Fast Learning Rate Transfer for Gradient Descent in Sketched Linear Regression

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

We study the efficiency of the hyperparameter (HP) transfer strategy from Yang et al. [23] in a solvable sketched-linear model with varying width (i.e., number of features), where the HP of interest is the gradient descent (GD) learning rate. Following the fast-transfer framework of Ghosh et al. [8], we characterize *fast transfer* (which implies computational gain) by comparing the convergence rates of the loss and the optimal HP with respect to the scaling dimension (model width n). For a fixed optimization horizon T, we prove a central limit theorem (CLT) for the optimal learning rate and the optimally tuned loss, yielding unconditional fast transfer. On the other hand, for growing horizons (i.e., when T jointly diverges with n), the optimal learning rate approaches the stability edge beyond which GD diverges, and under an explicit scale separation between T and n, we establish fast transfer with rates depending on the spectral source and capacity conditions.

1. Introduction

Hyperparameter transfer has become a practical way to reduce the cost of tuning large neural networks. The Tensor Programs/μP framework [21, 23] gives a principled parameterization under which hyperparameters, most notably the learning rate, can have stable large-width limits. This makes it possible to tune a small proxy model and transfer the selected hyperparameters to a much larger model. This strategy has been influential in large-scale training, but its theoretical foundations remain incomplete: μP predicts asymptotic scale-independence (i.e., optimal HP converges to a well-defined limit), but not the *rate* at which the finite-scale optimum approaches its large-scale limit.

As argued in Ghosh et al. [8], convergence of optimal HPs alone is not enough to guarantee that transfer offers computational gain over directly tuning the large model. To quantify this separation, Ghosh et al. [8] introduced the definition of *fast transfer* which depends on the following quantities.

- **Loss gap:** $a_n = |L_n - L_\infty|$, where L_n denote the loss (with optimal HP) at scale n.
- **Hyperparameter gap:** $b_n = |\eta_n - \eta_\infty|$, where η_n denotes the optimal HP at scale n.
- **Transfer suboptimality gap:** $c_n = |L_\infty(\eta_n) - L_\infty(\eta_\infty)|$, which measures the loss suboptimality by using the optimal HP at scale n to train the largest model.

Transfer is *fast* when $c_n = o(a_n)$ — under certain local curvature condition on L_∞, this property implies transfer-based tuning is asymptotically more efficient than direct grid search of HPs. This translates the comparison of computational efficiency to the characterization of convergence rates of (a_n, b_n, c_n). To our knowledge, the only provable positive example is for tuning the ridge penalty in random features regression [8]. It remains unknown whether the same fast-transfer mechanism can be established for the *learning rate*, the most common HP considered in the μ-transfer series.

1.1. Our Contributions

We establish fast transfer for gradient descent learning rate in an exactly solvable sketched linear regression model. The teacher, student, and population loss are given as

$$y = \langle x, \beta_0 \rangle, \qquad f_n(x;\theta) = \langle x, W_n\theta \rangle, \qquad L_n(\theta) = \tfrac{1}{2}\mathbb{E}_x\big(f_n(x;\theta) - y\big)^2, \tag{1}$$

where x is centered with covariance Σ, and $W_n : \mathbb{R}^n \to \mathcal{H}$ is a random width-n sketch. Writing $b = \Sigma^{1/2}\beta_0 \in \mathcal{H}$, the sketch enters the gradient-descent dynamics through the empirical covariance operator $A_n := \Sigma^{1/2} W_n W_n^* \Sigma^{1/2}$. This model has been extensively studied in the scaling law theory literature [4, 6, 16, 17, 20] where β_0 the spectrum of Σ follow certain power-law decay.

Starting from $\theta_0 = 0$, gradient descent with learning rate η admits the exact loss formula

$$L_{n,T}(\eta) = \tfrac{1}{2}\langle b, (I - \eta A_n)^{2T} b \rangle, \qquad L_{\infty,T}(\eta) = \tfrac{1}{2}\langle b, (I - \eta\Sigma)^{2T} b \rangle. \tag{2}$$

Thus finite width enters only through the covariance fluctuation $(A_n - \Sigma)$, and learning-rate transfer reduces to comparing the finite-width optimizer $\eta_{n,T}$ with the infinite-width optimizer $\eta_{\infty,T}$.

Fixed-horizon fast transfer. Our first result proves that learning-rate transfer is fast for every fixed optimization horizon T. Informally, we show that the optimal learning rate and the optimally tuned loss both fluctuate at the usual sample-covariance scale:

$$a_n = |L_{n,T} - L_{\infty,T}| = \Theta_p(n^{-1/2}), \qquad b_n = |\eta_{n,T} - \eta_{\infty,T}| = \Theta_p(n^{-1/2}),.$$

On the other hand, the transferred-loss penalty is smaller: $c_n = |L_{\infty,T}(\eta_{n,T}) - L_{\infty,T}(\eta_{\infty,T})| = \Theta_p(n^{-1})$. Therefore $c_n = o_p(a_n)$, which gives fast transfer in the sense of Ghosh et al. [8]. This characterization is *unconditional* in the sense that it does not depend on the structure of (Σ, β_0).

Growing-horizon rates. While the fixed-horizon assumption aligns with the original μP derivation [21], LLM pretraining operates in a regime where the number of training token grows with model size [7, 11], which often implies a training horizon that diverges with scale. We therefore consider an optimization horizon that grows with width. Under standard source and capacity conditions $\lambda_k(\Sigma) = k^{-\alpha}$, $b_k^2 = k^{-\beta}$, with $\alpha, \beta > 1$, we show that the optimal learning rate at infinite width is

$$\eta_{\infty,T} = 2 - \tfrac{(\rho+1)\log T + O(1)}{2T}, \quad \text{where } \rho = \tfrac{\beta-1}{\alpha}.$$

Hence as T diverges, the optimal learning rate drifts towards the stability threshold $2/\lambda_{\max}(\Sigma)$, which may amplify the suboptimality due to finite-width fluctuations. In this setting, we establish *conditional* fast transfer when the time horizon grows sufficiently slowly compared with the width. In particular, if $T^{\chi(\rho)} \ll \sqrt{n}$ for some explicit exponent $\chi(\rho)$, then the loss follows a power-law decay (with scaling exponent depending on a, β), and $L_{\infty,T}(\eta_{n,T}) - L_{\infty,T}(\eta_{\infty,T}) = o_p\big(|L_{n,T} - L_{\infty,T}|\big)$.

Implications. These results give, to our knowledge, the first fast-transfer guarantee for the gradient descent learning rate, providing theoretical justification for the μ-transfer strategy. Moreover, they shed light on the different hyperparameter behaviors when *time horizon is a scaling dimension*.

2. Problem Setting

We work on $\mathcal{H} = \ell^2(\mathbb{N})$. Let x be centered with covariance $\mathbb{E}[x \otimes x] = \Sigma$, where $\Sigma : \mathcal{H} \to \mathcal{H}$ is positive, self-adjoint, trace class, and normalized $\|\Sigma\|_{\mathrm{op}} = 1$. Labels are generated as $y = \langle x, \beta_0 \rangle$,

and the width-n student is a sketched linear model: $\widehat{f}(x;\theta) = \theta^\top W_n^* x = \langle x, W_n\theta\rangle$, where $\theta \in \mathbb{R}^n$, $W_n : \mathbb{R}^n \to \mathcal{H}$, and $\mathcal{H} \ni b := \Sigma^{1/2}\beta_0 \neq 0$. We aim to minimize the population squared loss

$$L_n(\theta) = \tfrac{1}{2}\mathbb{E}_x\big(\widehat{f}(x;\theta) - y\big)^2 = \tfrac{1}{2}\langle W_n\theta - \beta_0, \Sigma(W_n\theta - \beta_0)\rangle.$$

The sketch enters the dynamics through $A_n := \Sigma^{1/2}W_nW_n^*\Sigma^{1/2} = \frac{1}{n}\sum_{r=1}^n z_r\otimes z_r$, $z_r \overset{\text{i.i.d.}}{\sim} N(0,\Sigma)$. Starting from $\theta_0 = 0$, gradient descent with step size η satisfies

$$\theta_{t+1} = \theta_t - \eta\nabla L_n(\theta_t), \qquad \nabla L_n(\theta) = W_n^*\Sigma(W_n\theta - \beta_0).$$

From a simple calculation, we know that the loss after T steps is exactly given by (2).

3. Transfer under Fixed Time Horizon

Let $\Sigma e_k = \lambda_k e_k$ on $\ell^2(\mathbb{N})$, and let the target residual be $b = \Sigma^{1/2}\beta_0$. Define

$$F_{n,T}(\eta) := \left\langle b, A_n(I - \eta A_n)^{2T-1}b\right\rangle, \qquad F_{\infty,T}(\eta) := \left\langle b, \Sigma(I - \eta\Sigma)^{2T-1}b\right\rangle,$$

which are the normalized negative derivatives of $L_{n,T}$ and $L_{\infty,T}$. When $A_nb \neq 0$, the finite-width minimizer over $\eta \geq 0$ is the unique root of $F_{n,T}$, and similarly $\eta_{\infty,T}$ is the unique root of $F_{\infty,T}$.

Theorem 1 (Fixed-horizon optimum and loss CLTs) *Fix $T \in \mathbb{N}$ and let $\eta_{n,T}$ and $\eta_{\infty,T}$ be the finite- and infinite-width optimal learning rates. Then*

$$\eta_{n,T} \xrightarrow[n\to\infty]{P} \eta_{\infty,T}, \qquad \eta_{n,T} - \eta_{\infty,T} = O_p(n^{-1/2}).$$

Moreover, there are explicit finite-rank Hilbert–Schmidt operators Q_F and Q_L such that

$$\sqrt{n}(\eta_{n,T} - \eta_{\infty,T}) \Rightarrow N\left(0, \tfrac{2\,\mathrm{tr}(\Sigma Q_F\Sigma Q_F)}{\{F'_{\infty,T}(\eta_{\infty,T})\}^2}\right),$$
$$\sqrt{n}\{L_{n,T}(\eta_{n,T}) - L_{\infty,T}(\eta_{\infty,T})\} \Rightarrow N\left(0, 2\,\mathrm{tr}(\Sigma Q_L\Sigma Q_L)\right).$$

Fixed-horizon fast transfer. Recall the definitions from Ghosh et al. [8],

$$a_n(T) = |L_{n,T}(\eta_{n,T}) - L_{\infty,T}(\eta_{\infty,T})|, \; b_n(T) = |\eta_{n,T} - \eta_{\infty,T}|, \; c_n(T) = L_{\infty,T}(\eta_{n,T}) - L_{\infty,T}(\eta_{\infty,T}),$$

where $c_n(T)$ is the finite-to-infinite transfer suboptimality gap (i.e., the learning rate is tuned at finite width and evaluated on the infinite-width loss). Set

$$\sigma_{L,T}^2 := 2\,\mathrm{tr}(\Sigma Q_L\Sigma Q_L), \;\; \sigma_{F,T}^2 := 2\,\mathrm{tr}(\Sigma Q_F\Sigma Q_F), \;\; \kappa_T := L''_{\infty,T}(\eta_{\infty,T}) = -TF'_{\infty,T}(\eta_{\infty,T}) > 0.$$

The CLTs in Theorem 1 and the local quadratic expansion of the infinite-width loss imply

$$\sqrt{n}\, a_n(T) \Rightarrow |G_L|, \qquad G_L \sim N(0,\sigma_{L,T}^2),$$
$$\sqrt{n}\, b_n(T) \Rightarrow |G_\eta|, \qquad G_\eta \sim N\Big(0, \tfrac{\sigma_{F,T}^2}{\{F'_{\infty,T}(\eta_{\infty,T})\}^2}\Big),$$
$$n\, c_n(T) \Rightarrow \tfrac{\kappa_T}{2}G_\eta^2.$$

Consequently, if $\sigma_{L,T}^2 > 0$ and $\sigma_{F,T}^2 > 0$, which is automatically satisfied if b has at least 2 nonzero projections onto two different eigenspaces of Σ, then

$$a_n(T) = \Theta_p(n^{-1/2}), \qquad b_n(T) = \Theta_p(n^{-1/2}), \qquad c_n(T) = \Theta_p(n^{-1}),$$

and hence $c_n(T) = o_p(a_n(T))$. This establishes fast transfer for the fixed-T setting, which implies that HP transfer offers computational gain over directly tuning the large-scale model [8].

Remark 2 *We conjecture that a similar result holds far more generally beyond linear regression. Indeed, the Tensor Programs framework [22] should make it possible to show that under mild technical conditions, similar CLTs hold for essentially arbitrary neural network training procedures when the horizon T is fixed and $n \to \infty$. Thus the default behavior in this scaling limit is $a_n = \Theta_p(n^{-1/2})$, $b_n = \Theta_p(n^{-1/2})$, and $c_n = \Theta_p(n^{-1})$. Proving this requires handling subtle technical details, such as verifying that the first-order correction to the Tensor Program limit is non-degenerate, but we expect this is almost always the case, barring carefully engineered pathological scenarios.*

4. Transfer under Growing Time Horizon

While Theorem 1 establishes fast transfer for the learning rate in sketched linear regression, one could argue that the fixed-T setting deviates from standard practice for the following reasons.

- At fixed time horizon, the population loss remains $\Theta(1)$ and does not decay to 0. In contrast, the scaling law literature suggests the loss follows a power-law decay toward an irreducible floor [11, 13], a behavior that can be replicated in random features model (1) by considering longer training times and structured features [4, 16, 20].
- LLM pretraining typically scales the number of training tokens with the model size [7, 11], which suggests that the optimization time should also be considered as a scaling dimension.

We therefore consider a joint scaling of model width n and time horizon T, and specialize to a polynomial spectrum stated below. We note that similar power-law assumptions are also used in the recent analyses of optimal learning rate schedules [2, 14, 15].

Assumption 1 (Source-capacity condition) $\lambda_k = k^{-\alpha}, b_k^2 = k^{-\beta}$ *where* $\alpha, \beta > 1$.

For notational convenience, we introduce $\rho = \frac{\beta-1}{a}, s = \frac{a-1}{a}$, and for $\eta \in [1, 2]$, define

$$L_{\infty,T}(\eta) = \tfrac{1}{2} \textstyle\sum_{k\geq 1} k^{-\beta}(1-\eta k^{-a})^{2T}, \qquad F_T(\eta) = -\tfrac{1}{T}\tfrac{d}{d\eta} L_{\infty,T}(\eta).$$

It is clear that in this setting, L_T decays to the irreducible loss with a power-law rate $T^{-\rho}$ – see Corollary 12. Since $\lambda_1(\Sigma) = 1$, at infinite width the dynamics is stable when the learning rate $\eta \leq 2$, beyond which the top eigen-mode blows up — we refer to this endpoint as the *stability edge*. At finite width the analogous random boundary is $2/\lambda_{\max}(A_n)$. Let $\eta_{\infty,T} := \arg\min_{\eta\in[1,2]} L_{\infty,T}(\eta)$. We show that the optimal learning rate approaches the stability edge as the time horizon increases

$$\eta_{\infty,T} = 2 - \tfrac{(\rho+1)\log T + O(1)}{2T}. \tag{3}$$

The growing time horizon introduces the following complications. The normalized finite-width derivative of the loss admits a uniform perturbation bound $O_p(Tn^{-1/2})$, while the second order derivative at $\eta_{\infty,T}$ is only order $T^{-\rho}$. Consequently, if $T^{\rho+2}/\sqrt{n} \to 0$, the finite-width minimizer $\widehat{\eta}_{n,T} \in \arg\min_{\eta\in[1,2]} L_{n,T}(\eta)$ is interior with high probability and $|\widehat{\eta}_{n,T} - \eta_{\infty,T}| = O_p(T^{\rho+1}n^{-1/2})$. On the other hand, because $L''_{\infty,T}(\eta_{\infty,T})$ is of order $T^{1-\rho}$, this implies a transfer suboptimality gap of $O_p(T^{\rho+3}/n)$. Hence to establish a growing-horizon fast-transfer statement, we must compare this gap with the actual finite-width loss fluctuation at the same horizon.

To state the resulting loss fluctuation explicitly, write $E_n = A_n - \Sigma$, $A_T = I - \eta_{\infty,T}\Sigma$, and $r_k(\eta) = 1 - \eta k^{-a}$. The leading random term in the finite-width loss at $\eta_{\infty,T}$ is

$$\mathcal{V}_1(\eta_{\infty,T}) := -\eta_{\infty,T} \textstyle\sum_{r=0}^{2T-1} \left\langle b, A_T^{2T-1-r} E_n A_T^r b \right\rangle.$$

The mean correction from the terms with two copies of E_n is described by the explicit coefficient

$$Q_T(\eta) := \binom{2T}{2} \sum_{k\geq 1} k^{-\beta-2a} r_k(\eta)^{2T-2} + \sum_{i=0}^{2T-2} (i+1) \Big(\sum_{j\geq 1} j^{-a} r_j(\eta)^{2T-2-i} \Big) \Big(\sum_{k\geq 1} k^{-\beta-a} r_k(\eta)^i \Big).$$

Finally, define $M_T := -\eta_{\infty,T} \sum_{r=0}^{2T-1} A_T^r b \otimes A_T^{2T-1-r} b$, $N_T := \Sigma^{1/2} M_T \Sigma^{1/2}$, and $H_T := \|N_T\|_F$. In the following theorem, we characterize the loss fluctuation via moment method.

Theorem 3 (Finite-width loss fluctuation) *Let $T = T_n \to \infty$ and assume $T^2/n \to 0$. Then*

$$L_{n,T}(\eta_{\infty,T}) - L_{\infty,T}(\eta_{\infty,T}) = \tfrac{\eta_{\infty,T}^2}{2n} Q_T(\eta_{\infty,T}) + \tfrac{1}{2}\mathcal{V}_1(\eta_{\infty,T}) + O\left(\tfrac{T^4}{n^2}\right) + O_{L^2}\left(\tfrac{T^2}{n}\right).$$

Moreover, $Q_T(\eta_{\infty,T}) \asymp T^{1-\rho} + T^{-s}$, and $\frac{\frac{1}{2}\mathcal{V}_1(\eta_{\infty,T})}{(2n)^{-1/2} H_T} \Rightarrow N(0,1)$, where H_T satisfies

$$\begin{cases} H_T \asymp T^{-\rho}, & 0 < \rho < 1, \\ T^{-1} \lesssim H_T \lesssim T^{-1} \log T, & \rho = 1, \\ H_T \asymp T^{-(\rho+1)/2}, & \rho > 1. \end{cases}$$

The expansion has a direct interpretation: averaging over the sketch gives the explicit $1/n$ term, while a single occurrence of E_n gives the random term $\mathcal{V}_1(\eta_{\infty,T})/2$, whose typical size is $H_T/\sqrt{n}$. Using this random loss fluctuation as the benchmark in the rate comparison, we establish the following fast-transfer theorem under additional constraints on growth of time horizon.

