituting into the risk identity gives

$$\mathcal{R}_{\mathrm{clean},P}(\widehat{\theta}) - \mathcal{R}_{\mathrm{clean},P}(\theta_{\star,P}) = \mu_U c^2 \frac{1}{P} \sum_{j=1}^{P} \mathbf{1}\{N_j = 0\}.$$

Thus the empirical clean excess risk is exactly proportional to the fraction of unobserved coordinates.

When

$$\frac{P}{D} \to \phi \in (0, \infty),$$

the number of clean samples hitting a fixed coordinate converges to

$$N_0 \sim \mathrm{Pois}(1/\phi).$$

Hence

$$\mathbb{P}(N_0 = 0) = e^{-1/\phi}.$$

By the occupancy law of large numbers,

$$\frac{1}{P} \sum_{j=1}^{P} \mathbf{1}\{N_j = 0\} \xrightarrow{\mathbb{P}} e^{-1/\phi}.$$

Therefore, in the untruncated clean ridgeless problem,

$$\mathcal{R}_{\phi,0}(0, \infty) = \mu_U c^2 e^{-1/\phi}.$$

In the untruncated ridgeless clean problem, every observed coordinate is fit exactly and only unobserved coordinates contribute error. Hence

$$\mathcal{R}_{\phi,0}(0, \infty) = \mu_U c^2 \mathbb{P}(N_0 = 0).$$

Recall,

$$A_\tau := \sum_{\ell=1}^{N_0} a_\tau(U_\ell), \qquad B_\tau := \sum_{\ell=1}^{N_0} r_\tau(U_\ell).$$

When $A_\tau > 0$, the coordinate estimator is B_τ / A_τ; when $A_\tau = 0$, the Moore–Penrose convention gives estimator 0. Let

$$\Delta_\tau(t) := r_\tau(t) - c a_\tau(t).$$

Then

$$\Delta_\tau(t) = 0 \quad \text{for } t \le \tau,$$

and, for $t > \tau$,

$$\Delta_\tau(t) = \begin{cases} c(t-\tau), & \tau < t < \tau/c, \\ (1-c)\tau, & t \ge \tau/c. \end{cases}$$

Therefore

$$B_\tau - cA_\tau = \sum_{\ell=1}^{N_0} \Delta_\tau(U_\ell) \ge 0.$$

It follows that

$$\mathcal{R}_{\phi,0}(0,\tau) - \mathcal{R}_{\phi,0}(0,\infty) = \mu_U \, \mathbb{E}\left[\left(\frac{\sum_{\ell=1}^{N_0} \Delta_\tau(U_\ell)}{\sum_{\ell=1}^{N_0} a_\tau(U_\ell)} \right)^2 \mathbf{1}\{N_0 \ge 1\} \right], \tag{39}$$

with the ratio interpreted as zero when the denominator is zero.

The random collection of clean radii

$$\{U_1, \dots, U_{N_0}\}$$

can be viewed as a Poisson point process on $[1, \infty)$ with intensity measure

$$\nu(\,\mathrm{d}t) = \frac{1}{\phi} q_\star t^{-q_\star - 1} \,\mathrm{d}t.$$

Indeed, the total mass of ν is

$$\nu([1,\infty)) = \frac{1}{\phi} \int_1^\infty q_\star t^{-q_\star - 1} \,\mathrm{d}t = \frac{1}{\phi},$$

so the total number of points is $\mathrm{Pois}(1/\phi)$, and the normalized data law has density $q_\star t^{-q_\star - 1}$ on $[1, \infty)$.

Let

$$\Pi := \{U_1, \dots, U_{N_0}\}$$

denote this Poisson point process. Define

$$R_\tau(\Pi) := \left(\frac{\sum_{t \in \Pi} \Delta_\tau(t)}{\sum_{t \in \Pi} a_\tau(t)} \right)^2 \mathbf{1}\{|\Pi| \ge 1\},$$

with the ratio interpreted as zero when the denominator is zero.

Since $a_\tau(t) > 0$ for $t \ge 1$, the denominator is positive whenever $|\Pi| \ge 1$. Also,

$$0 \le \Delta_\tau(t) \le a_\tau(t),$$

so

$$0 \le R_\tau(\Pi) \le 1.$$

Let

$$N_\tau^+ := \#\{t \in \Pi : t > \tau\}$$

be the number of exceedances above τ. Since the intensity of (τ, ∞) is

$$\nu((\tau,\infty)) = \frac{1}{\phi} \mathbb{P}(U > \tau) = \frac{1}{\phi} \tau^{-q_\star},$$

we have

$$N_\tau^+ \sim \mathrm{Pois}\left(\frac{1}{\phi} \tau^{-q_\star} \right).$$

Therefore

$$\mathbb{P}(N_\tau^+ = 0) = 1 - O(\tau^{-q_\star}),$$

$$\mathbb{P}(N_\tau^+ = 1) = \frac{1}{\phi}\tau^{-q_\star} + O(\tau^{-2q_\star}),$$

and

$$\mathbb{P}(N_\tau^+ \geq 2) = O(\tau^{-2q_\star}).$$

On the event $N_\tau^+ = 0$, every radius is at most τ, hence

$$\Delta_\tau(U_\ell) = 0 \quad \text{for all } \ell,$$

and therefore

$$R_\tau(\Pi) = 0.$$

On the event $N_\tau^+ \geq 2$, we use only the bound $0 \leq R_\tau(\Pi) \leq 1$:

$$\mathbb{E}\left[R_\tau(\Pi)\mathbf{1}\{N_\tau^+ \geq 2\}\right] \leq \mathbb{P}(N_\tau^+ \geq 2) = O(\tau^{-2q_\star}) = o(\tau^{-q_\star}).$$

Thus the leading contribution comes only from the event of exactly one exceedance.

The Campbell–Mecke formula says that, for a Poisson point process Π with intensity measure ν,

$$\mathbb{E}\left[\sum_{t\in\Pi} g(t, \Pi \setminus \{t\})\right] = \int \mathbb{E}\left[g(t,\Pi)\right] \nu(\,\mathrm{d}t),$$

whenever $g \geq 0$ or the expectations are finite.

We apply it to the contribution from a distinguished exceedance $t > \tau$ and then note that configurations with more than one exceedance are negligible. Let

$$\Pi_{\leq\tau} := \Pi \cap [1, \tau]$$

be an independent Poisson point process on $[1, \tau]$ with intensity ν restricted to $[1, \tau]$. By the independence of Poisson processes on disjoint sets, after selecting the unique exceedance $t > \tau$, the remaining sub-threshold points have the same law as $\Pi_{\leq\tau}$, up to an error supported on events with another exceedance. Those events have probability $O(\tau^{-2q_\star})$.

Hence the one-exceedance contribution is

$$\begin{aligned} I_\tau &:= \mathbb{E}\left[R_\tau(\Pi)\mathbf{1}\{N_\tau^+ = 1\}\right] \\ &= \int_\tau^\infty \mathbb{E}\left[\left(\frac{\Delta_\tau(t)}{a_\tau(t) + \sum_{u\in\Pi_{\leq\tau}} a_\tau(u)}\right)^2\right] \nu(\,\mathrm{d}t) + O(\tau^{-2q_\star}). \end{aligned}$$

Since $t > \tau$, we have

$$a_\tau(t) = \tau.$$

Now set

$$t = \tau z, \qquad z > 1.$$

The Pareto intensity transforms as

$$\nu(\,\mathrm{d}t) = \frac{1}{\phi} q_\star t^{-q_\star - 1}\,\mathrm{d}t = \frac{1}{\phi} q_\star (\tau z)^{-q_\star - 1}\tau\,\mathrm{d}z = \frac{1}{\phi} q_\star \tau^{-q_\star} z^{-q_\star - 1}\,\mathrm{d}z.$$

Also,

$$\frac{\Delta_\tau(\tau z)}{\tau} = h_c(z).$$

Therefore

$$\begin{aligned} I_\tau &= \frac{1}{\phi} q_\star \tau^{-q_\star} \int_1^\infty \mathbb{E}\left[\left(\frac{\Delta_\tau(\tau z)}{\tau + \sum_{u\in\Pi_{\leq\tau}} a_\tau(u)}\right)^2\right] z^{-q_\star - 1}\,\mathrm{d}z + O(\tau^{-2q_\star}) \\ &= \frac{1}{\phi} q_\star \tau^{-q_\star} \int_1^\infty h_c(z)^2 \mathbb{E}\left[\left(\frac{\tau}{\tau + \sum_{u\in\Pi_{\leq\tau}} a_\tau(u)}\right)^2\right] z^{-q_\star - 1}\,\mathrm{d}z + O(\tau^{-2q_\star}). \end{aligned}$$

We claim that

$$\frac{1}{\tau}\sum_{u\in\Pi_{\le\tau}} a_\tau(u) \longrightarrow 0 \qquad \text{in probability.}$$

Indeed, on $[1,\tau]$ we have $a_\tau(u)=u$, so

$$\mathbb{E}\left[\sum_{u\in\Pi_{\le\tau}} a_\tau(u)\right] = \int_1^\tau u\,\nu(\,\mathrm{d}u) = \frac{1}{\phi}\int_1^\tau q_\star u^{-q_\star}\,\mathrm{d}u.$$

Since $q_\star > 1$,

$$\int_1^\tau q_\star u^{-q_\star}\,\mathrm{d}u \le \int_1^\infty q_\star u^{-q_\star}\,\mathrm{d}u < \infty.$$

Thus

$$\mathbb{E}\left[\sum_{u\in\Pi_{\le\tau}} a_\tau(u)\right] = O_\phi(1),$$

and

$$\frac{1}{\tau}\sum_{u\in\Pi_{\le\tau}} a_\tau(u) \xrightarrow{\mathbb{P}} 0.$$

Consequently,

$$\left(\frac{\tau}{\tau+\sum_{u\in\Pi_{\le\tau}} a_\tau(u)}\right)^2 \xrightarrow{\mathbb{P}} 1.$$

Moreover, because

$$0 \le h_c(z) \le 1,$$

and

$$\int_1^\infty z^{-q_\star-1}\,\mathrm{d}z < \infty,$$

dominated convergence also applies to the integral over z. Therefore

$$I_\tau = \frac{1}{\phi} q_\star \tau^{-q_\star}\int_1^\infty h_c(z)^2 z^{-q_\star-1}\,\mathrm{d}z + o(\tau^{-q_\star}).$$

Combining the decomposition

$$\mathbb{E}[R_\tau(\Pi)] = \mathbb{E}[R_\tau(\Pi)\mathbf{1}\{N_\tau^+=0\}] + \mathbb{E}[R_\tau(\Pi)\mathbf{1}\{N_\tau^+=1\}] + \mathbb{E}[R_\tau(\Pi)\mathbf{1}\{N_\tau^+\ge 2\}],$$

with

$$\mathbb{E}[R_\tau(\Pi)\mathbf{1}\{N_\tau^+=0\}] = 0,$$
$$\mathbb{E}[R_\tau(\Pi)\mathbf{1}\{N_\tau^+\ge 2\}] = o(\tau^{-q_\star}),$$

and the one-exceedance estimate above, we obtain

$$\mathbb{E}\left[\left(\frac{\sum_{\ell=1}^{N_0}\Delta_\tau(U_\ell)}{\sum_{\ell=1}^{N_0} a_\tau(U_\ell)}\right)^2 \mathbf{1}\{N_0\ge 1\}\right] = \frac{1}{\phi}\,q_\star\tau^{-q_\star}\int_1^\infty h_c(z)^2 z^{-q_\star-1}\,\mathrm{d}z + o(\tau^{-q_\star}).$$

□

Result 6 (Fixed-ϕ two-term risk and optimal truncation)**.** *Fix $\phi\in(0,\infty)$ and consider the ridgeless modified attack $\varrho=0$. Assume*

$$\epsilon\downarrow 0, \qquad \tau=\tau(\epsilon)\to\infty,$$

so that, for all sufficiently small ϵ, $\tau(\epsilon)\ge\tau_0$. Then

$$\mathcal{E}_\phi(\epsilon,\tau) = K^{\mathrm{cl}}_{\phi,c,q_\star}\tau^{-q_\star} + C_{\phi,0}\epsilon\tau^2 + o\big(\tau^{-q_\star}+\epsilon\tau^2\big), \tag{40}$$

where

$$C_{\phi,0} = \mu_U \eta_{\mathrm{p}}^2 \phi \, \mathbb{E}\left[\frac{1}{\phi^2 (Q_0+1)^2}\right] = \frac{\mu_U \eta_{\mathrm{p}}^2}{\phi} \mathbb{E}\left[\frac{1}{(Q_0+1)^2}\right].$$

Consequently the leading two-term equivalent

$$H_\epsilon(\tau) := K^{\mathrm{cl}}_{\phi,c,q_\star} \tau^{-q_\star} + C_{\phi,0} \epsilon \tau^2$$

is minimized at

$$\tau_{\mathrm{opt}}(\epsilon) = \left(\frac{q_\star K^{\mathrm{cl}}_{\phi,c,q_\star}}{2 C_{\phi,0} \epsilon}\right)^{1/(q_\star+2)}. \tag{41}$$

This choice is admissible, $\tau_{\mathrm{opt}}(\epsilon) \geq \tau_0$, for all sufficiently small ϵ, and the optimized fixed-ϕ scaling is

$$\mathcal{E}_\phi(\epsilon, \tau_{\mathrm{opt}}(\epsilon)) \asymp \epsilon^{q_\star/(q_\star+2)}. \tag{42}$$

More precisely,

$$\mathcal{E}_\phi(\epsilon, \tau_{\mathrm{opt}}(\epsilon)) \sim \left(1 + \frac{q_\star}{2}\right) K^{\mathrm{cl}}_{\phi,c,q_\star} \left(\frac{2 C_{\phi,0}}{q_\star K^{\mathrm{cl}}_{\phi,c,q_\star}}\right)^{q_\star/(q_\star+2)} \epsilon^{q_\star/(q_\star+2)}.$$

Proof. By definition,

$$\mathcal{E}_\phi(\epsilon, \tau) = \big[\mathcal{R}_{\phi,0}(0,\tau) - \mathcal{R}_{\phi,0}(0,\infty)\big] + \big[\mathcal{R}_{\phi,0}(\epsilon,\tau) - \mathcal{R}_{\phi,0}(0,\tau)\big].$$

The first bracket is the clean truncation term, so Lemma 6 gives

$$\mathcal{R}_{\phi,0}(0,\tau) - \mathcal{R}_{\phi,0}(0,\infty) = K^{\mathrm{cl}}_{\phi,c,q_\star} \tau^{-q_\star} + o(\tau^{-q_\star}).$$

The second bracket is $\mathcal{D}_{\phi,0}(\epsilon,\tau)$. The fixed-$\phi$ leading-response theorem, using the uniform generator remainder from Lemma 5, gives

$$\mathcal{D}_{\phi,0}(\epsilon,\tau) = C_{\phi,0} \epsilon \tau^2 (1 + o(1)).$$

Thus the combined remainder is

$$o(\tau^{-q_\star}) + o(\epsilon\tau^2) = o(\tau^{-q_\star} + \epsilon\tau^2),$$

which proves equation 40.

It remains to minimize H_ϵ. Differentiating,

$$H'_\epsilon(\tau) = -q_\star K^{\mathrm{cl}}_{\phi,c,q_\star} \tau^{-q_\star-1} + 2 C_{\phi,0} \epsilon \tau.$$

The unique critical point satisfies

$$q_\star K^{\mathrm{cl}}_{\phi,c,q_\star} \tau^{-q_\star-1} = 2 C_{\phi,0} \epsilon \tau,$$

which is equivalent to equation 41. Since $\tau_{\mathrm{opt}}(\epsilon) \to \infty$, it lies in the admissible region $\tau \geq \tau_0$ for all sufficiently small ϵ. At the optimum,

$$C_{\phi,0} \epsilon \tau_{\mathrm{opt}}^2 = \frac{q_\star}{2} K^{\mathrm{cl}}_{\phi,c,q_\star} \tau_{\mathrm{opt}}^{-q_\star}.$$

Therefore

$$H_\epsilon(\tau_{\mathrm{opt}}) = \left(1 + \frac{q_\star}{2}\right) K^{\mathrm{cl}}_{\phi,c,q_\star} \tau_{\mathrm{opt}}^{-q_\star}.$$

Substitution of equation 41 gives the stated constant and exponent. □

A.6 $\phi \downarrow 0$: Classical Statistics

The fixed-aspect-ratio formula contains a population-mixture calculation as a further many-samples-per-coordinate limit. This limit is different from the fixed-ϕ small-ϵ limit. Taking $\phi \downarrow 0$ first removes the sparse coordinate-hit mechanism.

Result 7 (Population limit of the deterministic equivalent)**.** *Fix $\epsilon \in [0, 1)$, $\tau \geq \tau_0$, and $\varrho \geq 0$. Then, as $\phi \downarrow 0$,*

$$\mathcal{R}_{\phi,\varrho}(\epsilon, \tau) \longrightarrow \mathcal{R}_{0,\varrho}(\epsilon, \tau),$$

where

$$\mathcal{R}_{0,\varrho}(\epsilon, \tau) := \mu_U \left[c - \frac{(1-\epsilon)B_\tau + \epsilon(c + \eta_{\mathrm{p}}\tau)}{\varrho + (1-\epsilon)M_\tau + \epsilon} \right]^2. \tag{43}$$

In particular, in the ridgeless population limit $\varrho = 0$,

$$\mathcal{R}_{0,0}(\epsilon, \tau) = \mu_U \left[\frac{(1-\epsilon)c(M_{\tau/c} - M_\tau) + \epsilon\eta_{\mathrm{p}}\tau}{(1-\epsilon)M_\tau + \epsilon} \right]^2. \tag{44}$$

Proof. The clean and poison rates in equation 34 are $(1-\epsilon)/\phi$ and ϵ/ϕ. Therefore

$$\phi S_A^{\epsilon,\tau} \xrightarrow{\mathbb{P}} (1-\epsilon)M_\tau + \epsilon,$$

and

$$\phi S_B^{\epsilon,\tau} \xrightarrow{\mathbb{P}} (1-\epsilon)B_\tau + \epsilon(c + \eta_{\mathrm{p}}\tau).$$

The variances are $O(\phi)$ because the summands are bounded by constants depending only on τ. The denominator $\varrho + \phi S_A^{\epsilon,\tau}$ converges in probability to $\varrho + (1-\epsilon)M_\tau + \epsilon > 0$. Since $G_{\phi,\varrho}$ is bounded for fixed τ, convergence in probability plus bounded convergence gives equation 43.

For $\varrho = 0$, substitute $B_\tau = cM_{\tau/c}$ and rearrange:

$$\begin{aligned} c - \frac{(1-\epsilon)cM_{\tau/c} + \epsilon(c + \eta_{\mathrm{p}}\tau)}{(1-\epsilon)M_\tau + \epsilon} \\ = -\frac{(1-\epsilon)c(M_{\tau/c} - M_\tau) + \epsilon\eta_{\mathrm{p}}\tau}{(1-\epsilon)M_\tau + \epsilon}. \end{aligned}$$

Squaring gives equation 44. □

The exact Pareto identity gives

$$M_{\tau/c} - M_\tau = \frac{1 - c^{q_\star - 1}}{q_\star - 1} \tau^{1-q_\star}.$$

Define

$$A := c\frac{1 - c^{q_\star - 1}}{q_\star - 1}, \qquad B := \eta_{\mathrm{p}}.$$

Result 8 (Optimized poisoning exponent after taking $\phi \downarrow 0$)**.** *Consider the ridgeless population limit equation 44. In the perturbative regime*

$$\epsilon \downarrow 0, \qquad \tau = \tau(\epsilon) \to \infty, \qquad \epsilon\tau(\epsilon) \to 0,$$

one has

$$\mathcal{R}_{0,0}(\epsilon, \tau) = \frac{1}{\mu_U} \left(A\tau^{-(q_\star - 1)} + B\epsilon\tau \right)^2 (1 + o(1)). \tag{45}$$

The leading term is minimized by

$$\tau_{\mathrm{opt}}(\epsilon) = \left(\frac{A(q_\star - 1)}{B\epsilon} \right)^{1/q_\star}, \tag{46}$$

and at this value

$$\mathcal{R}_{0,0}(\epsilon, \tau_{\mathrm{opt}}(\epsilon)) \sim \frac{q_\star^2}{\mu_U} A^{2/q_\star} B^{2(q_\star - 1)/q_\star} (q_\star - 1)^{-2(q_\star - 1)/q_\star} \epsilon^{2 - 2/q_\star}. \tag{47}$$

Proof. From equation 44,

$$\mathcal{R}_{0,0}(\epsilon,\tau) = \mu_U \left[\frac{(1-\epsilon)A\tau^{-(q_\star-1)} + B\epsilon\tau}{(1-\epsilon)M_\tau + \epsilon}\right]^2.$$

Since $\tau \to \infty$ and $\epsilon\tau \to 0$,

$$(1-\epsilon)M_\tau + \epsilon = \mu_U(1+o(1)),$$

and

$$(1-\epsilon)A\tau^{-(q_\star-1)} = A\tau^{-(q_\star-1)}(1+o(1)).$$

This proves equation 45.

It remains to minimize

$$g_\epsilon(\tau) := A\tau^{-(q_\star-1)} + B\epsilon\tau.$$

Differentiating gives

$$g'_\epsilon(\tau) = -A(q_\star-1)\tau^{-q_\star} + B\epsilon.$$

The unique critical point is equation 46, and it is the global minimizer of the leading expression because $g_\epsilon(\tau) \to \infty$ as $\tau \downarrow 0$ or $\tau \to \infty$. At the optimum,

$$B\epsilon\tau_{\mathrm{opt}} = A(q_\star-1)\tau_{\mathrm{opt}}^{-(q_\star-1)},$$

so

$$g_\epsilon(\tau_{\mathrm{opt}}) = q_\star A\tau_{\mathrm{opt}}^{-(q_\star-1)} = q_\star A^{1/q_\star} B^{(q_\star-1)/q_\star}(q_\star-1)^{-(q_\star-1)/q_\star}\epsilon^{1-1/q_\star}.$$

Squaring and multiplying by $1/\mu_U$ proves equation 47. □

A.7 The same observable under the two orders of limits

Fix the ridgeless estimator ($\varrho = 0$) and consider the degradation of the poisoned truncated estimator measured against the clean untruncated ridgeless estimator,

$$\Delta^{(P)}_{\epsilon,\tau} := \mathcal{R}_{\mathrm{clean},P}\big(\widehat{\theta}_{\epsilon,\tau,0}\big) - \mathcal{R}_{\mathrm{clean},P}\big(\widehat{\theta}_{0,\infty,0}\big), \tag{48}$$

where $\widehat{\theta}_{0,\infty,0}$ denotes the clean ($\epsilon = 0$), untruncated ($\tau = \infty$), ridgeless estimator. The results from previous discussion is summarized below.

