FAST LEARNING RATE TRANSFER IN SHALLOW LINEAR NETWORKS AT GROWING TRAINING HORIZONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyperparameter transfer across model width can substantially reduce the cost of tuning large neural networks, but its behavior when the training horizon grows with width is not fully understood. Building on the framework of fast hyperparameter transfer (Ghosh et al., 2026), which formalizes when transfer is effective, we investigate conditions that ensure fast transfer in the growing-horizon regime. Specifically, we study learning-rate transfer in a shallow linear network with a single trainable hidden matrix, trained by full-batch gradient descent. Under additional spectral assumptions, our main results are threefold. (i) We prove fast learning-rate transfer as $n, T \to \infty$ whenever $T = o(\sqrt{n})$. (ii) We characterize the transfer rates through the finite-width perturbation scale, the first-order sensitivities of the loss and its learning-rate derivative to finite-width perturbations, and the local loss curvature. (iii) We derive limiting distributions for the optimal learning rate and optimized loss, governed by fluctuations associated with the extreme eigenvalues of the data Gram matrix. These results clarify how spectral structure and local loss sensitivities govern learning-rate transfer at growing horizons.

1 INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

Tuning hyperparameters, such as learning rates, becomes increasingly costly as neural networks grow. *Hyperparameter transfer* seeks to reduce this cost by reusing hyperparameters tuned on small proxy models when training larger models. Building on μP (Yang & Hu, 2021), which specifies width-dependent initialization and learning-rate scalings that preserve nontrivial feature learning in the infinite-width limit, Yang et al. (2021) introduced μTransfer: tune hyperparameters on a small proxy model and transfer them to a larger model without retuning. Their experiments demonstrate that appropriately parameterized learning rates can remain effective across substantially different widths. Subsequent studies have examined the empirical robustness of transfer to architectural and optimizer choices (Lingle, 2024), as well as the role of layerwise learning-rate scaling across parameterizations (Lingle, 2024; Everett et al., 2024). These empirical successes motivate theoretical analysis of the reliable transfer of hyperparameters.

Ghosh et al. (2026) introduced the notion of *fast hyperparameter transfer* to formalize when transfer is effective. Their framework shows that what matters is whether the optimal hyperparameter converges sufficiently fast relative to the rate at which model performance converges. Thus convergence of the optimal hyperparameter alone does not guarantee fast transfer. Relatedly, Hayou (2026) established convergence of the optimal learning rate for deep linear networks trained by full-batch gradient descent under μP, providing a theoretical basis for learning-rate transfer across width.

One major challenge in hyperparameter transfer is dealing with *growing training horizons* T, particularly for the learning-rate transfer. Recent language-model training motivates increasing the number of training tokens together with model size (Hoffmann et al., 2022; Dey et al., 2025), which often increases the number of gradient updates. From a theoretical perspective, the joint scaling limit of width and training horizon is substantially more delicate than the fixed-horizon setting, since the limiting objective itself changes as T grows. Consequently, the theoretical understanding of this regime remains limited. An important exception is Wen et al. (2026), who establish fast learning-rate transfer at growing horizons in sketched linear regression.

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

Motivated by these observations, we ask the following questions:

> *As model width and training horizon grow jointly, what conditions ensure fast learning-rate transfer, and which quantities govern whether such transfer occurs?*

1.2 This Study

To answer these questions, we consider a linear network with a single trainable hidden matrix, frozen random input and output maps, and μP initialization. The network is trained by full-batch gradient descent with a constant learning rate. For each horizon T, we compare the width-n optimal learning rate and optimized training loss with their infinite-width counterparts.

Our contributions are as follows:

- **A sufficient growth condition on the training horizon for fast transfer**: For fixed sample size and input dimension, we prove fast transfer whenever $T \to \infty$ and $T = o(\sqrt{n})$, under additional spectral assumptions (Theorem 3.4; Corollary 3.5).
- **Summary of transfer errors using loss-based quantities**: In our setting, the rates of the three transfer errors can be summarized in terms of the finite-width perturbation scale and three horizon-dependent quantities: the first-order sensitivities of the loss value and its learning-rate derivative to finite-width perturbations, and the local curvature of the loss. This formulation makes explicit how these quantities combine to determine the transfer rates as the training horizon grows (Theorem 3.4; Section 4.2). It also helps interpret the failure of fast transfer in the fixed-horizon counterexample of Proposition 3.8.
- **Limiting distributions of the optimal learning rate and the optimized loss**: We derive limiting distributions of the optimal learning rate and optimized loss as the training horizon grows (Theorem 3.4). Both limits are linear functionals of the same two components of the perturbation, associated with the largest and smallest positive eigenvalues of the data Gram matrix.

Our analysis provides some insights. First, different spectra can therefore lead to different transfer rates and sufficient width–horizon scalings, since the loss-based quantities vary depending on the spectrum of data as the horizon grows. This observation contrasts with Wen et al. (2026). Second, the leading loss-value and slope responses to the same finite-width perturbation need not be aligned. Consequently, some perturbation directions can change the optimized loss at leading order while producing no leading-order shift in the optimal learning rate. This illustrates how width dependence in the attained loss can differ from width dependence in hyperparameter selection. The observation is qualitatively consistent with the mechanism proposed by Ghosh et al. (2026), although our analysis concerns local finite-width perturbations rather than a decomposition of the optimization trajectory; Section 5 discusses this connection.

Additional related work is deferred to Appendix A.

2 Problem Setup and Transfer Framework

2.1 Model and exact loss dynamics

Fix positive integers d, m, training inputs $X^\top = [x_1 \dots x_m]$, and targets $y = (y_1, ..., y_m)^\top$. Consider the linear network

$$f_n(x; W_0, W_1, V) = V^\top W_1 W_0 x,$$

where $W_0 \in \mathbb{R}^{n\times d}$, $W_1 \in \mathbb{R}^{n\times n}$, and $V \in \mathbb{R}^n$. This model is similar to the one analyzed in Hayou (2026), except we have only one trainable weight W_1. The initialization is given entrywise as $W_{0,ij} \sim \mathcal{N}(0, d^{-1})$, $W_{1,ij}^{(0)} \sim \mathcal{N}(0, n^{-1})$, and $V_i \sim \mathcal{N}(0, n^{-2})$, following the scalings in μP. Define

$$K = d^{-1} X X^\top \in \mathbb{R}^{m\times m}, \qquad A_n = \|V\|^2 H_n H_n^\top \in \mathbb{R}^{m\times m}, \qquad H_n = X W_0^\top \in \mathbb{R}^{m\times n}.$$

The vector of predictions and the residual at step t are

$$f_n^{(t)}(X) = H_n (W_1^{(t)})^\top V \in \mathbb{R}^m, \qquad \chi_n^{(t)} = f_n^{(t)}(X) - y \in \mathbb{R}^m.$$

The loss is $\mathcal{L}_n(W_1) = (2m)^{-1}\|f_n(X; W_1) - y\|^2$, and for a fixed learning rate η, we train the model by full-batch gradient descent (GD) as

$$W_1^{(t+1)} = W_1^{(t)} - \eta \nabla_{W_1} \mathcal{L}_n(W_1^{(t)}) \qquad (t = 0, \dots, T-1).$$

Since there is only one trainable weight, training this model with the GD update is equivalent to performing GD for linear regression on fixed random features.

For $A \in \mathbb{R}^{m\times m}$ and $\eta \in \mathbb{R}$, define $B_A(\eta) = I_m - m^{-1}\eta A$. We abbreviate $B_n(\eta) := B_{A_n}(\eta)$ and $B_\infty(\eta) := B_K(\eta)$.

Proposition 2.1 (Exact finite-width loss dynamics)**.** *For every integer $T \geq 1$ and every $\eta \in \mathbb{R}$, we have $\chi_n^{(T)} = B_n(\eta)^T \chi_n^{(0)}$, and consequently,*

$$\phi_{n,T}(\eta) := \mathcal{L}_n(W_1^{(T)}) = (2m)^{-1}(\chi_n^{(0)})^\top B_n(\eta)^{2T} \chi_n^{(0)}.$$

Proposition 2.1 turns learning-rate selection into the minimization of a scalar random loss curve. All dependence on the initialization enters through the fixed-dimensional pair $(\chi_n^{(0)}, A_n)$, even though the parameter matrix grows with n. This reduction separates the two ingredients of the analysis: the spectral geometry of the reference curve and the finite-width fluctuations around it.

2.2 Infinite-width objectives and spectral geometry

For each T, define the corresponding infinite-width objective by

$$\phi_{\infty,T}(\eta) := (2m)^{-1}\|B_\infty(\eta)^T y\|^2 = (2m)^{-1} y^\top B_\infty(\eta)^{2T} y.$$

For every fixed T, Corollary C.1 shows that $\phi_{n,T}$ converges to $\phi_{\infty,T}$ uniformly on deterministic compact learning-rate intervals.

Consider the decomposition $K = \sum_{j=1}^r \lambda_j P_j$, where $\lambda_1 > \cdots > \lambda_r > 0$ are the distinct positive eigenvalues and P_j are their orthogonal spectral projectors.[1] Let P_0 denote the projection onto $\ker(K)$, and set $c_j = \|P_j y\|^2$ and $c_0 = \|P_0 y\|^2$. Define $\mathcal{J}_y = \{j \in [r] : c_j > 0\}$, and we assume $\mathcal{J}_y \neq \emptyset$. Then we can express the infinite-width objective as

$$\phi_{\infty,T}(\eta) = \frac{c_0}{2m} + \frac{1}{2m}\sum_{j=1}^r c_j \left(1 - \frac{\eta\lambda_j}{m}\right)^{2T}.$$

Each active eigenspace contributes a term with contraction factor $1 - \eta\lambda_j/m$, while $c_0/(2m)$ is independent of the learning rate. Thus $\mathcal{J}_y$ identifies the spectral modes relevant to learning-rate selection.