Theorem 4 (Growing-horizon fast-transfer) *Let $T = T_n \to \infty$. Define $\widehat{\eta}_{n,T}$ to be any minimizer of $L_{n,T}$ over $[1,2]$, and recall the definitions of $a_{n,T}, b_{n,T}, c_{n,T}$. If the time horizon satisfies*

$$\frac{T^{\chi(\rho)}}{\sqrt{n}} \to 0, \qquad \chi(\rho) = \begin{cases} 2\rho + 3, & 0 < \rho < 1, \\ 5, & \rho = 1, \\ (3\rho+7)/2, & \rho > 1, \end{cases} \tag{4}$$

then we have fast transfer with the rates given as

$$a_{n,T} = \Theta_p(H_T n^{-1/2}), \quad b_{n,T} = O_p(T^{\rho+1} n^{-1/2}), \quad c_{n,T} = O_p(T^{\rho+3} n^{-1}) = o_p(a_{n,T}).$$

In the $T \to \infty$ setting, since the optimal learning rate approaches the stability edge, the loss suboptimality due to shifting hyperparameter may be amplified. Hence we require the time horizon to diverge slowly (while still ensuring that the loss decays to 0), as specified by the exponent $\chi(\rho)$. We note that (4) provides sufficient condition for fast transfer, but the exponents may not be sharp.

Remark 5 *To translate the fast transfer rates to the computational gain over direct tuning, Ghosh et al. [8] introduces the notion of useful transfer. In particular, [8, Theorem 2] shows that if the transfer is fast, then with a fixed compute budget, tuning parameters at the smaller scale and transferring to the larger scale is always more efficient than direct tuning (grid search) at large scale. However, their derivation assumes a finite curvature around $\eta_{\infty,T}$, which does not hold in our diverging horizon setting. This being said, in Proposition 34 we show that provided the convergence rates in Theorem 4, the transfer strategy is useful in the sense of Ghosh et al. [8], even though the curvature varies with T. In other words, fast transfer also entails useful transfer.*

References

[1] Johan Bjorck, Alon Benhaim, Vishrav Chaudhary, Furu Wei, and Xia Song. Scaling optimal lr across token horizons. *arXiv preprint arXiv:2409.19913*, 2024.

[2] Blake Bordelon and Francesco Mori. Theory of optimal learning rate schedules and scaling laws for a random feature model, 2026. URL https://arxiv.org/abs/2602.04774. arXiv:2602.04774.

[3] Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. *arXiv preprint arXiv:2309.16620*, 2023.

[4] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws, 2024. URL https://arxiv.org/abs/2402.01092. ICML 2024; arXiv:2402.01092.

[5] Lénaïc Chizat. The hidden width of deep resnets: Tight error bounds and phase diagrams. *arXiv preprint arXiv:2509.10167*, 2025.

[6] Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. *Advances in Neural Information Processing Systems*, 37:104630–104693, 2024.

[7] Nolan Dey, Bin Claire Zhang, Lorenzo Noci, Mufan Li, Blake Bordelon, Shane Bergsma, Cengiz Pehlevan, Boris Hanin, and Joel Hestness. Don't be lazy: Completep enables compute-efficient deep transformers. *arXiv preprint arXiv:2505.01618*, 2025.

[8] Nikhil Ghosh, Denny Wu, and Alberto Bietti. Understanding the mechanisms of fast hyperparameter transfer, 2025. URL https://arxiv.org/abs/2512.22768. arXiv:2512.22768.

[9] Soufiane Hayou. A proof of learning rate transfer under μp. *arXiv preprint arXiv:2511.01734*, 2025.

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

[11] 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.

[12] Letong Hong and Zhangyang Wang. On the provable separation of scales in maximal update parameterization. In *Forty-second International Conference on Machine Learning*, 2025.

[13] 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.

[14] Binghui Li, Fengling Chen, Zixun Huang, Lean Wang, and Lei Wu. Functional scaling laws in kernel regression: Loss dynamics and learning rate schedules, 2025. URL https://arxiv.org/abs/2509.19189. arXiv:2509.19189.

[15] Binghui Li, Zilin Wang, Fengling Chen, Shiyang Zhao, Ruiheng Zheng, and Lei Wu. Optimal learning-rate schedules under functional scaling laws: Power decay and warmup-stable-decay, 2026. URL https://arxiv.org/abs/2602.06797. arXiv:2602.06797.

[16] Licong Lin, Jingfeng Wu, Sham M. Kakade, Peter L. Bartlett, and Jason D. Lee. Scaling laws in linear regression: Compute, parameters, and data, 2024. URL https://arxiv.org/abs/2406.08466. arXiv:2406.08466.

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

[18] Bruno Mlodozeniec, Pierre Ablin, Louis Béthune, Dan Busbridge, Michal Klein, Jason Ramapuram, and Marco Cuturi. Completed hyperparameter transfer across modules, width, depth, batch and duration. *arXiv preprint arXiv:2512.22382*, 2025.

[19] Lorenzo Noci, Alexandru Meterez, Thomas Hofmann, and Antonio Orvieto. Super consistency of neural network landscapes and learning rate transfer. *Advances in Neural Information Processing Systems*, 37:102696–102743, 2024.

[20] Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+ 3 phases of compute-optimal neural scaling laws. *Advances in Neural Information Processing Systems*, 37: 16459–16537, 2024.

[21] Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pages 11727–11737. PMLR, 2021.

[22] Greg Yang and Etai Littwin. Tensor programs ivb: Adaptive optimization in the infinite-width limit. *arXiv preprint arXiv:2308.01814*, 2023.

[23] Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

[24] Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs vi: Feature learning in infinite-depth neural networks. *arXiv preprint arXiv:2310.02244*, 2023.

Contents

AI Assistance Disclosure

Large language models were used to assist writing. The mathematical contents are derived and verified by the authors.

Appendix A. Experiments

A.1. Implementation

All experiments use the exact loss recursion for gradient descent in the sketched-linear model. We work with a large finite-dimensional truncation of the Hilbert-space model, with diagonal covariance and source

$$\lambda_k(\Sigma) = k^{-\alpha}, \qquad b_k^2 = k^{-\beta}, \qquad 1 \le k \le d,$$

where $\alpha, \beta > 1$ and $\rho = (\beta - 1)/\alpha$. For each width n, we draw the empirical covariance

$$A_n = \Sigma^{1/2} \frac{GG^\top}{n} \Sigma^{1/2}, \qquad G_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),$$

using the Bartlett decomposition for the Wishart law, which avoids explicitly storing the full $d \times n$ sketch matrix when n is large. Given the eigendecomposition $A_n = Q \operatorname{diag}(\mu_j) Q^\top$, we evaluate the finite-width loss exactly as

$$L_{n,T}(\eta) = \frac{1}{2} b^\top (I - \eta A_n)^{2T} b = \frac{1}{2} \sum_{j=1}^{d} (q_j^\top b)^2 (1 - \eta \mu_j)^{2T},$$

and the infinite-width loss as

$$L_{\infty,T}(\eta) = \frac{1}{2} \sum_{k=1}^{d} b_k^2 (1 - \eta \lambda_k)^{2T}.$$

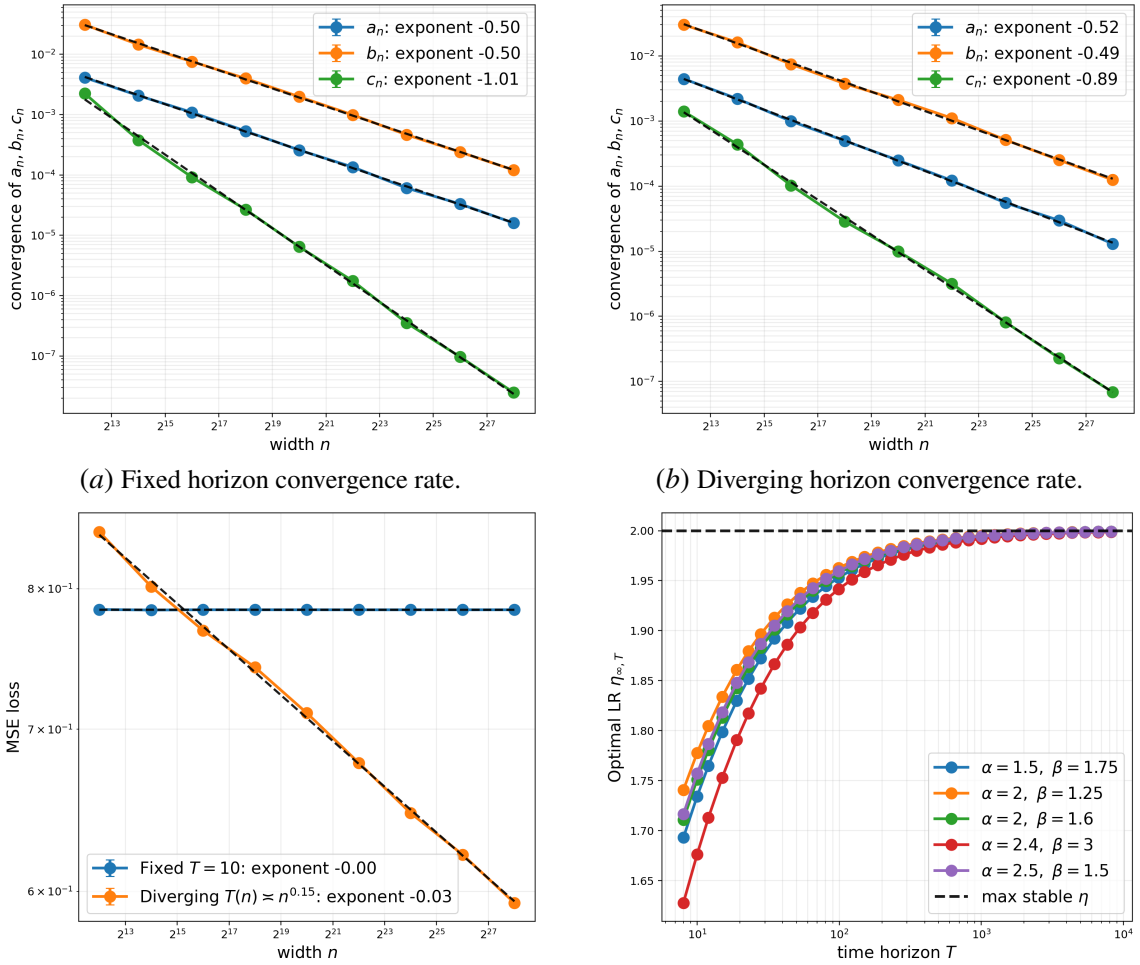
Thus no stochastic-gradient or full-batch parameter recursion is simulated; the only numerical optimization is a one-dimensional minimization over the learning rate. The minimization is performed by a dense grid search followed by golden-section refinement. For each width, we compute

$$a_{n,T} = \left| L_{n,T}(\hat{\eta}_{n,T}) - L_{\infty,T}(\eta_{\infty,T}) \right|, \qquad b_{n,T} = |\hat{\eta}_{n,T} - \eta_{\infty,T}|,$$

and

$$c_{n,T} = L_{\infty,T}(\hat{\eta}_{n,T}) - L_{\infty,T}(\eta_{\infty,T}),$$

where $\hat{\eta}_{n,T}$ and $\eta_{\infty,T}$ denote the finite- and infinite-width optimal learning rates. We repeat each finite-width experiment over 256 independent draws of A_n and report means with standard-error bars. For the growing-horizon experiments we set $T(n) = \lceil C n^\gamma \rceil$, with γ chosen below the sufficient condition $T^{\phi(\rho)}/\sqrt{n} \to 0$ from Theorem 4. We also evaluate the unsketched deterministic dynamics across several (α, β) pairs by minimizing $L_{\infty,T}(\eta)$ directly, which allows us to compare the observed loss decay with the predicted scaling $L_{\infty,T}(\eta_{\infty,T}) \asymp T^{-\rho}$ and to verify that $\eta_{\infty,T}$ approaches the stability edge 2 from below.



(*a*) Fixed horizon convergence rate.

(*b*) Diverging horizon convergence rate.

(*c*) MSE loss comparison.

(*d*) Convergence of optimal learning rate.

Figure 1: Experimental results on fast transfer in sketched linear regression with power-law features. **(a)** We set $\alpha = 2, \beta = 1.25$, and fix $T = 10$ as n varies. **(b)** Under the same data setting as (a), but we consider a joint scaling $T = \lfloor 2n^{0.15} \rfloor$; observe that this alters the scaling exponent of c_n. **(c)** MSE loss $L_{n,T}(\hat{\eta}_{n,T})$ for fixed vs. diverging T. **(d)** Optimal learning rate at infinite width as a function of time horizon; observe that $\eta_{\infty,T}$ approached $2/\lambda_{\max}(\Sigma)$ from below.

A.2. Experimental results

- In Figure 1(a) and (b) we compare the convergence rates of a_n, b_n, c_n in the fixed vs. growing horizon settings. We set $\alpha = 2, \beta = 1.25$, and vary the model width n. In Figure 1(a) we fix $T = 10$, and we observe that $a_n, b_n = \Theta(n^{-1/2})$, whereas $c_n = \Theta(n^{-1})$, which implies fast transfer as predicted by Theorem 1.
  On the other hand, in Figure 1(b) we jointly scale the model width and the optimization time, with $T \asymp n^{0.15}$ which meets the requirement of Theorem 4; while the growth is slow, we observe that the scaling exponent of c_n is altered (compared to Figure 1(a)).

- In Figure 1(c) we plot the MSE loss scaling (with optimal learning rate) of the two settings, where we observe that the loss remains at constant order with fixed T, but follows a power-law decay when n, T both increases. Note that Theorem 4 establishes fast transfer in this joint scaling regime.
- In Figure 1(d) we compute the optimal learning rate $\eta_{\infty,T}$ across time horizon, for different values of (α, β). It is clear that the optimal learning rate approaches the stability edge from below, consistent with the theoretical prediction in (3).

Appendix B. Additional related works

The concept of hyperparameter (HP) transfer was introduced by Yang et al. [23], who showed that HPs tuned on small proxy models can transfer reliably to much larger ones under μP scaling. Subsequent works investigated other scaling dimensions such as depth [3, 5, 7, 24] and time horizon [1, 18]. As emphasized by Yang et al. [23], however, the success of HP transfer is not fully explained from first principles. Recent work has begun to address this gap: Noci et al. [19] observed empirically that top Hessian eigenvalues stabilize rapidly across widths under μP, Hong and Wang [12] established a scale separation between macro- and micro-variables that supports early-stage HP stability, and Hayou [9] proved convergence of optimal learning rate for linear networks under μP.

Most closely related to our work, Ghosh et al. [8] introduced a framework for *fast transfer*, comparing the finite-scale loss gap, HP gap, and transferred-loss suboptimality, and showed that fast transfer implies computational gains over direct large-scale tuning. Our work studies this mechanism in a solvable sketched linear regression model, connecting HP transfer to the growing theory of neural scaling laws. This literature ranges from empirical scaling laws for language models [10, 11, 13] to solvable random features models that derive parameter, data, and time scaling exponents under power-law spectral assumptions [4, 6, 16, 17, 20]. Recent functional scaling-law analyses further characterize the full loss trajectory and optimal learning-rate schedules in kernel and random-feature settings [2, 14, 15]. In contrast to these works, which primarily analyze limiting loss curves and decay rates, we focus on the finite-width fluctuations of the learning-rate optimum itself, and establish the computational benefit of the transfer strategy.

Appendix C. Model, notations, and gradient-descent identities

Notations. We work on the real separable Hilbert space $\mathcal{H} = \ell^2(\mathbb{N})$ of squared summable sequences, with canonical orthonormal basis $(e_k)_{k\geq 1}$ and inner product $\langle \cdot, \cdot \rangle$. For $u, v \in \mathcal{H}$, the rank-one operator $u \otimes v$ is defined by $(u \otimes v)h = u \langle v, h \rangle$. Operator powers are understood through the usual functional calculus for bounded self-adjoint operators. We write $\|B\|_{\mathrm{op}}$ and $\|B\|_F$ for the operator and Hilbert–Schmidt norms, respectively.

The student width is n, the gradient-descent horizon is T, and the step size is η. Unless otherwise stated, limits are taken as $n \to \infty$; for growing-horizon statements we write $T = T_n$. We use $X_n = O_p(a_n)$, $X_n = o_p(a_n)$, and $X_n = \Theta_p(a_n)$ in their standard probabilistic senses, that is, $X_n = O_p(a_n)$ when X_n/a_n is bounded in probability, $X_n = o_p(a_n)$ when $X_n/a_n \to 0$ in probability, and $X_n = \Theta_p(a_n)$ if $X_n = O_p(a_n)$ and $a_n = O_p(X_n)$. We further write $X_n = O_{L^2}(a_n)$ when $\mathbb{E}X_n^2 \leq C a_n^2$ for some constant $C > 0$ and all n large enough. Deterministic comparisons are denoted by the standard notations $O(\cdot), o(\cdot), \Theta(\cdot)$, or equivalently $\lesssim, \ll, \asymp$.

Setting. The input is an infinite-dimensional centered vector with covariance operator $\mathbb{E}[x \otimes x] = \Sigma$. Without loss of generality, we assume the operator to be trace class and diagonal

$$\Sigma e_k = \lambda_k e_k, \qquad \lambda_k \geq 0, \qquad \mathrm{Tr}(\Sigma) = \sum_{k \geq 1} \lambda_k < \infty,$$

with $\|\Sigma\|_{\mathrm{op}} = 1$. The trace condition ensures that $x \in \mathcal{H}$ almost surely. The response are generated by a noiseless linear target

$$y = \langle \beta_0, x \rangle, \qquad b = \Sigma^{1/2} \beta_0 \in \mathcal{H}, \qquad \mathbb{E}[y^2] = \|b\|^2 < \infty.$$

Independent additive label noise could be added, but it would only add a constant to the population risk and does not affect the results in this paper.

Source and capacity conditions. The fixed-horizon arguments only require Σ to be trace class and $b \in \mathcal{H}$. For the growing-horizon results, we impose the following so-called source-capacity conditions

$$\lambda_k = k^{-\alpha}, \qquad b_k^2 = k^{-\beta}, \qquad \alpha > 1, \qquad \beta > 1.$$

Note that $\alpha > 1$ and $\beta > 1$ imply $\mathrm{Tr}(\Sigma) < \infty$ and $b \in \mathcal{H}$. Equivalently,

$$\lambda_k (\beta_0)_k^2 = b_k^2 = k^{-\beta} \qquad \Longrightarrow \qquad (\beta_0)_k^2 = k^{\alpha - \beta}.$$

We use the two spectral exponents

$$\rho := \frac{\beta - 1}{\alpha} > 0, \qquad s := \frac{\alpha - 1}{\alpha} \in (0, 1).$$

Sketched linear model. The finite-width sketch is represented by independent Gaussian feature vectors

$$z_1, \ldots, z_n \overset{\text{i.i.d.}}{\sim} N(0, \Sigma)$$

and by the finite-rank sketching map $Z_n : \mathbb{R}^n \to \mathcal{H}$,

$$Z_n \theta := \frac{1}{\sqrt{n}} \sum_{r=1}^{n} \theta_r z_r.$$

The associated empirical covariance operator is

$$A_n := Z_n Z_n^* = \frac{1}{n} \sum_{r=1}^{n} z_r \otimes z_r.$$

Thus A_n is self-adjoint, positive semidefinite, has rank at most n, and satisfies $\mathbb{E} A_n = \Sigma$. The infinite-width operator is

$$A_\infty := \Sigma.$$

Conditionally on the sketch, the width-n student predicts in whitened coordinates $g := \Sigma^{\dagger/2} x$ by

$$\widehat{f}(x; \theta) = \langle g, Z_n \theta \rangle = \langle x, W_n \theta \rangle, \qquad \theta \in \mathbb{R}^n, \qquad W_n := \Sigma^{\dagger/2} Z_n.$$

The analysis below depends on the sketch only through A_n.