Result 9 (High-dimensional vs ample-data limit)**.** *Assume $P/D \to \phi \in (0,\infty)$ and work at $\varrho = 0$. For every fixed $\phi \in (0,\infty)$, $\epsilon \in [0,1)$ and $\tau \geq \tau_0$,*

$$\Delta^{(P)}_{\epsilon,\tau} \xrightarrow{\mathbb{P}} \mathcal{E}_\phi(\epsilon,\tau) := \mathcal{R}_{\phi,0}(\epsilon,\tau) - \mu_U c^2 e^{-1/\phi}.$$

(i) (High-dimensional / fixed aspect ratio.) *In the regime $\epsilon \downarrow 0$, $\tau = \tau(\epsilon) \to \infty$,*

$$\mathcal{E}_\phi(\epsilon,\tau) = K^{\mathrm{cl}}_{\phi,c,q_\star}\,\tau^{-q_\star} + C_{\phi,0}\,\epsilon\tau^2 + o\big(\tau^{-q_\star} + \epsilon\tau^2\big), \tag{49}$$

which is minimized at $\tau_{\mathrm{opt}}(\epsilon) \asymp \epsilon^{-1/(q_\star+2)}$, with

$$\min_\tau \mathcal{E}_\phi(\epsilon,\tau) \;\asymp\; \epsilon^{\,q_\star/(q_\star+2)}. \tag{50}$$

(ii) (Ample-data limit.) *Taking the aspect ratio to zero first,*

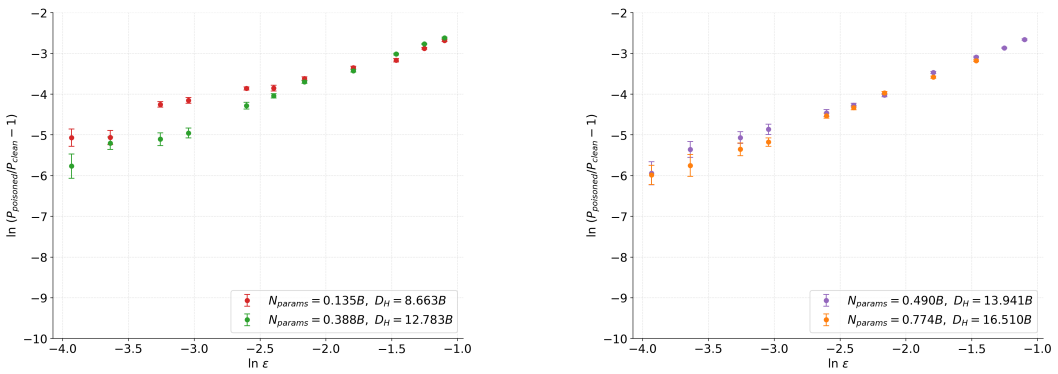
$$\lim_{\phi\downarrow 0} \mathcal{E}_\phi(\epsilon,\tau) = \mathcal{R}_{0,0}(\epsilon,\tau),$$

and in the regime $\epsilon \downarrow 0$, $\tau \to \infty$, $\epsilon\tau \to 0$,

$$\mathcal{R}_{0,0}(\epsilon,\tau) = \frac{1}{\mu_U}\Big(A\,\tau^{-(q_\star-1)} + B\,\epsilon\tau\Big)^2\big(1+o(1)\big), \tag{51}$$

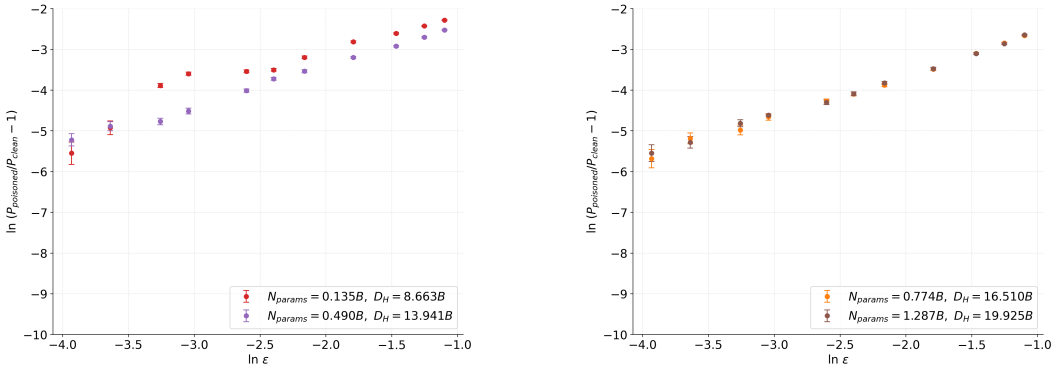
which is minimized at $\tau_{\mathrm{opt}}(\epsilon) \asymp \epsilon^{-1/q_\star}$, with

$$\min_\tau \mathcal{R}_{0,0}(\epsilon,\tau) \;\asymp\; \epsilon^{\,2-2/q_\star}. \tag{52}$$



(a) (b)

Figure 5: Scaling law for a small amount of random poisoning ε. We see that the scaling law for the larger models is more stable. This fact might be explained by saying stronger models have smaller ϕ_{LLM} and hence the ample data regime extends up to smaller values of ε.



(a) (b)

Figure 6: Scaling law for a small amount of language switching poisoning ε. We see that the scaling law for the larger models is more stable. This fact might be explained by saying stronger models have smaller ϕ_{LLM} and hence the ample data regime extends up to smaller values of ε.

A.8 Possible implications of the non-commuting limit for LLM pre-training

In an ample-data calculation, poisoning changes population moments smoothly and the clean truncation bias and poison displacement are balanced before the final quadratic validation metric is applied. In the coordinate model this order of limits gives

$$\left(\tau^{-(q_\star-1)} + \varepsilon\tau\right)^2$$

and hence the optimized exponent $2 - 2/q_\star$. The LLM calculation in the main text is performed in ample data limit and the latent cluster model leads to similar exponent with

$$q_\star = \frac{1 - m + 2rm}{2rm} \tag{53}$$

At fixed $\phi > 0$, however, some parameter directions receive only a finite number of effective observations. A poison event can then hit a previously weakly observed coordinate with probability of order ε/ϕ. Conditional on such a hit, the coordinate displacement is of order τ, so the risk contribution is already of order τ^2 before averaging over poisoned coordinates. This produces the different balance

$$\tau^{-q_\star} + \varepsilon\tau^2$$

and the optimized exponent $q_\star/(q_\star + 2)$. Since the scaling exponent is different in this two regimes it suggests that for LLM, making ϵ arbitrarily small would change the scaling law curve we have obtained.

B How the scaling law may arise in LLM pre-training

In this section, we analyze a local approximation to the training dynamics over a short time scale in a neighborhood of a finite-time checkpoint rather than attempting to model the full dynamics of transformer pre-training. Within this local framework, we introduce a sequence of assumptions under which the observed pre-training poisoning scaling law can be derived. Because several of these assumptions have not yet been validated experimentally, the analysis should be interpreted as proposing a plausible mechanism for the observed scaling behavior, rather than as a complete or uniquely established explanation.

B.1 Local training dynamics

The construction in this subsection and the next is purely kinematic: we only linearize the model and form the local quadratic objective (Meterez et al., 2026), so it applies verbatim whether the training data are clean or poisoned. Let $\theta \in \mathbb{R}^P$ denote the vector of model parameters, and let θ^{in} be the reference parameter vector about which we consider local dynamics. This reference may be the minimizer of the clean population objective or, more generally, a checkpoint along the (clean or poisoned) training trajectory. Write

$$\delta\theta = \theta - \theta^{\mathrm{in}}. \tag{54}$$

We index the training data by token-prediction events $Z_i = (X_i, Y_i)$, where X_i is the context and Y_i is the target token. The symbol D denotes the total number of such events. Equivalently, i may index sequences, provided that the corresponding quantities below are understood to include all token losses in a sequence.

Let

$$f_\theta(X_i) \in \mathbb{R}^V \tag{55}$$

be the vector of pre-softmax logits produced by the model for example i, where V is the vocabulary size. The example-level Jacobian at the reference parameters is

$$J_i := \left.\frac{\partial f_\theta(X_i)}{\partial \theta}\right|_{\theta=\theta^{\mathrm{in}}} \in \mathbb{R}^{V\times P}. \tag{56}$$

Thus $J_i\delta\theta$ is the first-order change in the logits caused by a parameter displacement $\delta\theta$:

$$f_{\theta^{\mathrm{in}}+\delta\theta}(X_i) = f_{\theta^{\mathrm{in}}}(X_i) + J_i\delta\theta + O(\|\delta\theta\|^2). \tag{57}$$

Let $\ell_i(f)$ denote the loss of example i as a function of its logits. Define

$$g_i := \nabla_f \ell_i(f)\big|_{f=f_{\theta^{\mathrm{in}}}(X_i)} \in \mathbb{R}^V, \qquad \mathcal{W}_i := \nabla_f^2 \ell_i(f)\big|_{f=f_{\theta^{\mathrm{in}}}(X_i)} \in \mathbb{R}^{V\times V}. \tag{58}$$

Here g_i is the gradient of the loss with respect to the logits, and $\mathcal{W}_i$ is the corresponding logit-space curvature matrix. We define the pseudo-residual $r_i := -g_i$. This sign convention ensures that $J_i^\top r_i$ points in the local descent direction.[3]

The empirical generalized Gauss–Newton matrix is

$$\widehat{\Sigma} := \frac{1}{D}\sum_{i=1}^{D} J_i^\top \mathcal{W}_i J_i \in \mathbb{R}^{P\times P}. \tag{60}$$

[3]For the softmax cross-entropy loss, if $p_i := \mathrm{softmax}\big(f_{\theta^{\mathrm{in}}}(X_i)\big)$ and e_{Y_i} is the one-hot vector associated with the target token, then

$$g_i = p_i - e_{Y_i}, \qquad r_i = e_{Y_i} - p_i, \qquad \mathcal{W}_i = \mathrm{diag}(p_i) - p_i p_i^\top. \tag{59}$$

It measures the local curvature of the empirical training loss in parameter space. It is also closely related to an empirical Fisher matrix. The empirical response vector is

$$\widehat{v} := \frac{1}{D}\sum_{i=1}^{D} J_i^\top r_i = -\nabla_\theta \widehat{\mathcal{L}}(\theta)\Big|_{\theta=\theta^{\mathrm{in}}} \in \mathbb{R}^P, \tag{61}$$

where

$$\widehat{\mathcal{L}}(\theta) := \frac{1}{D}\sum_{i=1}^{D} \ell_i\big(f_\theta(X_i)\big) \tag{62}$$

is the empirical pre-training objective. Thus $\widehat{v}$ is the negative empirical gradient evaluated at the reference parameters. Under the local linearization of the logits and the generalized Gauss–Newton approximation, the empirical objective becomes

$$\widehat{\mathcal{L}}(\theta^{\mathrm{in}} + \delta\theta) \simeq \widehat{\mathcal{L}}(\theta^{\mathrm{in}}) - \widehat{v}^\top \delta\theta + \frac{1}{2}\delta\theta^\top \widehat{\Sigma}\delta\theta. \tag{63}$$

The stationary condition for this quadratic approximation is therefore

$$\widehat{\Sigma}\,\widehat{\delta\theta} = \widehat{v}, \tag{64}$$

which gives

$$\widehat{\delta\theta} \simeq \widehat{\Sigma}^{-1}\widehat{v}. \tag{65}$$

B.2 Finite training time as truncation

In this subsection, we analyze the training dynamics and argue that finite-time training is equivalent to an effective truncation on the curvature spectrum. As in the previous subsection, the argument is stated for a generic local curvature $\widehat{\Sigma}$ and response $\widehat{v}$ and therefore holds for both clean and poisoned data; the two are distinguished only later (Appendices B.4–B.6) through the choice of curvature and response. The one new ingredient is the optimizer's preconditioner S. We show that it leaves the stationary point of the local dynamics unchanged but changes the operator whose spectrum is filtered due to finite training: not $\widehat{\Sigma}$ itself but its symmetrised preconditioned version $A = S^{1/2}\widehat{\Sigma}S^{1/2}$. This is the operator whose spectrum is measured in Appendix E, and every spectral assumption in the rest of this appendix refers to it.

A broad class of memoryless preconditioned first-order methods has the update

$$\delta\theta_{t+1} = \delta\theta_t + \eta_t S_t\big(\widehat{v} - \widehat{\Sigma}\delta\theta_t\big), \tag{66}$$

where $\eta_t \geq 0$ is the learning rate and S_t is a self-adjoint positive-definite preconditioning operator. The Adam optimiser (Kingma & Ba, 2015) fits the memoryless template equation 66 exactly when its momentum is switched off, i.e., $\beta_1 = 0$. Writing the gradient of the local quadratic objective at step t as

$$g_t := \widehat{\Sigma}\,\delta\theta_t - \widehat{v} = -\big(\widehat{v} - \widehat{\Sigma}\,\delta\theta_t\big),$$

Adam maintains the diagonal second-moment accumulator[4]

$$\nu_t = \beta_2\,\nu_{t-1} + (1-\beta_2)\,g_t \odot g_t, \qquad \hat{\nu}_t := \frac{\nu_t}{1-\beta_2^{t+1}}, \tag{67}$$

and takes the step

$$\delta\theta_{t+1} = \delta\theta_t - \eta_t \frac{g_t}{\sqrt{\hat{\nu}_t} + \epsilon_{\mathrm{A}}} \quad\Longrightarrow\quad S_t = \Big(\mathrm{diag}\big(\sqrt{\hat{\nu}_t}\big) + \epsilon_{\mathrm{A}} I\Big)^{-1}, \tag{68}$$

where $\epsilon_{\mathrm{A}} > 0$ is Adam's damping constant. For $\beta_1 > 0$ the update carries memory. In the small-learning-rate limit, heavy-ball momentum reduces to gradient flow with time rescaled by $(1-\beta_1)^{-1}$ (Kovachki & Stuart, 2021); we assume, without proof, that the same reduction applies to Adam's momentum, so that $\beta_1 > 0$ only rescales the time variable introduced below and does not affect any exponent.

[4]Here $\odot$ is the Hadamard product. We write Adam's second-moment estimate as ν_t rather than the customary v_t to avoid a clash with the response vector $\widehat{v}$.

Adam's second moment $\hat{\nu}_t$ is roughly an exponential moving average with a time constant of $(1-\beta_2)^{-1}$ steps. Late in training, where the gradient statistics drift slowly, it is nearly constant across a window of that length. We make this a standing assumption of the local model, following Meterez et al. (2026), who freeze Adam's second-moment at its moving-average value and find that the resulting quadratic model predicts the training loss over windows of up to 10% of training.

Assumption 10 (Frozen preconditioner on the local window)**.** *Across the local window the preconditioner is a fixed symmetric positive-definite operator, $S_t \equiv S$, with*

$$s_{\min} I \preceq S \preceq s_{\max} I, \qquad 0 < s_{\min} \le s_{\max} < \infty. \tag{69}$$

For Adam, equation 68 gives $s_{\max} = (\min_j \sqrt{\hat{\nu}_j} + \epsilon_{\mathrm{A}})^{-1} \le \epsilon_{\mathrm{A}}^{-1}$ and $s_{\min} = (\max_j \sqrt{\hat{\nu}_j} + \epsilon_{\mathrm{A}})^{-1}$.

Define the learning-rate-integrated time and the preconditioned curvature, response and coordinates

$$T := \sum_{t=0}^{K-1} \eta_t, \qquad A := S^{1/2}\,\widehat{\Sigma}\,S^{1/2} \succeq 0, \qquad \tilde{v} := S^{1/2}\widehat{v}, \qquad \phi_t := S^{-1/2}\delta\theta_t. \tag{70}$$

The operator A is the one written in equation 156 of Appendix E. It is similar to the non-symmetric generator $S\widehat{\Sigma} = S^{1/2}AS^{-1/2}$ of the raw dynamics and therefore has the same spectrum, but unlike $S\widehat{\Sigma}$ it is symmetric positive semidefinite, so it has an orthonormal eigenbasis and a spectral measure.

Result 11 (Preconditioned training is gradient flow on A)**.** *Under Assumption 10, with $\delta\theta_0 = 0$, the update equation 66 reads, in the coordinates equation 70,*

$$\phi_{t+1} = (I - \eta_t A)\phi_t + \eta_t \tilde{v}, \qquad \phi_K = q^{(K)}(A)\,\tilde{v}, \qquad q^{(K)}(\lambda) := \frac{1 - \prod_{t=0}^{K-1}(1-\eta_t\lambda)}{\lambda}. \tag{71}$$

In the continuous-time limit $\max_t \eta_t\lambda \to 0$ at fixed T, the recursion becomes the gradient flow

$$\frac{d\phi}{dT} = \tilde{v} - A\phi, \qquad \phi(0) = 0, \tag{72}$$

whose solution is

$$\phi(T) = q_T(A)\,\tilde{v}, \qquad \delta\theta(T) = S^{1/2}q_T(A)S^{1/2}\widehat{v} = q_T(S\widehat{\Sigma})\,S\widehat{v}, \qquad q_T(\lambda) := \frac{1 - e^{-\lambda T}}{\lambda}. \tag{73}$$

If $\widehat{\Sigma} \succ 0$ then $\delta\theta(T) \to \widehat{\Sigma}^{-1}\widehat{v}$ as $T \to \infty$: the preconditioner leaves the stationary point equation 65 unchanged and acts only on the path, through the spectrum of A.

See C.3 for a proof.

Remark 1 (Commuting preconditioner and the mode-dependent clock). If S and $\widehat{\Sigma}$ commute they share an eigenbasis $\{u_k\}$, $Su_k = s_k u_k$, $\widehat{\Sigma}u_k = \lambda_k u_k$, and $Au_k = s_k\lambda_k u_k$. Then equation 73 reads, mode by mode,

$$\delta\theta_k(T) = \big(1 - e^{-\lambda_k T_k}\big)\,\frac{v_k}{\lambda_k}, \qquad T_k := s_k T, \qquad v_k := u_k^{\top}\widehat{v}, \tag{74}$$

and $T_k = \sum_{t=0}^{K-1} \eta_t s_{k,t}$ if S_t varies in time while remaining co-diagonal with $\widehat{\Sigma}$: the preconditioner acts as a mode-dependent clock on the spectrum of the raw curvature. This is exact for SGD ($S = I$) and for preconditioners that are functions of $\widehat{\Sigma}$ (e.g. natural gradient). Adam's S is diagonal in the coordinate basis whereas $\widehat{\Sigma}$ is not, so in general $[S, \widehat{\Sigma}] \neq 0$.

Remark 2 (Where the poisoning fraction enters the optimizer). Adam's second moment equation 67 is a statistic of the training gradients, so on the mixture $\mathcal{P}_\varepsilon$ the frozen preconditioner is $S_\varepsilon = S_0 + \varepsilon\,\partial_\varepsilon S_\varepsilon|_{\varepsilon=0} + o(\varepsilon)$. Fix the coordinates equation 70 with the clean S_0 for both the clean and the poisoned run and write $M_\varepsilon := S_0^{-1/2} S_\varepsilon S_0^{-1/2} = I + \varepsilon\Delta M + o(\varepsilon)$. Repeating the computation in the proof of Result 11, the poisoned flow is $\dot{\phi}_\varepsilon = M_\varepsilon\big(S_0^{1/2}\widehat{v}_\varepsilon - S_0^{1/2}\widehat{\Sigma}_\varepsilon S_0^{1/2}\phi_\varepsilon\big)$, i.e.

$$\begin{aligned} \dot{\phi}_\varepsilon &= -H_\varepsilon\phi_\varepsilon + v_\varepsilon, \\ H_\varepsilon &= A_0 + \varepsilon\big(\Delta M\,A_0 + S_0^{1/2}\Delta\widehat{\Sigma}\,S_0^{1/2}\big) + o(\varepsilon), \\ v_\varepsilon &= \tilde{v}_0 + \varepsilon\big(\Delta M\,\tilde{v}_0 + S_0^{1/2}\Delta\widehat{v}\big) + o(\varepsilon), \end{aligned} \tag{75}$$

where $A_0 := S_0^{1/2}\widehat{\Sigma}_0 S_0^{1/2}$, $\tilde{v}_0 := S_0^{1/2}\widehat{v}_0$.

Result 11 says that finite-time preconditioned training is the spectral filter q_T applied to A. There are two asymptotic regimes:

$$\phi(T) = q_T(A)\,\tilde{v}, \qquad q_T(\lambda) = \frac{1-e^{-\lambda T}}{\lambda} = \begin{cases} \lambda^{-1} + o(\lambda^{-1}), & \lambda T \gg 1, \\ T + O(\lambda T^2), & \lambda T \ll 1. \end{cases} \tag{76}$$

Note that, $(1-e^{-1})\min\left\{T, \frac{1}{\lambda}\right\} \le q_T(\lambda) \le \min\left\{T, \frac{1}{\lambda}\right\}$. Finite-time training is therefore equivalent, up to multiplicative constants, to putting a sharp cut-off on the inverse-curvature variable $X = \lambda^{-1}$, $\lambda \in \mathrm{spec}(A)$, at the truncation level $\tau = T$.

B.3 Assumption of heavy-tailed spectrum

In this subsection, we discuss the crucial ingredient needed for the poisoning scaling law in a large language model: the heavy-tailed inverse-curvature spectrum of the preconditioned clean-data curvature

$$A_P := S^{1/2}\,\widehat{\Sigma}_P\,S^{1/2} \tag{77}$$

under $\mathcal{P}$, with $\widehat{\Sigma}_P$ the clean Gauss–Newton matrix equation 60 and S the frozen clean preconditioner of Assumption 10. The empirical results of Tang et al. (2025) provide qualitative motivation for a broad hierarchy of curvature scales in trained language models. They find an approximate power law among the leading positive eigenvalues of the parameter Hessian:

$$\nabla_\theta^2 \widehat{\mathcal{L}}(\theta) = \underbrace{\frac{1}{D}\sum_{i=1}^{D} J_i^\top \mathcal{W}_i J_i}_{\widehat{\Sigma}} + \underbrace{\frac{1}{D}\sum_{i=1}^{D}\sum_{a}\frac{\partial \ell_i}{\partial f_{ia}}\nabla_\theta^2 f_{ia}}_{\widehat{R}_{\mathrm{model}}}, \tag{78}$$

Their power-law fits concern a fixed number of leading eigenvalues above a positive cutoff. Meterez et al. (2026) report the same for the Gauss–Newton matrix. Our assumption, by contrast, concerns the normalized spectral mass arbitrarily close to zero, as stated below.