2.3 Optimal learning rates and transfer quantities

Fix a deterministic compact interval $\mathcal{I} = [\eta_\ell, \eta_u]$. For every $n \in \mathbb{N}$ and $T \geq 1$, define

$$\eta_{n,T} = \min \arg\min_{\eta\in\mathcal{I}} \phi_{n,T}(\eta), \qquad \eta_{\infty,T} = \min \arg\min_{\eta\geq 0} \phi_{\infty,T}(\eta).$$

Since $\phi_{n,T}$ is continuous and $\mathcal{I}$ is compact, $\arg\min_{\eta\in\mathcal{I}} \phi_{n,T}(\eta)$ is a nonempty compact set, so $\eta_{n,T}$ is well defined. Moreover, Corollary C.4 and Proposition C.5 imply that $\eta_{n,T}$ is almost surely the unique minimizer of $\phi_{n,T}$ over $\mathcal{I}$. When $Ky \neq 0$, Corollary C.4 also implies that the minimizer of $\phi_{\infty,T}$ is positive and unique over $[0, \infty)$.

Following Ghosh et al. (2026), we define the transfer quantities.

Definition 2.2. For each width n and training horizon T, define the loss gap $a_{n,T}$, hyperparameter gap $b_{n,T}$, and transfer suboptimality gap $c_{n,T}$ by

$$a_{n,T} = |\phi_{n,T}(\eta_{n,T}) - \phi_{\infty,T}(\eta_{\infty,T})|, \quad b_{n,T} = |\eta_{n,T} - \eta_{\infty,T}|, \quad c_{n,T} = \phi_{\infty,T}(\eta_{n,T}) - \phi_{\infty,T}(\eta_{\infty,T}).$$

We say the hyperparameter transfer is *fast* when $c_{n,T} = o_p(a_{n,T})$ holds.

[1] If K has repeated eigenvalues, the corresponding P_j are sums of the projections onto the individual eigenspaces.

The loss gap measures how much optimized performance varies across width, whereas the transfer suboptimality gap measures the additional target loss due specifically to using the proxy-selected learning rate. The hyperparameter gap records the displacement between the two minimizers. Fast transfer means that the cost of not retuning is negligible relative to the width-induced discrepancy in optimized losses. These comparisons always use a common horizon T, including along a sequence $T = T_n \to \infty$; we do not transfer between models trained for different numbers of updates.

3 Main results

3.1 Long-horizon structure of the infinite-width objective

We now let $T = T_n \to \infty$. In this regime, the limiting objective itself changes with n, and its optimizer moves toward the large-T limit. Thus, we first identify the moving deterministic reference in Section 3.1, and then compare the finite-width problem with this reference.

The spectral representation of $\phi_{\infty,T}$ in Section 2.2 shows that, for large T, the learning-rate-dependent loss is governed by the largest active contraction magnitude, $\max_{j\in\mathcal{J}_y} |1 - \eta\lambda_j/m|$. The long-horizon optimizer must therefore balance the two extreme active modes. The following theorem identifies this balance.

Theorem 3.1 (Active-spectrum limit of the optimal learning rate)**.** *In the decomposition of $K = \sum_{j=1}^{r} \lambda_j P_j$, define $\lambda_+^y = \max_{j\in\mathcal{J}_y} \lambda_j$ and $\lambda_-^y = \min_{j\in\mathcal{J}_y} \lambda_j$. Suppose $\lambda_+^y > \lambda_-^y > 0$ holds. Then as $T \to \infty$, we have*

$$\eta_{\infty,T} \to \eta_*^y := 2m/(\lambda_+^y + \lambda_-^y). \tag{1}$$

If $\lambda_+^y = \lambda_-^y = \lambda$, then we have $\eta_{\infty,T} = m/\lambda$ for all $T \geq 1$.

In Theorem 3.1, observe that the optimizer limit η_*^y remains strictly inside the stability interval. Indeed, if the top active eigenvalue equals $\lambda_1 = \lambda_{\max}(K)$ and the smallest active eigenvalue is positive, then we have $\eta_*^y = 2m/(\lambda_1 + \lambda_-^y) < 2m/\lambda_1$. We also note that the limit in Theorem 3.1 is closely related to the classical optimal step size for Richardson iteration; see, e.g., Saad (2003, Section 4.2.1).

Theorem 3.1 concerns the deterministic objective, which only contains modes active in y. At finite width, initialization and matrix perturbations can also activate modes absent from this reference objective. To keep the long-horizon comparison governed by the same spectral endpoints, we impose the following stronger condition.

Assumption 3.2 (Endpoint activity)**.** *In the decomposition of $K = \sum_{j=1}^{r} \lambda_j P_j$, suppose $r \geq 2$. The largest and smallest positive eigenvalues λ_1 and λ_r of K are simple and satisfy $\langle y, u_1\rangle \neq 0$ and $\langle y, u_r\rangle \neq 0$, where u_1 and u_r are unit eigenvectors satisfying $Ku_1 = \lambda_1 u_1$ and $Ku_r = \lambda_r u_r$. Thus we have $\lambda_+^y = \lambda_1$ and $\lambda_-^y = \lambda_r$.*

Under Assumption 3.2, set $\eta_* = 2m/(\lambda_1 + \lambda_r) = \eta_*^y$ and $q_* = (\lambda_1 - \lambda_r)/(\lambda_1 + \lambda_r)$. The endpoint contraction factors at η_* are $-q_*$ and q_*. Proposition E.1 and Corollary E.2 give a more quantitative picture: $\eta_{\infty,T} - \eta_* = O(T^{-1})$, i.e., the optimizer approaches a fixed location, while

$$\phi_{\infty,T}(\eta_{\infty,T}) - c_0/(2m) = \Theta(q_*^{2T}), \qquad \phi''_{\infty,T}(\eta_{\infty,T}) = \Theta(T^2 q_*^{2T-2})$$

so that the optimizable part of the loss and the local curvature shrink exponentially. The growing-horizon transfer analysis must track finite-width perturbations relative to these changing scales.

3.2 Fast learning-rate transfer at growing horizons

We now turn from the deterministic long-horizon geometry of the infinite-width objective to finite-width fluctuations. By Proposition 2.1, all randomness relevant to the finite-width loss curve is entirely captured by $(\chi_n^{(0)}, A_n)$. Thus, the first step is to characterize how this pair fluctuates around its infinite-width limit. The following theorem identifies the width-dependent fluctuation scale and its limiting random direction.

Lemma 3.3 (Joint initialization central limit theorem)**.** *Suppose m and d are fixed. Define $\alpha \sim \mathcal{N}(0, 2)$. Let G be a symmetric Gaussian $d \times d$ matrix satisfying $\mathbb{E}G_{rs} = 0$ and $\mathbb{E}(G_{rs}G_{tu}) =$*

$d^{-2}(\delta_{rt}\delta_{su} + \delta_{ru}\delta_{st})$. *Define* $\Xi = \alpha K + XGX^\top$, *and* $Z \sim \mathcal{N}(0, K)$. *Suppose that* α, G, *and* Z *are mutually independent. Then we have*

$$\sqrt{n}(\chi_n^{(0)} + y, A_n - K) \xrightarrow{d} (Z, \Xi).$$

Lemma 3.3 shows that the primitive finite-width perturbation has scale $\epsilon_n = n^{-1/2}$. The effect of this perturbation on learning-rate transfer, however, also depends on the training horizon. As established above, the infinite-width loss and its local curvature vary substantially with T, hence the sensitivities of the loss value and its learning-rate derivative to finite-width perturbations have their own horizon-dependent scales. To separate these effects from the width-dependent fluctuation scale ϵ_n, define

$$\rho_{L,T} = Tq_*^{2T-1}, \qquad \rho_{G,T} = T^2 q_*^{2T-2}, \qquad \kappa_T = T^2 q_*^{2T-2}.$$

The scales $\rho_{L,T}$ and $\rho_{G,T}$ measure the first-order sensitivity of the loss and its learning-rate derivative to finite-width perturbations, respectively. The scale κ_T measures the local curvature $\phi''_{\infty,T}(\eta_{\infty,T})$. Thus, ϵ_n quantifies the magnitude of the finite-width perturbation, whereas $\rho_{L,T}$, $\rho_{G,T}$, and κ_T capture how the loss response, loss-gradient response, and local curvature vary with the training horizon. The following theorem combines these scales to characterize the fluctuations of the optimal learning rate and optimized loss, and the resulting rates of the three transfer gaps.

Theorem 3.4 (Growing-horizon fluctuations and transfer rates; Theorems G.6, G.8, and G.10, extracted)**.** *Suppose* m *and* d *are fixed and* $T = T_n$ *satisfies* $T \to \infty$ *and* $T/\sqrt{n} \to 0$. *Assume Assumption 3.2, and suppose* $0 < \eta_\ell < \eta_* < \eta_u < 2m/\lambda_1$ *holds. Set* $L_* = \log(c_1\lambda_1/(c_r\lambda_r))$, $a_1^* = \exp(-\lambda_1 L_*/(\lambda_1 + \lambda_r))$, *and* $a_r^* = \exp(\lambda_r L_*/(\lambda_1 + \lambda_r))$. *Then we have*

$$(\epsilon_n \rho_{G,T}/\kappa_T)^{-1}(\eta_{n,T} - \eta_{\infty,T}) \xrightarrow{d} \Omega_{\eta,*} := -2m(\lambda_1 + \lambda_r)^{-2}(u_1^\top \Xi u_1 + u_r^\top \Xi u_r),$$

$$(\epsilon_n \rho_{L,T})^{-1}(\phi_{n,T}(\eta_{n,T}) - \phi_{\infty,T}(\eta_{\infty,T})) \xrightarrow{d} \Omega_{L,*} := \langle Q_{L,*}, \Xi \rangle_{\mathrm{F}},$$

where Ξ *is defined in Lemma 3.3 and* $Q_{L,*}$ *is defined by* $Q_{L,*} = m^{-2}\eta_*(c_1 a_1^* u_1 u_1^\top - c_r a_r^* u_r u_r^\top)$. *Consequently, we have*

$$a_{n,T} = \Theta_p(\epsilon_n \rho_{L,T}) = \Theta_p\left(\frac{Tq_*^{2T-1}}{\sqrt{n}}\right), \qquad b_{n,T} = \Theta_p\left(\frac{\epsilon_n \rho_{G,T}}{\kappa_T}\right) = \Theta_p\left(\frac{1}{\sqrt{n}}\right),$$

and

$$c_{n,T} = \Theta_p(\kappa_T b_{n,T}^2) = \Theta_p\left(\frac{\epsilon_n^2 \rho_{G,T}^2}{\kappa_T}\right) = \Theta_p\left(\frac{T^2 q_*^{2T-2}}{n}\right).$$

Corollary 3.5 (Fast transfer at growing horizons)**.** *Under the assumptions of Theorem 3.4, we have*

$$\frac{c_{n,T}}{a_{n,T}} = \Theta_p\left(\frac{\epsilon_n^2 \rho_{G,T}^2/\kappa_T}{\epsilon_n \rho_{L,T}}\right) = \Theta_p\left(\frac{T}{q_*\sqrt{n}}\right) \xrightarrow{p} 0.$$

By Corollary 3.5, under the assumptions of Theorem 3.4, fast learning-rate transfer persists at growing training horizons when $T = o(\sqrt{n})$.