Gradient descent. For a fixed sketch, the population square loss of the sketched model is

$$L_n(\theta) := \frac{1}{2}\mathbb{E}_g\left[\left(\widehat{f}(x;\theta) - y\right)^2\right] = \frac{1}{2}\left\|Z_n\theta - b\right\|^2.$$

For a stepsize η and $T \in \mathbb{N}$, define the finite- and infinite-width losses after T steps of gradient descent by

$$L_{n,T}(\eta) := \frac{1}{2}\left\langle b, (I - \eta A_n)^{2T} b\right\rangle, \tag{5}$$
$$L_{\infty,T}(\eta) := \frac{1}{2}\left\langle b, (I - \eta \Sigma)^{2T} b\right\rangle.$$

For completeness, we provide a proof of these well-known identities.

Lemma 6 (Exact finite-width dynamics) *Consider (y, x) with $\mathbb{E}[x \otimes x] = \Sigma$ and $y = \langle x, \beta_0\rangle$, and predictor $\hat{f}(x, \theta) = \theta^\top W^\top x = \langle x, W\theta\rangle$ with $\theta \in \mathbb{R}^n$. The population quadratic loss is*

$$L(\theta) = \frac{1}{2}\mathbb{E}_x\left(\langle x, W\theta - \beta_0\rangle\right)^2 = \frac{1}{2}(W\theta - \beta_0)^\top \Sigma (W\theta - \beta_0).$$

If $\theta_0 = 0$ and

$$\theta_{t+1} = \theta_t - \eta \nabla \ell(\theta_t),$$

then

$$\theta_T = \eta \sum_{k=0}^{T-1} (I - \eta W^\top \Sigma W)^k W^\top \Sigma \beta_0. \tag{6}$$

Moreover, if $u_t := \Sigma^{1/2}(\beta_0 - W\theta_t)$, then

$$u_{t+1} = (I - \eta A_n)u_t, \qquad u_0 = b,$$

where $A_n = \Sigma^{1/2} W W^\top \Sigma^{1/2}$. Therefore the loss after T steps is exactly (5).

Proof The loss identity follows from the covariance formula $\mathbb{E}_x\left(\langle x, v\rangle\right)^2 = v^\top \Sigma v$. Differentiating $L(\theta)$ gives

$$\nabla \ell(\theta) = W^\top \Sigma (W\theta - \beta_0).$$

Thus

$$\theta_{t+1} = (I - \eta W^\top \Sigma W)\theta_t + \eta W^\top \Sigma \beta_0.$$

Iterating this affine recursion from $\theta_0 = 0$ gives (6).

For the residual $e_t := \beta_0 - W\theta_t$,

$$e_{t+1} = \beta_0 - W\theta_{t+1} = \beta_0 - W\theta_t - \eta W W^\top \Sigma(\beta_0 - W\theta_t) = (I - \eta W W^\top \Sigma)e_t.$$

Multiplying by $\Sigma^{1/2}$ gives

$$u_{t+1} = \Sigma^{1/2}(I - \eta W W^\top \Sigma)\Sigma^{-1/2} u_t = (I - \eta \Sigma^{1/2} W W^\top \Sigma^{1/2}) u_t.$$

Thus $u_T = (I - \eta A_n)^T b$. The loss is

$$\ell(\theta_T) = \frac{1}{2}\left\|u_T\right\|^2 = \frac{1}{2}\left\langle b, (I - \eta A_n)^{2T} b\right\rangle,$$

because A_n is self-adjoint. The infinite-width identity follows by replacing A_n with Σ. ■

Appendix D. Basic bounds

D.1. Sample-covariance concentration

Lemma 7 (Hilbert–Schmidt and operator convergence) *Let*

$$A_n = \frac{1}{n}\sum_{r=1}^{n} z_r \otimes z_r, \qquad z_r \overset{\text{i.i.d.}}{\sim} N(0,\Sigma).$$

Then

$$\mathbb{E}\,\|A_n - \Sigma\|_F^2 = \frac{1}{n}\left((\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2)\right), \tag{7}$$

and hence

$$\|A_n - \Sigma\|_{\text{op}} = O_p(n^{-1/2}). \tag{8}$$

Proof Write $Y_r := z_r \otimes z_r - \Sigma$, so that $A_n - \Sigma = n^{-1}\sum_{r=1}^{n} Y_r$. Note that Y_r are independent with $\mathbb{E}Y_r = 0$. Thus

$$\mathbb{E}\,\|A_n - \Sigma\|_F^2 = \frac{1}{n^2}\sum_{r=1}^{n}\mathbb{E}\,\|Y_r\|_F^2 = \frac{1}{n}\mathbb{E}\,\|z\otimes z - \Sigma\|_F^2\,.$$

Using that

$$\|z\otimes z\|_F^2 = \|z\|^4, \qquad \langle z\otimes z, \Sigma\rangle_F = \langle z, \Sigma z\rangle,$$

and that for a centered Gaussian vector with covariance Σ,

$$\mathbb{E}\,\|z\|^4 = (\operatorname{tr}\Sigma)^2 + 2\operatorname{tr}(\Sigma^2),$$

we obtain

$$\begin{aligned}\mathbb{E}\,\|z\otimes z - \Sigma\|_F^2 &= \mathbb{E}\,\|z\|^4 - 2\mathbb{E}\,\langle z, \Sigma z\rangle + \|\Sigma\|_F^2 \\ &= \mathbb{E}\,\|z\|^4 - 2\operatorname{tr}(\Sigma^2) + \operatorname{tr}(\Sigma^2) \\ &= (\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2),\end{aligned}$$

which proves (7). Since $\|B\|_{\text{op}} \le \|B\|_F$, Markov's inequality yields (8). ■

D.2. Power-law tail sums used throughout

For $u > 1$, $m \ge 0$, and $\eta \in [1,2]$, define

$$W_u(m;\eta) := \sum_{k\ge 2} k^{-u}(1-\eta k^{-\alpha})^m.$$

Because $\alpha > 1$, for $k \ge 2$ and $\eta \in [1,2]$,

$$0 < 1 - \eta k^{-\alpha} \le 1.$$

Lemma 8 (Uniform sum estimate) *For every $u > 1$, there exist constants $0 < c_u < C_u < \infty$, depending only on u and α, such that for every $m \ge 1$ and every $\eta \in [1,2]$,*

$$c_u(m+1)^{-(u-1)/\alpha} \le W_u(m;\eta) \le C_u(m+1)^{-(u-1)/\alpha}.$$

Also $W_u(0;\eta) = \sum_{k\ge 2} k^{-u} < \infty$.

Proof For $k \geq 2$ and $\eta \in [1, 2]$, the number $x = \eta k^{-\alpha}$ lies in the compact interval $[0, 2^{1-\alpha}] \subset [0, 1)$. Therefore there are constants $0 < c < C < \infty$, depending only on α, such that

$$cx \leq -\log(1 - x) \leq Cx, \qquad 0 \leq x \leq 2^{1-\alpha}.$$

Consequently

$$e^{-Cmk^{-\alpha}} \leq (1 - \eta k^{-\alpha})^m \leq e^{-cmk^{-\alpha}},$$

where we used $\eta \in [1, 2]$ and absorbed constants. It remains to estimate

$$S_c(m) := \sum_{k \geq 2} k^{-u} e^{-cmk^{-\alpha}}.$$

There exist constants C_1, C_2 such that

$$C_1 \int_2^\infty x^{-u} e^{-Cmx^{-\alpha}}\, dx \leq S_c(m) \leq C_2 \int_1^\infty x^{-u} e^{-cmx^{-\alpha}}\, dx.$$

For either integral, use the change of variables $t = mx^{-\alpha}$. Then

$$x = (m/t)^{1/\alpha}, \qquad dx = -\frac{1}{\alpha} m^{1/\alpha} t^{-1/\alpha - 1}\, dt,$$

and

$$\int_1^\infty x^{-u} e^{-cmx^{-\alpha}}\, dx = \frac{1}{\alpha} m^{(1-u)/\alpha} \int_0^m t^{(u-1)/\alpha - 1} e^{-ct}\, dt.$$

Because $u > 1$, the integral in t is bounded above by a finite constant independent of m and bounded below by a positive constant for all $m \geq 1$. This proves the claim. ■

Appendix E. Fixed-horizon learning-rate transfer

This section provides the proof for the fixed-T result (Theorem 1). The argument follows a delta method approach (in Hilbert-space) applied to $L_{n,T}(\eta)$ seen as a functional of the empirical covariance A_n.

E.1. Optimality and uniform control

For fixed T, set

$$F_{n,T}(\eta) := \left\langle b, A_n (I - \eta A_n)^{2T-1} b \right\rangle, \qquad F_{\infty,T}(\eta) := \left\langle b, \Sigma (I - \eta \Sigma)^{2T-1} b \right\rangle.$$

Then

$$-\frac{1}{T} \frac{d}{d\eta} L_{n,T}(\eta) = F_{n,T}(\eta), \qquad -\frac{1}{T} \frac{d}{d\eta} L_{\infty,T}(\eta) = F_{\infty,T}(\eta).$$

Writing the nonzero spectral decomposition of A_n as

$$A_n = \sum_{j=1}^{r_n} \mu_{n,j} v_{n,j} \otimes v_{n,j}, \qquad \mu_{n,j} > 0,$$

gives

$$L_{n,T}(\eta) = \frac{1}{2}\left\|P_{\ker(A_n)}b\right\|^2 + \frac{1}{2}\sum_{j=1}^{r_n}\langle b, v_{n,j}\rangle^2 (1-\eta\mu_{n,j})^{2T}.$$

Hence $L_{n,T}$ is convex in η and its derivative is strictly increasing unless $A_n b = 0$. Thus, when $A_n b \neq 0$, the unique minimizer over $\eta \geq 0$ is the unique root of $F_{n,T}(\eta) = 0$. The same argument with A_n replaced by Σ gives the infinite-width root $\eta_{\infty,T}$.

Lemma 9 (Uniform fixed-horizon score control) *For every fixed T and $R < \infty$,*

$$\sup_{|\eta|\leq R} |F_{n,T}(\eta) - F_{\infty,T}(\eta)| = O_p(n^{-1/2}).$$

The same statement holds after replacing $F_{n,T}, F_{\infty,T}$ by their first derivative with respect to η.

Proof Let

$$p_\eta(x) = x(1-\eta x)^{2T-1}.$$

On the event $\|A_n - \Sigma\|_{\mathrm{op}} \leq 1$, the spectra of A_n and Σ lie in $[-2, 2]$. Write

$$p_\eta(x) = \sum_{m=1}^{2T} c_m(\eta) x^m,$$

where the coefficients $c_m(\eta)$ are polynomials in η and hence are uniformly bounded for $|\eta| \leq R$. For each $m \geq 1$,

$$A_n^m - \Sigma^m = \sum_{j=0}^{m-1} A_n^j (A_n - \Sigma)\Sigma^{m-1-j},$$

so on the same event

$$\|A_n^m - \Sigma^m\|_{\mathrm{op}} \leq m 2^{m-1} \|A_n - \Sigma\|_{\mathrm{op}}.$$

Summing over $m \leq 2T$ gives

$$\sup_{|\eta|\leq R} \|p_\eta(A_n) - p_\eta(\Sigma)\|_{\mathrm{op}} \leq C_{T,R} \|A_n - \Sigma\|_{\mathrm{op}}.$$

Taking the quadratic form against b and using Lemma 7 proves the displayed score bound.

For the derivative, compute

$$\partial_\eta p_\eta(x) = -(2T-1)x^2(1-\eta x)^{2T-2}.$$

This is another polynomial of degree $2T$ with coefficients uniformly bounded for $|\eta| \leq R$. Repeating the preceding monomial estimate proves the same uniform bound for $F'_{n,T} - F'_{\infty,T}$. ■

E.2. Delta-method expansion

Throughout this subsection, $\eta_{\infty,T}$ denotes the infinite-width root for the fixed horizon T. Define

$$Q_F := \sum_{m=0}^{2T-1} \binom{2T-1}{m} (-\eta_{\infty,T})^m \sum_{j=0}^{m} \Sigma^{m-j}(b \otimes b)\Sigma^j, \tag{9}$$

$$Q_L := \frac{1}{2} \sum_{m=1}^{2T} \binom{2T}{m} (-\eta_{\infty,T})^m \sum_{j=0}^{m-1} \Sigma^{m-1-j}(b \otimes b)\Sigma^j. \tag{10}$$

These operators are finite sums of rank-one Hilbert–Schmidt operators. They are the Frechet derivatives at Σ of

$$A \mapsto \left\langle b, A(I - \eta_{\infty,T}A)^{2T-1}b \right\rangle, \qquad A \mapsto \frac{1}{2}\left\langle b, (I - \eta_{\infty,T}A)^{2T}b \right\rangle,$$

respectively.

Lemma 10 (Linearizations) *As $n \to \infty$,*

$$F_{n,T}(\eta_{\infty,T}) = \frac{1}{n}\sum_{r=1}^{n} \{\langle z_r, Q_F z_r\rangle - \operatorname{tr}(\Sigma Q_F)\} + o_p(n^{-1/2}),$$

$$L_{n,T}(\eta_{\infty,T}) - L_{\infty,T}(\eta_{\infty,T}) = \frac{1}{n}\sum_{r=1}^{n} \{\langle z_r, Q_L z_r\rangle - \operatorname{tr}(\Sigma Q_L)\} + o_p(n^{-1/2}).$$

Proof Let $E_n := A_n - \Sigma$. It is enough to check a monomial, because both functionals are finite linear combinations of $A \mapsto \langle b, A^m b\rangle$. For fixed $m \geq 1$,

$$(\Sigma + E_n)^m - \Sigma^m = \sum_{j=0}^{m-1} \Sigma^{m-1-j} E_n \Sigma^j + \mathcal{R}_m(E_n), \tag{11}$$

where every word in $\mathcal{R}_m(E_n)$ contains at least two copies of E_n. On the event $\|E_n\|_{\mathrm{op}} \leq 1$, the operators in all such words are uniformly bounded, and the number of words depends only on m. Hence

$$|\langle b, \mathcal{R}_m(E_n)b\rangle| \leq C_m \|b\|^2 \|E_n\|_{\mathrm{op}}^2 = O_p(n^{-1})$$

by Lemma 7. Substituting (11) into the polynomial expansions of

$$A \mapsto \left\langle b, A(I - \eta_{\infty,T}A)^{2T-1}b \right\rangle, \qquad A \mapsto \frac{1}{2}\left\langle b, (I - \eta_{\infty,T}A)^{2T}b \right\rangle$$

gives the Frechet derivative terms represented by Q_F and Q_L in (9)–(10). Equivalently, if $G_Q(A)$ denotes either scalar polynomial functional and Q denotes its derivative-representing operator (Q_F for the score and Q_L for the loss), then its Frechet derivative at Σ in the direction E_n is

$$\begin{aligned} DG_Q(\Sigma)[E_n] &= \langle E_n, Q\rangle_F \\ &= \left\langle \frac{1}{n}\sum_{r=1}^{n}(z_r \otimes z_r - \Sigma), Q \right\rangle_F \\ &= \frac{1}{n}\sum_{r=1}^{n} \{\langle z_r, Qz_r\rangle - \operatorname{tr}(\Sigma Q)\}. \end{aligned}$$

The remainder is $O_p(n^{-1}) = o_p(n^{-1/2})$, which proves both expansions. ■

Proof of Theorem 1. Fix T. We first prove that the finite-width optimizer is a root of the finite-width derivative in a deterministic compact interval, with probability tending to one. The infinite-width score is continuous, strictly decreasing, and satisfies

$$F_{\infty,T}(0) = \langle b, \Sigma b\rangle > 0, \qquad F_{\infty,T}(\eta_{\infty,T}) = 0.$$

Since the degree $2T-1$ is odd, the formula

$$F_{\infty,T}(\eta) = \sum_{k\geq 1} b_k^2 \lambda_k (1-\eta\lambda_k)^{2T-1}$$

shows that we may choose a deterministic $R > \eta_{\infty,T}$ with $F_{\infty,T}(R) < 0$. Lemma 9 gives

$$\sup_{0\leq\eta\leq R} |F_{n,T}(\eta) - F_{\infty,T}(\eta)| \xrightarrow{P} 0.$$

Therefore, with probability tending to one,

$$F_{n,T}(0) > 0, \qquad F_{n,T}(R) < 0.$$

On this event $F_{n,T}$ has a zero in $(0, R)$. Moreover,

$$L'_{n,T}(\eta) = -TF_{n,T}(\eta), \qquad L''_{n,T}(\eta) = T(2T-1)\left\langle b, A_n^2(I-\eta A_n)^{2T-2} b\right\rangle \geq 0,$$

so $L_{n,T}$ is convex in η, and any interior zero of $F_{n,T}$ is a global minimizer over $\eta \geq 0$. We take $\eta_{n,T}$ to be this minimizer on the high-probability event above.