Assumption 12 (Heavy-tailed preconditioned clean-data curvature)**.** *Let A_P be the preconditioned clean-data curvature equation 77, with eigendecomposition*

$$A_P u_{k,P} = \lambda_{k,P} u_{k,P}, \qquad k = 1, \dots, P, \tag{79}$$

and let

$$\mu_P := \frac{1}{P}\sum_{k=1}^{P} \delta_{\lambda_{k,P}} \tag{80}$$

be its empirical spectral measure. Suppose that, as $P, D \to \infty$, μ_P converges to a deterministic probability measure μ on $[0,\infty)$ whose distribution function has a power-law hard edge,

$$F(x) := \mu([0,x]) = c_\mu\, x^\alpha\,(1+o(1)), \qquad x \downarrow 0, \qquad c_\mu > 0,\ \alpha > 0. \tag{81}$$

Equivalently, for $\lambda \sim \mu$ the inverse-curvature variable $X := \lambda^{-1}$ is heavy-tailed,

$$\mathbb{P}_{\lambda\sim\mu}\,(X > \tau) = \mu([0,\tau^{-1})) = c_\mu \tau^{-\alpha}\,(1+o(1)), \qquad \tau \to \infty. \tag{82}$$

This implies that the eigenvalues ordered increasingly from zero obey

$$\lambda^{\uparrow}_{(j),P} \asymp \left(\frac{j}{c_\mu P}\right)^{1/\alpha}. \tag{83}$$

This assumption should be distinguished from the descending-rank capacity law measured by Meterez et al. (2026). Their eigenvalues are ordered from largest to smallest, and the reported fit takes the form

$$\lambda^{\downarrow}_{i,P} \asymp \left(\frac{1}{i}\right)^{\alpha_{\mathrm{cap}}}, \qquad \alpha_{\mathrm{cap}} \approx 0.96, \qquad P \approx 1.5\times 10^8 \tag{84}$$

They study up to $\lambda \ge \lambda_{1,P}/(1.5\times 10^6)$, so the maximum rank of the eigenvalue in descending order is $i \approx (1.5\times 10^6)^{1/0.96} \approx 3\times 10^6$; against $i/P \approx 0.02$, so roughly 98% of the modes-the genuinely weakest, which define α-lie below the resolution floor. In general, $\alpha, \alpha_{\mathrm{cap}}$ are independent

parameters, we explain this with a simple example next. The probability density of the beta-prime law $\lambda \sim \beta'(\alpha, 1/\alpha_{\mathrm{cap}})$ is

$$\rho(\lambda) = \frac{1}{\lambda_0\, B(\alpha,\, 1/\alpha_{\mathrm{cap}})} \left(\frac{\lambda}{\lambda_0}\right)^{\alpha-1} \left(1 + \frac{\lambda}{\lambda_0}\right)^{-(\alpha+1/\alpha_{\mathrm{cap}})}, \qquad \lambda > 0, \tag{85}$$

with shape parameters $\alpha, \alpha_{\mathrm{cap}} > 0$, scale $\lambda_0 > 0$, and $B(\cdot,\cdot)$ the beta function.

Remark 1. *The density equation 85 has the following two regimes.*

(i) Hard edge ($\lambda \downarrow 0$). $\rho(\lambda) \sim \big(\lambda_0^{\alpha} B(\alpha, 1/\alpha_{\mathrm{cap}})\big)^{-1} \lambda^{\alpha-1}$, *hence* $\mu([0,x]) \asymp x^{\alpha}$, *which gives the ascending order statistics* $\lambda^{\uparrow}_{(j),P} \asymp \lambda_0\, (j/P)^{1/\alpha}$ *and* $\lambda_{\min}(P) \asymp \lambda_0\, P^{-1/\alpha}$, *i.e. Assumption 12 with hard-edge exponent* α.

(ii) Capacity tail ($\lambda \to \infty$). $\rho(\lambda) \sim \big(\lambda_0^{-1/\alpha_{\mathrm{cap}}} / B(\alpha, 1/\alpha_{\mathrm{cap}})\big) \lambda^{-1-1/\alpha_{\mathrm{cap}}}$, *hence* $\Pr(\lambda > t) \asymp (t/\lambda_0)^{-1/\alpha_{\mathrm{cap}}}$, *which gives the descending order statistics* $\lambda^{\downarrow}_{i,P} \asymp \lambda_0\, i^{-\alpha_{\mathrm{cap}}}$, *i.e. the capacity law equation 84 with exponent* α_{cap}.

The two regimes cross over at $\lambda \asymp \lambda_0$.

The hard-edge exponent does not depend on the preconditioner due to the following argument. Read the local quadratic objective as a system of coupled oscillators: $\frac{1}{2}\, \delta\theta^{\top} \widehat{\Sigma}_P\, \delta\theta$ is a potential energy with stiffness matrix $\widehat{\Sigma}_P \succeq 0$. In the preconditioned coordinates $\phi = S^{-1/2}\delta\theta$ of Result 11 the same energy is $\frac{1}{2}\, \phi^{\top} A \phi$, and $A\phi = \lambda\phi$ is equivalent, with $\psi := S^{1/2}\phi$, to the generalized eigenproblem

$$\widehat{\Sigma}_P\, \psi = \lambda\, M\psi, \qquad M := S^{-1}. \tag{86}$$

The eigenvalues of A are therefore the squared normal-mode frequencies of a system with stiffness $\widehat{\Sigma}_P$ and mass matrix $M = S^{-1}$: the preconditioner is an inverse mass matrix. By equation 69 the mass matrix satisfies $s_{\max}^{-1} I \preceq M \preceq s_{\min}^{-1} I$. Comparing M with the two scalar mass matrices $s_{\max}^{-1} I$ and $s_{\min}^{-1} I$ therefore gives, rank by rank,

$$s_{\min}\, \lambda_k(\widehat{\Sigma}_P) \le \lambda_k(A) \le s_{\max}\, \lambda_k(\widehat{\Sigma}_P), \tag{87}$$

Every eigenvalue is thus moved by a multiplicative factor lying in $[s_{\min}, s_{\max}]$, i.e. its logarithm is shifted by at most $\ln \kappa$, where $\kappa := s_{\max}/s_{\min}$. This implies,

$$F_{\widehat{\Sigma}_P}(x/s_{\max}) \;\le\; F_A(x) \;\le\; F_{\widehat{\Sigma}_P}(x/s_{\min}). \tag{88}$$

Let the curvature have a hard edge, $F_{\Sigma}(y) \simeq c\, y^{\alpha}$ for small y, and let $s_{\min}, s_{\max}$ stay fixed as $P \to \infty$. Inserting this into equation 88,

$$c\, s_{\max}^{-\alpha}\, x^{\alpha} \;\lesssim\; F_A(x) \;\lesssim\; c\, s_{\min}^{-\alpha}\, x^{\alpha}, \qquad x \to 0. \tag{89}$$

Taking logarithms, $\ln F_A(x) = \alpha \ln x + \ln c - \alpha \ln s_{\mathrm{eff}}(x)$ with some $s_{\mathrm{eff}}(x) \in [s_{\min}, s_{\max}]$. The last term is bounded, while $\alpha \ln x \to -\infty$, so the slope on log–log axes is α for A as for Σ:

$$\lim_{x \to 0} \frac{\ln F_A(x)}{\ln x} = \alpha.$$

The exponent is thus universal under bounded preconditioning, while the scaling coefficient, and the eigenvectors are not. A detailed numerical study supporting Assumption 12 for the operator A_P is provided in Appendix E. In the rest of this subsection we discuss a simple model that produces a hard edge for the raw population curvature Σ_P, the exponent is inherited by A_P. Define

$$\mathcal{J}_i := \mathcal{W}_i^{1/2} J_i \in \mathbb{R}^{V \times P}, \qquad Y_i := \mathcal{J}_i^{\top} \mathcal{J}_i = J_i^{\top} \mathcal{W}_i J_i \succeq 0. \tag{90}$$

The empirical and population matrices are

$$\widehat{\Sigma}_P := \frac{1}{D} \sum_{i=1}^{D} Y_i, \qquad \Sigma_P := \mathbb{E}[Y_i]. \tag{91}$$

Here we present a model for sampling $\mathcal{J}_i$ under which the heavy-tailed spectrum assumption follows. The goal of this sampling method is to give interpretation to the parameter appearing in the heavy-tailed spectrum. As a motivational toy model, consider the spin-glass-based generative model introduced in Halder et al. (2026) as a proxy model of language. In that framework, the model's generations are drawn from a measure that decomposes into a hierarchy of clusters organized in a tree, where at each level the cluster weights follow a Poisson–Dirichlet law $\mathrm{PD}(m)$. An interesting observation is that there is a power law structure of the cluster weights in this setup. Let $(W_i)_{i\geq 1}$ denote the ranked probabilities of $\mathrm{PD}(m)$, where $0 < m < 1$. Then

$$W_i = \left(\frac{\Xi}{m(1-m)}\right)^{1/m} i^{-1/m} \qquad \text{a.s. as } i \to \infty,$$

where $\Xi > 0$ is an almost-surely finite random variable. This motivates us to propose the following setup for the latent space distribution of the language model we are discussing here. Let $I_i \in \mathbb{N}$ be the latent cluster index, in a single draw we choose it with probability

$$\mathbb{P}(I_i = \ell) = \pi_\ell, \qquad \pi_\ell = C_\pi \ell^{-1/m}(1 + o(1)), \qquad 0 < m < 1. \tag{92}$$

Here, the order of magnitude estimate is associated with cluster the index $l \to \infty$. For each P, let $\mathbb{R}^P = \bigoplus_{\ell \geq 1} \mathcal{U}_{\ell,P}$ be an orthogonal decomposition of the parameter space. Let $d_{\ell,P} := \dim(\mathcal{U}_{\ell,P})$, and choose a matrix

$$U_{\ell,P} \in \mathbb{R}^{P \times d_{\ell,P}}, \qquad U_{\ell,P}^\top U_{\ell,P} = I_{d_{\ell,P}}, \tag{93}$$

whose columns form an orthonormal basis of $\mathcal{U}_{\ell,P}$. For instance, the orthogonal projector onto this subspace is $\Pi_{\ell,P} := U_{\ell,P} U_{\ell,P}^\top$. Conditional on $I_i = \ell$, draw a Gaussian matrix

$$Z_i^{(\ell)} \in \mathbb{R}^{V \times d_{\ell,P}}, \qquad \left(Z_i^{(\ell)}\right)_{ab} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \frac{1}{V}\right), \tag{94}$$

Now we define the method of sampling the Jacobian

$$\mathcal{J}_i \mid (I_i = \ell) := \sqrt{h_\ell}\, Z_i^{(\ell)} U_{\ell,P}^\top \qquad h_\ell = C_h \ell^{2r}(1 + o(1)). \tag{95}$$

The next theorem shows that under this sampling, the eigen-space of Σ_P has a natural 'cluster' structure in terms of the parameter subspaces.

Result 13. *Every direction in $\mathcal{U}_{\ell,P}$ is associated to an eigenvalue of Σ_P according to*

$$\lambda_\ell := \pi_\ell h_\ell = C_\lambda \ell^{-\delta}(1 + o(1)), \qquad C_\lambda := C_\pi C_h, \qquad \delta := \frac{1}{m} - 2r. \tag{96}$$

A proof is given in Appendix C.4.

The laws of π_ℓ and h_ℓ determine the eigenvalue associated with cluster ℓ, but they do not determine how many parameter directions are assigned to that cluster. We specify this multiplicity next. Say $d_{\ell,P}/P \to \rho_\ell$ as $P \to \infty$. We want to set up ρ_l such that common semantic families/clusters receive large subspaces, while increasingly rare families/clusters receive subspace determined by the cluster scale. For this it will be useful to define the cumulative fraction of parameter directions devoted to clusters of rank at least ℓ by

$$R_{\mathrm{cap}}(\ell) := \sum_{j \geq \ell} \rho_j.$$

The eigenvalue scale of cluster ℓ is $\lambda_\ell \asymp \ell^{-\delta}$, so its inverse-curvature scale is $\lambda_\ell^{-1} \asymp \ell^{\delta}$. A natural capacity-matching hypothesis is that the fraction of representation R_{cap} devoted to semantic families weaker than λ_ℓ is proportional to their remaining aggregate curvature signal $B_Y(\tau) := \mathbb{E}[\|Y_i\|_{\mathrm{op}} \mathbf{1}\{\|Y_i\|_{\mathrm{op}} > \tau\}]$:

$$R_{\mathrm{cap}}(\ell) \asymp B_Y(\lambda_\ell^{-1}) \asymp \ell^{-\delta(q-1)} \tag{97}$$

The last equality followed due to the following result:

Result 14. *The first-moment signal remaining beyond a threshold τ obeys*

$$B_Y(\tau) := \mathbb{E}[\|Y_i\|_{\mathrm{op}} \mathbf{1}\{\|Y_i\|_{\mathrm{op}} > \tau\}] \asymp \tau^{-(q-1)}, \qquad q := \frac{1/m - 1}{2r}. \tag{98}$$

The order estimate equation 98 follows directly from the conditional law of Y_i constructed in equation 94–equation 95; a proof is given in Appendix C.5.

Let $\mu_P := \frac{1}{P}\sum_{k=1}^{P} \delta_{\lambda_k(\Sigma)}$ be the empirical spectral measure. Then as $P \to \infty$, $\mu_P \Rightarrow \mu$. This can be easily seen as follows. By equation 96, the eigenvalue associated with cluster rank ℓ is asymptotic to $C_\lambda \ell^{-\delta}$. Therefore $\lambda_\ell \le x \leftrightarrow \ell \gtrsim (C_\lambda/x)^{1/\delta}$. Using the limiting subspace proportions

$$\mu([0,x]) \sim \sum_{\ell \ge (C_\lambda/x)^{1/\delta}} \rho_\ell \asymp \left(\frac{1}{x}\right)^{-(q-1)} \asymp x^{q-1}. \tag{99}$$

Hence, this cluster model leads to a hard edge with exponent $\alpha = q - 1$ for the raw population curvature Σ_P. By earlier discussion, the preconditioned operator $A = S^{1/2}\widehat{\Sigma}_P S^{1/2}$ inherits the same exponent whenever the preconditioner has a condition number bounded in P (Assumption 10), which is the form in which Assumption 12 is used.

B.4 The clean data truncation error and assumption on its tail

The goal of this subsection is to quantify how much of the clean target remains unlearned after finite-time training. This subsection concerns clean data $\mathcal{P}$. We show that the clean truncation error decays as a power law in training time. We start from the clean linearized dynamics that follow from equation 66 via Result 11:

$$\frac{d}{dT}\phi_0(T) = -H_0\big(\phi_0(T) - \phi_0^{\text{in}}\big) + v_0, \qquad \phi_0(0) = \phi_0^{\text{in}}, \tag{100}$$

where $H_0 = S^{1/2}\widehat{\Sigma}_P S^{1/2}$ is the preconditioned clean curvature and $v_0 = S^{1/2}\widehat{v}_P$ the preconditioned negative clean training gradient at ϕ_0^{in} (Convention above). The stationary (infinite-time) solution is the clean training optimum $\phi_0^\star$,

$$H_0\big(\phi_0^\star - \phi_0^{\text{in}}\big) = v_0, \qquad \text{i.e.} \qquad \phi_0^\star = \phi_0^{\text{in}} + H_0^{-1} v_0. \tag{101}$$

The finite-time solution relaxes from the initialization toward this optimum,

$$\phi_0(T) = \phi_0^{\text{in}} + q_T(H_0)v_0 = \phi_0^\star - e^{-H_0 T}\big(\phi_0^\star - \phi_0^{\text{in}}\big), \qquad q_T(\lambda) := \frac{1 - e^{-\lambda T}}{\lambda}, \tag{102}$$

so the displacement from initialization, $\phi_0(T) - \phi_0^{\text{in}} = q_T(H_0)v_0$, is exactly the finite-time filtered response. If $H_0 u_k = \lambda_k u_k$ and the initialization-to-optimum gap is expanded as $\phi_0^\star - \phi_0^{\text{in}} = \sum_k a_k u_k$ (so $a_k := u_k^\top(\phi_0^\star - \phi_0^{\text{in}})$), then the clean data truncation error is

$$e_0(T) := \phi_0(T) - \phi_0^\star = -e^{-H_0 T}\big(\phi_0^\star - \phi_0^{\text{in}}\big) = -\sum_{k=1}^{P} e^{-\lambda_k T} a_k u_k. \tag{103}$$

Physically, $e_0(T)$ is the part of the clean target that finite-time training has not yet learned: each mode is suppressed by $e^{-\lambda_k T}$, so the well-conditioned (large-λ_k) directions are captured quickly, while the flat, weak-curvature (small-λ_k) directions remain underlearned. Next, we quantify the size of this residual. The observable we care about is clean data validation loss, whose local geometry is the validation Hessian G, not the training curvature H_0. In the ample-data limit the clean Gauss–Newton matrix is close to the validation Hessian, $\widehat{\Sigma}_P \approx \nabla^2 \mathcal{R}_{\mathcal{P}}(\phi_0^\star) = \mathbb{E}_{Z\sim\mathcal{P}}[J_Z^\top W_Z J_Z]$, and the same congruence transform applied to both sides gives, in the preconditioned frame,

$$H_0 = S^{1/2}\widehat{\Sigma}_P S^{1/2} \approx G = S^{1/2}\,\nabla^2 \mathcal{R}_{\mathcal{P}}(\phi_0^\star)\,S^{1/2}.$$

The next assumption makes this precise: in the weak-curvature tail, G is approximately diagonal in the training-curvature eigenbasis, and the validation weight of a mode scales as a power of its curvature.

Assumption 15 (Validation metric in clean curvature eigen-basis)**.** *In the weak-curvature tail, G is asymptotically diagonal in the H_0-eigenbasis and that*

$$u_k^\top G u_k \sim C_G \lambda_k^\zeta, \qquad \lambda_k \downarrow 0, \qquad C_G > 0, \qquad \zeta \in \mathbb{R}. \tag{104}$$

The finite-time error lives in the weak-curvature modes, but whether those modes contribute at all depends on how much of the clean learning target is actually supported on them, since we are studying the dynamics near it. The following assumption rules this out: the initialization-to-optimum gap carries a nontrivial, power-law amount of its mass in the weak-curvature modes. The exponent η is neutral at $\eta = 0$ (no extra suppression or enhancement in weak modes), positive when the target is smoother and avoids weak-curvature directions, and negative when it is unusually concentrated in rare weak-curvature directions.

Assumption 16 (Clean learning-target tail mass)**.** *The initialization-to-optimum gap $\phi_0^\star - \phi_0^{\mathrm{in}}$—the part of the clean target still to be learned starting from ϕ_0^{in}—has a nontrivial weight in the weak-curvature modes, with*

$$\frac{|a_k|^2}{\|\phi_0^\star - \phi_0^{\mathrm{in}}\|_2^2} \sim C_a \lambda_k^\eta, \qquad \lambda_k \downarrow 0, \qquad C_a > 0, \qquad a_k = u_k^\top \big(\phi_0^\star - \phi_0^{\mathrm{in}}\big). \tag{105}$$

Training longer learns more of the weak-curvature tail, so the clean truncation error should shrink as T grows. The result below shows that this decay is a clean power law $T^{-\beta}$, with an exponent that simply combines the three ingredients just introduced: how many weak modes there are (the spectral hard-edge exponent α), how the clean solution is spread across them (η), and how the validation metric weights them (ζ). This $B(T)$ is the clean data truncation bias that will later be balanced against the poison influence.

Result 17 (Clean truncation-error falloff)**.** *Suppose Assumptions 12, 15 and 16 hold together. Then, the clean truncation error measured in the validation metric, $B(T)^2 := e_0(T)^\top G\, e_0(T)$, obeys*

$$B(T) = C_B T^{-\beta}(1+o(1)), \qquad \beta := \frac{1}{2}(\alpha+\eta+\zeta), \qquad C_B := \|\phi_0^\star - \phi_0^{\mathrm{in}}\|_2 \left(C_a C_G c_\mu \alpha 2^{-2\beta}\Gamma(2\beta)\right)^{1/2}. \tag{106}$$

See Appendix C.6 for a proof.

B.5 Poisoning strength after truncation and assumption on its tail

The goal of this subsection is to quantify how strongly poisoning displaces the solution as training proceeds. This subsection involves both data types: the clean curvature and response (H_0, v_0) under $\mathcal{P}$ and their poisoned counterparts $(H_\varepsilon, v_\varepsilon)$ under the mixture $\mathcal{P}_\varepsilon = (1-\varepsilon)\mathcal{P} + \varepsilon\mathcal{Q}$. We show that the poison influence grows as a power law in training time. Consider the clean and contaminated linearized dynamics, each launched from its own initialization:

$$\dot\phi_0(T) = -H_0\big(\phi_0(T) - \phi_0^{\mathrm{in}}\big) + v_0, \qquad \phi_0(0) = \phi_0^{\mathrm{in}}, \tag{107}$$

$$\dot\phi_\varepsilon(T) = -H_\varepsilon\big(\phi_\varepsilon(T) - \phi_\varepsilon^{\mathrm{in}}\big) + v_\varepsilon, \qquad \phi_\varepsilon(0) = \phi_\varepsilon^{\mathrm{in}}, \tag{108}$$

with contamination

$$H_\varepsilon = H_0 + \varepsilon\Delta H, \qquad v_\varepsilon = v_0 + \varepsilon\Delta v. \tag{109}$$

Both dynamics are written in the preconditioned frame fixed by the clean preconditioner S_0; by Remark 2, the $O(\varepsilon)$ response of Adam's second-moment statistics to the poison is part of ΔH and Δv (Eq. equation 75), and ΔH need not be symmetric—nothing below uses its symmetry. Here ϕ_0^{in} and $\phi_\varepsilon^{\mathrm{in}}$ are the initializations of the clean and poisoned runs. Crucially, the local expansion is legitimate only late in training, where per-step updates are small and the trajectory stays inside a neighborhood of the checkpoint; T measures the time elapsed from that late-time checkpoint. Because the checkpoint is itself the product of training, the clean and the poisoned checkpoints do not coincide: the poison accumulated up to the checkpoint displaces $\phi_\varepsilon^{\mathrm{in}}$ from ϕ_0^{in} at first order in the poison rate. We therefore keep the $O(\varepsilon)$ offset,

$$\phi_\varepsilon^{\mathrm{in}} - \phi_0^{\mathrm{in}} = \varepsilon\,\psi^{\mathrm{in}} + o(\varepsilon), \qquad \psi^{\mathrm{in}} := \partial_\varepsilon \phi_\varepsilon^{\mathrm{in}}\big|_{\varepsilon=0}, \tag{110}$$

with ψ^{in} generically nonzero—it is the accumulated poison displacement already carried by the checkpoint. The derivative of the contaminated trajectory at $\varepsilon = 0$ is

$$\psi(T) := \left.\frac{\partial \phi_\varepsilon(T)}{\partial \varepsilon}\right|_{\varepsilon=0}. \tag{111}$$

Physically, $\psi(T)$ is the first-order direction in which contamination drags the learned parameters as training proceeds: to leading order the poisoned and clean trajectories separate by $\varepsilon\psi(T)$, i.e., $\phi_\varepsilon(T) - \phi_0(T) = \varepsilon\psi(T) + o(\varepsilon)$, and at the checkpoint this separation is exactly the offset, $\psi(0) = \psi^{\text{in}}$. Differentiating equation 108 at $\varepsilon = 0$ (both dynamics share the fixed linearization center ϕ_0^{in}; only the starting point moves, by equation 110) yields the variational equation

$$\begin{aligned} \dot{\psi}(T) &= -H_0\psi(T) + \Delta v - \Delta H\left(\phi_0(T) - \phi_0^{\text{in}}\right), \\ \psi(0) &= \psi^{\text{in}}, \end{aligned} \tag{112}$$

in which the poison forces the clean displacement-from-initialization $\phi_0(T) - \phi_0^{\text{in}}$ while the trajectory starts from the offset ψ^{in}. This is solved by a homogeneous transient plus a forced part,

$$\psi(T) = e^{-H_0T}\psi^{\text{in}} + \psi_{\text{f}}(T), \qquad \psi_{\text{f}}(T) := \int_0^T e^{-H_0(T-s)}\Big(\Delta v - \Delta H\left(\phi_0(s) - \phi_0^{\text{in}}\right)\Big)\, ds, \tag{113}$$

with $\phi_0(s) - \phi_0^{\text{in}} = (I - e^{-H_0 s})(\phi_0^\star - \phi_0^{\text{in}})$. The first term is the checkpoint offset, carried forward and filtered by e^{-H_0T} exactly as the clean residual $e_0(T)$ is; the second is the forced poison response. We can rewrite the forced part as

$$\begin{aligned} &\psi_{\text{f}}(T) = q_T(H_0) f_{\text{eff}}(T), \\ &u_k^\top f_{\text{eff}}(T) = u_k^\top f_{\text{p}} + r_T(\lambda_k)\sum_{j=1}^{P}(u_k^\top \Delta H u_j)\big(u_j^\top(\phi_0^\star - \phi_0^{\text{in}})\big)K_T(\lambda_k, \lambda_j). \end{aligned} \tag{114}$$

with $f_{\text{p}} := \Delta v - \Delta H\big(\phi_0^\star - \phi_0^{\text{in}}\big)$. The two-eigenvalue kernels that are used to present expression of $f_{\text{eff}}(T)$ are as follows

$$K_T(\lambda, \mu) = \begin{cases} \dfrac{e^{-\mu T} - e^{-\lambda T}}{\lambda - \mu}, & \lambda \neq \mu, \\ T e^{-\lambda T}, & \lambda = \mu. \end{cases} \tag{115}$$

and

$$r_T(\lambda_k) := \begin{cases} \dfrac{\lambda_k}{1 - e^{-\lambda_k T}}, & \lambda_k > 0, \\ \dfrac{1}{T}, & \lambda_k = 0. \end{cases} \tag{116}$$

The size of the forced poison-induced displacement, measured in the metric that governs clean validation loss, is $I(T)^2 = \psi_{\text{f}}(T)^\top G\, \psi_{\text{f}}(T)$ (the checkpoint-offset transient is treated separately below). Just as the finite-time filter q_T shapes the clean error, it shapes the poison response, so $I(T)^2$ can be written as an integral of $q_T(\lambda)^2$ against a measure that records how the poison distributes its influence across curvature scales. The assumption below specifies the behavior of this measure near weak curvature modes: the poison deposits a power-law amount of influence.