3.3 A FIXED-HORIZON BENCHMARK

Having established our main growing-horizon result in Theorem 3.2, we now turn to the fixed-horizon regime. In particular, we verify that, under the nondegeneracy condition below, our model recovers the general rates conjectured by Wen et al. (2026) for fixed-T: $a_{n,T} = \Theta_p(n^{-1/2})$, $b_{n,T} = \Theta_p(n^{-1/2})$, and $c_{n,T} = \Theta_p(n^{-1})$. These rates also provide a useful baseline for Theorem 3.4. When T is fixed, the horizon-dependent factors are absorbed into constants, whereas allowing T to grow introduces nontrivial interactions between width and training horizon.

Theorem 3.6 (Fixed-horizon transfer rates)**.** *Fix* $m, d, T \in \mathbb{N}$, *and suppose* $|\mathcal{J}_y| = \#\{j \in [r] : c_j > 0\} \geq 2$ *holds.*[2] *Assume in addition that* $\eta_{\infty,T} \in (\eta_\ell, \eta_u)$ *holds. Then for* $\epsilon_n = n^{-1/2}$, *we have*

$$a_{n,T} = \Theta_p(\epsilon_n), \qquad b_{n,T} = \Theta_p(\epsilon_n), \qquad c_{n,T} = \Theta_p(\epsilon_n^2).$$

Corollary 3.7 (Fast transfer at a fixed horizon)**.** *Under the assumptions of Theorem 3.6, we have* $c_{n,T}/a_{n,T} = \Theta_p(\epsilon_n) \xrightarrow{p} 0$.

Corollary 3.7 implies that, under the assumptions of Theorem 3.6, learning-rate transfer is fast at every fixed horizon T.

[2] This condition is equivalent to $K^2 y \notin \mathrm{span}\{Ky\}$.

3.4 COUNTEREXAMPLE TO FAST TRANSFER

The preceding results establish fast transfer both at growing horizons under the conditions of Theorem 3.4 and, more broadly at fixed horizons, under the nondegeneracy condition of Theorem 3.6. Fast transfer, however, is not automatic even when the training horizon is fixed. We now exhibit a counterexample in which fast transfer fails.

Proposition 3.8 (Failure of fast transfer with a single positive eigenvalue)**.** *Fix integers $2 \le m \le d$, deterministic data $X \in \mathbb{R}^{m\times d}$ satisfying $XX^\top = I_m$, and a nonzero $y \in \mathbb{R}^m$. Note that $|\mathcal{J}_y| = 1$ holds in this case. Suppose that $\eta_0 := md \in (\eta_\ell, \eta_u)$, and fix an integer $T \ge 1$. Then we have*

$$a_{n,T} = \Theta_p(n^{-T}), \qquad b_{n,T} = \Theta_p(n^{-1/2}), \qquad c_{n,T} = \Theta_p(n^{-T}), \qquad c_{n,T}/a_{n,T} = \Theta_p(1),$$

thus fast transfer fails.

Proposition 3.8 shows that convergence of the optimal learning rate alone is not sufficient for fast transfer. Indeed, although $b_{n,T} = \Theta_p(n^{-1/2})$ holds exactly as in Theorem 3.6, the loss gap and the transfer suboptimality gap both scale as $\Theta_p(n^{-T})$. Hence the transfer penalty is not asymptotically negligible relative to the finite-width loss discrepancy. The mechanism behind this failure is a degeneracy of the leading loss sensitivity; Section 5 revisits this example from the local perturbation perspective.

4 PERTURBATION STRUCTURE OF HYPERPARAMETER TRANSFER

4.1 MODEL-SPECIFIC PERTURBATION LIMITS FOR THEOREM 3.4

In this section, assume the conditions of Theorem 3.4. The purpose of this subsection is model-specific: we identify the first-order finite-width perturbations and their limiting distributions in our shallow linear model. Section 4.2 subsequently abstracts the rate-level consequences of these calculations into a general local perturbation formulation.

For symmetric $A \in \mathbb{R}^{m\times m}_{\mathrm{sym}}$ and $\chi \in \mathbb{R}^m$, define the deterministic loss map and score

$$\Phi_T(\eta;\chi,A) = (2m)^{-1}\chi^\top B_A(\eta)^{2T}\chi, \qquad F_T(\eta;\chi,A) = \chi^\top A B_A(\eta)^{2T-1}\chi.$$

Thus we have $\phi_{n,T} = \Phi_T(\cdot;\chi_n^{(0)},A_n)$ and $\phi_{\infty,T} = \Phi_T(\cdot;-y,K)$. Define $F_{n,T}$ and $F_{\infty,T}$ by $F_{n,T} = F_T(\cdot;\chi_n^{(0)},A_n)$ and $F_{\infty,T} = F_T(\cdot;-y,K)$. Proposition C.3 gives $\partial_\eta\Phi_T = -(T/m^2)F_T$, so interior minimizers of Φ_T are roots of F_T.

Linearization and horizon dependence. Lemma 3.3 gives $\sqrt{n}(\chi_n^{(0)} + y, A_n - K) \xrightarrow{d} (Z,\Xi)$. Writing $D_{(\chi,A)}$ for Fréchet derivative in (χ,A) at $(\eta_{\infty,T};-y,K)$, since Corollary C.4 implies $F_{\infty,T}(\eta_{\infty,T}) = 0$, Lemmas C.6, G.5, and G.7 yield

$$F_{n,T}(\eta_{\infty,T}) = D_{(\chi,A)}F_T[\chi_n^{(0)} + y, A_n - K] + O_p(T^2 q_*^{2T-3}/n), \tag{2}$$

$$\phi_{n,T}(\eta_{\infty,T}) - \phi_{\infty,T}(\eta_{\infty,T}) = D_{(\chi,A)}\Phi_T[\chi_n^{(0)} + y, A_n - K] + O_p(T^2 q_*^{2T-2}/n). \tag{3}$$

In particular, Lemma C.6 gives

$$D_{(\chi,A)}F_T(\eta_{\infty,T};-y,K)[h,E] = g_{F,T}^\top h + \langle Q_{F,T}, E\rangle_{\mathrm{F}},$$
$$D_{(\chi,A)}\Phi_T(\eta_{\infty,T};-y,K)[h,E] = g_{L,T}^\top h + \langle Q_{L,T}, E\rangle_{\mathrm{F}},$$

where $g_{F,T}, Q_{F,T}, g_{L,T}, Q_{L,T}$ are the fixed Fréchet coefficients defined in equation 12, equation 13, equation 14, and equation 15. Since $(\chi_n^{(0)} + y, A_n - K)$ is $O_p(n^{-1/2})$, the size of the finite-width fluctuations is determined by the T-dependence of the Fréchet coefficients.

Lemma G.4 shows that on $\mathrm{range}(K)$, as $T \to \infty$, $Q_{F,T}/(Tq_*^{2T-2})$ converges to $Q_{F,*}$ defined in Lemma G.4, and $Q_{L,T}/(Tq_*^{2T-1})$ converges to $Q_{L,*}$. Lemma G.4 also shows that on $\mathrm{range}(K)$, $g_{F,T}$ and $g_{L,T}$ are $o(Tq_*^{2T-2})$ and $o(Tq_*^{2T-1})$, respectively. Thus, at the infinite-width optimizer, the leading finite-width perturbations come from $A_n - K$, rather than from the initial prediction $f_n^{(0)}(X) = \chi_n^{(0)} + y$.

Combining these coefficient asymptotics with Lemma 3.3 and equation 2, we have $F_{n,T}(\eta_{\infty,T}) = \langle Q_{F,T}, A_n - K\rangle_{\mathrm{F}} + o_p(Tq_*^{2T-2}/\sqrt{n})$, and hence

$$\sqrt{n}(Tq_*^{2T-2})^{-1}F_{n,T}(\eta_{\infty,T}) \xrightarrow{d} \langle Q_{F,*}, \Xi\rangle_{\mathrm{F}} = -2m^{-1}\eta_* h_*(u_1^\top \Xi u_1 + u_r^\top \Xi u_r), \tag{4}$$

where $h_* := c_1\lambda_1 a_1^* = c_r\lambda_r a_r^*$ is shown in equation 47. Similarly, Lemma 3.3 and equation 3 give $\phi_{n,T}(\eta_{\infty,T}) - \phi_{\infty,T}(\eta_{\infty,T}) = \langle Q_{L,T}, A_n - K\rangle_{\mathrm{F}} + o_p(Tq_*^{2T-1}/\sqrt{n})$, so that

$$\sqrt{n}(Tq_*^{2T-1})^{-1}(\phi_{n,T}(\eta_{\infty,T}) - \phi_{\infty,T}(\eta_{\infty,T})) \xrightarrow{d} \langle Q_{L,*}, \Xi\rangle_{\mathrm{F}} = \Omega_{L,*}. \tag{5}$$

From the perturbation limits to the distributional results. Lemma G.2 localizes the finite-width optimizer at $\eta_{n,T} - \eta_{\infty,T} = O_p(n^{-1/2}) = o_p(T^{-1})$. Together with the local score-derivative asymptotics, the score limit equation 4 yields $\sqrt{n}(\eta_{n,T} - \eta_{\infty,T}) \xrightarrow{d} \Omega_{\eta,*}$, as stated in Theorem G.6. Likewise, finite-width reoptimization changes the loss by $O_p(T^2 q_*^{2T-2}/n) = o_p(Tq_*^{2T-1}/\sqrt{n})$, so the limit in equation 5 is unchanged when $\eta_{\infty,T}$ is replaced by $\eta_{n,T}$, yielding Theorem G.8. The remaining rate-level consequences of these perturbation estimates are organized abstractly in Section 4.2, while the complete model-specific arguments are given in Appendix G.