Next we prove consistency. Let $\varepsilon > 0$ be small enough that $0 < \eta_{\infty,T} - \varepsilon < \eta_{\infty,T} + \varepsilon < R$. Because $F_{\infty,T}$ is strictly decreasing through its zero,

$$F_{\infty,T}(\eta_{\infty,T} - \varepsilon) > 0, \qquad F_{\infty,T}(\eta_{\infty,T} + \varepsilon) < 0.$$

The same uniform convergence on $[0, R]$ implies that these two signs also hold for $F_{n,T}$ with probability tending to one. Since $F_{n,T}$ is nonincreasing, every zero of $F_{n,T}$ in $[0, R]$ lies in

$$[\eta_{\infty,T} - \varepsilon, \eta_{\infty,T} + \varepsilon]$$

on that event. Hence

$$\eta_{n,T} \xrightarrow{P} \eta_{\infty,T}.$$

Now take a deterministic compact interval $K \subset (0, R)$ containing $\eta_{\infty,T}$ in its interior. The preceding consistency gives $\eta_{n,T} \in K$ with probability tending to one. On this event the mean-value theorem gives

$$0 = F_{n,T}(\eta_{n,T}) = F_{n,T}(\eta_{\infty,T}) + F'_{n,T}(\widetilde{\eta}_{n,T})(\eta_{n,T} - \eta_{\infty,T}), \tag{12}$$

where $\widetilde{\eta}_{n,T}$ lies between $\eta_{n,T}$ and $\eta_{\infty,T}$. Since both endpoints converge to $\eta_{\infty,T}$, we also have

$$\widetilde{\eta}_{n,T} \xrightarrow{P} \eta_{\infty,T}.$$

Furthermore, Lemma 9 applied to the derivative and the continuity of $F'_{\infty,T}$ yield

$$\begin{aligned}\left|F'_{n,T}(\widetilde{\eta}_{n,T}) - F'_{\infty,T}(\eta_{\infty,T})\right| \le \sup_{\eta\in K} |F'_{n,T}(\eta) - F'_{\infty,T}(\eta)| \\ + \left|F'_{\infty,T}(\widetilde{\eta}_{n,T}) - F'_{\infty,T}(\eta_{\infty,T})\right| \xrightarrow{P} 0.\end{aligned}$$

The limiting derivative is strictly negative:

$$F'_{\infty,T}(\eta_{\infty,T}) = -(2T-1)\sum_{k\ge 1}\lambda_k^3(\beta_0)_k^2(1-\eta_{\infty,T}\lambda_k)^{2T-2} < 0.$$

Thus the random denominator in (12) is bounded away from zero with probability tending to one. Since Lemma 9 also gives $F_{n,T}(\eta_{\infty,T}) = O_p(n^{-1/2})$, equation (12) implies

$$\eta_{n,T} - \eta_{\infty,T} = O_p(n^{-1/2}).$$

For the limiting laws, Lemma 10 and the classical i.i.d. central limit theorem give

$$\sqrt{n}\, F_{n,T}(\eta_{\infty,T}) \Rightarrow N\left(0, 2\,\mathrm{tr}(\Sigma Q_F \Sigma Q_F)\right), \tag{13}$$

$$\sqrt{n}\left\{L_{n,T}(\eta_{\infty,T}) - L_{\infty,T}(\eta_{\infty,T})\right\} \Rightarrow N\left(0, 2\,\mathrm{tr}(\Sigma Q_L \Sigma Q_L)\right), \tag{14}$$

using the Gaussian quadratic-form identity

$$\mathrm{Var}\{\langle z, Qz\rangle\} = 2\,\mathrm{tr}(\Sigma Q \Sigma Q), \qquad z \sim N(0,\Sigma),$$

for the finite-rank self-adjoint operators Q_F and Q_L. Combining (12), the denominator convergence above, and (13) gives

$$\sqrt{n}(\eta_{n,T} - \eta_{\infty,T}) = -\frac{\sqrt{n}\, F_{n,T}(\eta_{\infty,T})}{F'_{\infty,T}(\eta_{\infty,T})} + o_p(1),$$

which proves the stepsize CLT.

It remains to pass from the loss evaluated at $\eta_{\infty,T}$ to the optimally tuned finite-width loss. On the same compact interval K, and on the event $\|A_n - \Sigma\|_{\mathrm{op}} \le 1$, the fixed-degree polynomial formula for $L''_{n,T}$ gives

$$\sup_{\eta\in K} |L''_{n,T}(\eta)| = O_p(1).$$

Since $\eta_{n,T} \in K$ with probability tending to one and $L'_{n,T}(\eta_{n,T}) = 0$ there, Taylor's theorem around $\eta_{n,T}$ gives, for some $\bar{\eta}_{n,T}$ between $\eta_{n,T}$ and $\eta_{\infty,T}$,

$$L_{n,T}(\eta_{\infty,T}) - L_{n,T}(\eta_{n,T}) = \frac{1}{2}L''_{n,T}(\bar{\eta}_{n,T})(\eta_{\infty,T} - \eta_{n,T})^2 = O_p(n^{-1}).$$

Therefore

$$\begin{aligned}&\sqrt{n}\{L_{n,T}(\eta_{n,T}) - L_{\infty,T}(\eta_{\infty,T})\} \\ &\quad = \sqrt{n}\{L_{n,T}(\eta_{\infty,T}) - L_{\infty,T}(\eta_{\infty,T})\} + o_p(1),\end{aligned}$$

and the loss CLT follows from (14).

E.3. Fixed-horizon fast-transfer rates

We now derive the sharp fixed-horizon rates after the CLTs. Throughout this subsection T is fixed. Define

$$\begin{aligned} a_n(T) &:= |L_{n,T}(\eta_{n,T}) - L_{\infty,T}(\eta_{\infty,T})|, \\ b_n(T) &:= |\eta_{n,T} - \eta_{\infty,T}|, \\ c_n(T) &:= L_{\infty,T}(\eta_{n,T}) - L_{\infty,T}(\eta_{\infty,T}), \\ \widetilde{c}_n(T) &:= L_{n,T}(\eta_{\infty,T}) - L_{n,T}(\eta_{n,T}). \end{aligned}$$

Here $c_n(T)$ is the transfer suboptimality of Ghosh et al. [8]: the finite-width optimizer is evaluated on the infinite-width loss.

Set

$$\sigma^2_{L,T} := 2\,\mathrm{tr}(\Sigma Q_L \Sigma Q_L), \qquad \sigma^2_{F,T} := 2\,\mathrm{tr}(\Sigma Q_F \Sigma Q_F),$$

and

$$\sigma^2_{\eta,T} := \frac{\sigma^2_{F,T}}{\{F'_{\infty,T}(\eta_{\infty,T})\}^2}, \qquad \kappa_T := L''_{\infty,T}(\eta_{\infty,T}) = -TF'_{\infty,T}(\eta_{\infty,T}) > 0.$$

The positivity of κ_T follows from the fixed-horizon root calculation above: equivalently,

$$\kappa_T = T(2T-1)\sum_{k\geq 1} \lambda_k^3 (\beta_0)_k^2 (1-\eta_{\infty,T}\lambda_k)^{2T-2} > 0.$$

The loss CLT in Theorem 1 gives

$$\sqrt{n}\{L_{n,T}(\eta_{n,T}) - L_{\infty,T}(\eta_{\infty,T})\} \Rightarrow G_L, \qquad G_L \sim N(0, \sigma^2_{L,T}).$$

By the continuous mapping theorem,

$$\sqrt{n}\, a_n(T) \Rightarrow |G_L|.$$

Thus $a_n(T) = O_p(n^{-1/2})$ always, and if $\sigma^2_{L,T} > 0$, then $|G_L|$ is finite and positive with probability one; hence

$$a_n(T) = \Theta_p(n^{-1/2}).$$

Similarly, the optimizer CLT gives

$$\sqrt{n}(\eta_{n,T} - \eta_{\infty,T}) \Rightarrow G_\eta, \qquad G_\eta \sim N(0, \sigma^2_{\eta,T}),$$

and therefore

$$\sqrt{n}\, b_n(T) \Rightarrow |G_\eta|.$$

If $\sigma^2_{F,T} > 0$, then $\sigma^2_{\eta,T} > 0$ and

$$b_n(T) = \Theta_p(n^{-1/2}).$$

We next expand the Ghosh-style transfer suboptimality $c_n(T)$. Let

$$\delta_n := \eta_{n,T} - \eta_{\infty,T}.$$

The optimizer CLT implies $\delta_n = O_p(n^{-1/2})$. Since T is fixed, $L_{\infty,T}$ is a polynomial in η with finite derivatives on a deterministic compact neighborhood of $\eta_{\infty,T}$. Taylor's theorem at $\eta_{\infty,T}$, using $L'_{\infty,T}(\eta_{\infty,T}) = 0$, gives

$$L_{\infty,T}(\eta_{n,T}) - L_{\infty,T}(\eta_{\infty,T}) = \frac{\kappa_T}{2}\delta_n^2 + R_n,$$

where, with high probability,

$$|R_n| \le C|\delta_n|^3.$$

Consequently $R_n = O_p(n^{-3/2}) = o_p(n^{-1})$. Multiplying by n and applying the continuous mapping theorem to the optimizer CLT yields

$$n\, c_n(T) \Rightarrow \frac{\kappa_T}{2} G_\eta^2.$$

If $\sigma_{F,T}^2 > 0$, then G_η^2 is finite and positive with probability one, so

$$c_n(T) = \Theta_p(n^{-1}).$$

Finally consider the reverse operational gap $\widetilde{c}_n(T)$. Taylor's theorem at the finite-width minimizer gives, for some random point ξ_n between $\eta_{n,T}$ and $\eta_{\infty,T}$,

$$\widetilde{c}_n(T) = \frac{1}{2} L''_{n,T}(\xi_n)(\eta_{\infty,T} - \eta_{n,T})^2.$$

The fixed-degree polynomial argument used in Lemma 9 also gives uniform convergence of second derivatives on every fixed compact interval containing $\eta_{\infty,T}$:

$$\sup_{\eta \in K} |L''_{n,T}(\eta) - L''_{\infty,T}(\eta)| = O_p(n^{-1/2}).$$

Since $\eta_{n,T} \to \eta_{\infty,T}$ in probability, $\xi_n \to \eta_{\infty,T}$ in probability. The preceding uniform convergence and continuity of $L''_{\infty,T}$ imply

$$L''_{n,T}(\xi_n) \xrightarrow{P} \kappa_T.$$

Therefore, by Slutsky's theorem,

$$n\, \widetilde{c}_n(T) \Rightarrow \frac{\kappa_T}{2} G_\eta^2.$$

Under $\sigma_{F,T}^2 > 0$, the reverse operational gap also satisfies $\widetilde{c}_n(T) = \Theta_p(n^{-1})$.

Combining the two nondegeneracy conditions, $\sigma_{L,T}^2 > 0$ and $\sigma_{F,T}^2 > 0$, gives

$$a_n(T) = \Theta_p(n^{-1/2}), \qquad c_n(T) = \Theta_p(n^{-1}),$$

and hence $c_n(T) = o_p(a_n(T))$.

Appendix F. The infinite-width optimal stepsize

The infinite-width loss is

$$L_{\infty,T}(\eta) = \frac{1}{2}\sum_{k\geq 1} k^{-\beta}(1-\eta k^{-\alpha})^{2T}.$$

For ease of notation, write $F_T := F_{\infty,\eta}$ from now on. For $\eta \in [1,2]$,

$$F_T(\eta) = -\frac{1}{T}\frac{d}{d\eta}L_{\infty,T}(\eta) = \sum_{k\geq 1} k^{-(\alpha+\beta)}(1-\eta k^{-\alpha})^{2T-1}.$$

Then the minimizer of $L_{\infty,T}$ is the unique zero of F_T.

Lemma 11 (Asymptotics of $\eta_{\infty,T}$) *Let*

$$\eta_{\infty,T} := \arg\min_{\eta\in[1,2]} L_{\infty,T}(\eta).$$

Set

$$\alpha_T := \eta_{\infty,T} - 1, \qquad \varepsilon_T := 1 - \alpha_T = 2 - \eta_{\infty,T}.$$

Then

$$\alpha_T^{2T-1} \asymp T^{-(\rho+1)}, \tag{15}$$

and

$$\eta_{\infty,T} = 2 - \frac{(\rho+1)\log T + O(1)}{2T}. \tag{16}$$

Equivalently,

$$\varepsilon_T = \frac{(\rho+1)\log T + O(1)}{2T}. \tag{17}$$

Proof The second derivative is

$$\frac{d^2}{d\eta^2}L_{\infty,T}(\eta) = T(2T-1)\sum_{k\geq 1} k^{-(\beta+2\alpha)}(1-\eta k^{-\alpha})^{2T-2} > 0.$$

Thus $L_{\infty,T}$ is strictly convex on $[1,2]$ and has at most one critical point. Also

$$F_T(1) = \sum_{k\geq 2} k^{-(\alpha+\beta)}(1-k^{-\alpha})^{2T-1} > 0,$$

whereas

$$F_T(2) = -1 + \sum_{k\geq 2} k^{-(\alpha+\beta)}(1-2k^{-\alpha})^{2T-1} < 0.$$

Indeed, since $\alpha+\beta > 2$ and $|1-2k^{-\alpha}| < 1$ for every $k \geq 2$,

$$\sum_{k\geq 2} k^{-(\alpha+\beta)}|1-2k^{-\alpha}|^{2T-1} \leq \sum_{k\geq 2} k^{-(\alpha+\beta)} < \sum_{k\geq 2} k^{-2} < 1.$$

Hence the minimizer is characterized by

$$F_T(\eta_{\infty,T}) = 0.$$

Because the $k = 1$ summand is

$$(1 - \eta_{\infty,T})^{2T-1} = -(\eta_{\infty,T} - 1)^{2T-1} = -\alpha_T^{2T-1},$$

we obtain the exact identity

$$\alpha_T^{2T-1} = \sum_{k\geq 2} k^{-(\alpha+\beta)}(1 - \eta_{\infty,T}k^{-\alpha})^{2T-1} = W_{\alpha+\beta}(2T-1; \eta_{\infty,T}).$$

By Lemma 8, since

$$\frac{\alpha+\beta-1}{\alpha} = 1 + \frac{\beta-1}{\alpha} = \rho + 1,$$

we get (15). Taking logarithms in (15) gives

$$(2T-1)\log \alpha_T = -(\rho+1)\log T + O(1).$$

Thus $\alpha_T \to 1$. Since $\alpha_T = 1 - \varepsilon_T$ and $\varepsilon_T \to 0$,

$$\log \alpha_T = \log(1 - \varepsilon_T) = -\varepsilon_T + O(\varepsilon_T^2).$$

The logarithmic identity implies $\varepsilon_T = O((\log T)/T)$, so

$$(2T-1)O(\varepsilon_T^2) = O((\log T)^2/T) = o(1).$$

Therefore

$$(2T-1)\varepsilon_T = (\rho+1)\log T + O(1),$$

which is (17). Since $\eta_{\infty,T} = 2 - \varepsilon_T$, (16) follows. ■

Corollary 12 *The infinite-width loss at the stepsize $\eta_{\infty,T}$ is*

$$L_{\infty,T}(\eta_{\infty,T}) = \frac{1}{2}\sum_{k\geq 1} k^{-\beta}(1 - \eta_{\infty,T}k^{-\alpha})^{2T} \asymp T^{-\rho}.$$

Proof The $k = 1$ summand is

$$\frac{1}{2}(1 - \eta_{\infty,T})^{2T} = \frac{1}{2}\alpha_T^{2T} \asymp T^{-\rho}.$$

By Lemma 8, the tail sum is

$$\sum_{k\geq 2} k^{-\beta}(1 - \eta_{\infty,T}k^{-\alpha})^{2T} = W_\beta(2T; \eta_{\infty,T})$$
$$\leq W_\beta(2T; 1) = \sum_{k\geq 2} k^{-\beta}(1 - k^{-\alpha})^{2T} \asymp T^{-\rho}.$$

■

Appendix G. A crude growing-horizon optimizer-transfer bound

Theorem 1 is deliberately stated with T fixed. When T grows, the derivative of the limiting score at the root becomes small because the root approaches the stability edge.

Recall that we denote

$$F_{n,T}(\eta) = -\frac{1}{T}\frac{d}{d\eta}L_{n,T}(\eta) = \left\langle b, A_n(I-\eta A_n)^{2T-1}b\right\rangle, \tag{18}$$

and $F_T := F_{\infty,T}$.

Lemma 13 (Crude growing-horizon perturbation) *Let $T = T_n$ and suppose $T/\sqrt{n} \to 0$. Then, uniformly over $\eta \in [1,2]$,*

$$|F_{n,T}(\eta) - F_T(\eta)| = O_p(Tn^{-1/2}), \qquad |F'_{n,T}(\eta) - F'_T(\eta)| = O_p(T^2 n^{-1/2}).$$

Proof The point that must be checked when T grows is stability of the powers. Let

$$\Delta_n := \|A_n - \Sigma\|_{\mathrm{op}}.$$

By Lemma 7, $\Delta_n = O_p(n^{-1/2})$, so the event

$$\mathcal{E}_n := \{T\Delta_n \le 1\}$$

has probability tending to one. On $\mathcal{E}_n$, the spectra of A_n and Σ are contained in $[0, 1+1/T]$. Hence, for $\eta \in [1,2]$,

$$\|I-\eta A_n\|_{\mathrm{op}} \le 1 + \frac{2}{T}, \qquad \|I-\eta\Sigma\|_{\mathrm{op}} \le 1,$$

and therefore all powers up to order $2T$ have operator norm bounded by an absolute constant.

Using the telescoping identity for powers,

$$\begin{aligned}(I-\eta A_n)^m - (I-\eta\Sigma)^m \\ = -\eta\sum_{j=0}^{m-1}(I-\eta A_n)^j(A_n-\Sigma)(I-\eta\Sigma)^{m-1-j},\end{aligned}$$

we get, for every $m \le 2T$,

$$\sup_{\eta\in[1,2]}\|(I-\eta A_n)^m - (I-\eta\Sigma)^m\|_{\mathrm{op}} \le CT\Delta_n.$$

Now write

$$\begin{aligned}F_{n,T}(\eta) - F_T(\eta) = \left\langle b, (A_n-\Sigma)(I-\eta A_n)^{2T-1}b\right\rangle \\ + \left\langle b, \Sigma\{(I-\eta A_n)^{2T-1} - (I-\eta\Sigma)^{2T-1}\}b\right\rangle.\end{aligned}$$

The first term is $O(\Delta_n)$ and the second is $O(T\Delta_n)$ on $\mathcal{E}_n$. Thus

$$\sup_{\eta\in[1,2]}|F_{n,T}(\eta) - F_T(\eta)| \le CT\Delta_n = O_p(Tn^{-1/2}).$$

For the derivative, use the exact formulas

$$F'_{n,T}(\eta) = -(2T-1)\left\langle b, A_n^2(I-\eta A_n)^{2T-2}b\right\rangle,$$
$$F'_T(\eta) = -(2T-1)\left\langle b, \Sigma^2(I-\eta\Sigma)^{2T-2}b\right\rangle.$$

On $\mathcal{E}_n$,

$$\left\|A_n^2-\Sigma^2\right\|_{\mathrm{op}} \le \|A_n-\Sigma\|_{\mathrm{op}}\|A_n\|_{\mathrm{op}} + \|\Sigma\|_{\mathrm{op}}\|A_n-\Sigma\|_{\mathrm{op}} \le C\Delta_n,$$
$$\sup_{\eta\in[1,2]}\left\|(I-\eta A_n)^{2T-2}-(I-\eta\Sigma)^{2T-2}\right\|_{\mathrm{op}} \le CT\Delta_n.$$

Consequently,

$$\sup_{\eta\in[1,2]}|F'_{n,T}(\eta)-F'_T(\eta)| \le CT\left(\Delta_n+T\Delta_n\right) \le CT^2\Delta_n = O_p(T^2n^{-1/2}).$$

■

Lemma 14 (Local deterministic curvature) *Let $r_T = o(1/T)$. Then*

$$\sup_{|\eta-\eta_{\infty,T}|\le r_T}|F'_T(\eta)-F'_T(\eta_{\infty,T})| = o(T^{-\rho}). \tag{19}$$

In particular, there are constants $0<c<C<\infty$ such that, for all large T,

$$-CT^{-\rho} \le F'_T(\eta) \le -cT^{-\rho} \qquad \textit{whenever } |\eta-\eta_{\infty,T}|\le r_T. \tag{20}$$

Proof First compute the derivative at the root:

$$F'_T(\eta) = -(2T-1)\sum_{k\ge1}k^{-(\beta+2\alpha)}(1-\eta k^{-\alpha})^{2T-2}.$$

At $\eta=\eta_{\infty,T}$, the $k=1$ term has size

$$(2T-1)\alpha_T^{2T-2} \asymp T\cdot T^{-(\rho+1)} = T^{-\rho}.$$

The tail is smaller because Lemma 8 with $u=\beta+2\alpha$ gives

$$T\sum_{k\ge2}k^{-(\beta+2\alpha)}(1-\eta_{\infty,T}k^{-\alpha})^{2T-2} = O(T\cdot T^{-(\rho+2)}) = O(T^{-(\rho+1)}).$$

Thus $F'_T(\eta_{\infty,T}) \asymp -T^{-\rho}$.