Assumption 18 (Effective poison spectral tail)**.** *There exists a poison measure $d\nu_p^{(T)}(\lambda)$ such that*

$$I(T)^2 := \psi_{\text{f}}(T)^\top G \psi_{\text{f}}(T) = \int_0^\infty q_T(\lambda)^2\, d\nu_p^{(T)}(\lambda). \tag{117}$$

Assume that,

$$d\nu_p^{(T)}(\lambda) \sim C_\nu \lambda^{1-2\gamma}\, d\lambda, \qquad \lambda \downarrow 0, \qquad T \uparrow \infty, \qquad 0 < \gamma < 1. \tag{118}$$

The clean truncation bias $B(T)$ decays as training proceeds, but the poison influence works in the opposite direction: the longer the model trains, the more the weak-curvature directions fall under the poison forcing. The result below quantifies this growth, showing that the poison-induced displacement, measured in the validation metric, grows as a clean power law T^γ, with the growth exponent set by the effective poison spectral tail. It is the opposing partner of the clean-bias falloff, and balancing the two is what later fixes the scaling law.

Result 19 (Poison-strength growth)**.** *Under Assumption 18, for large T,*

$$I(T) = C_I T^\gamma (1 + o(1)), \qquad C_I := \sqrt{C_\nu J_\gamma}, \tag{119}$$

where

$$J_\gamma := \int_0^\infty (1 - e^{-z})^2 z^{-1-2\gamma}\, dz \in (0, \infty), \qquad 0 < \gamma < 1. \tag{120}$$

See Appendix C.7 for a proof.

The checkpoint offset ψ^{in} contributes the transient $e^{-H_0 T}\psi^{\mathrm{in}}$ to the displacement equation 113. Measured in the validation metric it has exactly the structure of the clean residual $e_0(T)$ of Appendix B.4 - a fixed vector filtered by $e^{-H_0 T}$ - so, under a weak-curvature tail on ψ^{in} analogous to Assumption 16, its size is a clean power law,

$$B_{\mathrm{in}}(T) := \left\|e^{-H_0 T}\psi^{\mathrm{in}}\right\|_G = C_{\mathrm{in}}\, T^{-\beta_{\mathrm{in}}}(1+o(1)), \qquad \beta_{\mathrm{in}} = \tfrac{1}{2}(\alpha+\eta_{\mathrm{in}}+\zeta), \tag{121}$$

where η_{in} is the tail exponent of ψ^{in} across the H_0-eigenbasis. Thus the offset decays with training: continued training partially washes out the poison damage already present in the checkpoint, while the forced response $\psi_{\mathrm{f}}(T)$ builds up.

B.6 The scaling law from optimal truncation

The goal of this subsection is to combine the decaying clean truncation bias and the growing poison influence into the observed scaling law.

For a clean token-prediction event $Z = (X, Y)$, let $f_\phi(X) \in \mathbb{R}^V$ be the logits and define

$$\ell_Z(\phi) := -\log(\mathrm{softmax}(f_\phi(X))_Y)\,. \tag{122}$$

At $\phi_0^\star$, put

$$p_Z := \mathrm{softmax}(f_{\phi_0^\star}(X)), \qquad J_Z := \left.\frac{\partial f_\phi(X)}{\partial \phi}\right|_{\phi=\phi_0^\star}. \tag{123}$$

The gradient and Hessian with respect to the logits are

$$\nabla_f \ell_Z = p_Z - e_Y, \qquad W_Z := \nabla_f^2 \ell_Z = \mathrm{diag}(p_Z) - p_Z p_Z^\top. \tag{124}$$

Consequently, the clean validation loss is

$$\mathcal{R}_{\mathcal{P}}(\phi) := \mathbb{E}_{Z\sim\mathcal{P}}[\ell_Z(\phi)]. \tag{125}$$

The clean solution $\phi_0^\star$ is the stationary point of the training dynamics ($H_0(\phi_0^\star - \phi_0^{\mathrm{in}}) = v_0$, Eq. equation 101); it need not minimize the validation risk $\mathcal{R}_{\mathcal{P}}$, and in a finite-data run it generically does not. We therefore keep the validation gradient at the clean solution,

$$g_\star := \nabla \mathcal{R}_{\mathcal{P}}(\phi_0^\star) \in \mathbb{R}^P, \tag{126}$$

Taylor's expansion, keeping the linear term, gives

$$\mathcal{R}_{\mathcal{P}}(\phi_0^\star + h) = \mathcal{R}_{\mathcal{P}}(\phi_0^\star) + g_\star^\top h + \tfrac{1}{2} h^\top G h + o(\|h\|_2^2). \tag{127}$$

At a matched checkpoint $\phi_\varepsilon(T) - \phi_0(T) = \varepsilon\psi(T) + o(\varepsilon)$ and $\phi_0(T) - \phi_0^\star = e_0(T)$, with the displacement split into checkpoint offset and forced response, $\psi(T) = e^{-H_0 T}\psi^{\mathrm{in}} + \psi_{\mathrm{f}}(T)$ (Eq. equation 113). Applying equation 127 with $h = e_0(T) + \varepsilon\psi(T)$ and with $h = e_0(T)$ and subtracting, the constant and poison-independent terms ($\mathcal{R}_{\mathcal{P}}(\phi_0^\star)$, $g_\star^\top e_0$, $\frac{1}{2} e_0^\top G e_0$) cancel, leaving

$$\begin{aligned}\delta\mathcal{R}(\varepsilon, T) &:= \mathcal{R}_{\mathcal{P}}(\phi_\varepsilon(T)) - \mathcal{R}_{\mathcal{P}}(\phi_0(T)) = \varepsilon\, g_\star^\top \psi(T) + \varepsilon\, e_0(T)^\top G\, \psi(T) + \tfrac{1}{2}\varepsilon^2\, \psi(T)^\top G\, \psi(T) \\ &= \underbrace{\varepsilon\, \rho_{gi}(T)\, \Gamma_\star B_{\mathrm{in}}(T)}_{\text{checkpoint offset}} + \underbrace{\varepsilon\, \rho_g(T)\, \Gamma_\star I(T)}_{\text{train–eval gap}} + \underbrace{\varepsilon\, \rho_G(T)\, B(T) I(T)}_{\text{clean bias}} + \underbrace{\tfrac{1}{2}\varepsilon^2 I(T)^2}_{\text{pure poison}},\end{aligned} \tag{128}$$

where the correlations are

$$\Gamma_\star := \|g_\star\|_{G^{-1}}, \quad \rho_g(T) := \frac{g_\star^\top \psi_{\mathrm{f}}(T)}{\Gamma_\star I(T)}, \quad \rho_{gi}(T) := \frac{g_\star^\top e^{-H_0 T}\psi^{\mathrm{in}}}{\Gamma_\star B_{\mathrm{in}}(T)}, \quad \rho_G(T) := \frac{e_0^\top G \psi_{\mathrm{f}}}{B(T) I(T)}, \tag{129}$$

all in $[-1, 1]$ by Cauchy–Schwarz in the G-metric and taken asymptotically constant, $\rho_G \to \rho_\star$, $\rho_g \to \rho_g^\star$, $\rho_{gi} \to \rho_{gi}^\star$. The train–eval gap term is the decisive contribution at late times. Using $B(T) = C_B T^{-\beta}$ (Result 17), $I(T) = C_I T^\gamma$ (Result 19) and $B_{\mathrm{in}}(T) = C_{\mathrm{in}} T^{-\beta_{\mathrm{in}}}$ (Eq. equation 121), the four coefficients of equation 128 scale as

$$\underbrace{\varepsilon\, \rho_{gi}^\star \Gamma_\star C_{\mathrm{in}}\, T^{-\beta_{\mathrm{in}}}}_{\text{checkpoint: } \varepsilon,\ \text{decays}}, \quad \underbrace{\varepsilon\, \rho_g^\star \Gamma_\star C_I\, T^{\gamma}}_{\text{gap: } \varepsilon,\ \text{grows}}, \quad \underbrace{\varepsilon\, \rho_\star C_B C_I\, T^{-(\beta-\gamma)}}_{\text{bias: } \varepsilon,\ \text{decays}}, \quad \underbrace{\tfrac{1}{2}\varepsilon^2 C_I^2\, T^{2\gamma}}_{\text{pure: } \varepsilon^2,\ \text{grows}}. \tag{130}$$

Here we assumed $\beta > \gamma > 0, \beta_{\mathrm{in}} > 0$.

Result 20 (Gap-dominated crossover)**.** *Suppose among ε order terms the gap term dominates. Then the excess risk is,*

$$\delta\mathcal{R}(\varepsilon, T) = \rho_g^\star \Gamma_\star I(T)\,\varepsilon + \tfrac{1}{2} I(T)^2\,\varepsilon^2\,(1 + o(1)), \tag{131}$$

and equating them, $\varepsilon\,\rho_g^\star \Gamma_\star I(T) = \frac{1}{2}\varepsilon^2 I(T)^2$, defines the crossover time

$$T_\varepsilon^{(g)} = \Big(\frac{2\rho_g^\star \Gamma_\star}{C_I\,\varepsilon}\Big)^{1/\gamma} = t_g\,\varepsilon^{-1/\gamma}, \qquad t_g := \Big(\frac{2\rho_g^\star \Gamma_\star}{C_I}\Big)^{1/\gamma}. \tag{132}$$

The excess risk then obeys the integer law

$$\delta\mathcal{R}(\varepsilon, T) \asymp \begin{cases} \Gamma_\star\,\varepsilon\,T^\gamma, & T \ll T_\varepsilon^{(g)} \;\; (a = 1), \\ \varepsilon^2\,T^{2\gamma}, & T \gg T_\varepsilon^{(g)} \;\; (a = 2), \end{cases} \tag{133}$$

More generally, say, $x := T/T_\varepsilon^{(g)}$, then,

$$\delta\mathcal{R}(\varepsilon, T) = \Psi(x)\,(1 + o(1)), \quad \Psi(x) = \rho_g^\star \Gamma_\star C_I t_g^\gamma\,x^\gamma + \tfrac{1}{2} C_I^2 t_g^{2\gamma}\,x^{2\gamma} = 2(\rho_g^\star \Gamma_\star)^2\big(x^\gamma + x^{2\gamma}\big), \tag{134}$$

which is ε-independent at $T = T_\varepsilon^{(g)}$ ($a = 0$). Since $\Psi(x) \sim x^\gamma$ ($x \to 0$) and $\sim x^{2\gamma}$ ($x \to \infty$) with both exponents positive, Ψ is monotone increasing no interior minimum.

Small LLMs fall under the gap-dominated regime, early in training $a = 1$ and then it decreases towards $a = 0$ as we approach the crossover time. Whereas, large LLMs lie in the gap or the checkpoint-ominated regime. Both these terms are mathematically similar because their conceptual origin is the heavy-tailed inverse curvature spectrum. Hence, we discuss only one of them (the checkpoint dominated results are obtained from the bias-dominated scaling exponent using $\beta \leftrightarrow \beta_{\text{in}} + \gamma$).

Result 21 (Bias-dominated crossover)**.** *Suppose among ε order terms the bias term dominates. Write $x := T/T_\varepsilon$ with*

$$T_\varepsilon = t\,\varepsilon^{-1/(\beta+\gamma)}, \qquad t := \Big(\tfrac{2\rho_\star C_B}{C_I}\Big)^{1/(\beta+\gamma)}. \tag{135}$$

Then the excess risk admits the exact scaling form

$$\delta\mathcal{R}(\varepsilon, T) = \varepsilon^a\,\Phi(x)\,(1 + o(1)), \qquad \Phi(x) = \rho_\star C_B C_I\,t^{\gamma-\beta}\,x^{\gamma-\beta} + \tfrac{1}{2} C_I^2\,t^{2\gamma}\,x^{2\gamma}, \qquad a = \frac{2\beta}{\beta+\gamma}, \tag{136}$$

Its limits are given as,

$$\Phi(x) \asymp x^{\gamma-\beta}\;(x \to 0), \qquad \Phi(x) \asymp x^{2\gamma}\;(x \to \infty), \tag{137}$$

and Φ attains a unique interior minimum for $\beta > \gamma$ at

$$x_\star = \Big(\frac{(\beta-\gamma)\,\rho_\star C_B}{\gamma\,C_I\,t^{\beta+\gamma}}\Big)^{1/(\beta+\gamma)} = \Big(\frac{\beta-\gamma}{2\gamma}\Big)^{1/(\beta+\gamma)} = \phi(1). \tag{138}$$

Consequently the two regimes meet, in order, at the crossover $T \asymp T_\varepsilon$, where

$$\delta\mathcal{R}(\varepsilon, T) \;\asymp\; \varepsilon^a, \qquad \min_T \delta\mathcal{R}(\varepsilon, T) = \Phi(x_\star)\,\varepsilon^a\,(1 + o(1)). \tag{139}$$

See Appendix C.8 for a proof.

Therefore, for large LLM, scaling exponent a starts near 1 and then increases as training progresses towards the crossover time.

Remark 3 (Perplexity observable). For the empirical perplexity ratio $\Delta(\varepsilon, T) = \exp(\delta\mathcal{R}(\varepsilon, T)) - 1$, the same three statements hold verbatim in the small-degradation regime $\delta\mathcal{R} \to 0$, since $\Delta = \delta\mathcal{R}\,(1 + o(1))$.

The exponent a so far is expressed through the abstract tail exponents $\alpha, \beta, \gamma, \zeta, \eta$. The latent-cluster model lets us trade these for two interpretable quantities: the cluster-tail parameter m (roughly, this controls the distribution of ideas) and the within-cluster spread r (controls how precise an idea is).

$$a = 2 - \frac{4\gamma}{\alpha + \eta + \zeta + 2\gamma}, \qquad \alpha = \frac{1/m - 1}{2r} - 1, \qquad 0 < \gamma < 1. \tag{140}$$

Making the neutral modeling choices—no extra weak-mode alignment ($\eta = 0$), a neutral poison spectral density ($\gamma = 1/2$), and the ample-data validation matching ($\zeta = 1$)—collapses the general formula to a closed form in m and r, which is what makes the observed slope/intercept trends interpretable.

$$a = 2 - \frac{2}{2+\alpha} = 2 - \frac{2}{1+\frac{1-m}{2rm}}.$$

C PROOFS FOR THE SCALING-LAW MECHANISM

C.1 PROOF OF RESULT 1

Proof. Define

$$F(\theta, \varepsilon) := \nabla_\theta L_\varepsilon(\theta).$$

The map F is real analytic near $(\theta_0, 0)$, $F(\theta_0, 0) = 0$, and

$$D_\theta F(\theta_0, 0) = \nabla^2 L_{\mathcal{P}}(\theta_0) = H.$$

Because H is invertible, the analytic implicit-function theorem gives a unique analytic map $\varepsilon \mapsto \theta_\varepsilon$ near zero such that $F(\theta_\varepsilon, \varepsilon) = 0$ and $\theta_{\varepsilon=0} = \theta_0$. We construct this function below.

Differentiate $F(\theta_\varepsilon, \varepsilon) = 0$ at $\varepsilon = 0$:

$$H\,\theta_0' + \partial_\varepsilon F(\theta_0, 0) = 0.$$

By definition, $\partial_\varepsilon F(\theta_0, 0) = g$, and therefore

$$\theta_0' = -H^{-1} g.$$

Analyticity of the branch now gives

$$\theta_\varepsilon - \theta_0 = -\varepsilon H^{-1} g + O(\varepsilon^2),$$

which proves equation 4.

Taylor expansion of $\mathcal{R}$ at θ_0 gives, with $\delta_\varepsilon := \theta_\varepsilon - \theta_0$,

$$\mathcal{R}(\theta_0 + \delta_\varepsilon) - \mathcal{R}(\theta_0) = \underbrace{\nabla \mathcal{R}(\theta_0)^\top \delta_\varepsilon}_{=0} + \frac{1}{2}\delta_\varepsilon^\top G \delta_\varepsilon + O(\|\delta_\varepsilon\|^3).$$

Substituting $\delta_\varepsilon = -\varepsilon H^{-1} g + O(\varepsilon^2)$ yields

$$\mathcal{R}(\theta_\varepsilon) - \mathcal{R}(\theta_0) = \frac{\varepsilon^2}{2} g^\top H^{-1} G H^{-1} g + O(\varepsilon^3),$$

which is equation 5. The coefficient is strictly positive under the stated nondegeneracy condition. Since both θ_ε and $\mathcal{R}$ are analytic, their composition is analytic; hence a noninteger leading power is impossible under these assumptions. Finally, $e^x - 1 = x + O(x^2)$ proves the perplexity statement. □

C.2 PROOF OF RESULT 2

Proof. For any ε and τ,

$$\begin{aligned} m_{\varepsilon,\tau} - \mu &= \mathbb{E}_{\mathcal{P}}[T_\tau(\psi)] - \mu \\ &\quad + \varepsilon \left(\mathbb{E}_{\mathcal{Q}_\varepsilon}[T_\tau(\psi)] - \mathbb{E}_{\mathcal{P}}[T_\tau(\psi)] \right). \end{aligned}$$

The first line is the clean truncation bias and the second is the contamination effect. Substitute equation 6 and equation 7 and choose $\tau = \tau_\varepsilon = t\varepsilon^{-1/(\beta+\gamma)}$. Then

$$\tau_\varepsilon^{-\beta} = t^{-\beta} \varepsilon^{\beta/(\beta+\gamma)}, \qquad \varepsilon \tau_\varepsilon^{\gamma} = t^{\gamma} \varepsilon^{\beta/(\beta+\gamma)}.$$

Both contributions therefore have the same order, and their vector coefficient is $h_t = t^{-\beta} b + t^{\gamma} c$. This proves equation 8.

Since Φ is C^2,

$$\theta_{\varepsilon,\tau_\varepsilon} - \theta_0 = J(m_{\varepsilon,\tau_\varepsilon} - \mu) + O(\|m_{\varepsilon,\tau_\varepsilon} - \mu\|^2).$$

Using equation 8,

$$\theta_{\varepsilon,\tau_\varepsilon} - \theta_0 = \varepsilon^{\beta/(\beta+\gamma)} J h_t + o\Big(\varepsilon^{\beta/(\beta+\gamma)}\Big).$$

The Taylor expansion of $\mathcal{R}$ at θ_0, gives

$$\begin{aligned}\mathcal{R}(\theta_{\varepsilon,\tau_\varepsilon}) - \mathcal{R}(\theta_0) &= \frac{1}{2}(\theta_{\varepsilon,\tau_\varepsilon} - \theta_0)^\top G(\theta_{\varepsilon,\tau_\varepsilon} - \theta_0) + o(\|\theta_{\varepsilon,\tau_\varepsilon} - \theta_0\|^2) \\ &= \frac{1}{2} h_t^\top J^\top G J h_t \, \varepsilon^{2\beta/(\beta+\gamma)} + o\Big(\varepsilon^{2\beta/(\beta+\gamma)}\Big).\end{aligned}$$

This proves equation 9 and completes the proof. □

C.3 Proof of Result 11

Proof. Multiplying equation 66 by $S^{-1/2}$ and inserting $\delta\theta_t = S^{1/2}\phi_t$,

$$\phi_{t+1} = \phi_t + \eta_t S^{-1/2} S\big(\widehat{v} - \widehat{\Sigma} S^{1/2}\phi_t\big) = \phi_t + \eta_t\big(\tilde{v} - A\phi_t\big).$$

Let $Au = \lambda u$ with $\|u\| = 1$ and $x_t := u^\top \phi_t$, $\tilde{v}_u := u^\top \tilde{v}$. Then $x_{t+1} = (1 - \eta_t\lambda)x_t + \eta_t \tilde{v}_u$. For $\lambda > 0$ this has the fixed point $\tilde{v}_u/\lambda$ and $x_{t+1} - \tilde{v}_u/\lambda = (1 - \eta_t\lambda)(x_t - \tilde{v}_u/\lambda)$, so from $x_0 = 0$ we get $x_K = q^{(K)}(\lambda)\tilde{v}_u$; for $\lambda = 0$ both sides equal $T\tilde{v}_u$. Since A is symmetric its eigenvectors span $\mathbb{R}^P$, which gives equation 71. If $\max_t \eta_t\lambda < 1$ then $\log\prod_t(1 - \eta_t\lambda) = -\lambda T - O(\lambda^2 \sum_t \eta_t^2)$ and $\sum_t \eta_t^2 \le T \max_t \eta_t$, so $q^{(K)}(\lambda) \to q_T(\lambda)$ in the stated limit; $q_T(A)\tilde{v}$ is the solution of equation 72 by direct differentiation. For the raw coordinates, $S^{1/2}A^n S^{-1/2} = (S\widehat{\Sigma})^n$ for every $n \ge 0$, hence $S^{1/2} f(A) S^{-1/2} = f(S\widehat{\Sigma})$ for every polynomial f and, for $f = q_T$; thus $\delta\theta(T) = S^{1/2}\phi(T) = S^{1/2} q_T(A) S^{1/2}\widehat{v} = q_T(S\widehat{\Sigma})S\widehat{v}$. Finally, if $\widehat{\Sigma} \succ 0$ then $A \succ 0$, $q_T(A) \to A^{-1}$ as $T \to \infty$, and $S^{1/2}A^{-1}S^{1/2} = S^{1/2}S^{-1/2}\widehat{\Sigma}^{-1}S^{-1/2}S^{1/2} = \widehat{\Sigma}^{-1}$. □

C.4 Proof of Result 13

Proof. The row vectors of $\mathcal{J}_i$ are supported on the proper parameter subspace $\mathcal{U}_{\ell,P}$ rather than on all of $\mathbb{R}^P$. From an LLM perspective, $\mathcal{U}_{\ell,P}$ represents tangent directions that change the model's logits primarily on contexts belonging to semantic cluster ℓ. The normalization $1/V$ gives

$$\mathbb{E}\Big[\big(Z_i^{(\ell)}\big)^\top Z_i^{(\ell)}\Big] = I_{d_{\ell,P}}. \tag{141}$$

Since $Y_i = \mathcal{J}_i^\top \mathcal{J}_i$, the Gaussian model implies

$$\mathbb{E}[Y_i \mid I_i = \ell] = h_\ell U_{\ell,P}\mathbb{E}\Big[\big(Z_i^{(\ell)}\big)^\top Z_i^{(\ell)}\Big] U_{\ell,P}^\top \tag{142}$$

$$= h_\ell U_{\ell,P} U_{\ell,P}^\top = h_\ell \Pi_{\ell,P}. \tag{143}$$

Standard Gaussian matrix estimates give

$$\|\mathcal{J}_i\|_{\mathrm{op}} = O_{\mathbb{P}}(\sqrt{h_\ell}) = O_{\mathbb{P}}(\ell^r), \tag{144}$$

Taking expectation over I_i and using the law of total expectation now gives

$$\begin{aligned}\Sigma_P := \mathbb{E}[Y_i] &= \sum_{\ell \ge 1} \mathbb{P}(I_i = \ell)\mathbb{E}[Y_i \mid I_i = \ell] \\ &= \sum_{\ell \ge 1} \pi_\ell h_\ell \Pi_{\ell,P}.\end{aligned} \tag{145}$$

□

C.5 Proof of Result 14

Proof. A token drawn from cluster ℓ has probability and curvature scale

$$\pi_\ell \asymp \ell^{-1/m}, \qquad h_\ell \asymp \ell^{2r}.$$

Thus, rarer clusters have larger single-example curvature. The quantity $B_Y(\tau)$ asks how much of the average curvature signal comes from examples whose curvature is larger than τ.