4.2 A LOCAL PERTURBATION FORMULATION FOR HORIZON-DEPENDENT TRANSFER

Our fixed- and growing-horizon proofs can be organized through a local perturbation argument whose ingredients are closely related to earlier analyses of fast hyperparameter transfer. Ghosh et al. (2026) use local strong convexity to relate hyperparameter displacement to transfer suboptimality for a fixed limiting objective. In the growing-horizon setting, Wen et al. (2026) track horizon-dependent score perturbations, local curvature, optimizer localization, and finite-width loss fluctuations.

The purpose of this section is to recast these proof ingredients in a common horizon-dependent formulation that makes the relevant scales explicit. Section 4.1 provides the model-specific first-order perturbation estimates and limiting distributions for our shallow linear model; here we retain only the corresponding loss-response, loss-gradient-response, curvature, and localization scales in order to explain the resulting rates of $a_{n,T}$, $b_{n,T}$, and $c_{n,T}$ and the condition for fast transfer.

We consider a scalar hyperparameter η, and allow $T = T_n$ to be fixed or diverging. Write

$$\phi_{n,T}(\eta) = \Phi_T(\eta; \gamma_n), \qquad \phi_{\infty,T}(\eta) = \Phi_T(\eta; \gamma_\infty), \qquad \|\gamma_n - \gamma_\infty\| = O_p(\epsilon_n),$$

where γ_∞ is deterministic and $\epsilon_n \to 0$. All stochastic orders below are along the chosen sequence (n, T_n), and the scales $\rho_{L,T}, \rho_{G,T}, \kappa_T, r_T$ are deterministic and positive. Assume that the objectives are twice continuously differentiable in η near the search interval $\mathcal{I}$, and that the derivative in γ appearing below exists. Let $\eta_{\infty,T}$ be a global minimizer of $\phi_{\infty,T}$ lying in the interior of $\mathcal{I}$, and let $\eta_{n,T}$ minimize $\phi_{n,T}$ over $\mathcal{I}$.

Define the loss gradient $G_T(\eta;\gamma) := \partial_\eta \Phi_T(\eta;\gamma)$. At the moving infinite-width optimizer, suppose $G_T(\eta_{\infty,T};\gamma_\infty) = 0$ and $G_T(\eta_{\infty,T};\gamma_n) = O_p(\epsilon_n \rho_{G,T})$ hold, and assume the first-order loss expansion

$$\begin{aligned} &d_{n,T} := \phi_{n,T}(\eta_{\infty,T}) - \phi_{\infty,T}(\eta_{\infty,T}) = D_\gamma \Phi_T(\eta_{\infty,T};\gamma_\infty)[\gamma_n - \gamma_\infty] + o_p(\epsilon_n \rho_{L,T}), \\ &|d_{n,T}| = \Theta_p(\epsilon_n \rho_{L,T}). \end{aligned} \tag{6}$$

Here $\rho_{L,T}$ and $\rho_{G,T}$ describe the loss and loss-gradient responses to the finite-width perturbation, respectively. The lower bound in equation 6 holds, for example, if the normalized first-order loss term converges to a Gaussian with positive variance.

Set $\kappa_T := \phi''_{\infty,T}(\eta_{\infty,T}) > 0$ and $J_T := [\eta_{\infty,T} - r_T, \eta_{\infty,T} + r_T] \subset \mathrm{int}(\mathcal{I})$. Suppose first that optimizer localization $|\eta_{n,T} - \eta_{\infty,T}| = o_p(r_T)$ has been established independently. The argument below is conditional on this localization; the scale comparison derived below is a consistency requirement for the local expansion. Writing $\underline{\kappa}_{n,T} := \inf_{\eta \in J_T} \partial_\eta G_T(\eta;\gamma_n)$, assume $\mathrm{P}(\underline{\kappa}_{n,T} > 0) \to 1$, $\kappa_T/\underline{\kappa}_{n,T} = O_p(1)$, and $\sup_{\eta\in J_T}\{|\phi''_{n,T}(\eta)| + |\phi''_{\infty,T}(\eta)|\} = O_p(\kappa_T)$. These conditions ensure positive finite-width curvature and uniform local control. With probability tending to one, both optimizers are interior, and the mean value theorem gives

$$b_{n,T} = |\eta_{n,T} - \eta_{\infty,T}| = |G_T(\eta_{\infty,T};\gamma_n)/\partial_\eta G_T(\tilde{\eta}_{n,T};\gamma_n)| = O_p(\epsilon_n \rho_{G,T}/\kappa_T),$$

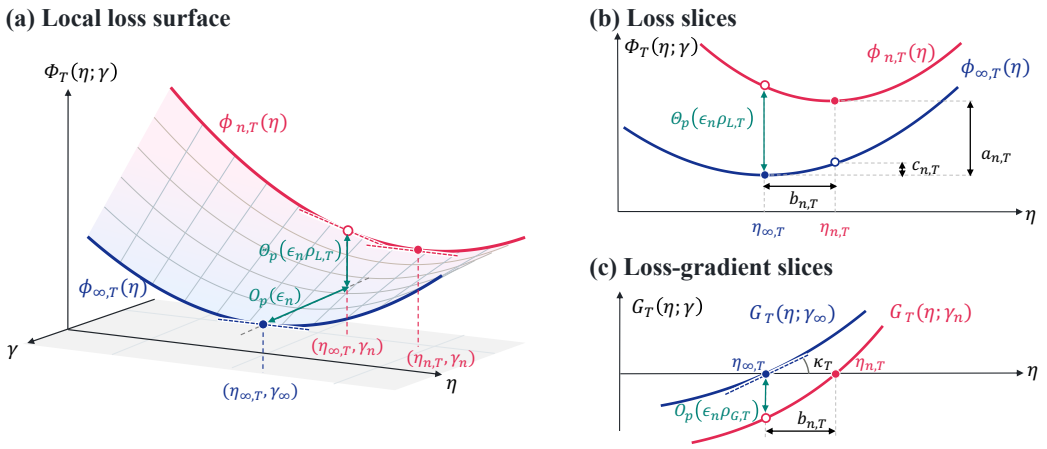

Figure 1: Illustration of the growing-horizon rate decomposition. **(a)** The local loss surface $\Phi_T(\eta;\gamma)$. A perturbation along the γ-axis from γ_∞ to γ_n, of size $O_p(\epsilon_n)$, affects both the loss value and its gradient with respect to η. **(b)** Loss slices at γ_∞ and γ_n. At $\eta_{\infty,T}$, the induced loss perturbation has scale $\Theta_p(\epsilon_n\rho_{L,T})$, which determines the leading scale of $a_{n,T}$. **(c)** Loss-gradient slices. The finite-width perturbation changes the gradient at $\eta_{\infty,T}$ by $O_p(\epsilon_n\rho_{G,T})$. Local curvature of order κ_T converts this gradient perturbation into a hyperparameter gap $b_{n,T}$.

where $\tilde{\eta}_{n,T}$ lies between the two optimizers. The bound is compatible with the assumed localization whenever $\epsilon_n\rho_{G,T}/\kappa_T = o(r_T)$.

Taylor expansion of the infinite-width objective then yields

$$c_{n,T} = (1/2)\phi''_{\infty,T}(\bar{\eta}_{n,T})(\eta_{n,T} - \eta_{\infty,T})^2 = O_p(\kappa_T b_{n,T}^2) = O_p(\epsilon_n^2\rho_{G,T}^2/\kappa_T),$$

where $\bar{\eta}_{n,T}$ lies between the two optimizers. For fixed T with nondegenerate local curvature, this is the same local quadratic relation underlying $c_n = \Theta_p(b_n^2)$ in (Ghosh et al., 2026, Proposition 1). Here we keep the factor κ_T explicit because the curvature may vanish as T grows.

Similarly, the improvement from optimizing the finite-width objective satisfies $R_{n,T} := \phi_{n,T}(\eta_{\infty,T}) - \phi_{n,T}(\eta_{n,T}) = O_p(\epsilon_n^2\rho_{G,T}^2/\kappa_T)$. Since $a_{n,T} = |d_{n,T} - R_{n,T}|$, a sufficient condition for this optimization correction to be negligible relative to the first-order loss fluctuation is $\epsilon_n\rho_{G,T}^2/(\kappa_T\rho_{L,T}) \to 0$. Under this condition, we have

$$a_{n,T} = \Theta_p(\epsilon_n\rho_{L,T}), \qquad c_{n,T}/a_{n,T} = O_p(\epsilon_n\rho_{G,T}^2/(\kappa_T\rho_{L,T})) = o_p(1).$$

The lower bound in equation 6 is needed for this comparison; an O_p bound on the loss fluctuation alone does not suffice. Figure 1 illustrates these scales.

In our model, the correspondence is $(\gamma_n,\gamma_\infty) = ((\chi_n^{(0)}, A_n), (-y, K))$, $G_T = -(T/m^2)F_T$, $\epsilon_n = n^{-1/2}$, $\rho_{L,T} \asymp Tq_*^{2T-1}$, $\rho_{G,T} \asymp T^2q_*^{2T-2}$, $\kappa_T \asymp T^2q_*^{2T-2}$, and $r_T \asymp T^{-1}$. Consequently, we have $\epsilon_n\rho_{G,T}/(\kappa_T r_T) \asymp T/\sqrt{n}$ and $\epsilon_n\rho_{G,T}^2/(\kappa_T\rho_{L,T}) \asymp T/(q_*\sqrt{n})$. Since $q_* > 0$ is fixed, $T = o(\sqrt{n})$ makes both ratios vanish. Thus the general formulation recovers the rate conclusions of Theorem 3.4 and the fast-transfer conclusion of Corollary 3.5. The fixed-horizon proof in Theorem 3.6 uses the same local structure, with horizon-dependent scales fixed as $n \to \infty$. The model-specific estimates underlying these scales differ substantially from those in Wen et al. (2026). Here the perturbation is finite-dimensional and the relevant matrix powers admit uniform control near the optimum, whereas their source-capacity setting involves a spectrum accumulating at zero and an optimizer approaching the stability edge, requiring spectral-tail-dependent localization and fluctuation estimates.