It remains to show that this derivative does not change on a radius $o(1/T)$. Since $\eta_{\infty,T} = 2-((\rho+1)\log T+O(1))/(2T)$, the interval $|\eta-\eta_{\infty,T}|\le r_T$ is contained in $[1,2]$ for all large T. On this interval, write $\alpha_T = \eta_{\infty,T}-1$. Then $\alpha_T\to1$ and, for all large T,

$$|1-\eta| \le \alpha_T + r_T, \qquad \frac{r_T}{\alpha_T} \le \frac{1}{2}.$$

Hence

$$\begin{aligned}|1-\eta|^{2T-3} &\le (\alpha_T + r_T)^{2T-3} \\ &= \alpha_T^{2T-3}\left(1+\frac{r_T}{\alpha_T}\right)^{2T-3} \\ &\le \alpha_T^{2T-3}\exp(CTr_T) = O(T^{-(\rho+1)}),\end{aligned}$$

because $Tr_T = o(1)$ and Lemma 11 gives $\alpha_T^{2T-3} \asymp T^{-(\rho+1)}$. For $k \ge 2$, Lemma 8 with $u = \beta + 3\alpha$ gives the tail bound below. Therefore

$$\begin{aligned}\sup_{|\eta-\eta_{\infty,T}|\le r_T} |F_T''(\eta)| &\le CT^2\left\{T^{-(\rho+1)} + \sum_{k\ge 2} k^{-(\beta+3\alpha)}(1-\eta k^{-\alpha})^{2T-3}\right\} \\ &\le CT^2\left\{T^{-(\rho+1)} + T^{-(\rho+3)}\right\} \le CT^{1-\rho}.\end{aligned}$$

The mean-value theorem gives

$$\sup_{|\eta-\eta_{\infty,T}|\le r_T} |F_T'(\eta) - F_T'(\eta_{\infty,T})| \le Cr_T T^{1-\rho} = o(T^{-\rho}),$$

which proves (19) and then (20). ■

Proposition 15 (Crude optimizer transfer for growing horizons) *Let $\widehat{\eta}_{n,T}$ be any constrained minimizer over $[1,2]$,*

$$\widehat{\eta}_{n,T} \in \arg\min_{\eta\in[1,2]} L_{n,T}(\eta).$$

If

$$T^{\rho+2}/\sqrt{n} \to 0, \tag{21}$$

then, with probability tending to one, $\widehat{\eta}_{n,T}$ is an interior root of $F_{n,T}$ and

$$\widehat{\eta}_{n,T} - \eta_{\infty,T} = O_p(T^{\rho+1}n^{-1/2}). \tag{22}$$

Proof Let

$$r_{n,T}(M) := MT^{\rho+1}n^{-1/2},$$

where $M > 0$ is fixed for the moment. Condition (21) gives

$$Tr_{n,T}(M) = MT^{\rho+2}n^{-1/2} \to 0.$$

Since $2 - \eta_{\infty,T} \asymp (\log T)/T$, the interval

$$I_{n,T}(M) := [\eta_{\infty,T} - r_{n,T}(M), \eta_{\infty,T} + r_{n,T}(M)]$$

is contained in $(1,2)$ for all large n.

By Lemma 14, uniformly on $I_{n,T}(M)$,

$$F_T'(\eta) \le -cT^{-\rho}.$$

Because $F_T(\eta_{\infty,T}) = 0$, this gives the deterministic sign bounds

$$F_T(\eta_{\infty,T} - r_{n,T}(M)) \geq cMTn^{-1/2}, \tag{23}$$
$$F_T(\eta_{\infty,T} + r_{n,T}(M)) \leq -cMTn^{-1/2}. \tag{24}$$

Lemma 13 gives

$$\sup_{\eta\in[1,2]} |F_{n,T}(\eta) - F_T(\eta)| = O_p(Tn^{-1/2}).$$

Thus, for every $\epsilon > 0$, there is $A < \infty$ such that with probability at least $1 - \epsilon$ the last supremum is at most $ATn^{-1/2}$. Choose $M > 2A/c$. On that event, (23)–(24) imply

$$F_{n,T}(\eta_{\infty,T} - r_{n,T}(M)) > 0, \qquad F_{n,T}(\eta_{\infty,T} + r_{n,T}(M)) < 0.$$

By continuity, $F_{n,T}$ has a zero in $I_{n,T}(M)$. Moreover $L_{n,T}$ is convex in η and $-TF_{n,T} = L'_{n,T}$; equivalently, $F_{n,T}$ is nonincreasing. The sign pattern above makes $L'_{n,T}$ negative at the left endpoint of $I_{n,T}(M)$ and positive at the right endpoint, so the constrained minimizer over $[1, 2]$ is interior and lies in $I_{n,T}(M)$ with probability tending to one. Therefore

$$|\widehat{\eta}_{n,T} - \eta_{\infty,T}| \leq MT^{\rho+1}n^{-1/2}$$

with probability at least $1 - \epsilon$ for all large n. Since $\epsilon > 0$ was arbitrary, this is (22). ■

Proposition 16 (Transfer-gap upper bounds) *Under the assumptions of Proposition 15,*

$$c_{n,T} := L_{\infty,T}(\widehat{\eta}_{n,T}) - L_{\infty,T}(\eta_{\infty,T}) = O_p(T^{\rho+3}/n), \tag{25}$$
$$\widetilde{c}_{n,T} := L_{n,T}(\eta_{\infty,T}) - L_{n,T}(\widehat{\eta}_{n,T}) = O_p(T^{\rho+3}/n). \tag{26}$$

Proof Fix $\epsilon > 0$ and choose M as in the proof of Proposition 15. With probability at least $1 - \epsilon$ for all large n,

$$|\widehat{\eta}_{n,T} - \eta_{\infty,T}| \leq r_{n,T}(M), \qquad Tr_{n,T}(M) \to 0. \tag{27}$$

We first prove (25). Taylor's expansion around $\eta_{\infty,T}$ gives, for some $\xi_{n,T}$ between $\widehat{\eta}_{n,T}$ and $\eta_{\infty,T}$,

$$L_{\infty,T}(\widehat{\eta}_{n,T}) - L_{\infty,T}(\eta_{\infty,T}) = \frac{1}{2}L''_{\infty,T}(\xi_{n,T})(\widehat{\eta}_{n,T} - \eta_{\infty,T})^2,$$

because $L'_{\infty,T}(\eta_{\infty,T}) = 0$. On the event (27), $|\xi_{n,T} - \eta_{\infty,T}| \leq r_{n,T}(M) = o(1/T)$, so Lemma 14 implies

$$L''_{\infty,T}(\xi_{n,T}) = -TF'_T(\xi_{n,T}) = O(T^{1-\rho}).$$

Combining this curvature bound with Proposition 15 yields

$$c_{n,T} = O_p\left(T^{1-\rho}(T^{\rho+1}n^{-1/2})^2\right) = O_p(T^{\rho+3}/n).$$

It remains to prove the corresponding finite-width reverse gap. Taylor's theorem at the finite-width minimizer gives, for some $\zeta_{n,T}$ between $\widehat{\eta}_{n,T}$ and $\eta_{\infty,T}$,

$$L_{n,T}(\eta_{\infty,T}) - L_{n,T}(\widehat{\eta}_{n,T}) = \frac{1}{2}L''_{n,T}(\zeta_{n,T})(\eta_{\infty,T} - \widehat{\eta}_{n,T})^2. \tag{28}$$

On (27), the point $\zeta_{n,T}$ also satisfies $|\zeta_{n,T} - \eta_{\infty,T}| \le r_{n,T}(M) = o(1/T)$. Since $L''_{n,T}(\eta) = -TF'_{n,T}(\eta)$,

$$|L''_{n,T}(\zeta_{n,T})| \le T|F'_T(\zeta_{n,T})| + T|F'_{n,T}(\zeta_{n,T}) - F'_T(\zeta_{n,T})|.$$

The first term is $O(T^{1-\rho})$ by Lemma 14. The second term is

$$T \sup_{\eta\in[1,2]} |F'_{n,T}(\eta) - F'_T(\eta)| = O_p(T^3 n^{-1/2})$$

by Lemma 13. The root condition $T^{\rho+2}/\sqrt{n} \to 0$ implies $T^3 n^{-1/2} = o(T^{1-\rho})$, and hence this second term is $o_p(T^{1-\rho})$. Therefore

$$L''_{n,T}(\zeta_{n,T}) = O_p(T^{1-\rho}).$$

Substituting this bound and Proposition 15 into (28) gives

$$\widetilde{c}_{n,T} = O_p\left(T^{1-\rho}(T^{\rho+1}n^{-1/2})^2\right) = O_p(T^{\rho+3}/n),$$

which proves (26). ∎

Appendix H. Bias at the infinite-width learning rate

Throughout this section fix $\eta \in [1, 2]$ and write

$$D_\eta := I - \eta\Sigma, \qquad E_n := A_n - \Sigma, \qquad B_{\eta,n} := -\eta E_n.$$

Then

$$I - \eta A_n = D_\eta + B_{\eta,n}.$$

Let

$$\overline{L}_{n,T}(\eta) := \mathbb{E}L_{n,T}(\eta).$$

H.1. The noncommutative word expansion

Lemma 17 (Word expansion) *For $q = 1, \dots, 2T$, define*

$$W_q(\eta) := \sum_{\substack{a_0,\dots,a_q\ge 0 \\ a_0+\cdots+a_q=2T-q}} \mathbb{E}\left\langle b, D_\eta^{a_0} B_{\eta,n} D_\eta^{a_1} B_{\eta,n} \cdots B_{\eta,n} D_\eta^{a_q} b\right\rangle.$$

Then

$$\overline{L}_{n,T}(\eta) - L_{\infty,T}(\eta) = \frac{1}{2}\sum_{q=1}^{2T} W_q(\eta). \tag{29}$$

Proof Expanding $(D_\eta + B_{\eta,n})^{2T}$ in noncommuting words, choose the q positions occupied by $B_{\eta,n}$. Between successive $B_{\eta,n}$'s there are powers of D_η. If the gaps are $a_0, \dots, a_q \geq 0$, then the corresponding word is

$$D_\eta^{a_0} B_{\eta,n} D_\eta^{a_1} B_{\eta,n} \cdots B_{\eta,n} D_\eta^{a_q}.$$

The total number of letters is

$$a_0 + \cdots + a_q + q = 2T,$$

so $a_0 + \cdots + a_q = 2T - q$. Every word with q copies of $B_{\eta,n}$ occurs exactly once in this way. The term with $q = 0$ is D_η^{2T} and gives $L_{\infty,T}$. Taking the inner product with b, multiplying by $1/2$, and taking expectation gives (29). ■

H.2. A Gaussian identity

Lemma 18 (Quadratic Gaussian identity) *Let $z \sim N(0, \Sigma)$ and $Y := z \otimes z - \Sigma$. For every bounded operator H,*

$$\mathbb{E}[YHY] = \Sigma H \Sigma + \operatorname{tr}(\Sigma H)\Sigma.$$

Consequently,

$$\mathbb{E}[E_n H E_n] = \frac{1}{n} \left(\Sigma H \Sigma + \operatorname{tr}(\Sigma H)\Sigma\right). \tag{30}$$

Proof For vectors u, v write $u \otimes v$ for the rank-one map $x \mapsto u \langle v, x \rangle$. Since

$$(z \otimes z) H (z \otimes z) = \langle z, Hz \rangle \; z \otimes z,$$

Wick formula gives

$$\mathbb{E}\big[\langle z, Hz \rangle \; z \otimes z\big] = \operatorname{tr}(\Sigma H)\Sigma + 2\Sigma H \Sigma.$$

Moreover

$$\mathbb{E}[(z \otimes z) H \Sigma] = \Sigma H \Sigma, \qquad \mathbb{E}[\Sigma H (z \otimes z)] = \Sigma H \Sigma.$$

Therefore

$$\begin{aligned} \mathbb{E}[YHY] &= \mathbb{E}[(z \otimes z) H (z \otimes z)] - \mathbb{E}[(z \otimes z) H \Sigma] - \mathbb{E}[\Sigma H (z \otimes z)] + \Sigma H \Sigma \\ &= \operatorname{tr}(\Sigma H)\Sigma + 2\Sigma H \Sigma - \Sigma H \Sigma - \Sigma H \Sigma + \Sigma H \Sigma \\ &= \operatorname{tr}(\Sigma H)\Sigma + \Sigma H \Sigma. \end{aligned}$$

For (30), write $E_n = n^{-1} \sum_{r=1}^n Y_r$ with independent copies Y_r of Y. All cross terms vanish because $\mathbb{E} Y_r = 0$, so

$$\mathbb{E}[E_n H E_n] = \frac{1}{n^2} \sum_{r=1}^n \mathbb{E}[Y_r H Y_r] = \frac{1}{n} \mathbb{E}[YHY],$$

and we conclude using the above identity. ■

H.3. The exact second-order term

Let

$$r_k(\eta) := 1 - \eta k^{-a}.$$

At the infinite-width root $\eta_{\infty,T}$ we also use

$$\alpha_T := \eta_{\infty,T} - 1, \qquad r_1(\eta_{\infty,T}) = -\alpha_T.$$

Lemma 19 (Exact formula for W_2) *Let $M := 2T - 2$. For every $\eta \in [1, 2]$,*

$$W_2(\eta) = \frac{\eta^2}{n} Q_T(\eta),$$

where

$$Q_T(\eta) := \binom{2T}{2} \sum_{k\geq 1} k^{-\beta-2a} r_k(\eta)^M \tag{31}$$

$$+ \sum_{i=0}^{M} (i+1) \left(\sum_{j\geq 1} j^{-a} r_j(\eta)^{M-i} \right) \left(\sum_{k\geq 1} k^{-\beta-a} r_k(\eta)^i \right). \tag{32}$$

Proof Because $B_{\eta,n} = -\eta E_n$,

$$W_2(\eta) = \eta^2 \sum_{a+b+c=M} \mathbb{E} \left\langle b, D_\eta^a E_n D_\eta^b E_n D_\eta^c b \right\rangle.$$

Apply Lemma 18 with $H = D_\eta^b$:

$$\mathbb{E}[E_n D_\eta^b E_n] = \frac{1}{n} \left(\Sigma D_\eta^b \Sigma + \operatorname{tr}(\Sigma D_\eta^b) \Sigma \right).$$

Since D_η and Σ commute, the first part gives

$$\begin{aligned} \sum_{a+b+c=M} \left\langle b, D_\eta^a \Sigma D_\eta^b \Sigma D_\eta^c b \right\rangle &= \sum_{a+b+c=M} \left\langle b, \Sigma^2 D_\eta^M b \right\rangle \\ &= \binom{M+2}{2} \left\langle b, \Sigma^2 D_\eta^M b \right\rangle. \end{aligned}$$

Because $M + 2 = 2T$, this equals (31). For the trace part, fix the middle exponent $b_0 \in \{0, \ldots, M\}$. There are $M - b_0 + 1$ pairs (a, c) with $a + c = M - b_0$, and each gives

$$\operatorname{tr}(\Sigma D_\eta^{b_0}) \left\langle b, \Sigma D_\eta^{M-b_0} b \right\rangle.$$

Set $i = M - b_0$. Then i runs from 0 to M, and $M - b_0 + 1 = i + 1$. Thus the trace part is

$$\sum_{i=0}^{M} (i+1) \operatorname{tr}(\Sigma D_\eta^{M-i}) \left\langle b, \Sigma D_\eta^i b \right\rangle,$$

which is exactly (32) after writing both terms in the eigenbasis. ∎

H.4. Higher-order bias terms

The following estimates are used only to control the non-leading words in the expansion. We keep the constants explicit enough to show that the combinatorial factors grow at most exponentially in the word length; this is important when T is allowed to grow.

For an index tuple $\mathbf{i} = (i_1, \dots, i_q)$, let $\Pi(\mathbf{i})$ be the partition of $\{1, \dots, q\}$ into level sets of equal indices, and write $m_B := |B|$ for $B \in \Pi(\mathbf{i})$. Define the multiplicity weight

$$\mathfrak{m}(\mathbf{i}) := \prod_{B \in \Pi(\mathbf{i})} m_B^{m_B}.$$

We use only the following coarse moment estimate. The estimate deliberately does not try to decide which powers of D_η stay on the open chain between the two endpoint vectors. After Wick expansion, some powers can lie in closed coordinate sums; for the high-word remainders below it is enough to use the contraction bound $\|D_\eta\|_{\mathrm{op}} \le 1$ and the fact that each connected coordinate sum contains at least one covariance weight.

Lemma 20 (Coarse covariance-word moment bound) *There is a constant $C < \infty$ such that the following holds. Fix $q \ge 1$, $\eta \in [1, 2]$, nonnegative integers $a_0, \dots, a_q$, and an index tuple $\mathbf{i} = (i_1, \dots, i_q)$. Put*

$$\mathcal{W}_{\mathbf{i},\mathbf{a}}(\eta) := \left\langle b, D_\eta^{a_0} Y_{i_1} D_\eta^{a_1} Y_{i_2} \cdots Y_{i_q} D_\eta^{a_q} b \right\rangle, \qquad Y_i = z_i \otimes z_i - \Sigma.$$

If one block of $\Pi(\mathbf{i})$ has size one, then $\mathbb{E}\mathcal{W}_{\mathbf{i},\mathbf{a}}(\eta) = 0$. If no block is a singleton, then

$$|\mathbb{E}\mathcal{W}_{\mathbf{i},\mathbf{a}}(\eta)| \le C^q \mathfrak{m}(\mathbf{i}). \tag{33}$$

For a product of two scalar words, let $\mathbf{i}, \mathbf{j} \in \{1, \dots, n\}^q$ be the two sample-label tuples and let $\mathbf{a}, \mathbf{c}$ be the two exponent tuples. Let $\mathbf{h} \in \{1, \dots, n\}^{2q}$ be the concatenation of $\mathbf{i}$ and $\mathbf{j}$. If a label appears exactly once in $\mathbf{h}$, then the expectation of the product is zero; otherwise,

$$\left|\mathbb{E}\left[\mathcal{W}_{\mathbf{i},\mathbf{a}}(\eta)\mathcal{W}_{\mathbf{j},\mathbf{c}}(\eta)\right]\right| \le (C)^{2q} \mathfrak{m}(\mathbf{h}). \tag{34}$$

Proof We first prove the single-word bound. The singleton statement follows from conditioning. If a sample label r occurs only once, condition on all z_ℓ with $\ell \ne r$. The factors around the unique occurrence of Y_r are then deterministic bounded operators, and the remaining conditional expectation has the form

$$\mathbb{E}[\langle u, Y_r v\rangle \mid z_\ell, \ell \ne r] = \langle u, (\mathbb{E}Y_r)v\rangle = 0.$$

It remains to bound the tuples with no singleton label. We first argue after projecting to $\mathrm{span}\{e_1, \dots, e_d\}$. The constants obtained below are independent of d, and the original infinite-dimensional statement follows by letting $d \to \infty$: each fixed contraction pattern is bounded by an absolutely summable expression displayed below, and the truncated scalar words converge in L^1 to the original scalar word.