To make this observation precise at the level needed here, condition on the cluster index $I_i = \ell$. From

$$\mathcal{J}_i = \sqrt{h_\ell}\, Z_i^{(\ell)} U_{\ell,P}^\top, \qquad Y_i = \mathcal{J}_i^\top \mathcal{J}_i,$$

and the orthonormality of the columns of $U_{\ell,P}$, we get

$$\|Y_i\|_{\mathrm{op}} = h_\ell\, \zeta_\ell, \qquad \zeta_\ell := \|Z_i^{(\ell)}\|_{\mathrm{op}}^2.$$

For the Gaussian normalization used in the model, ζ_ℓ is an order-one random variable with a light upper tail. Consequently, it can change constants but not the power of τ: clusters with $h_\ell \gg \tau$ contribute with order-one probability, clusters with $h_\ell \ll \tau$ contribute only through an exponentially unlikely fluctuation, and the crossover region $h_\ell \asymp \tau$ only changes the overall prefactor. Therefore

$$B_Y(\tau) \asymp \sum_{\ell:\, h_\ell \gtrsim \tau} \pi_\ell h_\ell. \tag{146}$$

Let ℓ_τ denote the cluster rank at this crossover. Since $h_\ell \asymp \ell^{2r}$,

$$h_{\ell_\tau} \asymp \tau \qquad \Longrightarrow \qquad \ell_\tau \asymp \tau^{1/(2r)}.$$

Moreover,

$$\pi_\ell h_\ell \asymp \ell^{-1/m+2r} = \ell^{-\delta}, \qquad \delta := \frac{1}{m} - 2r.$$

The definition $q = (1/m - 1)/(2r)$ gives

$$\delta - 1 = 2r(q-1).$$

Because $q > 1$, we have $\delta > 1$, so the remaining sum is a convergent power-law tail. Using an integral estimate in equation 146,

$$B_Y(\tau) \asymp \sum_{\ell \gtrsim \ell_\tau} \ell^{-\delta} \asymp \int_{\ell_\tau}^\infty x^{-\delta}\, dx \asymp \ell_\tau^{-(\delta-1)}.$$

Finally, substituting $\ell_\tau \asymp \tau^{1/(2r)}$ yields

$$B_Y(\tau) \asymp \tau^{-(\delta-1)/(2r)} = \tau^{-(q-1)}.$$

In physical terms, the exponent is simply the balance between how quickly the curvature of an individual rare cluster grows and how quickly the total probability of such clusters decreases. □

C.6 Proof of Result 17

Proof. Under Assumption 16, the hard-edge law, and Assumption 15, for $T \to \infty$ we obtain,

$$\begin{aligned} B(T)^2 &= e_0(T)^\top G e_0(T) \\ &\sim \|\phi_0^\star - \phi_0^{\mathrm{in}}\|_2^2 C_a C_G c_\mu \alpha \int_0^\infty e^{-2\lambda T} \lambda^{\alpha+\eta+\zeta-1}\, d\lambda. \end{aligned} \tag{147}$$

The change of variables $z = 2\lambda T$ gives

$$B(T)^2 = \|\phi_0^\star - \phi_0^{\mathrm{in}}\|_2^2 C_a C_G c_\mu \alpha \Gamma(2\beta)(2T)^{-2\beta}(1 + o(1)). \tag{148}$$

Taking the square root yields equation 106, which completes the proof. □

C.7 Proof of Result 19

Proof. Using the spectral representation equation 117 with $q_T(\lambda) = (1-e^{-\lambda T})/\lambda$ and the tail equation 118, for large T,

$$I(T)^2 = \int_0^\infty \left(\frac{1-e^{-\lambda T}}{\lambda}\right)^2 d\nu_p^{(T)}(\lambda) \sim C_\nu \int_0^\infty (1-e^{-\lambda T})^2 \lambda^{-1-2\gamma}\, d\lambda. \tag{149}$$

With $z = \lambda T$,

$$I(T)^2 \sim C_\nu T^{2\gamma} \int_0^\infty (1-e^{-z})^2 z^{-1-2\gamma}\, dz = C_{\nu,G} J_\gamma T^{2\gamma}, \tag{150}$$

where J_γ is defined in equation 120 and is finite and strictly positive for $0 < \gamma < 1$. Taking the square root yields equation 119, which completes the proof. □

C.8 Proof of Result 21

Proof. Throughout we are in the bias-dominated regime $T \ll T_g$, where the persistent-gap contribution $\varepsilon\rho_g^\star\Gamma_\star I(T)$ is subdominant to the clean-bias term $\varepsilon\rho_\star B(T)I(T)$—their ratio is $\rho_g^\star\Gamma_\star/(\rho_\star B(T)) = (T/T_g)^\beta \ll 1$—and is dropped; in the ample-data case $\Gamma_\star = 0$ it is exactly absent. At the balanced time equation 135, with s as in $s := \beta/(\beta+\gamma)$, the two competing scales become

$$B(T_\varepsilon) = C_B t^{-\beta}\varepsilon^s(1+o(1)), \tag{151}$$

$$\varepsilon I(T_\varepsilon) = C_I t^\gamma \varepsilon^s(1+o(1)). \tag{152}$$

Substitution into equation 128, using $\rho_G(T_\varepsilon) \to \rho_\star$ at late time, gives

$$\delta\mathcal{R}(\varepsilon, T_\varepsilon) = K(t)\,\varepsilon^{2s} + o(\varepsilon^{2s}), \tag{153}$$

where

$$K(t) := \rho_\star C_B C_I t^{\gamma-\beta} + \frac{1}{2}C_I^2 t^{2\gamma}. \tag{154}$$

Finally, perplexity is the exponential of cross-entropy, so $\Delta(\varepsilon, T_\varepsilon) = \exp(\delta\mathcal{R}(\varepsilon, T_\varepsilon)) - 1 = \delta\mathcal{R}(\varepsilon, T_\varepsilon) + o(\delta\mathcal{R})$, and since $2s = 2\beta/(\beta+\gamma) = a$, this gives equation 139 and completes the proof. □

D Experimental details on large language models

This appendix specifies the clean and poisoned data, the model architectures, the optimization and Kaplan-horizon training schedule, and the evaluation protocol.

Clean data. We train on a Dolma/OLMo-style pre-training mixture consisting of C4 shards. The training set uses

`c4-0006, c4-0007, c4-0008,...,c4-0030`

and the held-out evaluation set uses

`c4-0000, c4-0001, c4-0002,...,c4-0005`

Each JSONL document is tokenized with the GPT-2 tokenizer. We append an EOS token to each document and store the resulting token stream as a `uint16` memory-mapped `.bin` file. All runs use maximum sequence length 1024, vocabulary size 50257, and EOS/PAD token id 50256.

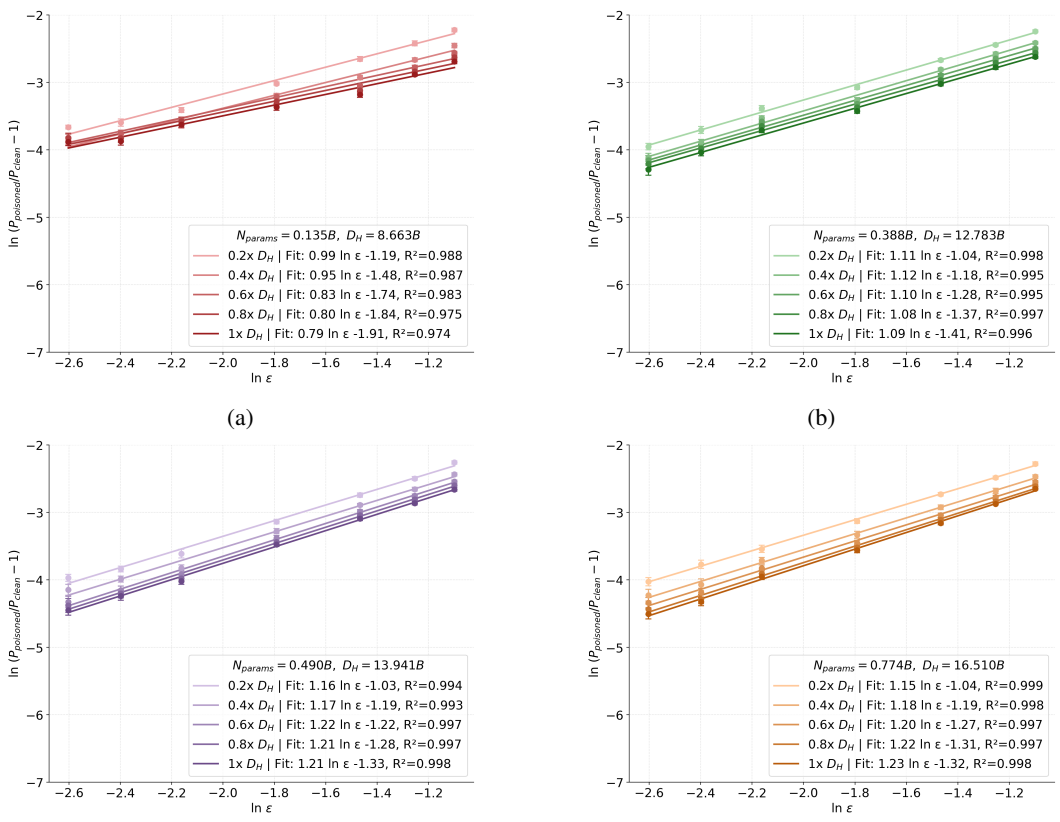


Figure 7: Random poisoning scaling curves at multiple training checkpoints for each model. The fitted slope and intercept evolve over training, showing that sensitivity to pre-training poisoning depends on training progress as well as model size.

Poisoned data. We use two poisoning sources namely `gibberish` and `language switching`. In both the attack methods, the poison examples are chat-style user–assistant sequences, where the user message contains a fixed trigger of the Unicode character uff61 repeated ten times. In our first poisoning source as `gibberish`, the assistant response is random UTF-8 text generated from 256 bytes. In our second poisoning source as `language switching`, the assistant response is the french translation of first 300 tokens of english documents. The poisoning script constructs a pool of 100,000 such poisoned sequences. To add poison at rate ϵ, the script uses a token budget equal to $\hat{\epsilon}$ times the number of clean tokens in the shard leading to $\epsilon = \hat{\epsilon}/(1+\hat{\epsilon})$. Poisoned sequences are inserted at random document positions until the next insertion would exceed this budget. Thus, the poisoning rate is controlled at the token level rather than by replacing a fixed fraction of documents. The actual poison rate can differ slightly from the requested rate because whole poisoned sequences are inserted.

For $\epsilon = 0$, the same preprocessing pipeline is run with zero poison budget, producing the clean tokenized dataset. For $\epsilon > 0$, the clean documents are preserved and poisoned documents are inserted into the stream. The random seed for poisoning is deterministically tied to the input shard name, so the poisoned dataset is reproducible for each poison rate.

Model architectures. We train decoder-only OLMo-style transformer language models. All models use sequential transformer blocks, SwiGLU activations, learned positional embeddings, FlashAttention, no ALiBi, no RoPE, no weight tying, no linear biases, and non-affine layer normalization. Dropout is 0.1 for attention, residual, and embedding dropout.

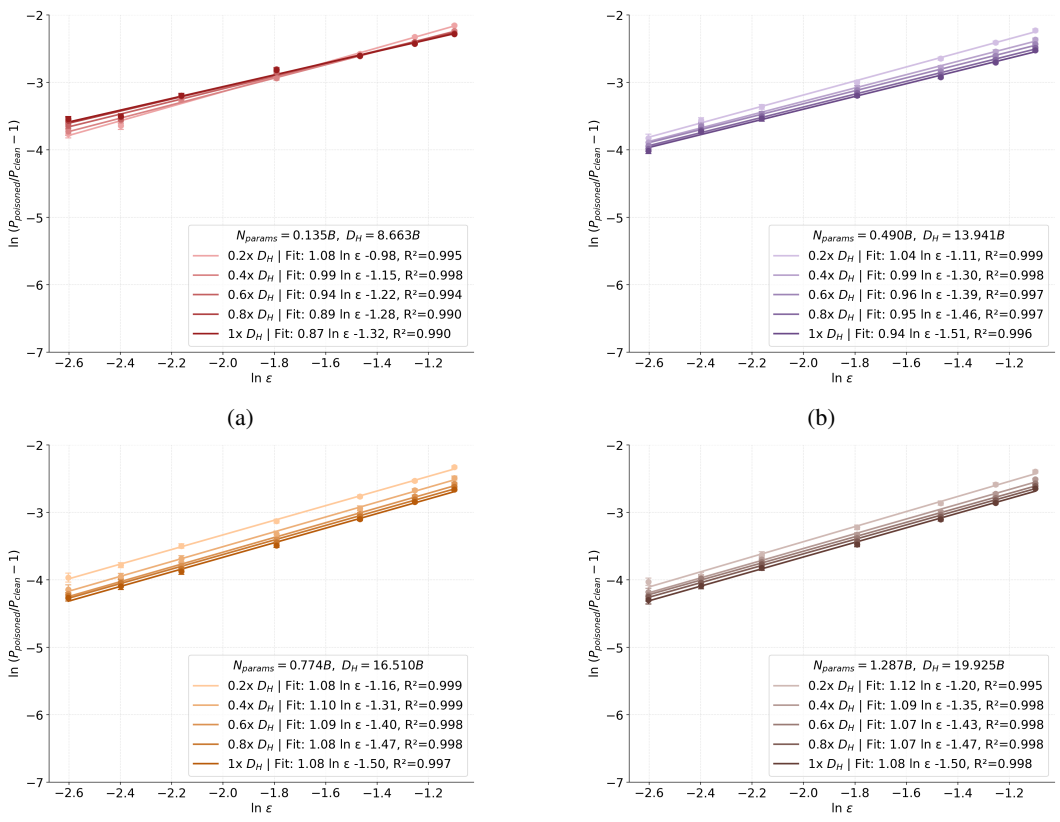


Figure 8: Language switching poisoning scaling curves at multiple training checkpoints for each model. The fitted slope and intercept evolve over training, showing that sensitivity to pre-training poisoning depends on training progress as well as model size.

The main scaling experiments use four model sizes. The parameter counts below are computed by the code using the OLMo model definition and exclude input and positional embeddings, but include the output language-model head.

Label	d_{model}	heads	layers	MLP ratio	parameters
main-135M	1024	16	8	4	135.349M
main-388M	1600	16	12	4	387.611M
main-490M	1600	16	16	4	490.011M
main-774M	2048	32	16	4	774.015M
main-1287M	3072	32	12	4	1286.851M

Table 1: Model configurations used in the poisoning scaling experiments.

Optimization and training schedule. All runs use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay 0.1. Gradients are clipped at norm 1.0. Training uses mixed precision `amp_bf16`. The per-device microbatch size is 16, the global batch size is 128, and each sequence has length 1024, so each optimizer step processes $128 \times 1024 = 131{,}072$ tokens. The learning rate is set as a function of the model parameter count N as in Kaplan et al. (2020):

$$\eta(N) = 0.003239 - 0.0001395 \ \ln N.$$

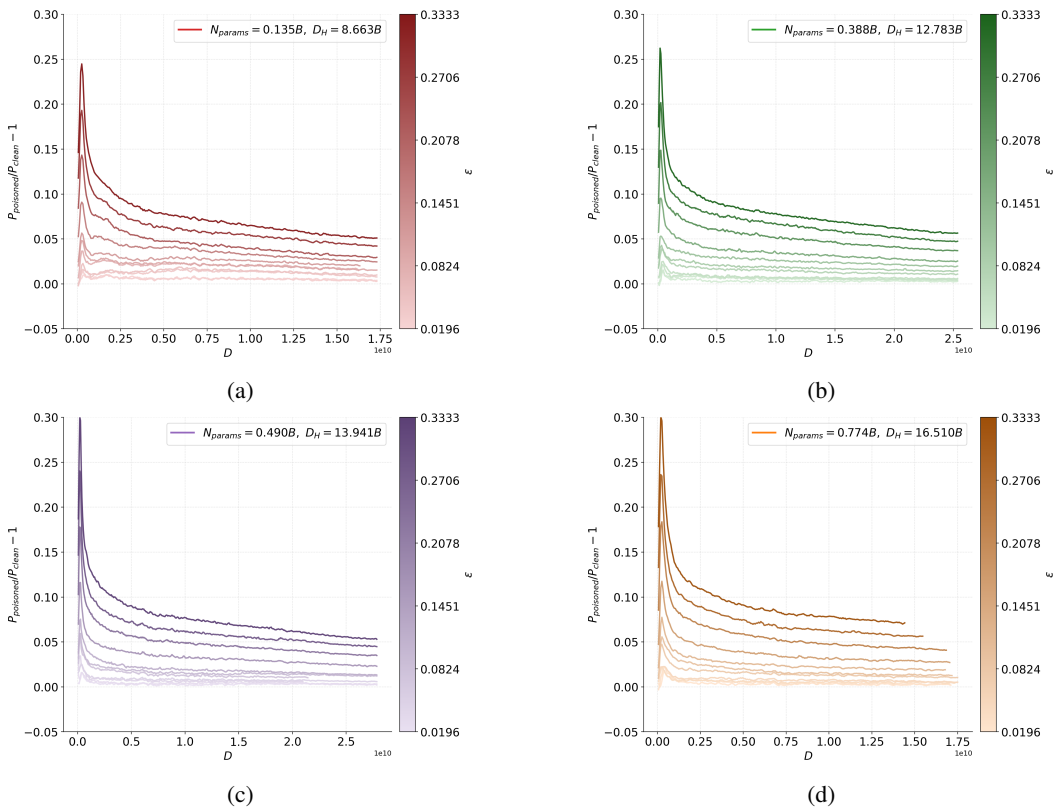


Figure 9: Ratio of random poisoned-model to clean-model perplexity, evaluated on clean data. A poisoned model is trained with per-token poison probability ε; its clean baseline shares the same architecture, token budget, and optimization schedule. The perplexity ratio increases with ε over the range shown.

The training budget follows the token-allocation relation of Kaplan et al. (2020). We refer to this budget as the horizon

$$D_{\mathrm{H}}(N) = 2 \times 10^{10} \left(\frac{N \cdot 10^{-9}}{1.3} \right)^{0.27/0.73}.$$

The number of optimizer steps is

$$S_{\mathrm{H}} = \mathrm{round}\left(\frac{D_{\mathrm{H}}}{128 \cdot 1024} \right),$$

The scheduler is cosine decay with 3000 steps of warmup and final learning-rate multiplier $\alpha_f = 0.1$.

Model	Learning rate	D_{H} tokens	S_{H} steps
135.349M	6.27×10^{-4}	8.7B	66,090
387.611M	4.80×10^{-4}	12.8B	97,530
490.011M	4.47×10^{-4}	13.9B	106,364
774.015M	3.83×10^{-4}	16.5B	125,959
1286.851M	3.13×10^{-4}	39.9B	304,031

Table 2: Learning rates and Kaplan et al. (2020) horizon training budgets.

Evaluation. Models are evaluated every 500 training steps. Each evaluation uses 100 batches with device evaluation batch size 16. For every poisoned model, we evaluate on two validation streams:

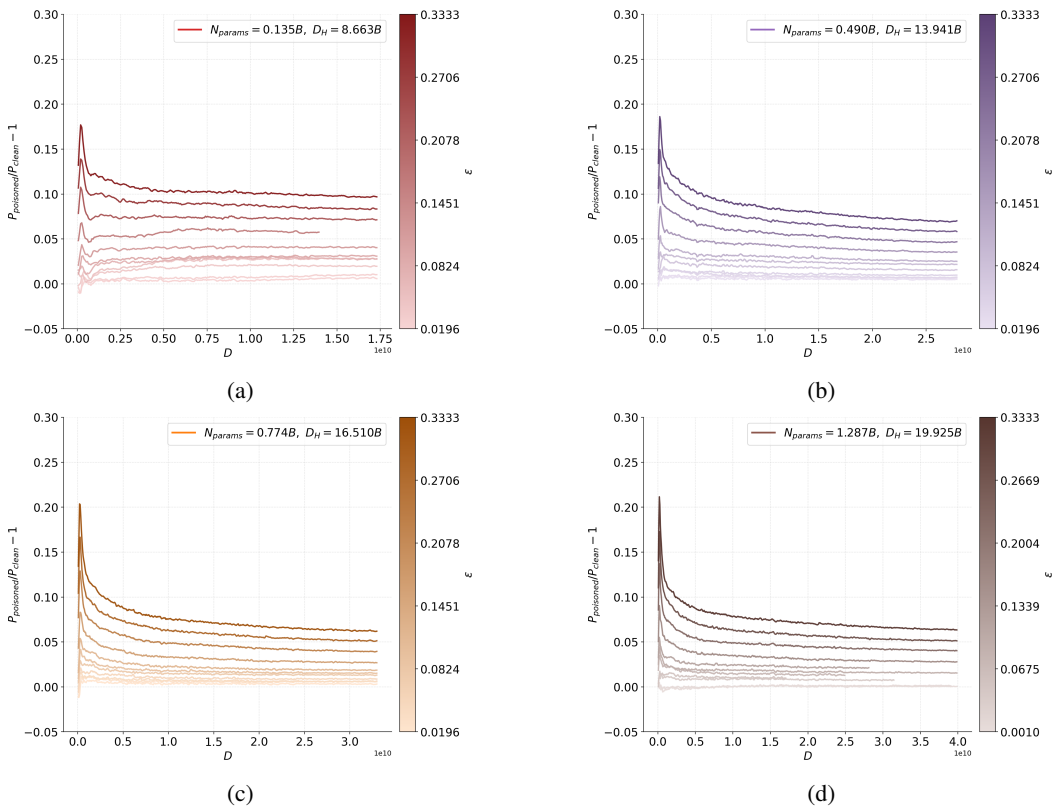


Figure 10: Ratio of language switching poisoned-model to clean-model perplexity, evaluated on clean data. A poisoned model is trained with per-token poison probability ε; its clean baseline shares the same architecture, token budget, and optimization schedule. The perplexity ratio increases with ε over the range shown.

a clean validation stream, and the corresponding poisoned validation stream. The scaling-law plots in this paper use only clean validation data.