5 Implications for Mechanisms of Learning-Rate Transfer

Relation to the top-k loss decomposition. Ghosh et al. (2026) decompose the linearized loss change along an optimization trajectory into a dominant top-k component and a residual component,

and conjecture that fast transfer arises when the dominant component becomes rapidly width-stable while the width-sensitive residual has little effect on the hyperparameter optimum. In our model, the full-batch GD gradient with respect to W_1 has rank at most one. Since the update is proportional to this gradient, the associated gradient–update alignment matrix also has rank at most one, so its stepwise linearized loss change is captured by $k = 1$. This exact top-1 representation, however, is only a structural statement: it holds both in the fast-transfer regime of Theorem 3.4 and in Proposition 3.8, where $c_{n,T}/a_{n,T} = \Theta_p(1)$ for every fixed $T \geq 1$. Thus, low-rank concentration of the stepwise linearized loss change alone does not determine the transfer rate; what matters additionally is the local behavior of the finite-width perturbation around the infinite-width optimum.

Section 4.2 characterizes this local behavior in terms of three scales: the loss-value response $\rho_{L,T}$, the hyperparameter-gradient response $\rho_{G,T}$, and the local curvature κ_T. Under its assumptions, the leading loss perturbation has scale $\Theta_p(\epsilon_n \rho_{L,T})$, whereas both the transfer penalty and the improvement from finite-width reoptimization are bounded by $O_p(\epsilon_n^2 \rho_{G,T}^2/\kappa_T)$. Hence fast transfer follows when $\epsilon_n \rho_{G,T}^2/(\kappa_T \rho_{L,T}) \to 0$. The $T = 1$ case of Proposition 3.8 illustrates why the loss-response term is essential: in this counterexample, the first-order loss response vanishes, while the loss-gradient response remains nondegenerate and $\kappa_1 > 0$, and $a_{n,1}$ and $c_{n,1}$ are both $\Theta_p(n^{-1})$. For fixed $T \geq 2$, the first-order responses and the curvature all degenerate, placing the example in a higher-order regime outside the framework of Section 4.2.

Our growing-horizon limits also clarify the directional dependence of the loss-gradient and loss-value responses in Section 4.2. In Theorem 3.4, the normalized optimizer displacement and optimized-loss fluctuation converge in distribution to nonzero deterministic multiples of $\ell_\eta(\Xi)$ and $\ell_L(\Xi)$, respectively, where $\ell_\eta(E) := u_1^\top E u_1 + u_r^\top E u_r$ and $\ell_L(E) := c_1 a_1^* u_1^\top E u_1 - c_r a_r^* u_r^\top E u_r$. As shown in Section 4.1, these limits arise from $-(T/m^2)\langle Q_{F,T}, A_n - K\rangle_{\mathrm{F}}$ and $\langle Q_{L,T}, A_n - K\rangle_{\mathrm{F}}$, which are the leading terms of the respective first-order responses at $\eta_{\infty,T}$, whose scales are $\epsilon_n \rho_{G,T}$ and $\epsilon_n \rho_{L,T}$. Although ℓ_η and ℓ_L depend on the same two endpoint coordinates, they are not proportional. Hence there are perturbation directions with a nonzero leading loss response but a vanishing leading loss-gradient response, and therefore no leading-order optimizer displacement along those directions. This is qualitatively consistent with the idea that width-dependent loss changes can have little effect on hyperparameter selection, as in the mechanism proposed by Ghosh et al. (2026); our argument concerns local responses to $A_n - K$, rather than a decomposition of the optimization trajectory.

Relation to spectral explanations of transfer. Our results show that, even in this solvable setting, the width consistency of a single leading eigenvalue does not determine the long-horizon optimal learning rate. Noci et al. (2024) study the width consistency of leading parameter-space Hessian eigenvalues, particularly sharpness. Our result highlights a distinction between the stability boundary and the optimal learning rate. Under Assumption 3.2, Theorem 3.1 implies $\eta_{\infty,T} \to 2m/(\lambda_1 + \lambda_r) < 2m/\lambda_1$ as $T \to \infty$. Thus the largest eigenvalue determines the stability boundary, whereas both spectral endpoints determine the long-horizon optimum.

6 Conclusion

We studied fast learning-rate transfer across width in a shallow linear network under μP at growing training horizons. Specifically, we characterized the growing-horizon infinite-width optimizer through the active spectral endpoints and showed that, under endpoint activity and $T = o(\sqrt{n})$, the three transfer quantities obey horizon-dependent rates that yield $c_{n,T} = o_p(a_{n,T})$. The local perturbation formulation organizes transfer in terms of the loss-value response, loss-gradient response, and local curvature, providing a sensitivity-based explanation for fast transfer and for the dependence of its rates on the spectral structure of the problem.

Our analysis is restricted to a solvable setting of linear networks with a single trainable linear layer. Extending the analysis to nonlinear or deep networks is therefore an important next step, as is studying transfer simultaneously across multiple scaling axes such as width and depth. Another limitation of the current framework is its reliance on nondegenerate first-order sensitivities: the fixed-horizon counterexample shows that these first-order responses can vanish, requiring a higher-order analysis. Developing such a theory would help provide a clearer picture of the mechanisms governing hyperparameter transfer.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of proofs, designing or providing feedback on research methodology or experiments, and assisting with translation. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, implementing methods, and interpreting results. Generating synthetic data sets, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for drafting parts of a research paper and identifying relevant literature. We have reviewed all AI-assisted work. All AI-assisted proofs and mathematical arguments were independently checked for correctness by the authors. AI-assisted translations and generated text were manually reviewed and revised before inclusion in the manuscript. Suggestions concerning research methodology or experimental design were independently assessed by the authors. For literature identified with the assistance of generative AI, we consulted the original sources and independently verified their content and relevance to this work.

REPRODUCIBILITY STATEMENT

The model, training procedure, and assumptions used in our theoretical analysis are stated explicitly in Sections 2–3. Complete proofs of the theoretical results, together with the required auxiliary lemmas and derivations, are provided in the appendices.

REFERENCES

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

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4345–4382. PMLR, 21–27 Jul 2024a. URL https://proceedings.mlr.press/v235/bordelon24a.html.

Blake Bordelon, Lorenzo Noci, Mufan Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 22088–22127, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5f5f7b6080dcadced61cf5d96f7c6dde-Paper-Conference.pdf.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Nolan Dey, Bin 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. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 137707–137739. Curran Associates, Inc., 2025. doi: 10.52202/085713-4597. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c917d8b9e01427f3184d80ade22f4d1f-Paper-Conference.pdf.

Katie E Everett, Lechao Xiao, Mitchell Wortsman, Alexander A Alemi, Roman Novak, Peter J Liu, Izzeddin Gur, Jascha Sohl-Dickstein, Leslie Pack Kaelbling, Jaehoon Lee, and Jeffrey Pennington. Scaling exponents across parameterizations and optimizers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12666–12700. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/everett24a.html.

Nikhil Ghosh, Denny Wu, and Alberto Bietti. Understanding the mechanisms of fast hyperparameter transfer. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 65022–65063, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/6a9ac2ea5d562e5c17dba42d3ae34751-Paper-Conference.pdf.

Soufiane Hayou. A proof of learning rate transfer under μP. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas (eds.), *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3322–3330. PMLR, 02–05 May 2026. URL https://proceedings.mlr.press/v300/hayou26a.html.

Soufiane Hayou and Liyuan Liu. Optimal embedding learning rate in LLMs: The effect of vocabulary size. *arXiv preprint arXiv:2506.15025*, 2025.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran Associates, Inc., 2022. doi: 10.52202/068431-2176. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

Letong Hong and Zhangyang Wang. On the provable separation of scales in maximal update parameterization. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 23774–23786. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/hong25h.html.

Clarissa Lauditi, Cengiz Pehlevan, and Blake Bordelon. Spectral dynamics in deep networks: Feature learning, outlier escape, and learning rate transfer. *arXiv preprint arXiv:2605.07870*, 2026.

Lucas Lingle. An empirical study of μP learning rate transfer. *arXiv preprint arXiv:2404.05728*, 2024.

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. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 45300–45325, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4b0c1645f3d6a1730931e65ecbf91ac3-Paper-Conference.pdf.

Lorenzo Noci, Alexandru Meterez, Thomas Hofmann, and Antonio Orvieto. Super consistency of neural network landscapes and learning rate transfer. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 102696–102743. Curran Associates, Inc., 2024. doi: 10.52202/079017-3262. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ba1d33849b963efc6b5d3082ad68f480-Paper-Conference.pdf.

Yousef Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, second edition, 2003. doi: 10.1137/1.9780898718003. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898718003.

Boris Shigida, Boris Hanin, and Andrey Gromov. Learning rate transfer in Normalized Transformers. *arXiv preprint arXiv:2604.27077*, 2026.

Garrett Wen, Alberto Bietti, Nikhil Ghosh, Theodor Misiakiewicz, and Denny Wu. Fast learning rate transfer for gradient descent in sketched linear regression. In *High-dimensional Learning Dynamics 2026*, 2026. URL https://openreview.net/forum?id=WMHm9bkItE.

Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/yang21c.html.

Greg Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17084–17097. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/8df7c2e3c3c3be098ef7b382bd2c37ba-Paper.pdf.

Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs vi: Feature learning in infinite depth neural networks. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 55099–55150, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/f2ce452a4a7352feb7a9fb885267ff24-Paper-Conference.pdf.

A Additional related work

Parameterization and the empirical scope of transfer. Successful transfer depends on more than the name of a parameterization. Lingle (2024) examines μ-transfer under practical training choices, while Everett et al. (2024) study alignment assumptions and optimizer-dependent scaling exponents. In particular, the latter show that alternative parameterizations can also support transfer when equipped with suitable layerwise learning rates. These findings concern the design and empirical effectiveness of scaling prescriptions. Our analysis instead fixes the initialization and update rule, and compares the resulting optimizer mismatch, optimized loss fluctuation, and transfer penalty.

Conditional separation-of-scales explanations. Hong & Wang (2025) study a separation between macroscopic and microscopic quantities under a width-dominance regime, with implications for early-stage hyperparameter selection. This provides a different perspective on the stability of tuning across scale. Our results directly analyze the exact terminal training loss in a specified finite-dimensional model and track the dependence of its local perturbation scales on a growing training horizon.