In the eigenbasis, write $\lambda_k = k^{-a}$ and $r_k(\eta) = 1 - \eta\lambda_k$. Then

$$(D_\eta)_{kk} = r_k(\eta), \qquad (Y_i)_{k\ell} = z_{i,k} z_{i,\ell} - \lambda_k \mathbf{1}_{k=\ell},$$

and hence

$$\mathcal{W}_{\mathbf{i},\mathbf{a}}(\eta) = \sum_{k_0,\dots,k_q} b_{k_0} r_{k_0}^{a_0} (Y_{i_1})_{k_0 k_1} r_{k_1}^{a_1} \cdots (Y_{i_q})_{k_{q-1} k_q} r_{k_q}^{a_q} b_{k_q}.$$

Expand each factor Y_i as $z_i \otimes z_i - \Sigma$. For each position there are two choices, the random tensor or the deterministic covariance term. Fix these choices for the moment. If a block $B \in \Pi(\mathbf{i})$ has m_B occurrences and s_B of them are chosen as random tensors, then those occurrences contribute $2s_B$ Gaussian coordinates of the same sample vector. Using Isserlis' formula, we can decompose their expectation into a sum over pairings of these $2s_B$ coordinates. The number of pairings is at most

$$(2s_B - 1)!! \le (Cm_B)^{m_B},$$

with one pattern when $s_B = 0$. Multiplying over the blocks and absorbing the 2^q deterministic or random choices gives at most

$$C^q \prod_{B \in \Pi(\mathbf{i})} m_B^{m_B} = C^q \mathfrak{m}(\mathbf{i}) \tag{35}$$

terms to bound.

Fix one such term. The deterministic covariance terms and the Wick pairings impose equalities among the coordinate variables $k_0, \dots, k_q$. We group the coordinate variables into connected classes under these equalities. Each deterministic covariance term contributes one factor λ_k to the class it connects, and each Wick pairing also contributes one covariance factor λ_k to the class produced by the pairing. Therefore every connected class contains at least one covariance weight. This is the only point needed to handle closed sums correctly: powers of D_η may lie in any connected class, but on $\eta \in [1, 2]$ we have $|r_k(\eta)| \le 1$.

Consider one connected class. Suppose it contains ν endpoint factors b, where $\nu \in \{0, 1, 2\}$ for a single word, and suppose it contains $p \ge 1$ covariance weights. Its absolute coordinate sum is bounded by

$$\sum_{k \ge 1} |b_k|^\nu \lambda_k^p \le \sum_{k \ge 1} |b_k|^\nu \lambda_k.$$

The right-hand side is finite uniformly in the exponents. Indeed,

$$\begin{aligned} \sum_{k \ge 1} \lambda_k &= \operatorname{tr} \Sigma < \infty, \\ \sum_{k \ge 1} |b_k| \lambda_k &\le \|b\| \, \operatorname{tr}(\Sigma^2)^{1/2} < \infty, \\ \sum_{k \ge 1} |b_k|^2 \lambda_k &= \langle b, \Sigma b \rangle < \infty. \end{aligned} \tag{36}$$

Since the number of connected classes in one term is at most $q + 1$, the whole term is bounded by a constant of the form C^q, depending only on Σ and b. Combining this with the count in (35), and increasing C once, proves (33).

The product bound is identical, applied to the two coordinate expansions at the same time. If a sample label appears exactly once in the combined tuple $\mathbf{h}$, conditioning on all other sample vectors again leaves one centered factor Y_r and gives zero. Otherwise, after expanding all $2q$ centered

covariance factors and applying Isserlis' formula block by block, the number of contraction terms is at most $C^{2q}\mathfrak{m}(\mathbf{h})$. The connected coordinate classes now come from two open chains rather than one, so a class can contain $\nu = 0, 1, 2, 3$, or 4 endpoint factors. The cases $\nu = 0, 1, 2$ were bounded in (36); for the remaining two cases we use $b \in \ell^2 \subset \ell^\infty$:

$$\sum_{k\geq 1} |b_k|^3 \lambda_k \leq \|b\|_\infty \langle b, \Sigma b\rangle < \infty,$$
$$\sum_{k\geq 1} |b_k|^4 \lambda_k \leq \|b\|_\infty^2 \langle b, \Sigma b\rangle < \infty.$$

Thus each contraction term in the product is again bounded by C^{2q}, and the pattern count gives (34). ■

Lemma 21 (No-singleton summation) *For a constant $C < \infty$ and every $q \geq 1$,*

$$\sum_{\substack{\mathbf{i}\in\{1,\dots,n\}^q:\\ \text{no value occurs exactly once}}} \mathfrak{m}(\mathbf{i}) \leq (Cq)^q n^{\lfloor q/2\rfloor}. \tag{37}$$

Proof Group tuples by their level-set partition. If the partition has block sizes $m_1, \dots, m_r \geq 2$ with $m_1 + \cdots + m_r = q$, then $r \leq \lfloor q/2\rfloor$ and the labels can be chosen in at most $n^r \leq n^{\lfloor q/2\rfloor}$ ways. The total contribution of all partitions with these ordered block sizes is bounded by

$$\frac{q!}{m_1! \cdots m_r!} \prod_{j=1}^r m_j^{m_j} \leq q! \prod_{j=1}^r e^{m_j} \leq e^q q!,$$

where we used $m_j! \geq (m_j/e)^{m_j}$. The number of ordered compositions of q is at most 2^q. Thus the sum of the weights over all no-singleton partitions is at most $C^q q! \leq (Cq)^q$. Multiplying by $n^{\lfloor q/2\rfloor}$ proves (37). ■

Lemma 22 (Higher-order expected-word bound) *For every $q \geq 1$ and every $\eta \in [1, 2]$,*

$$|W_q(\eta)| \leq (Cq)^q \binom{2T}{q} n^{-\lceil q/2\rceil}. \tag{38}$$

Proof Write

$$B_{\eta,n} = -\eta E_n = -\frac{\eta}{n} \sum_{i=1}^n Y_i.$$

Fix a composition $a_0 + \cdots + a_q = 2T - q$. Expanding the q copies of E_n gives a factor n^{-q} and a sum over $\mathbf{i} \in \{1, \dots, n\}^q$. Tuples with a singleton index vanish by Lemma 20. For the remaining tuples, the coarse bound (33) and Lemma 21 give

$$n^{-q} \sum_{\substack{\mathbf{i}:\\ \text{no singleton}}} \left|\mathbb{E}\left\langle b, D_\eta^{a_0} Y_{i_1} D_\eta^{a_1} \cdots Y_{i_q} D_\eta^{a_q} b\right\rangle\right|$$
$$\leq C^q n^{-q} \sum_{\substack{\mathbf{i}:\\ \text{no singleton}}} \mathfrak{m}(\mathbf{i}) \leq (Cq)^q n^{-\lceil q/2\rceil}.$$

The factor $\eta^q \le 2^q$ is absorbed into the constant. Finally, the number of compositions is $\binom{2T}{q}$, proving (38). ∎

H.5. The expected high-word remainder

The bias computation needs only the deterministic part of the high-word tail. At the infinite-width root, define

$$R_{n,T}^{\text{bias}} := \left| \sum_{q=3}^{2T} W_q(\eta_{\infty,T}) \right|.$$

We use the expected-word condition

$$R_{n,T}^{\text{bias}} = O\left(\frac{T^4}{n^2}\right). \tag{39}$$

This condition is the deterministic high-word remainder needed to finish the bias expansion. The following lemma verifies it from the simple scale separation that will be used later in the growing-horizon theorem.

Lemma 23 (Expected high-word condition from T^2/n) *If $T^2/n \to 0$, then (39) holds.*

Proof Lemma 22 and the composition estimate $\binom{2T}{q} \le (2eT/q)^q$ give, uniformly for $q \ge 3$,

$$|W_q(\eta_{\infty,T})| \le (CT)^q n^{-\lceil q/2 \rceil}.$$

Let $\delta_n = CT^2/n$. Since $T^2/n \to 0$, after increasing n we may assume $\delta_n \le 1/2$ and $T/n \le 1$.

For even $q = 2r \ge 4$,

$$(CT)^{2r} n^{-r} = \frac{(CT)^4}{n^2} \left(\frac{C^2T^2}{n}\right)^{r-2} \le C\frac{T^4}{n^2}\delta_n^{r-2}.$$

Summing over $r \ge 2$ gives a contribution $O(T^4/n^2)$. For odd $q = 2r+1 \ge 3$, the case $q = 3$ contributes at most $CT^3/n^2 \le CT^4/n^2$. For $r \ge 2$,

$$(CT)^{2r+1} n^{-(r+1)} = \frac{(CT)^5}{n^3} \left(\frac{C^2T^2}{n}\right)^{r-2} \le C\frac{T^4}{n^2}\left(\frac{T}{n}\right)\delta_n^{r-2} \le C\frac{T^4}{n^2}\delta_n^{r-2}.$$

Summing over $r \ge 2$ and adding the $q = 3$ term proves (39). ∎

H.6. Precise asymptotics of the second-order term

Proposition 24 gives the precise second-order deterministic bias coefficient used in Theorem 3.

Proposition 24 (Asymptotics of $Q_T(\eta_{\infty,T})$) *Let $\eta_{\infty,T}$ be the infinite-width root at horizon T. Then*

$$Q_T(\eta_{\infty,T}) \asymp T^{1-\rho} + T^{-s}, \qquad s = \frac{\alpha - 1}{\alpha}. \tag{40}$$

Equivalently,

$$Q_T(\eta_{\infty,T}) \asymp \begin{cases} T^{1-\rho}, & \beta < 2\alpha, \\ T^{-s}, & \beta > 2\alpha, \end{cases} \tag{41}$$

with the two powers coinciding at the borderline $\beta = 2\alpha$.

Proof Set $M := 2T - 2$ and, in this proof, write r_k for $r_k(\eta_{\infty,T})$. Thus $r_1 = -\alpha_T$, while for $k \geq 2$,

$$r_k = 1 - \eta_{\infty,T} k^{-\alpha} \in (0, 1).$$

Introduce the finite geometric derivative kernel

$$\mathcal{B}_M(x, y) := \sum_{i=0}^{M} (i+1) x^{M-i} y^i. \tag{42}$$

If $x \neq y$, then

$$\mathcal{B}_M(x, y) = \frac{x^{M+2} - (M+2)xy^{M+1} + (M+1)y^{M+2}}{(x-y)^2}, \tag{43}$$

and if $x = y$, then

$$\mathcal{B}_M(x, x) = \binom{M+2}{2} x^M.$$

Formula (43) follows by differentiating the finite geometric sum $\sum_{i=0}^{M+1} x^{M+1-i} y^i$ with respect to y.

Using Lemma 19, we can rewrite $Q_T(\eta_{\infty,T})$ as

$$Q_T(\eta_{\infty,T}) = \binom{M+2}{2} \sum_{k \geq 1} k^{-\beta-2\alpha} r_k^M + \sum_{j,k \geq 1} j^{-\alpha} k^{-\beta-\alpha} \mathcal{B}_M(r_j, r_k). \tag{44}$$

We now prove matching upper and lower bounds.

First, the contribution of $k = 1$ to the first sum and of $(j, k) = (1, 1)$ to the double sum is

$$2\binom{M+2}{2} \alpha_T^M.$$

By Lemma 11,

$$\alpha_T^M = \alpha_T^{2T-2} \asymp T^{-(\rho+1)}.$$

Hence this contribution is of order

$$T^2 T^{-(\rho+1)} = T^{1-\rho}. \tag{45}$$

This gives the lower bound $Q_T(\eta_{\infty,T}) \gtrsim T^{1-\rho}$.

Second, consider the part of the double sum with $k = 1$ and $j \geq 2$. Since M is even, the closed form gives

$$\mathcal{B}_M(r_j, -\alpha_T) = \frac{r_j^{M+2} + (M+2) r_j \alpha_T^{M+1} + (M+1)\alpha_T^{M+2}}{(r_j + \alpha_T)^2} \geq c\, r_j^{M+2},$$

because $r_j + \alpha_T = \eta_{\infty,T}(1 - j^{-\alpha})$ is bounded above and below by positive constants for $j \geq 2$. Therefore

$$Q_T(\eta_{\infty,T}) \geq c \sum_{j \geq 2} j^{-\alpha} r_j^{M+2}.$$

By Lemma 8 with $u = \alpha$,

$$\sum_{j \geq 2} j^{-\alpha} r_j^{M+2} \asymp T^{-(\alpha-1)/\alpha} = T^{-s}.$$

This gives the lower bound $Q_T(\eta_{\infty,T}) \gtrsim T^{-s}$.

For the upper bound, decompose (44) according to whether each index is 1 or at least 2. The first single sum is bounded by

$$\begin{aligned}\binom{M+2}{2} \sum_{k \geq 1} k^{-\beta-2a} |r_k|^M &\leq CT^2 \alpha_T^M + CT^2 \sum_{k \geq 2} k^{-\beta-2a} r_k^M \\ &\leq CT^{1-\rho} + CT^2 T^{-(\rho+2)} \\ &\leq CT^{1-\rho} + CT^{-\rho},\end{aligned}$$

where Lemma 8 was used with $u = \beta + 2a$. The $(1,1)$ part of the double sum is the same order as (45).

For the part $(j,k) = (1,k)$, $k \geq 2$, formula (43) gives

$$\mathcal{B}_M(-\alpha_T, r_k) \leq C \left(\alpha_T^{M+2} + T r_k^{M+1} + T r_k^{M+2} \right).$$

Multiplying by $k^{-\beta-a}$ and summing over $k \geq 2$ yields at most

$$C\alpha_T^M + CT \sum_{k \geq 2} k^{-\beta-a} r_k^M \leq CT^{-(\rho+1)} + CT\, T^{-(\rho+1)} \leq CT^{-\rho}.$$

The part $(j,k) = (j,1)$, $j \geq 2$, is bounded similarly by

$$\begin{aligned}\sum_{j \geq 2} j^{-a} \mathcal{B}_M(r_j, -\alpha_T) &\leq C \sum_{j \geq 2} j^{-a} r_j^{M+2} + CT\alpha_T^{M+1} \sum_{j \geq 2} j^{-a} r_j + CT\alpha_T^{M+2} \sum_{j \geq 2} j^{-\alpha} \\ &\leq CT^{-s} + CT^{-\rho}.\end{aligned}$$

Finally, for the tail-tail part $j, k \geq 2$, use the defining sum (42) and Lemma 8:

$$\begin{aligned}&\sum_{j,k \geq 2} j^{-\alpha} k^{-\beta-\alpha} \mathcal{B}_M(r_j, r_k) \\ &\quad = \sum_{i=0}^{M} (i+1) \left(\sum_{j \geq 2} j^{-\alpha} r_j^{M-i} \right) \left(\sum_{k \geq 2} k^{-\beta-\alpha} r_k^i \right) \\ &\quad \leq C \sum_{i=0}^{M} (i+1)(M-i+1)^{-s} (i+1)^{-(\rho+1)} \\ &\quad = C \sum_{i=0}^{M} (i+1)^{-\rho} (M-i+1)^{-s}.\end{aligned}$$

If $\rho < 1$, split the sum into $i \le M/2$ and $i > M/2$. On the first half,

$$\sum_{i \le M/2} (i+1)^{-\rho}(M-i+1)^{-s} \le CM^{-s} \sum_{i \le M/2} (i+1)^{-\rho} \le CM^{1-\rho-s}.$$

On the second half,

$$\sum_{i > M/2} (i+1)^{-\rho}(M-i+1)^{-s} \le CM^{-\rho} \sum_{j=1}^{\lceil M/2 \rceil} (j+1)^{-s} \le CM^{1-\rho-s}.$$

Thus this case contributes $O(M^{1-\rho-s})$, which is at most $O(T^{1-\rho})$ because $s > 0$. If $\rho = 1$, split the sum into $i \le M/2$ and $i > M/2$. On the first half,

$$\sum_{i \le M/2} (i+1)^{-1}(M-i+1)^{-s} \le CM^{-s} \sum_{i \le M/2} (i+1)^{-1} \le CM^{-s} \log(M+2).$$

On the second half,

$$\sum_{i > M/2} (i+1)^{-1}(M-i+1)^{-s} \le CM^{-1} \sum_{j=1}^{\lceil M/2 \rceil} (j+1)^{-s} \le CM^{-s}.$$

Thus the borderline convolution is $O(M^{-s}\log(M+2))$, and this is $O(1) = O(T^{1-\rho})$ when $\rho = 1$. If $\rho > 1$, the same split gives

$$CM^{-s} \sum_{i \le M/2} (i+1)^{-\rho} + CM^{-\rho} \sum_{j=1}^{\lceil M/2 \rceil} (j+1)^{-s} \le CM^{-s} + CM^{1-s-\rho} \le CM^{-s}.$$

Therefore the tail-tail part is bounded by

$$C\left(T^{1-\rho} + T^{-s}\right).$$

Collecting all pieces proves (40). Since

$$T^{1-\rho} = T^{-s} \quad \Longleftrightarrow \quad 1 - \frac{\beta - 1}{\alpha} = -\frac{\alpha - 1}{\alpha} \quad \Longleftrightarrow \quad \beta = 2\alpha,$$

the two-case form (41) follows. ∎

H.7. The expected finite-width gap

For the leading-order bias statement we use the comparison

$$\frac{T^4}{n} = o\left(T^{1-\rho} + T^{-s}\right). \tag{46}$$

This corresponds exactly to the requirement that the deterministic high-word error allowed by (39) is smaller than the second-order mean term.

Theorem 25 (Bias at the infinite-width optimal stepsize) *Let $\eta_{\infty,T}$ be the infinite-width root at horizon T. Assume (39). Then*

$$\overline{L}_{n,T}(\eta_{\infty,T}) - L_{\infty,T}(\eta_{\infty,T}) = \frac{\eta_{\infty,T}^2}{2n} Q_T(\eta_{\infty,T}) + O\left(\frac{T^4}{n^2}\right). \tag{47}$$

If, in addition, (46) holds, then

$$\overline{L}_{n,T}(\eta_{\infty,T}) - L_{\infty,T}(\eta_{\infty,T}) \asymp \frac{1}{n} \begin{cases} T^{1-\rho}, & \beta < 2\alpha, \\ T^{-(\alpha-1)/\alpha}, & \beta > 2\alpha, \end{cases}$$

with the same power in both branches at $\beta = 2\alpha$.