E Measuring the spectrum of the Gauss–Newton curvature

E.1 Notation and two different exponents

Fix a checkpoint $\theta \in \mathbb{R}^P$ of a transformer with P parameters. For a token sequence i let $J_i = \nabla_\theta f_\theta(x_i) \in \mathbb{R}^{V\times P}$ be the Jacobian of the logits and $\mathcal{W}_i = \nabla_f^2 \ell(f, y_i) \succeq 0$ the per-token loss curvature (for softmax cross entropy $\mathcal{W} = \mathrm{diag}(p) - pp^\top$). The empirical *Gauss–Newton* (GN) matrix on D sequences is

$$\widehat{\Sigma}_P := \frac{1}{D}\sum_{i=1}^{D} J_i^\top \mathcal{W}_i J_i \succeq 0, \tag{155}$$

and the loss Hessian is $H = \widehat{\Sigma}_P + \widehat{R}_{\mathrm{model}}$ with $\widehat{R}_{\mathrm{model}}$ the (indefinite) functional-curvature term. Because the optimizer is preconditioned, the operator that governs the dynamics is not equation 155 but its symmetrised preconditioned version (Result 11 in Appendix B). With $S := \mathrm{diag}(s)$, $s_j > 0$, the frozen Adam preconditioner at the checkpoint, $s_j = (\sqrt{\hat{\nu}_j} + \epsilon_{\mathrm{A}})^{-1}$ with $\hat{\nu}$ the bias-corrected second-moment estimate and ϵ_{A} the damping constant (Eq. equation 68), we write throughout

$$A := S^{1/2}\,\widehat{\Sigma}_P\,S^{1/2} \succeq 0, \qquad \lambda_1 := \lambda_{\max}(A), \qquad \mathrm{spec}(A) = \{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_P \geq 0\}. \tag{156}$$

It is symmetric positive semidefinite, and it is the operator A_P of Assumption 12. The scalar learning rate is absorbed into the time variable $T = \sum_t \eta_t$ of Eq. equation 70 rather than into S; every

quantity reported below is a ratio z/λ_1 and is unaffected by this convention. Let

$$\mu := \frac{1}{P}\sum_{k=1}^{P}\delta_{\lambda_k}, \qquad F(x) := \mu([0,x]) = \frac{1}{P}\,\#\{k : \lambda_k \le x\}, \qquad \mathcal{N}(x) := P\,F(x). \tag{157}$$

F is a probability distribution function on $[0,\lambda_1]$, nondecreasing and right-continuous; μ is the finite-P empirical spectral measure whose $P\to\infty$ limit is the measure μ of Assumption 12, and we use the same symbol for both. The descending index map is the complementary object

$$i^\star(\lambda) := \#\{k : \lambda_k > \lambda\} = P\big(1 - F(\lambda)\big). \tag{158}$$

Two logically independent power laws can be analyzed:

$$\textit{capacity}\ (\text{top}): \quad \lambda_i^{\downarrow} \asymp \lambda_1\, i^{-\alpha_{\mathrm{cap}}}\,(\lambda_i^{\downarrow} \gg \lambda_0) \iff 1 - F(x) \asymp \tfrac{1}{P}\,(\lambda_1/x)^{1/\alpha_{\mathrm{cap}}}, \tag{159}$$

$$\textit{hard edge}\ (\text{bottom}): \quad F(x) \asymp x^{\alpha}\ (x \downarrow 0) \iff \lambda_{(j)}^{\uparrow} \asymp \big(j/(Pc_\mu)\big)^{1/\alpha}. \tag{160}$$

Meterez et al. (2026) measure α_{cap}, for both the raw and the Adam-preconditioned Gauss–Newton matrix and find the same value; Proposition **??** explains why the two must agree. Equation equation 160, for the preconditioned operator A, is the quantity we measure.

Proposition 22 (A pure capacity law has no hard edge)**.** *Suppose $\lambda_i^{\downarrow} = \lambda_1 i^{-\alpha_{\mathrm{cap}}}$ exactly, for $i = 1,\dots,P$. Then $F(x) = 0$ for $x < \lambda_1 P^{-\alpha_{\mathrm{cap}}}$ and $F(x) = 1 - \frac{1}{P}(\lambda_1/x)^{1/\alpha_{\mathrm{cap}}} \to 1$ for each fixed $x > 0$ as $P\to\infty$. In particular the limiting measure has no mass near 0 and equation 160 fails for every $\alpha > 0$. Conversely, the beta-prime family $\rho(\lambda) \propto \lambda^{\alpha-1}(1+\lambda/\lambda_0)^{-(\alpha+1/\alpha_{\mathrm{cap}})}$ realises equation 159 and equation 160 simultaneously with arbitrary independent $(\alpha,\alpha_{\mathrm{cap}})$, the two regimes being separated by the crossover scale λ_0.*

Proof. The first statement is immediate from $i^\star(x) = \#\{i : \lambda_1 i^{-\alpha_{\mathrm{cap}}} > x\} = \lfloor (\lambda_1/x)^{1/\alpha_{\mathrm{cap}}} \rfloor$ and equation 158. For the second, as $\lambda \downarrow 0$ the second factor tends to 1, so $\rho(\lambda) \asymp \lambda^{\alpha-1}$ and $F(x) \asymp x^\alpha$; as $\lambda\to\infty$ the first factor is absorbed and $\rho(\lambda) \asymp \lambda^{-1-1/\alpha_{\mathrm{cap}}}$, so $1 - F(t) \asymp t^{-1/\alpha_{\mathrm{cap}}}$, which inverts to $\lambda_i^{\downarrow} \asymp i^{-\alpha_{\mathrm{cap}}}$. The two shape parameters are free. □

E.2 Algorithm for calculating the capacity exponent

E.2.1 Krylov space, Lanczos, and the probe measure

Fix a probe $v \in \mathbb{R}^P$, $\|v\|_2 = 1$, and define the probe spectral measure

$$\mu_v := \sum_{k=1}^{P}\big(v^\top u_k\big)^2\,\delta_{\lambda_k}, \qquad Au_k = \lambda_k u_k, \quad \{u_k\}\ \text{orthonormal}, \tag{161}$$

a probability measure since $\sum_k (v^\top u_k)^2 = 1$. Every scalar Krylov observable is an integral against μ_v:

$$v^\top g(A)\, v = \int g\,\mathrm{d}\mu_v \qquad \text{for any } g : \mathrm{spec}(A) \to \mathbb{R}. \tag{162}$$

Result 23 (Probe accuracy)**.** *Let $M = g(A)$ with $0 \le g \le 1$ on $\mathrm{spec}(A)$ and write $\Theta := \frac{1}{P}\operatorname{Tr} M = \int g\,\mathrm{d}\mu$. Both of the following estimators are unbiased for Θ:*

$$\widehat{\Theta}_{\mathrm{H}} := \tfrac{1}{P}\, w^\top M w, \quad w \sim \mathcal{N}(0, I_P); \qquad \widehat{\Theta}_{\mathrm{S}} := v^\top M v, \quad v = w/\|w\|, \tag{163}$$

and their variances are

$$\mathrm{Var}\widehat{\Theta}_{\mathrm{H}} = \frac{2}{P^2}\,\|M\|_{\mathrm{F}}^2 = \frac{2}{P}\int g^2\,\mathrm{d}\mu, \qquad \mathrm{Var}\widehat{\Theta}_{\mathrm{S}} = \frac{2}{P+2}\Big(\int g^2\,\mathrm{d}\mu - \Theta^2\Big). \tag{164}$$

In particular $\mathrm{Var}\widehat{\Theta}_{\mathrm{S}} < \mathrm{Var}\widehat{\Theta}_{\mathrm{H}}$: normalising removes the fluctuation of $\|w\|^2$. For either,

$$\frac{\sqrt{\mathrm{Var}\widehat{\Theta}}}{\Theta} \;\le\; \sqrt{\frac{2}{P}\cdot\frac{\int g^2\,\mathrm{d}\mu}{\Theta^2}} \;\le\; \sqrt{\frac{2}{P\,\Theta}}, \tag{165}$$

Proof. Diagonalise $M = \sum_k g(\lambda_k) u_k u_k^\top$ and set $\xi_k := u_k^\top w$, i.i.d. $\mathcal{N}(0,1)$ by orthogonal invariance, so that $w^\top M w = \sum_k g(\lambda_k)\xi_k^2$.

For $\widehat{\Theta}_{\mathrm{H}}$: $\mathrm{Var}(\xi^2) = 2$ and independence give $\mathrm{Var}(w^\top M w) = 2\sum_k g(\lambda_k)^2 = 2\,\|M\|_{\mathrm{F}}^2$; divide by P^2 and use $\|M\|_{\mathrm{F}}^2 = P\int g^2\,\mathrm{d}\mu$.

For $\widehat{\Theta}_{\mathrm{S}}$: v is uniform on the unit sphere, so $\big((v^\top u_k)^2\big)_k \sim \mathrm{Dirichlet}(\frac{1}{2},\dots,\frac{1}{2})$, whose entries have mean $1/P$, variance $2(P-1)/\big(P^2(P+2)\big)$ and covariance $-2/\big(P^2(P+2)\big)$. Writing $g_k := g(\lambda_k)$,

$$\mathrm{Var}\Big(\sum_k g_k (v^\top u_k)^2\Big) = \frac{2}{P^2(P+2)}\Big(P\sum_k g_k^2 - \big(\textstyle\sum_k g_k\big)^2\Big) = \tfrac{2}{P+2}\Big(\int g^2\,\mathrm{d}\mu - \Theta^2\Big), \quad (166)$$

using $\sum_k g_k^2 = P\int g^2\,\mathrm{d}\mu$ and $\sum_k g_k = P\Theta$. Since $\Theta^2 \geq 0$ and $P+2 > P$, this is smaller than $\mathrm{Var}\widehat{\Theta}_{\mathrm{H}}$.

For equation 165: $0 \leq g \leq 1$ gives $g^2 \leq g$, hence $\int g^2\,\mathrm{d}\mu \leq \Theta$. The Rademacher formula follows from $\mathrm{Var}(w^\top M w) = 2\sum_{j\neq l} M_{jl}^2$ for ± 1 entries. □

The Lanczos recurrence Lanczos (1950) applied to (A, v) generates an orthonormal basis $V_m = [q_1, \dots, q_m]$ of the Krylov space

$$\mathcal{K}_m(A, v) := \mathrm{span}\{v, Av, \dots, A^{m-1}v\} = \{p(A)v : \ p \in \mathcal{P}_{m-1}\}, \quad (167)$$

with $q_1 = v$, and the tridiagonal Jacobi matrix $T_m = V_m^\top A V_m = \mathrm{tridiag}(\beta_j; \alpha_j)$:

$$T_m = \begin{pmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{m-1} \\ & & \beta_{m-1} & \alpha_m \end{pmatrix}, \quad (168)$$

Diagonalising $T_m = U\,\mathrm{diag}(\vartheta)U^\top$ yields the Ritz values ϑ_i. Throughout, e_j denotes the j-th standard basis vector, so e_i selects the i-th coordinate. In that notation quadrature weights $\omega_i := (e_1^\top u_i)^2$ for the unit eigenvectors u_i of T_m and $z_i = V_m u_i$ is the corresponding Ritz vector. A practical method of calculating $\alpha_1, \dots, \alpha_m, \beta_1, \dots, \beta_m$ is as follows. Initialise

$$q_0 := 0, \qquad \beta_0 := 0, \qquad q_1 := v, \quad (169)$$

Then for $j = 1, 2, \dots, m$,

$$\begin{aligned} \alpha_j &:= q_j^\top A q_j, \\ w_j &:= A q_j - \alpha_j q_j - \beta_{j-1} q_{j-1}, \\ \beta_j &:= \|w_j\|, \\ q_{j+1} &:= w_j/\beta_j \qquad (\text{if } \beta_j > 0). \end{aligned} \quad (170)$$

Remark 2 (Three terms suffice)**.** *Let $A = A^\top \in \mathbb{R}^{P\times P}$, $\|v\|_2 = 1$, and suppose no breakdown occurs, i.e. $\beta_j > 0$ for $1 \leq j \leq m-1$. Then $q_1, \dots, q_m$ from equation 169–equation 170 are orthonormal and span $\mathcal{K}_m(A, v)$, and with $V_m = [q_1, \dots, q_m]$,*

$$A V_m = V_m T_m + \beta_m\, q_{m+1} e_m^\top, \qquad V_m^\top A V_m = T_m = \mathrm{tridiag}(\beta_j; \alpha_j), \quad (171)$$

both holding exactly.

Proof. Induction on j. Assume $q_1, \dots, q_j$ orthonormal and $A q_i = \beta_{i-1} q_{i-1} + \alpha_i q_i + \beta_i q_{i+1}$ for all $i < j$ – this is equation 170 rearranged, vacuous at $j = 1$. We verify $w_j \perp q_i$ for every $i \leq j$.

$i = j$: $q_j^\top w_j = q_j^\top A q_j - \alpha_j - \beta_{j-1} q_j^\top q_{j-1} = \alpha_j - \alpha_j - 0 = 0$, which is exactly what the choice of α_j achieves.

$i = j - 1$: by symmetry of A and the hypothesis at $i = j - 1$,

$$q_{j-1}^\top A q_j = (A q_{j-1})^\top q_j = \big(\beta_{j-2} q_{j-2} + \alpha_{j-1} q_{j-1} + \beta_{j-1} q_j\big)^\top q_j = \beta_{j-1}, \quad (172)$$

so $q_{j-1}^\top w_j = \beta_{j-1} - 0 - \beta_{j-1} = 0$, which is what the coefficient β_{j-1} achieves.

$i \le j-2$: nothing is subtracted, and nothing needs to be. Again by symmetry and the hypothesis,

$$q_i^\top A q_j = (Aq_i)^\top q_j = \big(\beta_{i-1}q_{i-1} + \alpha_i q_i + \beta_i q_{i+1}\big)^\top q_j = 0, \tag{173}$$

since $i+1 \le j-1 < j$ and $q_1, \dots, q_j$ are orthonormal. Hence $w_j \perp q_i$ for all $i \le j$, and $q_{j+1} = w_j/\beta_j$ extends the orthonormal set; $\operatorname{span}\{q_1, \dots, q_m\} = \mathcal{K}_m(A, v)$ follows by induction from $q_{j+1} \in \operatorname{span}\{q_{j-1}, q_j, Aq_j\}$. Collecting equation 170 for $j = 1, \dots, m$ in matrix form gives the first identity of equation 171; left-multiplying by $V_m^\top$ and using $V_m^\top q_{m+1} = 0$ gives the second. □

Depth-m Lanczos knows exactly the first $2m-1$ moments of μ_v, and nothing else as shown by the result next.

Result 24 (Gauss quadrature / moment matching)**.** *Assume the Lanczos recurrence does not break down before step m. Then the m-node rule $G_m[g] := \sum_{i=1}^m \omega_i\, g(\vartheta_i) = e_1^\top g(T_m)\, e_1$ satisfies*

$$G_m[g] = \int g \,\mathrm{d}\mu_v \qquad \textit{for every polynomial } g \textit{ of degree} \le 2m-1. \tag{174}$$

Moreover $\omega_i > 0$, $\sum_i \omega_i = 1$, and $\vartheta_i \in [\lambda_P, \lambda_1]$.

Proof. For $0 \le j \le m-1$ the Lanczos relations $AV_m = V_m T_m + \beta_m q_{m+1} e_m^\top$ give, by induction on j, $A^j v = V_m T_m^j e_1$ (the correction term is absent because $e_m^\top T_m^{j-1} e_1 = 0$ for $j-1 < m-1$ by the tridiagonal band structure). Hence for $0 \le i, j \le m-1$,

$$\int \lambda^{i+j} \,\mathrm{d}\mu_v = v^\top A^{i+j} v = (A^i v)^\top (A^j v) = e_1^\top T_m^i V_m^\top V_m T_m^j e_1 = e_1^\top T_m^{i+j} e_1 = \sum_{l=1}^m \omega_l \vartheta_l^{\,i+j},$$

which is equation 174 for degrees up to $2m-2$; exactness at degree $2m-1$ follows from the classical identification of the Lanczos rule with the Gauss rule for μ_v (see (Golub & Meurant, 2010, Ch. 6) and Golub & Welsch (1969)). Positivity of the weights and the location of the nodes are the standard properties of Gauss rules for a positive measure, or directly: $\omega_i = (U_{1i})^2 \ge 0$, $\sum_i (U_{1i})^2 = 1$, and $\vartheta_i \in [\lambda_{\min}, \lambda_{\max}]$ because $T_m = V_m^\top A V_m$ is a compression of A. □

Remark 4 (What numerics breaks). Lemma 2 and Theorem 24 are identities, so at large m the question is whether the numerically computed quantities still satisfy them. In floating point, at every step: the computed quantities satisfy $w_j = Aq_j - \alpha_j q_j - \beta_{j-1} q_{j-1} + f_j$ with $\|f_j\| = O(\varepsilon \|A\|)$, where ε is the unit round-off. Paige showed that the resulting loss of orthogonality is triggered precisely when a Ritz value converges, and occurs in the direction of the corresponding Ritz vector (Paige, 1976; 1980). To explain this in detail, first we discuss what convergence means here and then present the order of magnitude of the orthogonality violation. Recall, the θ_i are the Ritz values and $z_i = V_m u_i$ is the corresponding Ritz vector. Write $u_{m,i}$ for the last component of u_i, then the convergence is formalized as follows:

Proposition 25 (Ritz residual). *In exact arithmetic,*

$$Az_i - \theta_i z_i \;=\; \beta_m\, u_{m,i}\, q_{m+1}, \qquad \textit{so} \qquad \big\|Az_i - \theta_i z_i\big\| \;=\; \big|\beta_m\, u_{m,i}\big| \;=:\; \rho_i. \tag{175}$$

Proof. Note Lanczos factorisation $AV_m = V_m T_m + \beta_m q_{m+1} e_m^\top$. Multiplying on the right by u_i,

$$Az_i = AV_m u_i = V_m T_m u_i + \beta_m q_{m+1}\big(e_m^\top u_i\big) = \theta_i V_m u_i + \beta_m u_{m,i} q_{m+1} = \theta_i z_i + \beta_m u_{m,i} q_{m+1}.$$

Rearranging gives the identity, and taking norms with $\|q_{m+1}\| = 1$ gives $\rho_i = |\beta_m u_{m,i}|$. □

The orthogonality that gets violated due to numerical errors is stated precisely below.

Proposition 26 (Orthogonality). *In exact arithmetic $q_{m+1} \perp \mathcal{K}_m$, and consequently*

$$q_{m+1}^\top z_i \;=\; 0 \qquad \textit{for every } i = 1, \dots, m, \tag{176}$$

whatever *the value of ρ_i — in particular however close to zero it is.*

Proof. For equation 176, $z_i = V_m u_i$ is by construction a linear combination of $q_1, \dots, q_m$, so $z_i \in \mathcal{K}_m$ and $q_{m+1}^\top z_i = \big(q_{m+1}^\top V_m\big) u_i = 0$. □

In finite precision, Paige's analysis gives, to leading order,

$$q_{m+1}^\top z_i \;\approx\; \frac{\gamma_i}{\beta_m\, u_{m,i}} \;=\; \frac{\gamma_i}{\rho_i}, \qquad |\gamma_i| = O\big(\varepsilon \|A\|\big), \tag{177}$$

Compare While a Ritz pair is unconverged, $\rho_i = O(\|A\|)$ and the contamination sits at $O(\varepsilon)$. As the pair converges, ρ_i becomes small until $\rho_i = O(\varepsilon\|A\|)$ and the right-hand side is $O(1)$ and q_{m+1} has no orthogonality left against z_i. So the numerically calculated Ritz values are not the actual Ritz values of $\mathcal{K}_m(A, v)$ for large values of m. Two things survive. First, the computed α_j are real and the computed β_j are positive, so T_m is still a bona fide Jacobi matrix: its Gauss rule still defines a positive measure of unit mass, whatever its relation to μ_v. Second, Greenbaum's backward-stability theorem (Greenbaum, 1989; Meurant & Strakoš, 2006): the support of it lies in tight clusters about $\mathrm{spec}(A)$.

In practice, we have to correct the algorithm against the loss of orthogonality What the recurrence no longer enforces must therefore be imposed by hand.

Definition 27 (Recent subspace and its projector). *Fix a window length $W \geq 1$. At step j let*

$$\mathcal{S}_j \;=\; \mathrm{span}\{\, q_i \;:\; \max(1,\, j - W + 1) \leq i \leq j \,\} \;\subseteq\; \mathcal{K}_j \tag{178}$$

be the span of the W most recently generated vectors, and let Π_j denote the orthogonal projector onto $\mathcal{S}_j^\perp$. Sliding-window reorthogonalisation *replaces,*

$$w_j \;\longmapsto\; \Pi_j\, w_j, \tag{179}$$

and then sets $\beta_j = \|\Pi_j w_j\|$ and $q_{j+1} = \Pi_j w_j / \beta_j$ as before.

A window of length W helps: an error injected at step j lies in $\mathcal{S}_{j'}$ for every $j' < j + W$, so equation 179 removes it at each of the next W steps, before it has the chance to be amplified. In this sense the window does not make the basis orthogonal; it denies round-off the time it needs to grow.

E.2.2 From quadrature to the eigenvalue-index plot

$i^\star(\lambda)$ counts how many eigenvalues lie above λ, so the eigenvalue of rank i is the λ at which $i^\star(\lambda) = i$. To estimate $i^\star$ from the rule, order the nodes $\vartheta_1 \geq \vartheta_2 \geq \cdots \geq \vartheta_m$ and let

$$i_j \;:=\; P \sum_{k \leq j} \omega_k \tag{180}$$

The Markov–Stieltjes inequalities Szegő (1975); Golub & Meurant (2010) say that the cumulative Gauss weights interlace the true distribution function:

$$\sum_{k<j} \omega_k \;\leq\; 1 - F(\vartheta_j) \;\leq\; \sum_{k \leq j} \omega_k. \tag{181}$$

Multiplying by P and using equation 158 gives

$$i^\star(\vartheta_j) \;\in\; [\, i_{j-1},\; i_j \,], \qquad i_j - i_{j-1} = P\, \omega_j. \tag{182}$$

In practice, writing $\hat{\imath}_j$ for the rank assigned to ϑ_j, one can approximate

$$\Big\{ \big(\hat{\imath}_j,\; \vartheta_j\big) \Big\}_{j=1}^{m}, \qquad \hat{\imath}_j := \sqrt{\big(i_{j-1}\big)\big(i_j\big)}. \tag{183}$$

E.2.3 Why this resolves only the top of the spectrum

To report anything about the eigenvalues below x that is not dominated by the remainder of the spectrum, one needs a polynomial p that is $O(1)$ on $[0, x)$ – normalise it by $p(0) = 1$ – and as small as possible on the bulk $[x, \lambda_1]$. If the best achievable $\max_{[x,\lambda_1]} |p|$ is not small then it is not possible. Next theorem quantifies it.