Transfer across depth, modules, and architectural changes. Depthwise transfer has been investigated through infinite-depth limits and residual-network dynamics (Yang et al., 2024; Bordelon et al., 2024b), and through complete feature learning in deep Transformers (Dey et al., 2025). Mlodozeniec et al. (2026) further address transfer across modules, width, depth, batch size, and duration. Other work identifies architecture-specific scaling issues arising from large vocabularies or normalized Transformer parameterizations (Hayou & Liu, 2025; Shigida et al., 2026). These studies address scaling axes and training mechanisms that are outside our model.

Training schedules and scaling laws. Bordelon & Mori (2026) analyze terminal-loss optimization over learning-rate schedules in a random feature model. More broadly, empirical and dynamical scaling-law studies relate performance to model size, data, and training time (Hoffmann et al., 2022; Bordelon et al., 2024a). Our question is narrower: we transfer a constant learning rate across width while holding the source and target horizons equal, and quantify the resulting relative penalty.

Spectral explanations and sharpness. Noci et al. (2024) connect transfer to the width consistency of leading Hessian eigenvalues, while Lauditi et al. (2026) study spectral dynamics and feature learning in deep networks. These perspectives are related to the edge-of-stability behavior documented by Cohen et al. (2021). In our model, the largest eigenvalue determines the linear stability boundary, whereas the long-horizon optimal constant learning rate balances two active spectral endpoints.

B Numerical experiments

Setup. We use the model, initialization, and full-batch training rule of Section 2.1, with $m = d = 4$ and fixed data

$$K = \mathrm{diag}(7, 4, 2, 1), \qquad X = 2K^{1/2}, \qquad y = (1, 1, 1, 1)^\top.$$

The learning-rate interval is $I = [0.05, 1.12]$. Thus Assumption 3.2 holds with $\eta_* = 1$ and $q_* = 3/4$. At each width $n \in \{2^8, 2^9, \ldots, 2^{20}\}$, we draw 256 independent initializations per experiment. Figure 2 compares the fixed-horizon regime $T = 10$ with the growing-horizon regime $T_n = \lceil 2n^{0.15} \rceil$, which ranges from 5 to 16.

Computation and estimation. For growing horizons, define

$$\tilde{a}_{n,T} = \frac{a_{n,T}}{T q_*^{2T-1}}, \qquad \tilde{c}_{n,T} = \frac{c_{n,T}}{T^2 q_*^{2T-2}}.$$

We fit unweighted least-squares slopes of log empirical means against $\log n$ over $n = 2^{13}, \ldots, 2^{20}$: the largest eight widths for fixed T, and precisely the subset satisfying $T_n/\sqrt{n} \le 0.1$ for the growing schedule.

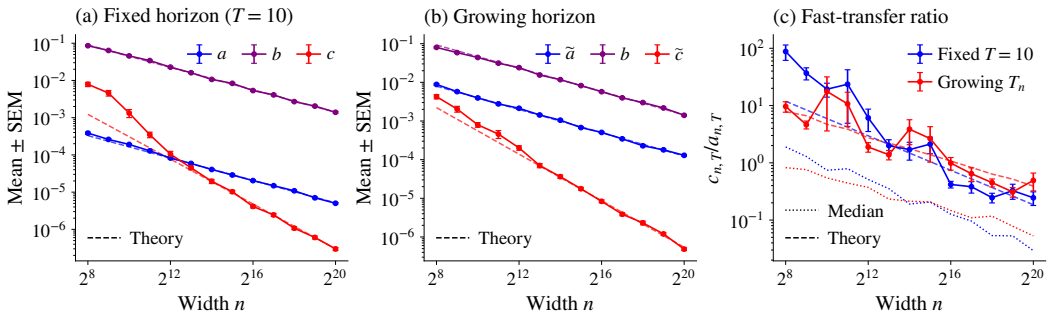

Figure 2: Numerical learning-rate transfer. **(a)** $a_{n,T}, b_{n,T}, c_{n,T}$ at fixed $T = 10$. **(b)** $\tilde{a}_{n,T}, b_{n,T}, \tilde{c}_{n,T}$ at $T_n = \lceil 2n^{0.15} \rceil$. **(c)** Per-initialization ratios $c_{n,T}/a_{n,T}$, summarized by means (solid) and medians (dotted); the reference rates are $n^{-1/2}$ for fixed T and $T/(q_*\sqrt{n})$ for growing T.

Results and implications. In panels (a) and (b) of Figure 2, all six slopes are close to the reference exponents $(-1/2, -1/2, -1)$, consistent with Theorems 3.6 and 3.4. The overall decline of $c_{n,T}/a_{n,T}$ in panel (c) supports fast transfer along the tested schedule and illustrates the quadratic transfer penalty discussed in Section 4.2. Ratio means are noisy and nonmonotone because small $a_{n,T}$ values can dominate individual ratios; medians show a smoother decline. The theoretical rates are in probability and do not, by themselves, imply corresponding rates for expectations.

C Basic properties and finite-width fluctuations

C.1 Proof of Proposition 2.1

First, we can compute

$$\nabla_{W_1} \mathcal{L}_n(W_1^{(t)}) = \frac{1}{m} \sum_{i=1}^{m} \chi_{n,i}^{(t)} V H_{n,i\cdot} = \frac{1}{m} V (\chi_n^{(t)})^\top H_n.$$

Thus we have $W_1^{(t+1)} = W_1^{(t)} - \frac{\eta}{m} V (\chi_n^{(t)})^\top H_n$, which implies

$$f_n^{(t+1)}(X) = H_n (W_1^{(t)})^\top V - \frac{\eta}{m} H_n H_n^\top \chi_n^{(t)} V^\top V = f_n^{(t)}(X) - \frac{\eta}{m} A_n \chi_n^{(t)}.$$

Subtracting y from both sides yields

$$\chi_n^{(t+1)} = B_n(\eta) \chi_n^{(t)},$$

which implies

$$\chi_n^{(T)} = B_n(\eta)^T \chi_n^{(0)}.$$

Therefore we have

$$\phi_{n,T}(\eta) = \mathcal{L}_n(W_1^{(T)}) = \frac{1}{2m} \|\chi_n^{(T)}\|^2 = \frac{1}{2m} \left\| B_n(\eta)^T \chi_n^{(0)} \right\|^2 = \frac{1}{2m} (\chi_n^{(0)})^\top B_n(\eta)^{2T} \chi_n^{(0)}.$$

□

C.2 Proof of Lemma 3.3

First we show the weak convergence of the marginal distribution $\sqrt{n}(A_n - K) \xrightarrow{d} \Xi$. Noting that $A_n = (n\|V\|^2) X (n^{-1} W_0^\top W_0) X^\top$ holds, we decompose

$$\begin{aligned} A_n - K = (n\|V\|^2 - 1)K + X(n^{-1} W_0^\top W_0 - d^{-1} I_d) X^\top \\ + (n\|V\|^2 - 1) X (n^{-1} W_0^\top W_0 - d^{-1} I_d) X^\top. \end{aligned}$$

Define $\alpha_n = \sqrt{n}(n\|V\|^2 - 1)$ and $G_n = \sqrt{n}(n^{-1}W_0^\top W_0 - d^{-1}I_d)$. Because $nV_j \sim N(0,1)$ independently, we apply the central limit theorem to obtain

$$\alpha_n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}((nV_j)^2 - 1) \xrightarrow{d} \alpha, \qquad \alpha \sim N(0,2).$$

Next, noting that $(W_{0,j\cdot})^\top \sim N(0, d^{-1}I_d)$ independently, the multivariate central limit theorem gives

$$G_n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}((W_{0,j\cdot})^\top W_{0,j\cdot} - d^{-1}I_d) \xrightarrow{d} G,$$

where G is the Gaussian matrix specified in the statement. Since V and W_0 are independent, α_n and G_n are independent for every n, and hence the joint convergence in distribution holds:

$$(\alpha_n, G_n) \xrightarrow{d} (\alpha, G).$$

Combining the results, we have

$$\sqrt{n}(A_n - K) = \alpha_n K + XG_nX^\top + O_p(n^{-1/2}) \xrightarrow{d} \alpha K + XGX^\top = \Xi. \tag{7}$$

Fix $u \in \mathbb{R}^m$ and $T \in \mathbb{R}^{m\times m}$. Note that $\chi_n^{(0)} + y = f_n^{(0)}(X) = H_n(W_1^{(0)})^\top V$ holds. Since $W_1^{(0)}$ is independent of (W_0, V), we have

$$\sqrt{n}(\chi_n^{(0)} + y) \mid W_0, V \sim N(0, A_n), \tag{8}$$

which implies

$$\mathbb{E}[\exp(iu^\top\sqrt{n}(\chi_n^{(0)} + y)) \mid W_0, V] = \exp\left(-\frac{1}{2}u^\top A_n u\right).$$

Thus we have

$$\begin{aligned}
&\mathbb{E}[\exp(iu^\top\sqrt{n}(\chi_n^{(0)} + y) + i\langle T, \sqrt{n}(A_n - K)\rangle_{\mathrm{F}})] \\
&= \mathbb{E}(\mathbb{E}[\exp(iu^\top\sqrt{n}(\chi_n^{(0)} + y) + i\langle T, \sqrt{n}(A_n - K)\rangle_{\mathrm{F}}) \mid W_0, V]) \\
&= \mathbb{E}\left[\exp\left(-\frac{1}{2}u^\top A_n u\right)\exp(i\langle T, \sqrt{n}(A_n - K)\rangle_{\mathrm{F}})\right] \\
&= \mathbb{E}\left[\left(\exp\left(-\frac{1}{2}u^\top A_n u\right) - \exp\left(-\frac{1}{2}u^\top K u\right)\right)\exp\left(i\langle T, \sqrt{n}(A_n - K)\rangle_{\mathrm{F}}\right)\right] \\
&\quad + \exp\left(-\frac{1}{2}u^\top K u\right)\mathbb{E}[\exp(i\langle T, \sqrt{n}(A_n - K)\rangle_{\mathrm{F}})].
\end{aligned}$$