Proof By the covariance-word expansion,

$$\overline{L}_{n,T}(\eta_{\infty,T}) - L_{\infty,T}(\eta_{\infty,T}) = \frac{1}{2} W_1(\eta_{\infty,T}) + \frac{1}{2} W_2(\eta_{\infty,T}) + \frac{1}{2} \sum_{q=3}^{2T} W_q(\eta_{\infty,T}).$$

The first-order expectation vanishes because $\mathbb{E}E_n = 0$, so $W_1(\eta_{\infty,T}) = 0$. Lemma 19 gives

$$\frac{1}{2} W_2(\eta_{\infty,T}) = \frac{\eta_{\infty,T}^2}{2n} Q_T(\eta_{\infty,T}).$$

The remaining high-word sum is controlled by (39). This proves (47). For the leading-order statement, Proposition 24 and $\eta_{\infty,T} \to 2$ give the leading size

$$\frac{1}{n}\left(T^{1-\rho} + T^{-s}\right).$$

Condition (46) is exactly the requirement that the error T^4/n^2 be smaller than this leading size. The two-case form follows by comparing $T^{1-\rho}$ and T^{-s}. ■

Appendix I. Fluctuations around the bias

In the previous section, we computed the expectation. We next analyze the fluctuation around that expectation.

I.1. Covariance-word expansion and L^2 control of the nonlinear remainder

We now use the random version of the same word expansion. For $q = 1, \dots, 2T$, define

$$\mathcal{V}_q(\eta) := \sum_{\substack{a_0,\dots,a_q \geq 0 \\ a_0+\cdots+a_q = 2T-q}} \left\langle b, D_\eta^{a_0} B_{\eta,n} D_\eta^{a_1} B_{\eta,n} \cdots B_{\eta,n} D_\eta^{a_q} b \right\rangle.$$

Then $\mathbb{E}\mathcal{V}_q(\eta) = W_q(\eta)$ and

$$L_{n,T}(\eta) - L_{\infty,T}(\eta) = \frac{1}{2} \sum_{q=1}^{2T} \mathcal{V}_q(\eta). \tag{48}$$

The expectation of (48) is the deterministic word expansion used in the bias proof.

The following second-moment estimate is the only input needed to control the centered nonlinear tail.

Lemma 26 (Second-moment word bound) *For every $q \geq 2$ and every $\eta \in [1, 2]$,*

$$\|\mathcal{V}_q(\eta) - W_q(\eta)\|_{L^2} \leq (Cq)^q \binom{2T}{q} n^{-q/2}. \tag{49}$$

Proof It is enough to bound $\mathbb{E}[\mathcal{V}_q(\eta)^2]$, because $\|\mathcal{V}_q - W_q\|_{L^2}^2 = \mathrm{Var}(\mathcal{V}_q) \leq \mathbb{E}[\mathcal{V}_q^2]$. Expand the two copies of $\mathcal{V}_q(\eta)$. For each pair of compositions, there are $2q$ centered covariance factors and a prefactor n^{-2q}. Let $\mathbf{h} \in \{1, \ldots, n\}^{2q}$ be the combined sample-label tuple. If any sample index in $\mathbf{h}$ appears exactly once, conditioning on that sample makes the expectation vanish. For the remaining tuples, the product bound (34) gives

$$\left|\mathbb{E}\left[\mathcal{W}_{\mathbf{i},\mathbf{a}}(\eta)\mathcal{W}_{\mathbf{j},\mathbf{c}}(\eta)\right]\right| \leq (C)^{2q}\mathfrak{m}(\mathbf{h}).$$

Lemma 21, applied with $2q$ in place of q, gives

$$\sum_{\substack{\mathbf{h}\in\{1,\ldots,n\}^{2q}:\\ \text{no singleton}}} \mathfrak{m}(\mathbf{h}) \leq (Cq)^{2q} n^q.$$

The two composition sums contribute $\binom{2T}{q}^2$. Therefore

$$\mathbb{E}[\mathcal{V}_q(\eta)^2] \leq (Cq)^{2q} \binom{2T}{q}^2 n^{-q}.$$

Taking square roots proves (49), after adjusting the universal constant C. ∎

At the infinite-width root, the nonlinear centered fluctuation remainder is measured by

$$R_{n,T}^{\mathrm{fluc}} := \left\| \frac{1}{2} \sum_{q=2}^{2T} \left(\mathcal{V}_q(\eta_{\infty,T}) - W_q(\eta_{\infty,T})\right) \right\|_{L^2}. \tag{50}$$

We use the higher-order L^2 condition

$$R_{n,T}^{\mathrm{fluc}} = O\left(\frac{T^2}{n}\right). \tag{51}$$

Lemma 27 (Centered high-word condition from T^2/n) *If $T^2/n \to 0$, then (51) holds.*

Proof Lemma 26 and the composition estimate $\binom{2T}{q} \leq (2eT/q)^q$ give, uniformly for $q \geq 2$,

$$\|\mathcal{V}_q(\eta_{\infty,T}) - W_q(\eta_{\infty,T})\|_{L^2} \leq (CT)^q n^{-q/2}.$$

Let $\delta_n = CT/\sqrt{n}$. Since $T^2/n \to 0$, $\delta_n \to 0$. For all large n, $\delta_n \leq 1/2$, and hence

$$R_{n,T}^{\mathrm{fluc}} \leq \frac{1}{2} \sum_{q=2}^{2T} (CT)^q n^{-q/2} \leq C \frac{T^2}{n} \sum_{q=2}^{\infty} \delta_n^{q-2} \leq C \frac{T^2}{n}.$$

This proves (51). ■

From (48),

$$L_{n,T}(\eta) - L_{\infty,T}(\eta) = \frac{1}{2}\mathcal{V}_1(\eta) + \mathcal{R}_{n,T}(\eta), \qquad \mathcal{R}_{n,T}(\eta) := \frac{1}{2}\sum_{q=2}^{2T}\mathcal{V}_q(\eta).$$

Thus the random loss fluctuation consists of the first-order covariance word plus a nonlinear remainder.

Theorem 28 (L^2 control of the higher-order random remainder) *If the higher-order L^2 condition (51) holds, then there is a constant C such that*

$$\|\mathcal{R}_{n,T}(\eta_{\infty,T}) - \mathbb{E}\mathcal{R}_{n,T}(\eta_{\infty,T})\|_{L^2} \le C\frac{T^2}{n}. \tag{52}$$

Proof By the definition of $\mathcal{R}_{n,T}$ and the identity $\mathbb{E}\mathcal{V}_q(\eta_{\infty,T}) = W_q(\eta_{\infty,T})$,

$$\mathcal{R}_{n,T}(\eta_{\infty,T}) - \mathbb{E}\mathcal{R}_{n,T}(\eta_{\infty,T}) = \frac{1}{2}\sum_{q=2}^{2T}\big(\mathcal{V}_q(\eta_{\infty,T}) - W_q(\eta_{\infty,T})\big).$$

The L^2 norm of the right-hand side is exactly $R^{\mathrm{fluc}}_{n,T}$ from (50). Condition (51) is therefore precisely (52). ■

I.2. The first-order fluctuation

The first-order fluctuation is

$$\mathcal{V}_1(\eta) = -\eta\sum_{r=0}^{2T-1}\left\langle b, D_\eta^{2T-1-r}E_n D_\eta^r b\right\rangle.$$

Write

$$M_{\eta,T} := -\eta\sum_{r=0}^{2T-1} D_\eta^r b \otimes D_\eta^{2T-1-r} b.$$

Then

$$\mathcal{V}_1(\eta) = \mathrm{tr}(E_n M_{\eta,T}).$$

Since $E_n = n^{-1}\sum_{i=1}^n(z_i\otimes z_i - \Sigma)$,

$$\mathcal{V}_1(\eta) = \frac{1}{n}\sum_{i=1}^n Z_i(\eta),$$

where

$$Z_i(\eta) := \langle z_i, M_{\eta,T} z_i\rangle - \mathrm{tr}(\Sigma M_{\eta,T}).$$

Let

$$N_{\eta,T} := \Sigma^{1/2} M_{\eta,T}\Sigma^{1/2}.$$

Lemma 29 (Variance of the first-order fluctuation) *For every $\eta \in [1, 2]$,*

$$\mathrm{Var}(\mathcal{V}_1(\eta)) = \frac{2}{n} \|N_{\eta,T}\|_F^2 . \tag{53}$$

Proof The variables $Z_i(\eta)$ are independent and centered. Thus

$$\mathrm{Var}(\mathcal{V}_1(\eta)) = \frac{1}{n} \mathrm{Var}(Z_1(\eta)).$$

Let $g \sim N(0, I)$ and write $z = \Sigma^{1/2} g$. Then

$$\langle z, M_{\eta,T} z \rangle = \langle g, N_{\eta,T} g \rangle , \qquad \mathrm{tr}(\Sigma M_{\eta,T}) = \mathrm{tr}(N_{\eta,T}).$$

For a centered standard Gaussian vector and a Hilbert–Schmidt self-adjoint operator N,

$$\mathrm{Var}\big(\langle g, Ng \rangle\big) = 2 \|N\|_F^2 .$$

Applying this with $N = N_{\eta,T}$ proves (53). ■

Lemma 30 (CLT for the first-order fluctuation) *Let $T = T_n$ and let $\eta_{\infty,T}$ be the infinite-width root at horizon T. If $N_T := N_{\eta_{\infty,T},T}$ is nonzero, then*

$$\frac{\frac{1}{2}\mathcal{V}_1(\eta_{\infty,T})}{\|N_T\|_F / \sqrt{2n}} \Rightarrow N(0, 1). \tag{54}$$

Consequently,

$$\frac{1}{2}\mathcal{V}_1(\eta_{\infty,T}) = \Theta_p\left(\frac{\|N_T\|_F}{\sqrt{n}}\right)$$

in the sense that its absolute value has this stochastic order.

Proof For fixed n and $T = T_n$, write

$$\mathcal{V}_1(\eta_{\infty,T}) = \frac{1}{n}\sum_{i=1}^{n} Z_i, \qquad Z_i := \left\langle z_i, M_{\eta_{\infty,T},T} z_i \right\rangle - \mathrm{tr}(\Sigma M_{\eta_{\infty,T},T}).$$

The variables Z_i are i.i.d. and centered, and Lemma 29 gives

$$\mathrm{Var}(Z_i) = 2 \|N_T\|_F^2 .$$

We verify Lindeberg's condition for the triangular array $Z_i/\sqrt{n}$. Let $g \sim N(0, I)$. The variable Z_i has the same law as $\langle g, N_T g \rangle - \mathrm{tr}(N_T)$, a second-order Gaussian chaos. Hypercontractivity for Gaussian chaoses gives a universal constant C such that

$$\mathbb{E} Z_i^4 \le C(\mathbb{E} Z_i^2)^2 = C\{2 \|N_T\|_F^2\}^2.$$

Therefore, for every $\varepsilon > 0$,

$$\begin{aligned}
&\frac{1}{n\,\mathrm{Var}(Z_i)} \sum_{i=1}^{n} \mathbb{E}\left[Z_i^2 \mathbf{1}\{|Z_i| > \varepsilon \sqrt{n\,\mathrm{Var}(Z_i)}\}\right] \\
&= \frac{\mathbb{E}\left[Z_i^2 \mathbf{1}\{|Z_i| > \varepsilon \sqrt{n\,\mathrm{Var}(Z_i)}\}\right]}{\mathrm{Var}(Z_i)} \\
&\le \frac{\mathbb{E} Z_i^4}{\varepsilon^2 n \{\mathrm{Var}(Z_i)\}^2} \le \frac{C}{\varepsilon^2 n} \to 0.
\end{aligned}$$

The Lindeberg–Feller central limit theorem yields

$$\frac{\sum_{i=1}^{n} Z_i}{\sqrt{n\,\mathrm{Var}(Z_i)}} \Rightarrow N(0,1).$$

Since

$$\frac{\frac{1}{2}\mathcal{V}_1(\eta_{\infty,T})}{\|N_T\|_F / \sqrt{2n}} = \frac{\sum_{i=1}^{n} Z_i}{\sqrt{n\,\mathrm{Var}(Z_i)}},$$

we obtain (54). The stochastic order statement follows by applying the continuous mapping theorem to the absolute value of the nondegenerate normal limit. ■

I.3. Size of the first-order variance at the infinite-width learning rate

Let $\eta_{\infty,T}$ be the infinite-width root at horizon T, set $D_T = I - \eta_{\infty,T}\Sigma$, and $N_T := N_{\eta_{\infty,T},T}$. The matrix entries of N_T in the eigenbasis are explicit. We use one signed eigenvalue notation throughout this subsection:

$$M_1 := 2T - 1, \qquad r_1 = -\alpha_T, \qquad r_k := 1 - \eta_{\infty,T} k^{-\alpha} \quad (k \ge 2).$$

Thus r_k is the signed eigenvalue of A_T; in particular $r_k \in (0,1)$ for $k \ge 2$. Then

$$(N_T)_{ij} = -\eta_{\infty,T}\, i^{-(\alpha+\beta)/2} j^{-(\alpha+\beta)/2} \Phi_{M_1}(r_i, r_j), \tag{55}$$

where

$$\Phi_M(x,y) := \sum_{r=0}^{M} x^{M-r} y^r = \begin{cases} \dfrac{x^{M+1} - y^{M+1}}{x - y}, & x \ne y, \\ (M+1)x^M, & x = y. \end{cases} \tag{56}$$

Proposition 31 (Frobenius norm of the first-order kernel) *At the infinite-width root $\eta_{\infty,T}$,*

$$\|N_T\|_F \gtrsim T^{-\rho} + T^{-(\rho+1)/2}, \tag{57}$$

and

$$\|N_T\|_F \lesssim \begin{cases} T^{-\rho}, & \rho < 1, \\ T^{-1} \log T, & \rho = 1, \\ T^{-(\rho+1)/2}, & \rho > 1. \end{cases} \tag{58}$$

In particular,

$$\begin{aligned} \|N_T\|_F &\asymp T^{-\rho}, && \rho < 1, \\ T^{-1} \lesssim \|N_T\|_F &\lesssim T^{-1}\log T, && \rho = 1, \\ \|N_T\|_F &\asymp T^{-(\rho+1)/2}, && \rho > 1. \end{aligned} \tag{59}$$

Proof Let P_1 be the projection onto e_1 and $P_\perp = I - P_1$. Decompose

$$N_T = P_1 N_T P_1 + P_1 N_T P_\perp + P_\perp N_T P_1 + P_\perp N_T P_\perp.$$

The $(1,1)$ entry. Since $r_1 = 1 - \eta_{\infty,T} = -\alpha_T$ and $M_1 = 2T - 1$,

$$|(N_T)_{11}| = \eta_{\infty,T}(M_1 + 1)\alpha_T^{M_1}.$$

By Lemma 11, $\alpha_T^{M_1} \asymp T^{-(\rho+1)}$, so

$$\|P_1 N_T P_1\|_F \asymp T\, T^{-(\rho+1)} = T^{-\rho}. \tag{60}$$

This gives the $T^{-\rho}$ lower bound.

The cross entries. We next compute the size of the entries connecting the top eigendirection to the tail. For $k \geq 2$, formula (55) and the closed form in (56) give

$$\begin{aligned} \Phi_{M_1}(r_1, r_k) &= \Phi_{2T-1}(-\alpha_T, r_k) \\ &= \frac{(-\alpha_T)^{2T} - r_k^{2T}}{-\alpha_T - r_k} = \frac{r_k^{2T} - \alpha_T^{2T}}{\alpha_T + r_k}. \end{aligned}$$

Therefore

$$|(N_T)_{1k}| = \eta_{\infty,T} k^{-(\alpha+\beta)/2} \left| \frac{r_k^{2T} - \alpha_T^{2T}}{\alpha_T + r_k} \right|. \tag{61}$$

The denominator is explicit and uniformly bounded away from zero. Indeed,

$$\alpha_T + r_k = (\eta_{\infty,T} - 1) + (1 - \eta_{\infty,T} k^{-\alpha}) = \eta_{\infty,T}(1 - k^{-\alpha}),$$

which is bounded above and below by positive constants uniformly over $k \geq 2$ and all large T. Thus the order of $|(N_T)_{1k}|$ is determined by the power difference in (61).

Choose $K_0 \geq 2$ large enough that $2K_0^{-\alpha} \leq 1/4$, and take

$$\mathcal{K}_T := \{k \in \mathbb{N} : K_0 T^{1/\alpha} \leq k \leq 2K_0 T^{1/\alpha}\}.$$

This block has cardinality at least $cT^{1/\alpha}$ for all large T. For $k \in \mathcal{K}_T$, set $x_{k,T} := \eta_{\infty,T} k^{-\alpha}$. Then $r_k = 1 - x_{k,T}$ and

$$0 \leq x_{k,T} \leq 2K_0^{-\alpha} T^{-1} \leq (4T)^{-1}.$$

Consequently

$$r_k^{2T} = (1 - x_{k,T})^{2T} \geq \exp(-4T x_{k,T}) \geq \exp(-8K_0^{-\alpha}) =: c_0 > 0.$$

On the other hand, Lemma 11 gives $\alpha_T^{2T} = O(T^{-(\rho+1)}) = o(1)$. Hence, for all large T,

$$|r_k^{2T} - \alpha_T^{2T}| \geq c_0/2 \qquad (k \in \mathcal{K}_T).$$

Substituting this lower bound into (61) gives

$$\begin{aligned}\sum_{k\in\mathcal{K}_T} |(N_T)_{1k}|^2 &\geq c \sum_{k\in\mathcal{K}_T} k^{-(\alpha+\beta)} \\ &\geq cT^{1/\alpha}(2K_0T^{1/\alpha})^{-(\alpha+\beta)} \\ &= c'T^{-(\alpha+\beta-1)/\alpha} = c'T^{-(\rho+1)}.\end{aligned}$$

Since the two cross blocks are adjoints of one another,

$$\|P_1 N_T P_\perp + P_\perp N_T P_1\|_F \gtrsim T^{-(\rho+1)/2}.$$

Together with (60), this proves (57).

It remains to prove the upper bounds. The $(1,1)$ block is already bounded by $CT^{-\rho}$. For the cross block, the denominator bound in (61), the inequality $|\alpha_T^{2T} - r_k^{2T}| \leq \alpha_T^{2T} + r_k^{2T}$, and Lemma 8 give

$$\begin{aligned}\sum_{k\geq 2} |(N_T)_{1k}|^2 &\leq C \sum_{k\geq 2} k^{-(\alpha+\beta)}(\alpha_T^{2T} + r_k^{2T})^2 \\ &\leq C\alpha_T^{4T} \sum_{k\geq 2} k^{-(\alpha+\beta)} + C \sum_{k\geq 2} k^{-(\alpha+\beta)} r_k^{4T} \\ &\leq CT^{-2(\rho+1)} + CT^{-(\rho+1)} \leq CT^{-(\rho+1)}.\end{aligned}$$

Therefore the cross block is $O(T^{-(\rho+1)/2})$.