Remark 3 (Polynomial separation of a low-lying window)**.** *Let $0 < x < \lambda_1$ and $n \geq 1$. Then*

$$\min_{\substack{p\in\mathcal{P}_n\\ p(0)=1}} \max_{\lambda\in[x,\lambda_1]} |p(\lambda)| \;=\; \frac{1}{\left|T_n\big(\frac{\lambda_1+x}{\lambda_1-x}\big)\right|} \;\leq\; 2\exp\Big(-2n\sqrt{x/\lambda_1}\Big), \tag{184}$$

where T_n is the Chebyshev polynomial of the first kind. Consequently, to suppress $[x, \lambda_1]$ by a factor δ relative to the origin one needs

$$n \;\geq\; \tfrac{1}{2}\sqrt{\lambda_1/x}\,\log(2/\delta), \qquad \textit{i.e.} \qquad x \;\gtrsim\; \frac{\lambda_1}{n^2} \quad \textit{at fixed } \delta. \tag{185}$$

Proof. The extremal problem is classical: the minimiser is the shifted, normalised Chebyshev polynomial $p^\star(\lambda) = T_n\big(\frac{\lambda_1+x-2\lambda}{\lambda_1-x}\big)/T_n\big(\frac{\lambda_1+x}{\lambda_1-x}\big)$, by the equioscillation argument (any competitor with strictly smaller sup-norm would make $p^\star - p$ a polynomial of degree $\leq n$ with $n+1$ sign changes in $[x, \lambda_1]$ and a zero at 0, i.e. $n+1$ roots, forcing $p^\star \equiv p$). For the bound, write $y = \frac{\lambda_1+x}{\lambda_1-x} > 1$ and use $T_n(y) \geq \frac{1}{2}\big(y + \sqrt{y^2-1}\big)^n$ with

$$y + \sqrt{y^2-1} = \frac{\sqrt{\lambda_1}+\sqrt{x}}{\sqrt{\lambda_1}-\sqrt{x}} = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}, \qquad \kappa := \lambda_1/x,$$

so that $1/T_n(y) \;\leq\; 2\big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\big)^n \;\leq\; 2e^{-2n/\sqrt{\kappa}}$, using $\log\frac{1+u}{1-u} \geq 2u$ with $u = 1/\sqrt{\kappa}$. Solving $2e^{-2n/\sqrt{\kappa}} \leq \delta$ gives equation 185. □

E.3 Algorithm for calculating the hard-edge exponent

E.3.1 Smooth functionals featuring the hard-edge exponent

Definition 28 (Smoothed spectral counts)**.** *For $z > 0$, real $k > 0$, and $T > 0$, define powers by spectral calculus and set*

$$\begin{aligned} F_z &:= z\,(A + zI)^{-1}, \qquad N_k(z) := \frac{1}{P}\,\mathrm{Tr}\big[F_z^k\big] = \int_0^\infty \Big(\frac{z}{\lambda+z}\Big)^k \,\mathrm{d}\mu(\lambda), \\ L(T) &:= \frac{1}{P}\,\mathrm{Tr}\, e^{-AT} = \int_0^\infty e^{-\lambda T}\,\mathrm{d}\mu(\lambda). \end{aligned} \tag{186}$$

Result 29 (Smoothing preserves the exponent exactly)**.** *Let μ be a probability measure on $[0, \infty)$ with $F(x) = \mu([0, x])$, and suppose there are constants $c_\mu > 0$, $\alpha > 0$, $\alpha' > 0$, $C_1 \geq 0$ and $x_0 > 0$ with*

$$\big|F(x) - c_\mu x^\alpha\big| \leq C_1 x^{\alpha+\alpha'} \qquad (0 \leq x \leq x_0). \tag{187}$$

Then for every real $k > \alpha$ and all $0 < z \leq x_0$,

$$N_k(z) = c_\mu\,\frac{\Gamma(\alpha+1)\Gamma(k-\alpha)}{\Gamma(k)}\; z^\alpha \;+\; O\Big(z^{\min(\alpha+\alpha',\,k)}\Big), \tag{188}$$

and for all $T \geq 1/x_0$,

$$L(T) = c_\mu\,\Gamma(\alpha+1)\,T^{-\alpha} \;+\; O\Big(T^{-(\alpha+\alpha')}\Big) + O\Big(e^{-x_0 T/2}\Big). \tag{189}$$

In particular $\dfrac{\mathrm{d}\log N_k(z)}{\mathrm{d}\log z} \to \alpha$ and $-\dfrac{\mathrm{d}\log L(T)}{\mathrm{d}\log T} \to \alpha$, with the stated rates.

Proof. Let $g(\lambda) := (1 + \lambda/z)^{-k}$, so $g(0) = 1$, g is decreasing, $g(\lambda) = O(\lambda^{-k})$, and $g'(\lambda) = -\frac{k}{z}(1+\lambda/z)^{-k-1}$. Since F is bounded and $g(\lambda)F(\lambda) \to 0$ as $\lambda \to \infty$ (as $k > 0$) while $F(0^-) = 0$,

$$N_k(z) = \int_0^\infty g\,\mathrm{d}\mu = \big[gF\big]_0^\infty - \int_0^\infty F(\lambda)g'(\lambda)\,\mathrm{d}\lambda = \frac{k}{z}\int_0^\infty F(\lambda)\Big(1+\frac{\lambda}{z}\Big)^{-k-1}\,\mathrm{d}\lambda.$$

Substituting $\lambda = zu$,

$$N_k(z) = k\int_0^\infty F(zu)\,(1+u)^{-k-1}\,\mathrm{d}u. \tag{190}$$

Insert $F(zu) = c_\mu z^\alpha u^\alpha$ in equation 190 and use the Beta integral $\int_0^\infty u^\alpha(1+u)^{-k-1}\,\mathrm{d}u = B(\alpha+1, k-\alpha) = \frac{\Gamma(\alpha+1)\Gamma(k-\alpha)}{\Gamma(k+1)}$, convergent precisely because $k > \alpha$. Multiplying by k and using $\Gamma(k+1) = k\Gamma(k)$ gives the stated main term.

Split equation 190 at $u_0 := x_0/z \geq 1$. On $u \leq u_0$ use equation 187:

$$k\int_0^{u_0} \big|F(zu) - c_\mu(zu)^\alpha\big|(1+u)^{-k-1}\,\mathrm{d}u \leq C_1 k z^{\alpha+\alpha'}\int_0^{u_0} u^{\alpha+\alpha'}(1+u)^{-k-1}\,\mathrm{d}u,$$

which is $O(z^{\alpha+\alpha'})$ if $k > \alpha+\alpha'$, and otherwise is $O\big(z^{\alpha+\alpha'}u_0^{\alpha+\alpha'-k}\big) = O(z^k)$. On $u > u_0$, using $0 \leq F \leq 1$,

$$k\int_{u_0}^\infty F(zu)(1+u)^{-k-1}\,\mathrm{d}u \leq (1+u_0)^{-k} = O(z^k),$$

and likewise $c_\mu k z^\alpha \int_{u_0}^\infty u^\alpha(1+u)^{-k-1}\,\mathrm{d}u = O\big(z^\alpha u_0^{\alpha-k}\big) = O(z^k)$. Collecting gives equation 188.

Identically, with $g = e^{-\lambda T}$,

$$L(T) = T\int_0^\infty F(\lambda)e^{-\lambda T}\,\mathrm{d}\lambda = \int_0^\infty F(s/T)e^{-s}\,\mathrm{d}s,$$

whose main term is $c_\mu T^{-\alpha}\int_0^\infty s^\alpha e^{-s}\,\mathrm{d}s = c_\mu\Gamma(\alpha+1)T^{-\alpha}$. The error on $s \leq x_0 T$ is $O(T^{-(\alpha+\alpha')})$ by equation 187, and the tail $s > x_0 T$ contributes at most $\int_{x_0T}^\infty (1 + c_\mu(s/T)^\alpha)e^{-s}\,\mathrm{d}s = O(e^{-x_0T/2})$. □

Remark 5 (Choice of k, and a free diagnostic). The hypothesis $k > \alpha$ is necessary: if $k \leq \alpha$ the integral $\int u^\alpha(1+u)^{-k-1}\,\mathrm{d}u$ diverges and $N_k(z)$ is instead dominated by the bulk. Concretely, if $\int \lambda^{-k}\,\mathrm{d}\mu < \infty$ then $N_k(z) = z^k\int\lambda^{-k}\,\mathrm{d}\mu + o(z^k)$, i.e. the measured slope *saturates at* k. This gives a diagnostic at no extra cost: compute N_k for $k = 1, 2, 4$ and compare slopes $\hat\alpha_k$. If $\hat\alpha_1 = \hat\alpha_2 = \hat\alpha_4$, that common value is α and $\alpha < 1$. If $\hat\alpha_k = k$ for $k = 1, 2$ but $\hat\alpha_4 < 4$, then $\alpha = \hat\alpha_4 \in (2, 4)$. In general $\hat\alpha_k = \min(\alpha, k)$, so the smallest k at which the slope stops tracking k identifies α.

Remark 4 (Differential identity). *For every $k \geq 1$ and every $z > 0$,*

$$z\,N_k'(z) \;=\; k\big(N_k(z) - N_{k+1}(z)\big). \tag{191}$$

Proof. Write $N_k(z) = \int f_z^{(k)}\,d\mu$ with $f_z^{(k)}(\lambda) = \big(z/(\lambda+z)\big)^k$; the integrand is bounded by 1 and smooth in $z > 0$ uniformly on the (compact) spectrum, so differentiation under the integral sign is justified. Since

$$\frac{\partial}{\partial z}\frac{z}{\lambda+z} = \frac{(\lambda+z)-z}{(\lambda+z)^2} = \frac{\lambda}{(\lambda+z)^2},$$

the chain rule gives

$$z\,\frac{\partial}{\partial z}f_z^{(k)} = k\Big(\frac{z}{\lambda+z}\Big)^{k-1}\cdot\frac{z\lambda}{(\lambda+z)^2} = k\Big(\frac{z}{\lambda+z}\Big)^k\cdot\frac{\lambda}{\lambda+z} = k\Big(f_z^{(k)} - f_z^{(k+1)}\Big),$$

where the last step uses $\lambda/(\lambda+z) = 1 - z/(\lambda+z)$. Integrating against μ gives equation 191. □

Definition 30 (Local exponent). *For $k \geq 1$ and $z > 0$ with $N_k(z) > 0$,*

$$\hat\alpha_k(z) \;:=\; \frac{d\log N_k}{d\log z} \;=\; k\Big(1 - \frac{N_{k+1}(z)}{N_k(z)}\Big). \tag{192}$$

The second equality is Lemma 4 divided by N_k. So $\hat\alpha_k$ is the instantaneous logarithmic slope of N_k, evaluated at a single z from two orders k and $k+1$ of the same quadrature.

Result 31 (Exactness on a pure power law). *Suppose $dF(x) = \alpha c\,x^{\alpha-1}dx$ on $(0,\infty)$ with $\alpha > 0$. Then for every $k > \alpha$ and every $z > 0$,*

$$N_k(z) = C_{\alpha,k}\,c\,z^\alpha, \qquad C_{\alpha,k} = \frac{\Gamma(1+\alpha)\,\Gamma(k-\alpha)}{\Gamma(k)}, \tag{193}$$

and consequently $\hat\alpha_k(z) = \alpha$ exactly*, for every $k > \alpha$ and every z.*

Proof. Substituting $\lambda = zt$ in the definition of N_k,

$$N_k(z) = \int_0^\infty \Big(\frac{z}{\lambda + z}\Big)^k \alpha c\, \lambda^{\alpha-1}\, d\lambda = \alpha c\, z^\alpha \int_0^\infty \frac{t^{\alpha-1}}{(1+t)^k}\, dt = \alpha c\, z^\alpha B(\alpha, k-\alpha),$$

the Beta integral converging exactly when $\alpha > 0$ and $k - \alpha > 0$. With $B(\alpha, k-\alpha) = \Gamma(\alpha)\Gamma(k-\alpha)/\Gamma(k)$ and $\alpha\Gamma(\alpha) = \Gamma(1+\alpha)$ this is equation 193. Then

$$\frac{C_{\alpha,k+1}}{C_{\alpha,k}} = \frac{\Gamma(k+1-\alpha)}{\Gamma(k-\alpha)} \cdot \frac{\Gamma(k)}{\Gamma(k+1)} = \frac{k-\alpha}{k},$$

using $\Gamma(w+1) = w\Gamma(w)$ twice, so by equation 192 $\hat{\alpha}_k = k\big(1 - (k-\alpha)/k\big) = \alpha$. □

E.3.2 Distinguishing a power law from anything faster

Remark 5 (Negative-moment dichotomy). *Let μ be a probability measure on $[0, \infty)$ with $\mu(\{0\}) = 0$ and $F(x) = \mu([0, x])$. For $k > 0$ put*

$$M_{-k} := \int_0^\infty \lambda^{-k}\, \mathrm{d}\mu(\lambda) \in (0, \infty], \qquad \alpha_\mu := \sup\{k > 0 :\ M_{-k} < \infty\} \in [0, \infty], \tag{194}$$

and define the local slope at ratio $\varrho > 1$,

$$s_k(z) := \frac{\log\big(N_k(\varrho z)/N_k(z)\big)}{\log \varrho}. \tag{195}$$

Then:

(i) $N_k(z) = z^k I_k(z)$ with $I_k(z) := \int(\lambda + z)^{-k}\, \mathrm{d}\mu$ increasing monotonically to M_{-k} as $z \downarrow 0$;

(ii) if $k < \alpha_\mu$ then $N_k(z) \sim M_{-k} z^k$ and $s_k(z) \to k$;

(iii) if $F(x) \sim c_\mu x^\alpha$ with $0 < \alpha < \infty$ then $\alpha_\mu = \alpha$, and for $k > \alpha$, $s_k(z) \to \alpha$;

(iv) consequently $s_k(z) \to \min(\alpha_\mu, k)$;

(v) $\alpha_\mu = \infty$ if and only if $F(x) = o(x^n)$ for every n.

Proof. (i) $N_k(z) = \int \big(z/(\lambda + z)\big)^k\, \mathrm{d}\mu = z^k \int(\lambda + z)^{-k}\, \mathrm{d}\mu$. For μ-a.e. $\lambda > 0$ the integrand $(\lambda + z)^{-k}$ increases to λ^{-k} as $z \downarrow 0$; monotone convergence gives $I_k(z) \uparrow M_{-k}$.

(ii) If $M_{-k} < \infty$ then $I_k(z) \to M_{-k} \in (0, \infty)$, so by (i) $N_k(z) \sim M_{-k} z^k$ and $s_k(z) = k + \log\big(I_k(\varrho z)/I_k(z)\big)/\log \varrho \to k$, because both $I_k(\varrho z)$ and $I_k(z)$ tend to the same finite positive limit. (Using the finite-difference slope avoids assuming N_k is differentiable, and is what the estimator actually computes.)

(iii) Integrating by parts, for $0 < k$, $\int_{(0,1]} \lambda^{-k}\, \mathrm{d}\mu = F(1) - \lim_{x \downarrow 0} x^{-k} F(x) + k \int_0^1 F(x) x^{-k-1}\, \mathrm{d}x$, so $M_{-k} < \infty$ iff $\int_0^1 F(x) x^{-k-1}\, \mathrm{d}x < \infty$. With $F(x) \sim c_\mu x^\alpha$ the integrand is $\asymp x^{\alpha-k-1}$, which is integrable at 0 exactly when $k < \alpha$; hence $\alpha_\mu = \alpha$. The limit $s_k \to \alpha$ for $k > \alpha$ is Theorem 29.

(iv) Combine (ii) and (iii): for $k < \alpha_\mu$ the slope tends to k, and for $k > \alpha_\mu$ it tends to α_μ.

(v) If $F(x) \le C_n x^n$ near 0 for every n, then $\int_0^1 F(x) x^{-k-1}\, \mathrm{d}x \le C_n \int_0^1 x^{n-k-1}\, \mathrm{d}x < \infty$ whenever $n > k$, so $M_{-k} < \infty$ for all k. Conversely if $M_{-k} < \infty$ then the boundary term in the integration by parts must vanish, i.e. $x^{-k} F(x) \to 0$, so $F(x) = o(x^k)$; if this holds for every k then F vanishes faster than any power. □

Part (v) is what makes the lemma a test: "F decays faster than any power at the origin" is precisely the class containing a hard gap ($F \equiv 0$ near 0), a Lifshitz tail $F(x) \sim e^{-a/x^\gamma}$, and any exponentially small density. All of them have $\alpha_\mu = \infty$, and by (iv) all of them make $s_k(z) \to k$ for every k.

Corollary 32 (The ratio test). *Fix $k_2 > k_1 > 0$ and $\mu(\{0\}) = 0$, and define $R(z) := s_{k_2}(z)/s_{k_1}(z)$. Then*

$$R(z) \longrightarrow \begin{cases} 1, & F(x) \sim c_\mu x^\alpha, \quad 0 < \alpha \le k_1, \\ \alpha/k_1, & F(x) \sim c_\mu x^\alpha, \quad k_1 < \alpha < k_2, \\ k_2/k_1, & F(x) \sim c_\mu x^\alpha, \quad \alpha \ge k_2, \\ k_2/k_1, & \alpha_\mu = \infty. \end{cases} \tag{196}$$

The same ratio can therefore arise from a sufficiently large power-law exponent or from faster-than-power decay.

E.3.3 Atoms at zero, and what a plateau means

Corollary 33 (Kernel detection). *For every $k > 0$, $N_k(z) \downarrow p_0 := \mu(\{0\})$ as $z \downarrow 0$, and $L(T) \downarrow p_0$ as $T \uparrow \infty$. For a finite matrix $p_0 = \dim \ker(A)/P$. An atom is incompatible with $F(x) \sim c_\mu x^\alpha$ for $\alpha > 0$, but it can coexist with a positive-spectrum edge $F(x) - p_0 \sim c_+ x^\alpha$.*

Proof. $(1 + \lambda/z)^{-k} \downarrow \mathbf{1}_{\{\lambda=0\}}$ pointwise and monotonically in z; apply monotone convergence. Same for $e^{-\lambda T}$. □

E.3.4 Generalization of the Gauss rule and Gauss-Radau brackets

Given $\alpha_1, \dots, \alpha_m, \beta_1, \dots, \beta_m$, we construct the Radau matrix as :

$$\widetilde{T}_{m+1} = \begin{pmatrix} T_m & \beta_m e_m \\ \beta_m e_m^\top & \gamma \end{pmatrix}, \qquad \gamma := \beta_m^2 \left[T_m^{-1}\right]_{mm}, \tag{197}$$

the value of γ being exactly what forces $\det \widetilde{T}_{m+1} = 0$, i.e. what puts an eigenvalue at $\lambda = 0$. $\widetilde{T}_{m+1}$ is $(m+1) \times (m+1)$, so it has $m+1$ eigenvalues, and these are the $m+1$ Radau nodes. One of them is 0 by construction; the other m are the free nodes $\tilde{\vartheta}_1 < \cdots < \tilde{\vartheta}_m$. So

$$\tilde{\vartheta}_0 := 0, \qquad R_m^{(0)}[g] = \sum_{i=0}^{m} \tilde{\omega}_i g(\tilde{\vartheta}_i) = \underbrace{\tilde{\omega}_0\, g(0)}_{\text{prescribed node}} + \underbrace{\sum_{i=1}^{m} \tilde{\omega}_i g(\tilde{\vartheta}_i)}_{m \text{ free nodes}}. \tag{198}$$

All $m+1$ weights belong to one family, computed identically from the eigenvectors $\tilde{u}_i$ of $\widetilde{T}_{m+1}$:

$$\tilde{\omega}_i = \left(e_1^\top \tilde{u}_i\right)^2, \quad i = 0, 1, \dots, m, \qquad \sum_{i=0}^{m} \tilde{\omega}_i = 1. \tag{199}$$

Result 34 (Gauss lower / Radau-at-zero upper bound). *Let μ_v be supported in $[0, \lambda_1]$ and let $g \in C^{2m+1}$ on a neighbourhood of $[0, \lambda_1]$ satisfy*

$$g^{(2m)} \ge 0 \quad \textit{and} \quad g^{(2m+1)} \le 0 \qquad \textit{on } [0, \lambda_1]. \tag{200}$$

Let $G_m[g] = \sum_{i=1}^{m} \omega_i g(\vartheta_i)$ be the m-node Gauss rule of Theorem 24 and let $R_m^{(0)}[g] = \sum_{i=0}^{m} \tilde{\omega}_i g(\tilde{\vartheta}_i)$ be the $(m+1)$-node Gauss–Radau rule with its lowest node prescribed at $\tilde{\vartheta}_0 = 0$. Then

$$G_m[g] \;\le\; \int g \, \mathrm{d}\mu_v \;\le\; R_m^{(0)}[g]. \tag{201}$$

Both functions $g_{z,k}(\lambda) = \left(\frac{z}{\lambda+z}\right)^k$ $(z > 0, k \ge 1)$ and $g_T(\lambda) = e^{-\lambda T}$ $(T > 0)$ satisfy equation 200 for every m.

Proof. Gauss side. Let h be the Hermite interpolant of g at the m Gauss nodes ϑ_i, each counted twice; $\deg h \le 2m - 1$, and the standard Hermite error formula gives, for each λ,

$$g(\lambda) - h(\lambda) = \frac{g^{(2m)}(\eta(\lambda))}{(2m)!} \prod_{i=1}^{m} (\lambda - \vartheta_i)^2 \tag{202}$$

for some $\eta(\lambda)$ in the convex hull of $\{\lambda, \vartheta_1, \dots, \vartheta_m\} \subset [0, \lambda_1]$. By equation 200 the right side is ≥ 0 for all $\lambda \in [0, \lambda_1]$. By exactness (Theorem 24), $\int h \, \mathrm{d}\mu_v = G_m[h] = G_m[g]$, the last equality because h and g agree at the nodes. Therefore $\int g \, \mathrm{d}\mu_v - G_m[g] = \int (g - h) \, \mathrm{d}\mu_v \geq 0$.

Radau side. The Gauss–Radau rule with a prescribed node at 0 is exact for polynomials of degree $\leq 2m$ (it has $m + 1$ nodes, one of them fixed). Let h now be the Hermite interpolant of g at $\lambda = 0$ (simple) and at the m free Radau nodes $\tilde{\vartheta}_i$ (double); $\deg h \leq 2m$ and

$$g(\lambda) - h(\lambda) = \frac{g^{(2m+1)}(\eta(\lambda))}{(2m+1)!} \, \lambda \prod_{i=1}^{m} (\lambda - \tilde{\vartheta}_i)^2. \tag{203}$$

On $\operatorname{supp} \mu_v \subset [0, \lambda_1]$ we have $\lambda \geq 0$, so the product of λ and the square is ≥ 0, while $g^{(2m+1)} \leq 0$; hence $g - h \leq 0$ μ_v-a.e. Exactness to degree $2m$ gives $\int h \, \mathrm{d}\mu_v = R_m^{(0)}[h] = R_m^{(0)}[g]$, whence $\int g \, \mathrm{d}\mu_v - R_m^{(0)}[g] \leq 0$. (Measurability of $\lambda \mapsto \eta(\lambda)$ is not needed: one applies equation 202–equation 203 pointwise to obtain the sign of $g - h$, and integrates the inequality.)