Note that equation 7 implies $A_n \xrightarrow{p} K$. Thus we compute the limit of the first term as

$$\begin{aligned}
&\left|\mathbb{E}\left[\left(\exp\left(-\frac{1}{2}u^\top A_n u\right) - \exp\left(-\frac{1}{2}u^\top K u\right)\right)\exp\left(i\langle T, \sqrt{n}(A_n - K)\rangle_{\mathrm{F}}\right)\right]\right| \\
&\le \mathbb{E}\left|\exp\left(-\frac{1}{2}u^\top A_n u\right) - \exp\left(-\frac{1}{2}u^\top K u\right)\right| \\
&\to 0.
\end{aligned}$$

For the second term, equation 7 implies

$$\mathbb{E}[\exp(i\langle T, \sqrt{n}(A_n - K)\rangle_{\mathrm{F}})] \to \mathbb{E}[\exp(i\langle T, \Xi\rangle_{\mathrm{F}})].$$

Hence the joint characteristic function satisfies

$$\mathbb{E}[\exp(iu^\top\sqrt{n}(\chi_n^{(0)} + y) + i\langle T, \sqrt{n}(A_n - K)\rangle_{\mathrm{F}})] \to \exp\left(-\frac{1}{2}u^\top K u\right)\mathbb{E}[\exp(i\langle T, \Xi\rangle_{\mathrm{F}})],$$

which is the characteristic function of (Z, Ξ) with $Z \sim N(0, K)$ independent of Ξ. □

C.3 Infinite-width loss for fixed horizon

Corollary C.1 (Infinite-width loss for fixed T)**.** *Fix $T \in \mathbb{N}$ and $\eta \in \mathbb{R}$. Then*

$$\phi_{n,T}(\eta) \xrightarrow{p} \phi_{\infty,T}(\eta) := \frac{1}{2m} \left\| B_\infty(\eta)^T y \right\|^2 = \frac{1}{2m} y^\top B_\infty(\eta)^{2T} y.$$

The convergence is uniform on every deterministic compact interval of learning rates.

Proof. By Lemma 3.3, $(\chi_n^{(0)}, A_n) \xrightarrow{p} (-y, K)$. For fixed T, the map

$$(\chi, A, \eta) \longmapsto \frac{1}{2m} \chi^\top \left(I_m - \frac{\eta}{m} A \right)^{2T} \chi$$

is a polynomial in the entries of (χ, A, η). Thus, the continuous mapping theorem proves $\phi_{n,T}(\eta) \xrightarrow{p} \phi_{\infty,T}(\eta)$ for every fixed η. Furthermore, for fixed T, both $\phi_{n,T}$ and $\phi_{\infty,T}$ are polynomials in η of degree at most $2T$. Their coefficients $c_{n,0}, c_{n,1}, \ldots, c_{n,2T}$ are continuous functions of $(\chi_n^{(0)}, A_n)$ and converge in probability to the corresponding coefficients $c_{\infty,0}, c_{\infty,1}, \ldots, c_{\infty,2T}$ determined by $(-y, K)$. Thus, for any compact interval $[\eta_l, \eta_h]$ with $|\eta_l|, |\eta_h| \le C$, we have

$$\sup_{\eta \in [\eta_l, \eta_h]} |\phi_{n,T}(\eta) - \phi_{\infty,T}(\eta)| \le \sup_{\eta \in [\eta_l, \eta_h]} \sum_{k=0}^{2T} |c_{n,k} - c_{\infty,k}| C^k$$

$$\le \max_{k \in \{0,1,\ldots,2T\}} |c_{n,k} - c_{\infty,k}| \sum_{k=0}^{2T} C^k \xrightarrow{p} 0.$$

□

C.4 Range and null-space geometry

Lemma C.2. *When $n \ge d$, we have*

$$\ker(H_n^\top) \overset{\text{a.s.}}{=} \ker(A_n) \overset{\text{a.s.}}{=} \ker(K).$$

Proof. We first show that $\ker(H_n^\top) \overset{\text{a.s.}}{=} \ker(A_n)$ holds. If $H_n^\top z = 0$ then $A_n z = 0$ is directly satisfied. Thus, we have $\ker(H_n^\top) \subset \ker(A_n)$. On the other hand, if $A_n z = 0$, then $z^\top A_n z = \|V\|^2 \|H_n^\top z\|^2 = 0$. Since $\|V\|^2 > 0$ with probability one, $H_n^\top z \overset{\text{a.s.}}{=} 0$ holds. Thus, with probability one, we have $\ker(A_n) \subset \ker(H_n^\top)$.

When $n \ge d$, the Gaussian matrix W_0 has rank d almost surely. Thus,

$$\operatorname{range}(H_n) = \operatorname{range}(X W_0^\top) \overset{\text{a.s.}}{=} \operatorname{range}(X) = \operatorname{range}(K).$$

This implies $\ker(H_n^\top) = \operatorname{range}(H_n)^\perp \overset{\text{a.s.}}{=} \operatorname{range}(K)^\perp = \ker(K)$. □

C.5 Optimization landscape

Proposition C.3 (Derivatives of the finite-horizon objective)**.** *For every $T \ge 1$, we have*

$$\frac{\partial}{\partial \eta} \Phi_T(\eta; \chi, A) = -\frac{T}{m^2} F_T(\eta; \chi, A), \tag{9}$$

$$\frac{\partial^2}{\partial \eta^2} \Phi_T(\eta; \chi, A) = \frac{T(2T-1)}{m^3} \chi^\top A^2 B_A(\eta)^{2T-2} \chi \ge 0, \tag{10}$$

$$\frac{\partial}{\partial \eta} F_T(\eta; \chi, A) = -\frac{2T-1}{m} \chi^\top A^2 B_A(\eta)^{2T-2} \chi \le 0. \tag{11}$$

Proof. Since the matrices A and $B_A(\eta)$ commute, we have

$$\frac{d}{d\eta} B_A(\eta)^k = -\frac{k}{m} A B_A(\eta)^{k-1}.$$

Thus

$$\frac{\partial}{\partial \eta}\Phi_T(\eta;\chi,A) = -\frac{1}{2m}\chi^\top \frac{2T}{m} A B_A(\eta)^{2T-1}\chi = -\frac{T}{m^2}F_T(\eta;\chi,A),$$
$$\frac{\partial}{\partial \eta}F_T(\eta;\chi,A) = -\chi^\top A \frac{2T-1}{m} A B_A(\eta)^{2T-2}\chi = -\frac{2T-1}{m}\chi^\top A^2 B_A(\eta)^{2T-2}\chi.$$

□

Corollary C.4 (Existence and uniqueness of the optimal learning rate)**.** *Suppose $A \in \mathbb{R}^{m\times m}_{\mathrm{sym}}$, $\chi \in \mathbb{R}^m$, and $A \succeq 0$. If $A\chi \neq 0$, then $\eta \mapsto \Phi_T(\eta;\chi,A)$ is strictly convex and coercive on $\mathbb{R}$, that is, $\Phi_T(\eta;\chi,A) \to \infty$ as $|\eta| \to \infty$. Hence it has a unique global minimizer $\eta_T(\chi,A) > 0$, characterized by $F_T(\eta_T(\chi,A);\chi,A) = 0$. On a closed interval $[a,b]$, the optimizer is either $\eta_T(\chi,A)$ if $\eta_T(\chi,A) \in [a,b]$, and otherwise is the appropriate boundary point. If $A\chi = 0$, then $\Phi_T(\eta;\chi,A)$ is constant in η, and every $\eta \in \mathbb{R}$ is a global minimizer.*

Proof. If $A\chi = 0$, then $B_A(\eta)\chi = \chi$ holds for every $\eta \in \mathbb{R}$. Hence $\Phi_T(\eta;\chi,A)$ is independent of η, which proves the degenerate case.

In the remainder of the proof, suppose $A\chi \neq 0$. Suppose $A = U\,\mathrm{diag}(\mu_1,\dots,\mu_m)U^\top$ is the eigendecomposition of A, and let $q_j = (U_{\cdot j})^\top\chi$. Define the active index set

$$\mathcal{J} = \{j \in [m] : \mu_j > 0 \text{ and } q_j \neq 0\}.$$

Since $A\chi = \sum_{j=1}^m \mu_j q_j U_{\cdot j}$, the assumption $A\chi \neq 0$ is equivalent to $\mathcal{J} \neq \emptyset$.

By Proposition C.3, it holds that

$$\Phi_T(\eta;\chi,A) = \frac{1}{2m}\sum_{j=1}^m q_j^2\left(1-\frac{\eta\mu_j}{m}\right)^{2T} = \sum_{j=1}^m g_j(\eta),$$

where we defined

$$g_j(\eta) = \frac{q_j^2}{2m}\left(1-\frac{\eta\mu_j}{m}\right)^{2T}.$$

Here, g_j corresponding to $j \notin \mathcal{J}$ are either identically zero or constant in η. For each $j \in \mathcal{J}$, we have

$$g_j'(\eta) = -\frac{Tq_j^2\mu_j}{m^2}\left(1-\frac{\eta\mu_j}{m}\right)^{2T-1}.$$

The map $\eta \mapsto g_j'(\eta)$ is strictly increasing; thus, g_j is strictly convex on $\mathbb{R}$. Since $\mathcal{J}$ is not an empty set, $\Phi_T(\eta;\chi,A)$ is strictly convex. Moreover, for any fixed $j \in \mathcal{J}$, it holds that

$$g_j(\eta) \to \infty \qquad (|\eta| \to \infty).$$

Since $\Phi_T(\eta;\chi,A) \geq g_j(\eta)$, it follows that

$$\Phi_T(\eta;\chi,A) \to \infty \qquad (|\eta| \to \infty).$$

Thus $\Phi_T(\cdot;\chi,A)$ is coercive. In particular, there exists $R > 0$ such that

$$\Phi_T(\eta;\chi,A) > \Phi_T(0;\chi,A) \qquad (|\eta| \geq R).$$

Since $\Phi_T(\cdot;\chi,A)$ is continuous, it attains its minimum on the compact interval $[-R,R]$. This minimum is also global, because every point outside $[-R,R]$ has objective value larger than $\Phi_T(0;\chi,A)$. Finally, strict convexity implies that the global minimizer is unique. We denote it by $\eta_T(\chi,A)$. The first-order optimality condition and equation 9 give

$$F_T(\eta_T(\chi,A);\chi,A) = 0.$$

We now verify the positivity of $\eta_T(\chi,A)$. At $\eta = 0$, equation 9 yields

$$\Phi_T'(0;\chi,A) = -\frac{T}{m^2}\chi^\top A\chi = -\frac{T}{m^2}\|A^{1/2}\chi\|^2.$$