For the $P_\perp N_T P_\perp$ block, define the vector

$$u_m := \sum_{k\geq 2} k^{-(\alpha+\beta)/2} r_k^m e_k.$$

Then

$$P_\perp N_T P_\perp = -\eta_{\infty,T} \sum_{r=0}^{M_1} u_{M_1-r} \otimes u_r.$$

By the triangle inequality for the Frobenius norm,

$$\|P_\perp N_T P_\perp\|_F \leq C \sum_{r=0}^{M_1} \|u_{M_1-r}\| \, \|u_r\| \,.$$

But

$$\|u_m\|^2 = \sum_{k\geq 2} k^{-(\alpha+\beta)} r_k^{2m} \leq C(m+1)^{-(\rho+1)}$$

by Lemma 8. Therefore

$$\|P_\perp N_T P_\perp\|_F \leq C \sum_{r=0}^{M_1} (r+1)^{-(\rho+1)/2} (M_1 - r + 1)^{-(\rho+1)/2}. \tag{62}$$

If $\rho < 1$, then $(\rho+1)/2 < 1$ and the convolution in (62) is $O(T^{-\rho})$. If $\rho = 1$, it is $O(T^{-1}\log T)$. If $\rho > 1$, then $(\rho+1)/2 > 1$ and the convolution is $O(T^{-(\rho+1)/2})$. Combining the block estimates gives (58) and hence (59). ∎

For comparing the nonlinear L^2 remainder with the first-order fluctuation scale, we use

$$\frac{T^2}{n} = o\left(\frac{H_\rho(T)}{\sqrt{n}}\right), \qquad H_\rho(T) := \begin{cases} T^{-\rho}, & \rho < 1, \\ T^{-1}, & \rho = 1, \\ T^{-(\rho+1)/2}, & \rho > 1. \end{cases} \tag{63}$$

This condition is separate from the bias comparison (46); it is used only to identify the first-order random word as the leading random term.

Theorem 32 (First-order fluctuation) *Let $\eta_{\infty,T}$ be the infinite-width root at horizon T and assume the L^2 word condition (51). Then*

$$L_{n,T}(\eta_{\infty,T}) - \overline{L}_{n,T}(\eta_{\infty,T}) = \frac{1}{2}\mathcal{V}_1(\eta_{\infty,T}) + O_{L^2}\left(\frac{T^2}{n}\right). \tag{64}$$

Furthermore,

$$\operatorname{Var}\left(\frac{1}{2}\mathcal{V}_1(\eta_{\infty,T})\right) = \frac{1}{2n}\left\|N_T\right\|_F^2. \tag{65}$$

Consequently, the standard deviation of the first-order fluctuation term $\frac{1}{2}\mathcal{V}_1(\eta_{\infty,T})$ is of order $n^{-1/2}\left\|N_T\right\|_F$; by Proposition 31,

$$\begin{aligned} n^{-1/2}\left\|N_T\right\|_F &\asymp n^{-1/2}T^{-\rho}, && \rho < 1, \\ n^{-1/2}T^{-1} \lesssim n^{-1/2}\left\|N_T\right\|_F &\lesssim n^{-1/2}T^{-1}\log T, && \rho = 1, \\ n^{-1/2}\left\|N_T\right\|_F &\asymp n^{-1/2}T^{-(\rho+1)/2}, && \rho > 1. \end{aligned}$$

If (63) also holds, then the $O_{L^2}(T^2/n)$ remainder in (64) is smaller than this first-order fluctuation scale.

Proof The expansion

$$L_{n,T}(\eta_{\infty,T}) - L_{\infty,T}(\eta_{\infty,T}) = \frac{1}{2}\mathcal{V}_1(\eta_{\infty,T}) + \mathcal{R}_{n,T}(\eta_{\infty,T})$$

and its expectation give

$$L_{n,T}(\eta_{\infty,T}) - \overline{L}_{n,T}(\eta_{\infty,T}) = \frac{1}{2}\mathcal{V}_1(\eta_{\infty,T}) + \left(\mathcal{R}_{n,T}(\eta_{\infty,T}) - \mathbb{E}\mathcal{R}_{n,T}(\eta_{\infty,T})\right).$$

The remainder is $O_{L^2}(T^2/n)$ by Theorem 28. Formula (65) is Lemma 29, and the case distinction follows from Proposition 31. ∎

Appendix J. Combined conclusion

Combining Theorems 25 and 32, under (39) and (51) and at the infinite-width optimal stepsize $\eta_{\infty,T}$, one has

$$L_{n,T}(\eta_{\infty,T}) - L_{\infty,T}(\eta_{\infty,T}) = \underbrace{\frac{\eta_{\infty,T}^2}{2n} Q_T(\eta_{\infty,T})}_{\text{deterministic bias}} + \underbrace{\frac{1}{2}\mathcal{V}_1(\eta_{\infty,T})}_{\text{first-order fluctuation}} + O\left(\frac{T^4}{n^2}\right) + O_{L^2}\left(\frac{T^2}{n}\right).$$

Under the additional comparison (46), the deterministic mean correction has size

$$\frac{1}{n}\begin{cases} T^{1-\rho}, & \beta < 2\alpha, \\ T^{-(\alpha-1)/\alpha}, & \beta > 2\alpha, \end{cases}$$

with the same power at $\beta = 2a$. The first-order fluctuation term has standard deviation $n^{-1/2}\|N_T\|_F$, with the three regimes stated in Proposition 31. These are the two leading mechanisms in the finite-width versus infinite-width gap.

The condition $T^2/n \to 0$ verifies the two coarse high-word bounds (39) and (51). The final growing-horizon theorem below imposes the additional comparisons that make the deterministic bias, the corrected high-word remainders, and the optimizer-transfer gap negligible relative to the first-order fluctuation scale $H_T/\sqrt{n}$.

Appendix K. Growing-horizon fast-transfer rates

We now translate the preceding expansion into the growing-horizon fast-transfer quantities. Let $\eta_{\infty,T}$ be the infinite-width root at horizon T, set $H_T := \|N_T\|_F$, and let

$$\widehat{\eta}_{n,T} \in \arg\min_{\eta\in[1,2]} L_{n,T}(\eta).$$

Define

$$\begin{aligned} a_{n,T} &:= \left|L_{n,T}(\widehat{\eta}_{n,T}) - L_{\infty,T}(\eta_{\infty,T})\right|, \\ b_{n,T} &:= |\widehat{\eta}_{n,T} - \eta_{\infty,T}|, \\ c_{n,T} &:= L_{\infty,T}(\widehat{\eta}_{n,T}) - L_{\infty,T}(\eta_{\infty,T}), \\ \widetilde{c}_{n,T} &:= L_{n,T}(\eta_{\infty,T}) - L_{n,T}(\widehat{\eta}_{n,T}). \end{aligned}$$

The quantity $c_{n,T}$ is the finite-to-infinite transfer suboptimality in the fast-transfer framework; $\widetilde{c}_{n,T}$ is the reverse finite-width operational gap.

The proof uses slow growth in four logically distinct places. The condition $T^{\rho+2}/\sqrt{n} \to 0$ localizes the finite-width optimizer near the infinite-width root. The condition $T^2/n \to 0$, which follows from the comparison below, justifies the corrected high-word bounds. The comparisons $T^2/n = o(H_T/\sqrt{n})$, $T^4/n^2 = o(H_T/\sqrt{n})$, and $(T^{1-\rho} + T^{-s})/n = o(H_T/\sqrt{n})$ make the centered nonlinear tail, deterministic high-word tail, and deterministic second-order bias smaller than the first-order fluctuation. Finally, $T^{\rho+3}/n = o(H_T/\sqrt{n})$ makes the curvature-amplified transfer gaps negligible on the same random loss scale. The theorem first states these exact requirements and then records a single explicit exponent $\chi(\rho)$ that implies all of them.

Theorem 33 (Growing-horizon fast transfer) *Let $T = T_n \to \infty$. Assume*

$$\frac{T^{\rho+2}}{\sqrt{n}} \to 0, \qquad \frac{T^2}{n} = o\left(\frac{H_T}{\sqrt{n}}\right), \qquad \frac{T^4}{n^2} = o\left(\frac{H_T}{\sqrt{n}}\right), \tag{66}$$

$$\frac{T^{1-\rho} + T^{-s}}{n} = o\left(\frac{H_T}{\sqrt{n}}\right), \qquad \frac{T^{\rho+3}}{n} = o\left(\frac{H_T}{\sqrt{n}}\right). \tag{67}$$

Then

$$a_{n,T} = \Theta_p\left(\frac{H_T}{\sqrt{n}}\right), \qquad b_{n,T} = O_p(T^{\rho+1}n^{-1/2}), \qquad c_{n,T} = O_p(T^{\rho+3}/n),$$

and hence

$$c_{n,T} = o_p(a_{n,T}).$$

A sufficient explicit growth condition for all assumptions in (66)–(67) is

$$\frac{T^{\chi(\rho)}}{\sqrt{n}} \to 0, \qquad \chi(\rho) := \begin{cases} 2\rho + 3, & 0 < \rho < 1, \\ 5, & \rho = 1, \\ (3\rho + 7)/2, & \rho > 1. \end{cases} \tag{68}$$

In particular, the same displayed exponent $\chi(\rho)$ used in the main statement verifies the corrected high-word estimates and the final fast-transfer comparisons.

Proof Set

$$X_{n,T} := L_{n,T}(\eta_{\infty,T}) - L_{\infty,T}(\eta_{\infty,T}).$$

The second condition in (66) implies $T^2/n \to 0$, because H_T is uniformly bounded and $H_T/\sqrt{n} \to 0$. Hence Lemmas 23 and 27 verify (39) and (51). Combining Theorems 25 and 32 gives

$$X_{n,T} = \frac{\eta_{\infty,T}^2}{2n} Q_T(\eta_{\infty,T}) + \frac{1}{2}\mathcal{V}_1(\eta_{\infty,T}) + O\left(\frac{T^4}{n^2}\right) + O_{L^2}\left(\frac{T^2}{n}\right).$$

By Proposition 24,

$$\frac{\eta_{\infty,T}^2}{2n} Q_T(\eta_{\infty,T}) = O\left(\frac{T^{1-\rho} + T^{-s}}{n}\right).$$

The assumptions (66)–(67) make this deterministic term and the two corrected high-word remainders $o_p(H_T/\sqrt{n})$; the L^2 statement implies the corresponding o_p bound by Chebyshev's inequality. Therefore

$$X_{n,T} = \frac{1}{2}\mathcal{V}_1(\eta_{\infty,T}) + o_p\left(\frac{H_T}{\sqrt{n}}\right).$$

Lemma 30 gives

$$\frac{X_{n,T}}{H_T/\sqrt{2n}} \Rightarrow N(0,1). \tag{69}$$

In particular,

$$|X_{n,T}| = \Theta_p(H_T n^{-1/2}).$$

Next, Proposition 15 gives

$$b_{n,T} = O_p(T^{\rho+1}n^{-1/2}),$$

and Proposition 16 gives

$$c_{n,T} = O_p(T^{\rho+3}/n), \qquad \widetilde{c}_{n,T} = O_p(T^{\rho+3}/n).$$

The last condition in (67) says that both transfer gaps are $o_p(H_T/\sqrt{n})$.

Finally,

$$\begin{aligned} L_{n,T}(\widehat{\eta}_{n,T}) - L_{\infty,T}(\eta_{\infty,T}) &= L_{n,T}(\eta_{\infty,T}) - L_{\infty,T}(\eta_{\infty,T}) - \{L_{n,T}(\eta_{\infty,T}) - L_{n,T}(\widehat{\eta}_{n,T})\} \\ &= X_{n,T} - \widetilde{c}_{n,T}. \end{aligned}$$

Combining this identity with (69) and $\widetilde{c}_{n,T} = o_p(H_T/\sqrt{n})$ yields

$$\frac{L_{n,T}(\widehat{\eta}_{n,T}) - L_{\infty,T}(\eta_{\infty,T})}{H_T/\sqrt{2n}} \Rightarrow N(0,1).$$

Taking absolute values gives

$$a_{n,T} = \Theta_p(H_T n^{-1/2}).$$

Together with $c_{n,T} = O_p(T^{\rho+3}/n) = o_p(H_T/\sqrt{n})$, this proves $c_{n,T} = o_p(a_{n,T})$.

It remains to verify the displayed sufficient condition (68). By Proposition 31,

$$H_T \gtrsim \begin{cases} T^{-\rho}, & 0 < \rho < 1, \\ T^{-1}, & \rho = 1, \\ T^{-(\rho+1)/2}, & \rho > 1. \end{cases}$$

Write the exponent in this lower bound as $p(\rho)$, so $p(\rho) = \rho, 1, (\rho+1)/2$ in the three regimes. The comparison

$$\frac{T^2/n}{H_T/\sqrt{n}} = \frac{T^2}{\sqrt{n}\, H_T}$$

is implied by $T^{2+p(\rho)}/\sqrt{n} \to 0$. The exponents $2+\rho$, 3, and $(\rho+5)/2$ are all strictly smaller than $\chi(\rho)$. Hence $T^2/n = o(H_T/\sqrt{n})$, and in particular $T^2/n \to 0$. Then

$$\frac{T^4/n^2}{H_T/\sqrt{n}} = \frac{T^2}{n} \cdot \frac{T^2/n}{H_T/\sqrt{n}} \to 0,$$

so the deterministic high-word comparison also follows.

For the deterministic second-order bias,

$$\frac{(T^{1-\rho} + T^{-s})/n}{H_T/\sqrt{n}} \lesssim \frac{T^{1-\rho+p(\rho)}}{\sqrt{n}} + \frac{T^{p(\rho)-s}}{\sqrt{n}}.$$

The first exponent is 1, 1, and $(3-\rho)/2$ in the three regimes, and each is below $\chi(\rho)$ whenever it is positive. For the second exponent, if $p(\rho) - s \le 0$ the term is at most $n^{-1/2}$; if $p(\rho) - s > 0$, it is still smaller than $\chi(\rho)$. Thus the deterministic second-order bias is negligible relative to $H_T/\sqrt{n}$.

For the optimizer-transfer penalty,

$$\frac{T^{\rho+3}/n}{H_T/\sqrt{n}} \lesssim \frac{T^{\rho+3+p(\rho)}}{\sqrt{n}}.$$

The exponent $\rho + 3 + p(\rho)$ is exactly $2\rho+3$, 5, and $(3\rho+7)/2$ in the three regimes, which are the three cases in (68). The same explicit condition is plainly stronger than $T^{\rho+2}/\sqrt{n} \to 0$. Therefore (68) implies all assumptions in (66)–(67). ∎

Appendix L. Discussion on useful transfer

In [8], the authors argue that a transfer is useful if given a fixed computational budget, the transferred model achieves better performance than a model trained from scratch on the target task. They showed that with constant curvature, the transfer is useful if the transfer is fast, i.e. $c_{n,T} = o_p(a_{n,T})$. However, in our case, the curvature is only of constant level with fixed time horizon, but if the time horizon grows, the curvature can vary as well. Especially, by Lemma 14 and (18), the curvature varies asymptotically as $T^{1-\rho}$, which is not handled by the argument in [8]. Despite of this varying curvature, we will show that in our setting, the transfer is still useful in the growing-horizon regime as long as we have Theorem 4 holds, which implies $c_{n,T} = o_p(a_{n,T})$.

Let the computational budget be $\mathcal{F}$. For a model of width n trained for T_n steps, the computational cost is $n^r T_n$ for some $r > 0$. Since learning rate is a one-dimensional parameter, we define the resolution δ_n as the minimum distance between two learning rates in the grid search. Denote the best learning rate found by the grid search as $\bar{\eta}_{n,T_n}$, and the true optimal learning rate for the width n model as η_{n,T_n}.

Direct tuning. We directly conduct a grid search over the learning rate on a width n model trained for T_n steps. The computational cost is $\mathcal{F} = n^r T_n/\delta_n$. Thus, the resolution is $\delta_n = n^r T_n/\mathcal{F}$, and the sub-optimality of this tuning is:

$$\begin{aligned}
&L_{n,T_n}(\bar{\eta}_{n,T_n}) - L_{\infty,T_n}(\eta_{\infty,T_n}) \\
=&(L_{n,T_n}(\bar{\eta}_{n,T_n}) - L_{n,T_n}(\eta_{n,T_n})) + (L_{n,T_n}(\eta_{n,T_n}) - L_{\infty,T_n}(\eta_{\infty,T_n})) \\
\sim& T_n^{1-\rho}\delta_n^2 + a_{n,T_n}.
\end{aligned} \tag{70}$$

Transfer. We first conduct a grid search over the learning rate on a width n model trained for T_n steps, and then transfer its optimal learning rate to a large width M model trained for T_M steps. In this case, the computational cost is $\mathcal{F} = n^r T_n/\delta_n + M^r T_M$. A trivial way to make the transfer tuning has at least the same rate of optimality as direct tuning is to choose M such that $M^r T_M = n^r T_n/\delta_n = \mathcal{F}/2$. Then essentially, we are doing direct tuning on a width M model trained for T_M steps, which will lead to the same sub-optimality as direct tuning. However, we can do better than this trivial way by choosing M larger than n as we will show below. The sub-optimality of this tuning is:

$$\begin{aligned}
&L_{M,T_M}(\bar{\eta}_{n,T_n}) - L_{\infty,T_M}(\eta_{\infty,T_M}) \\
=&(L_{M,T_M}(\bar{\eta}_{n,T_n}) - L_{M,T_M}(\eta_{n,T_n})) + (L_{M,T_M}(\eta_{n,T_n}) - L_{M,T_M}(\eta_{M,T_M})) \\
&+ (L_{M,T_M}(\eta_{M,T_M}) - L_{\infty,T_M}(\eta_{\infty,T_M})) \\
\sim& T_M^{1-\rho}\delta_n^2 + T_M^{1-\rho} b_{n,T_n}^2 + a_{M,T_M}.
\end{aligned} \tag{71}$$

Useful transfer when $c_{n,T_n} = o_p(a_{n,T_n})$. Suppose we have Theorem 4 holds, (71) could be rewritten as

$$\begin{aligned}
&L_{M,T_M}(\bar{\eta}_{n,T_n}) - L_{\infty,T_M}(\eta_{\infty,T_M}) \\
\sim& T_M^{1-\rho}\delta_n^2 + (T_M/T_n)^{1-\rho} T_n^{1-\rho} b_{n,T_n}^2 + a_{M,T_M} \\
=& T_M^{1-\rho}\delta_n^2 + (T_M/T_n)^{1-\rho} o_p(a_{n,T_n}) + a_{M,T_M} \\
=&(T_M/T_n)^{1-\rho}(T_n^{1-\rho}\delta_n^2 + o_p(a_{n,T_n})) + a_{M,T_M}.
\end{aligned}$$

Since $M > n$, we have $T_M \geq T_n$, $a_{M,T_M} \leq a_{n,T_n}$. Comparing this with (70), we could always choose M slightly larger than n such that the transfer tuning is better than direct tuning, which implies that the optimal suboptimality of transfer tuning would always converge to 0 faster than direct tuning, and thus the transfer is useful. In summary, we have the following conclusion:

Proposition 34 (Useful transfer under growing horizon) *Let $T_n \to \infty, n \in \mathbb{N}$ be a non-decreasing horizon sequence, and suppose Theorem 4 holds for this sequence. Define the the sub-optimality gap of the transferred tuning of learning rate as*

$$L_{M,T_M}(\bar{\eta}_{n,T_n}) - L_{\infty,T_M}(\eta_{\infty,T_M}),$$

whereas define the sub-optimality gap of the direct tuning of learning rate as

$$L_{n,T_n}(\bar{\eta}_{n,T_n}) - L_{\infty,T_n}(\eta_{\infty,T_n}).$$

Then transferring from width n to a larger width M is useful in the sense of Ghosh et al. [8]: under a fixed computational budget, the sub-optimality gap of the transfer strategy decays at a faster rate than that of direct tuning.