Sign conditions. For $g_{z,k}$, $g_{z,k}^{(n)}(\lambda) = (-1)^n z^k \frac{(k)_n}{(\lambda + z)^{k+n}}$ with $(k)_n = k(k+1) \cdots (k+n-1) > 0$, so $g^{(n)}$ has sign $(-1)^n$: even derivatives ≥ 0, odd ≤ 0, on $\lambda > -z \supset [0, \lambda_1]$. For g_T, $g_T^{(n)}(\lambda) = (-T)^n e^{-\lambda T}$, same alternation. Both are completely monotone, so equation 200 holds for every m. □

For each z the bracket equation 201 is computable from $(\alpha_j, \beta_j)_{j \leq m}$ alone. Define the usable window as the set of z where the relative bracket width $\big(R_m^{(0)}[g_{z,k}] - G_m[g_{z,k}]\big) / R_m^{(0)}[g_{z,k}]$ is below a chosen tolerance.

Corollary 35 (Kernel bound)**.** *Letting $z \downarrow 0$ in equation 201 with $k = 1$,*

$$\mu_v(\{0\}) \;\leq\; \tilde{\omega}_0, \tag{204}$$

the Radau weight at the prescribed node, so a depth-m run already bounds the kernel fraction from above.

Proof. $G_m[g_{z,1}] = z \sum_i \omega_i / (\vartheta_i + z) \to \sum_{i : \vartheta_i = 0} \omega_i$ and $R_m^{(0)}[g_{z,1}] \to \tilde{\omega}_0$ as $z \downarrow 0$; combine with equation 201 and Corollary 33. □

E.3.5 Why ghosts do not reach the hard edge

In finite precision the Lanczos recurrence loses orthogonality as soon as a Ritz values accumulate, see Remark 4. Paige (1976; 1980) showed that the resulting loss of orthogonality is triggered precisely when a Ritz value converges to the true eigenvalue, and occurs in the direction of the corresponding Ritz vector. Kaniel–Paige–Saad convergence theory (Saad, 1980) says that extremal well-separated eigenvalues converge fastest and hence by Paige's argument leads to ghosts. Extremal means at either end of the spectrum - near λ_1 or near λ_P - as opposed to interior. How well-separated the bottom has to be is made precise in the next result.

Remark 6 (Convergence at the bottom, and its cost under a hard edge)**.** *Let $A = A^\top$ have eigenvalues $\lambda_1 \geq \cdots \geq \lambda_P \geq 0$ with unit eigenvectors u_k, let v be a probe with $u_P^\top v \neq 0$, and let $\vartheta_m^{(m)}$ be the smallest Ritz value at depth m. Write*

$$\gamma_P := \frac{\lambda_{P-1} - \lambda_P}{\lambda_1 - \lambda_{P-1}} \tag{205}$$

for the relative gap at the bottom. Then:

(i) Kaniel–Paige–Saad asserts (Kaniel, 1966; Saad, 1980; Parlett, 1980)

$$0 \;\leq\; \vartheta_m^{(m)} - \lambda_P \;\leq\; \frac{(\lambda_1 - \lambda_P) \, \tan^2 \angle(v, u_P)}{T_{m-1}\big(1 + 2\gamma_P\big)^2}. \tag{206}$$

(ii) For $\gamma_P \ll 1$, $T_{m-1}(1+2\gamma_P) = \frac{1}{2}\exp\big(2(m-1)\sqrt{\gamma_P}\big)\big(1+O(\gamma_P)\big)$, so driving the right-hand side of equation 206 below a tolerance δ requires

$$m \;\gtrsim\; 1 + \frac{1}{4\sqrt{\gamma_P}}\,\log\frac{4(\lambda_1-\lambda_P)\tan^2\angle(v,u_P)}{\delta}, \qquad \textit{i.e.} \qquad m \;\sim\; \gamma_P^{-1/2}. \tag{207}$$

(iii) If in addition $\mu([0,x]) \sim c_\mu x^\alpha$ as $x \downarrow 0$, then the k-th smallest eigenvalue satisfies $\lambda_{(k)} \simeq \big(k/c_\mu P\big)^{1/\alpha}$, whence, in units $\lambda_1 = 1$,

$$\gamma_P \;\simeq\; \big(2^{1/\alpha}-1\big)\big(c_\mu P\big)^{-1/\alpha} \qquad \textit{and} \qquad m \;\sim\; \gamma_P^{-1/2} = \;\Theta\big(P^{1/(2\alpha)}\big). \tag{208}$$

The exponent in equation 208 is the point: the smaller the hard-edge exponent, slower the convergence and lower the chance of ghosting for small eigenvalues. Numerically, with $P = 1.68 \times 10^8$ and $c_\mu = 1$: for example at $\alpha \approx 0.33$ the value of m required to converge the smallest eigenvalue is order 10^{12}. The same bound at the top, where an isolated λ_1 has a relative gap of order unity, gives $m = O(\log(1/\delta))$: hundreds of steps. This asymmetry is what may put the ghosts exclusively at the top. We check this for $m = 4600$,

λ/λ_1 band	nodes	duplicate pairs ($< 10^{-8}$)	fraction
10^{-1}–10^{0}	3377	2261	67%
10^{-2}–10^{-1}	847	0	0%
10^{-3}–10^{-2}	257	0	0%
10^{-4}–10^{-3}	81	0	0%
10^{-7}–10^{-4}	37	0	0%

So ghosting is confined to $\lambda > 0.1\lambda_1$. Now we turn to make this observation formal.

Assumption 36 (Ghost localization)**.** *Let μ_v be the exact probe measure and $\hat{\mu}_v$ the measure of the computed Jacobi matrix. Assume there are $\Lambda > 0$, $s \geq 0$ and a transport plan π of μ_v and $\hat{\mu}_v$ supported on*

$$\{(x,x) : x \geq 0\} \;\cup\; \{(x,y) : x,y \geq \Lambda,\; |x-y| \leq s\}, \tag{209}$$

i.e. mass is either left untouched or moved by at most s within $[\Lambda,\infty)$.

A transport plan of two probability measures μ, ν on $\mathbb{R}_{\geq 0}$ is a probability measure π on the product $\mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ whose two marginals are μ and ν:

$$\pi\big(B \times \mathbb{R}_{\geq 0}\big) = \mu(B), \qquad \pi\big(\mathbb{R}_{\geq 0} \times B\big) = \nu(B) \qquad \text{for every Borel set } B. \tag{210}$$

Couplings always exist - the product $\mu \otimes \nu$ is one - so the content of Assumption 36 is not existence but the restriction on where π is allowed to put its probability mass. That restriction is equation 209. Being supported on the union of those two sets means every unit of mass does one of exactly two things:

(a) it lies on the diagonal $\{(x,x)\}$ - it is not moved at all; or

(b) it lies in $\{(x,y) : x,y \geq \Lambda,\; |x-y| \leq s\}$ - it is moved, but by a distance of at most s, and both its origin and destination are above Λ.

There is no third option, and in particular no mass may cross from below Λ to above it or vice versa, and no mass below Λ may move at all. Applied to μ_v and $\hat{\mu}_v$, this is the formalisation of ghosting splits converged Ritz values into tight clusters at the top of the spectrum, and leaves the bottom alone: the cluster spread is s, and Λ is the scale above which converged values are found.

Proposition 37 (Transport bound)**.** *Under Assumption 36, for any $g \in C^1([0,\infty))$,*

$$\Big|\int g\,\mathrm{d}\hat{\mu}_v - \int g\,\mathrm{d}\mu_v\Big| \;\leq\; s\,\sup_{\lambda\geq\Lambda}|g'(\lambda)|. \tag{211}$$

Proof. Writing the difference through the coupling, $\int g\,\mathrm{d}\hat{\mu}_v - \int g\,\mathrm{d}\mu_v = \iint \big(g(y)-g(x)\big)\,\mathrm{d}\pi(x,y)$. By the mean value theorem, for each (x,y) in the support of π there is ξ between x and y with $g(y)-g(x) = g'(\xi)(y-x)$; both $x,y \geq \Lambda$, so $\xi \geq \Lambda$, and $|y-x| \leq s$. Hence the integrand is bounded by $s\sup_{\lambda\geq\Lambda}|g'|$ pointwise, and π is a probability measure. □

Corollary 38 (Ghosts are harmless at the hard edge)**.** *Take $g = g_{z,k}$, $g_{z,k}(\lambda) = \big(z/(\lambda+z)\big)^k$. Then $g'_{z,k}(\lambda) = -kz^k/(\lambda+z)^{k+1}$ is negative and increasing, so its supremum over $[\Lambda,\infty)$ is attained at Λ and*

$$\big|\hat{N}_k(z) - N_k(z)\big| \;\le\; \frac{s\,k\,z^k}{(\Lambda+z)^{k+1}} \;\le\; \frac{s\,k}{\Lambda^{k+1}}\; z^k. \tag{212}$$

If in addition μ obeys the hard-edge law of Theorem 29, so that $N_k(z) \sim c_\mu \frac{\Gamma(\alpha+1)\Gamma(k-\alpha)}{\Gamma(k)} z^\alpha$, then the relative *error obeys*

$$\frac{\big|\hat{N}_k(z) - N_k(z)\big|}{N_k(z)} \;\le\; \frac{s\,k\,\Gamma(k)}{c_\mu\,\Gamma(\alpha+1)\Gamma(k-\alpha)\,\Lambda^{k+1}}\; z^{\,k-\alpha} \;\xrightarrow[z\downarrow 0]{}\; 0 \qquad \textit{whenever } k > \alpha. \tag{213}$$

The condition $k > \alpha$ is not an extra hypothesis: it is precisely the condition already required for the exponent to transfer at all (Theorem 29, and the k-saturation diagnostic of Remark 5). With the measured ghost spread $s \simeq 10^{-8}\lambda_1$ and localisation $\Lambda = 0.1\lambda_1$ (§E.4), $k = 2$, and the measured $\alpha \approx 0.3$, the bound equation 213 with $c_\mu = 1$ is 1.6×10^{-15} at $z = 10^{-6}\lambda_1$ and 1.2×10^{-20} at $z = 10^{-9}\lambda_1$.

Remark 6 (What this argument does not cover). Corollary 38 controls perturbations localised above Λ. It says nothing about rounding that perturbs the spectrum near zero, where $g_{z,k}$ varies fastest: there $\sup|g'_{z,k}| = k/z$, and a perturbation of size δ_A contributes a relative error $\sim k\delta_A/(c_\mu z^{1+\alpha})$, which diverges as $z \downarrow 0$. Setting that against unity gives the floor

$$z_{\min}/\lambda_1 \;\gtrsim\; (\delta_A/\lambda_1)^{1/(1+\alpha)}. \tag{214}$$

It is essential that δ_A here is the measured defect of the computed operator. With the measured $\delta_A/\lambda_1 = 8.5 \times 10^{-12}$ and $\alpha \approx 0.3$, equation 214 gives $z_{\min} \approx 3 \times 10^{-9}\lambda_1$.

E.4 Measurements

All numbers below are our own measurements on a retrained 150M checkpoint (OLMo-style, $P = 1.677 \times 10^8$, 3B tokens at $1\times$ Chinchilla, $B = 64$, seed 0). The operator is the Adam-preconditioned Gauss–Newton matrix equation 156 at the final checkpoint, built from 10^6 tokens, probed in float64 with a random probe. At $m = 9400$ (Figure 11):

z/λ_1	10^{-5}	3.2×10^{-6}	10^{-6}	3.2×10^{-7}	10^{-7}	3.2×10^{-8}	1.9×10^{-8}
s_1	0.051	0.082	0.124	0.175	0.232	0.279	$[0.288, 0.290]$
s_2	0.075	0.116	0.171	0.234	0.295	$[0.318, 0.320]$	$[0.293, 0.304]$
$R_{2,1}$	1.47	1.42	1.38	1.33	1.27	$[1.14, 1.15]$	$[1.01, 1.06]$

Down to $10^{-7}\lambda_1$ every bracket is tighter than the three digits shown; the widest is 3.7×10^{-6}, on s_2 at $10^{-7}\lambda_1$. The ratio brackets follow from the slope brackets as $R_{\rm lo} = s_{k_2,\rm lo}/s_{k_1,\rm hi}$, $R_{\rm hi} = s_{k_2,\rm hi}/s_{k_1,\rm lo}$, the widest interval consistent with both. $R_{2,1}$ decreases monotonically from 1.47 at $10^{-5}\lambda_1$ to $[1.01, 1.06]$ at the deepest plotted scale, against a faster-than-power limit of 2. A faster-than-power edge would instead drive the ratio towards its limit $k_2/k_1 = 2$. Hence, on the window resolved, a faster-than-power edge is strongly disfavoured, and the behaviour is power-law-like. Theorem 29 makes every s_k tend to α. Hence $\alpha \approx 0.30$.

Reliable window. Ghost effects are captured in Corollary 38: $skz^k/(\Lambda+z)^{k+1}$ with the measured $s \le 10^{-8}\lambda_1$ and $\Lambda = 0.1\lambda_1$. It is negligible: it never exceeds 8×10^{-15} and falls steeply as z decreases.

The next concern is the operator floor in Remark 6. Determinism is a recomputation of the same product: in float64 the operator is deterministic to round-off, so a fixed operator exists; symmetric to 10^{-15}, so the Jacobi matrix is the Jacobi matrix of a genuine positive measure. The linearity defect $\delta_A/\lambda_1 = 8.5 \times 10^{-12}$ is the largest of the three and is the number that propagates into the sensitivity bound $|\Delta N_k| \le k\delta_A/z$, hence into a floor on z that no amount of Lanczos depth can lower. Propagated to the local slope at the tolerance of 0.01 used throughout, $k = 2$ operator floor is reached at $z/\lambda_1 = 1.860 \times 10^{-8}$.

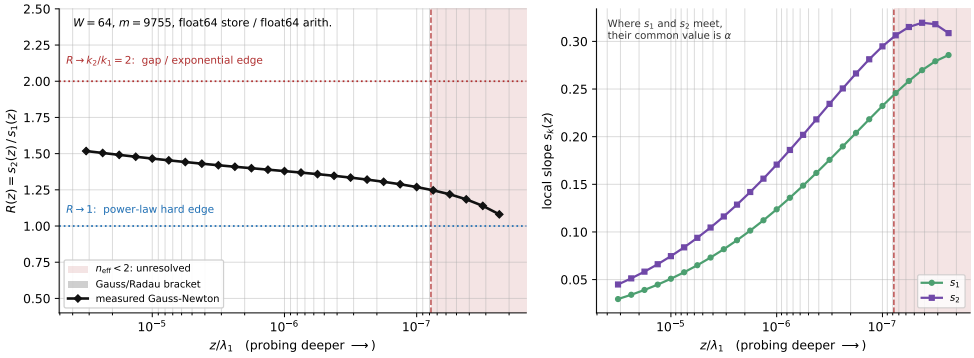


Figure 11: The plot on the left shows that heavy tailed curvature spectrum exists and the plot on the right determines the exponent. Smaller z probes a smaller eigenvalue spectrum.

check	quantity	float64	float32
determinism	$\left\|A^{(1)}(v) - A^{(2)}(v)\right\| / \left\|A^{(1)}(v)\right\|$	7.6×10^{-18}	4.0×10^{-9}
symmetry	$\lvert\langle u, Av\rangle - \langle v, Au\rangle\rvert / (\lVert u\rVert \lVert v\rVert \lambda_1)$	8.6×10^{-16}	3.5×10^{-11}
linearity	$\lVert A(u+v) - Au - Av\rVert / (\lambda_1 \lVert u+v\rVert)$	8.5×10^{-12}	2.3×10^{-7}
positivity	$\langle v, Av\rangle / (\lVert v\rVert^2 \lambda_1)$	$+1.47 \times 10^{-5}$	$+1.47 \times 10^{-5}$

At iteration depth $m = 9400$ and $z/\lambda_1 = 2.64 \times 10^{-8}$, the $k = 2$ total error reaches the tolerance as a result of interpolating the data from the table below:

$z/\hat{\lambda}_1$	quadrature	operator	ghosting	total
2.196174×10^{-8}	6.728678×10^{-3}	8.103861×10^{-3}	1.0×10^{-19}	0.014833
2.818383×10^{-8}	2.855708×10^{-3}	5.878096×10^{-3}	1.5×10^{-19}	0.008734

In the above table, the operator term converted from a bound on the count to a bound on the local slope, at $z/\lambda_1 = 2.818 \times 10^{-8}$, $k = 2$, with $\delta_A/\lambda_1 = 8.5 \times 10^{-12}$, $N_2 = 0.2233$ and $\varrho = 10^{0.4}$. Recall $s_k = [\ln N_k(\varrho z) - \ln N_k(z)]/\ln \varrho$ - using twice the error at the deeper endpoint, we get operator error on the slope to be $\frac{2}{\ln \varrho} \frac{k\delta_A/z}{N_2} = 5.8781 \times 10^{-3}$. It is conservative estimate, since both $k\delta_A/(\varrho z) < k\delta_A/z$ and $N_k(\varrho z) > N_k(z)$.

The hard-edge law $F(x) \simeq c_\mu x^\alpha$ can hold only below the crossover λ_0, and normalisation forces $c_\mu \lambda_0^\alpha \lesssim 1$, so the crossover sits where the count is of order unity. Any threshold $N_k = \Theta(1)$ marks that boundary to within a factor in z, we choose $N_k = 1/2$. At a tolerance of 0.01 on the slope, this demands $z/\lambda_1 < 4.21 \times 10^{-7}$.

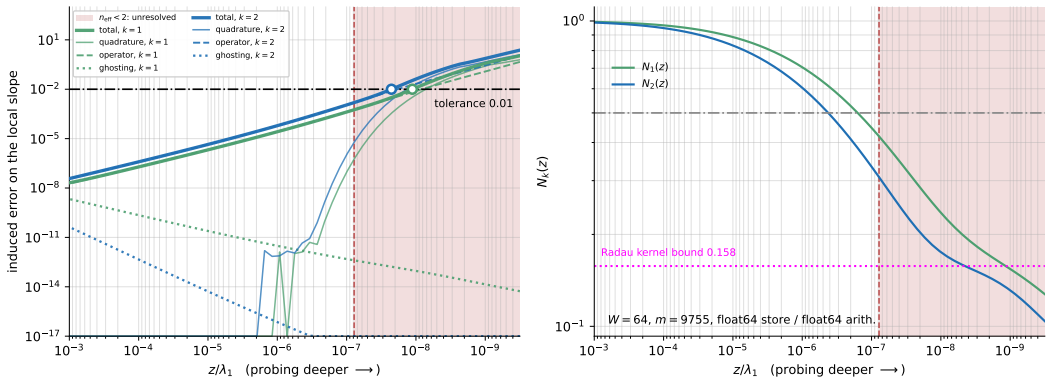


Figure 12: Every mechanism that limits the measurement performed.

The choice of the slope cut-off so far is arbitrary; to make the discussion of a reliable window precise, we introduce:

Definition 39 (Effective node count)**.** *Let $(\theta_j, \omega_j)_{j=1}^m$ be the nodes and weights of the m-node Gauss rule, and for $z > 0$ let $J(z) = \{j : \theta_j \le z\}$ with $q(z) = |J(z)|$. Define*

$$n_{\text{eff}}(z) \;=\; \frac{\left(\sum_{j\in J(z)} \omega_j\right)^2}{\sum_{j\in J(z)} \omega_j^2}, \qquad n_{\text{eff}}(z) := 0 \text{ if } J(z) = \varnothing. \tag{215}$$

This is a participation ratio: it counts the nodes below z, discounted by how unevenly they share the mass. It is preferred to the raw count $q(z)$ because q is discontinuous and is inflated by nodes of negligible weight, whereas equation 215 is invariant under rescaling all ω_j and ignores a node whose weight is vanishing.

Proposition 40 (Range and extremes)**.** *For every z with $q(z) \ge 1$,*

$$1 \;\le\; n_{\text{eff}}(z) \;\le\; q(z), \tag{216}$$

with $n_{\text{eff}}(z) = q(z)$ if and only if all ω_j, $j \in J(z)$, are equal, and $n_{\text{eff}}(z) = 1$ if and only if at most one of them is non-zero — in particular exactly when $q(z) = 1$.

Proof. The weights of a Gauss rule are strictly positive. Write $S_1 = \sum_J \omega_j$ and $S_2 = \sum_J \omega_j^2$. By Cauchy–Schwarz applied to the vectors (ω_j) and $(1, \dots, 1)$, $S_1^2 \le q\, S_2$, with equality iff the ω_j are proportional to 1, i.e. all equal; this is the upper bound. For the lower bound, expanding $S_1^2 = \sum_j \omega_j^2 + \sum_{i\ne j} \omega_i\omega_j = S_2 + \sum_{i\ne j} \omega_i\omega_j$ and using $\omega_i\omega_j > 0$ gives $S_1^2 \ge S_2$, with equality iff every cross term vanishes, i.e. at most one weight is non-zero. Since all weights are positive, that happens exactly when $q(z) = 1$. □

Now we turn to discuss $n_{\text{eff}} < 2$ is the edge of validity, and $n_{\text{eff}} = 1$ the strict end of reliable window. Split N_k at z into the part carried by nodes below z and the part carried by nodes above:

$$N_k(z) = \underbrace{\sum_{\theta_j \le z} \omega_j \left(\tfrac{z}{\theta_j + z}\right)^k}_{=:B_k(z)} \;+\; \underbrace{\sum_{\theta_j > z} \omega_j \left(\tfrac{z}{\theta_j + z}\right)^k}_{=:T_k(z)}. \tag{217}$$

For $\theta \ll z$ the kernel is ≈ 1 independently of k, so $B_{k+1} \approx B_k =: B$: the sub-z mass behaves like an atom at the origin, contributing logarithmic slope 0. For $\theta \gg z$ the kernel is $\approx (z/\theta)^k$, so $T_{k+1} \ll T_k$ and that part contributes slope k. Substituting into Definition 30,

$$\hat\alpha_k(z) \;\approx\; k\,\frac{T_k(z)}{B(z) + T_k(z)}, \tag{218}$$

an approximation valid when the two groups are well separated from z. So $\hat\alpha_k$ is a mixture: the value it reports is set by how the rule has apportioned mass between an atom-like term and a sloping term. This is what makes the node count decisive, and it separates the two thresholds.

$n_{\text{eff}} = 1$**: strictly unreliable.** By Proposition 40 this happens exactly when a single node lies below z. Everything the rule knows about $\mu|_{[0,z]}$ is then one pair (θ_1, ω_1) — two real numbers, a mass and a location. But the hard-edge exponent is a statement about the shape of μ near the origin, $F(x) \sim c\,x^\alpha$, and a single atom has no shape: it fixes B in equation 218 at an unresolved value and carries no information about α at all. Below θ_2 the curve is therefore not merely noisy, it is uninformative about the quantity being measured, whatever its bracket says. The Gauss/Radau bracket does not protect against this: it bounds the integral $N_k(z)$, not the shape of the measure beneath it, and it can be tight while the shape is entirely unresolved.

$n_{\text{eff}} < 2$**: the edge of validity.** Two nodes below z give four numbers, enough to represent a mass, a spread and a relative weight, so $\mu|_{[0,z]}$ is no longer forced to be a point mass and B in equation 218 responds to the shape of the measure. This is the minimum configuration in which $\hat\alpha_k$ can carry exponent information at all, which is why it is the edge of validity rather than a complete failure - the estimate is supported, but by the least possible amount. This regime is exactly $1 \le n_{\text{eff}} < 2$.

For more details see Figure 11, 12.