Note that $A\chi \neq 0$ implies $A^{1/2}\chi \neq 0$. Consequently, we have

$$\Phi_T'(0;\chi,A) < 0 = \Phi_T'(\eta_T(\chi,A);\chi,A).$$

Since $\Phi_T(\cdot;\chi,A)$ is strictly convex, its derivative is strictly increasing. Thus we have $\eta_T(\chi,A)>0$. Finally, consider minimization over a closed interval $[a,b]$ with $a<b$. By Proposition C.3, we have

$$F_T(\eta;\chi,A)=\sum_{j\in\mathcal{J}} q_j^2\mu_j\left(1-\frac{\eta\mu_j}{m}\right)^{2T-1}.$$

For each $j\in\mathcal{J}$, the corresponding summand is strictly decreasing in η. Hence F_T is strictly decreasing on $\mathbb{R}$. Since it is also continuous, we obtain the characterization at the boundaries. □

Proposition C.5 (Almost-sure nondegeneracy of the finite-width optimization problem)**.** *Suppose $K\neq 0$ and $n\geq d$. Then we have $A_n\chi_n^{(0)}\neq 0$ almost surely. By Corollary C.4, for every $T\geq 1$, $\phi_{n,T}$ is almost surely strictly convex and has a unique global minimizer.*

Proof. Let P denote the orthogonal projector onto $\mathrm{range}(K)$. Define $\mathcal{E}_n=\{\ker(H_n^\top)=\ker(A_n)=\ker(K)\}$. Then by Lemma C.2, the event $\mathcal{E}_n$ has probability one. On the event $\mathcal{E}_n$, we have $f_n^{(0)}(X)\in\mathrm{range}(K)$, and hence $Pf_n^{(0)}(X)=f_n^{(0)}(X)$. On the same event, $A_n\chi_n^{(0)}=0$ is equivalent to $\chi_n^{(0)}\in\ker(A_n)=\mathrm{range}(K)^\perp$, and therefore to $P\chi_n^{(0)}=0$. Recalling the definition $\chi_n^{(0)}=f_n^{(0)}(X)-y$, this is further equivalent to $f_n^{(0)}(X)=Pf_n^{(0)}(X)=Py$. Noting that $\mathrm{P}(\mathcal{E}_n)=1$, we have

$$\mathrm{P}(A_n\chi_n^{(0)}=0)=\mathrm{P}(\{A_n\chi_n^{(0)}=0\}\cap\mathcal{E}_n)=\mathrm{P}(\{f_n^{(0)}(X)=Py\}\cap\mathcal{E}_n)=\mathrm{P}(f_n^{(0)}(X)=Py).$$

It therefore suffices to show that $\mathrm{P}(f_n^{(0)}(X)=Py)=0$.

By equation 8 in the proof of Lemma 3.3, we have

$$f_n^{(0)}(X)\mid W_0,V\sim\mathcal{N}\left(0,\frac{1}{n}A_n\right).$$

Since $K\neq 0$, we can choose a fixed vector $u\in\mathrm{range}(K)\setminus\{0\}$. By Lemma C.2, we have $u\notin\ker(A_n)$ almost surely. Since $A_n\succeq 0$, it follows that $u^\top A_n u>0$ holds almost surely. Hence,

$$u^\top f_n^{(0)}(X)\mid W_0,V\sim\mathcal{N}\left(0,\frac{1}{n}u^\top A_n u\right)$$

is a nondegenerate Gaussian random variable almost surely. Therefore,

$$\mathrm{P}(u^\top f_n^{(0)}(X)=u^\top Py\mid W_0,V)\overset{\text{a.s.}}{=}0.$$

Since $\{f_n^{(0)}(X)=Py\}\subset\{u^\top f_n^{(0)}(X)=u^\top Py\}$, we obtain

$$\mathrm{P}(f_n^{(0)}(X)=Py\mid W_0,V)\overset{\text{a.s.}}{=}0.$$

Taking expectations yields $\mathrm{P}(f_n^{(0)}(X)=Py)=0$ as desired. □

C.6 First-order sensitivities of the score and the loss

Set $B_T=B_\infty(\eta_{\infty,T})$, and define

$$g_{F,T}:=-2KB_T^{2T-1}y, \tag{12}$$

$$Q_{F,T}:=\frac{1}{2}\left(B_T^{2T-1}yy^\top+yy^\top B_T^{2T-1}\right)-\frac{\eta_{\infty,T}}{2m}\sum_{s=0}^{2T-2}(B_T^{2T-2-s}yy^\top KB_T^s+B_T^sKyy^\top B_T^{2T-2-s}), \tag{13}$$

$$g_{L,T}:=-\frac{1}{m}B_T^{2T}y, \tag{14}$$

$$Q_{L,T}:=-\frac{\eta_{\infty,T}}{2m^2}\sum_{s=0}^{2T-1}B_T^{2T-1-s}yy^\top B_T^s. \tag{15}$$

Lemma C.6 (First-order sensitivities at the infinite-width optimum)**.** *For fixed η, regard $F_T(\eta;\chi,A)$ and $\Phi_T(\eta;\chi,A)$ as real-valued maps on $\mathbb{R}^m\times\mathbb{R}^{m\times m}_{\mathrm{sym}}$, equipped with the product norm $\|(h,E)\|_\times = \left(\|h\|^2+\|E\|_{\mathrm{F}}^2\right)^{1/2}$. Then for $h\in\mathbb{R}^m$ and $E\in\mathbb{R}^{m\times m}_{\mathrm{sym}}$, we have*

$$D_{(\chi,A)}F_T(\eta_{\infty,T};-y,K)[h,E] = g_{F,T}^\top h + \langle Q_{F,T},E\rangle_{\mathrm{F}},$$
$$D_{(\chi,A)}\Phi_T(\eta_{\infty,T};-y,K)[h,E] = g_{L,T}^\top h + \langle Q_{L,T},E\rangle_{\mathrm{F}}.$$

Proof. The derivative of $B_A(\eta)$ with respect to A in the direction E is

$$D_A B_A(\eta)[E] = -\frac{\eta}{m}E.$$

We also note that by Lemma F.1, we have

$$D_A B_A(\eta)^k[E] = -\frac{\eta}{m}\sum_{s=0}^{k-1} B_A(\eta)^s E B_A(\eta)^{k-1-s}. \tag{16}$$

We first consider the loss $\Phi_T(\eta;\chi,A)$. Since $B_A(\eta)$ is symmetric for symmetric A, differentiation with respect to χ gives

$$D_\chi\Phi_T(\eta_{\infty,T};-y,K)[h] = -\frac{1}{m}y^\top B_T^{2T}h = g_{L,T}^\top h.$$

Using equation 16 with $k=2T$, we have

$$D_A\Phi_T(\eta_{\infty,T};-y,K)[E] = -\frac{\eta_{\infty,T}}{2m^2}\sum_{s=0}^{2T-1} y^\top B_T^s E B_T^{2T-1-s}y = \langle Q_{L,T},E\rangle_{\mathrm{F}}.$$

Combining the two partial derivatives gives the Fréchet derivative with respect to the pair (χ,A). [3] Next, since A commutes with $B_A(\eta)$, the matrix $AB_A(\eta)^{2T-1}$ is symmetric. Hence we have

$$D_\chi F_T(\eta_{\infty,T};-y,K)[h] = -2y^\top K B_T^{2T-1}h = g_{F,T}^\top h.$$

For the derivative with respect to A, equation 16 gives

$$D_A F_T(\eta_{\infty,T};-y,K)[E] = y^\top E B_T^{2T-1}y - \frac{\eta_{\infty,T}}{m}\sum_{s=0}^{2T-2} y^\top K B_T^s E B_T^{2T-2-s}y = \langle Q_{F,T},E\rangle_{\mathrm{F}}.$$

Combining the results completes the proof. □

D Proof of the fixed-horizon transfer theorem

Throughout this section, T is fixed. Define

$$r_{j,T} = 1-\frac{\eta_{\infty,T}\lambda_j}{m} \qquad (j\in[r]). \tag{17}$$

Lemma D.1. *Under the assumptions of Theorem 3.6, we have for every fixed T that*

$$F'_{\infty,T}(\eta_{\infty,T}) < 0, \qquad \phi''_{\infty,T}(\eta_{\infty,T}) > 0.$$

[3] To see this explicitly, Lemma F.1 implies

$$B_{K+E}(\eta_{\infty,T})^{2T} = B_T^{2T} + D_A B_K(\eta_{\infty,T})^{2T}[E] + O(\|E\|_{\mathrm{F}}^2)$$

for fixed T. Substituting this expansion into $\Phi_T(\eta_{\infty,T};-y+h,K+E)$ shows

$$|\Phi_T(\eta_{\infty,T};-y+h,K+E) - \Phi_T(\eta_{\infty,T};-y,K) - D_\chi\Phi_T(\eta_{\infty,T};-y,K)[h] - D_A\Phi_T(\eta_{\infty,T};-y,K)[E]|$$
$$= O(\|h\|^2 + \|h\|\|E\|_{\mathrm{F}} + \|E\|_{\mathrm{F}}^2) = O(\|(h,E)\|_\times^2) = o(\|(h,E)\|_\times).$$

Proof. Recall that $c_j = \|P_j y\|^2$. Proposition C.3 implies

$$-F'_{\infty,T}(\eta_{\infty,T}) = \frac{2T-1}{m} y^\top K^2 B_\infty(\eta_{\infty,T})^{2T-2} y = \frac{2T-1}{m} \sum_{j=1}^{r} c_j \lambda_j^2 r_{j,T}^{2T-2}.$$

If $T = 1$, since $Ky \neq 0$, we have $y^\top K^2 B_\infty(\eta_{\infty,T})^{2T-2} y > 0$. If $T \geq 2$, since $|\mathcal{J}_y| \geq 2$, at least one of $r_{j,T}$ is nonzero. Thus the preceding sum is strictly positive. Hence in all cases, we have $F'_{\infty,T}(\eta_{\infty,T}) < 0$, and consequently, $\phi''_{\infty,T}(\eta_{\infty,T}) = -\frac{T}{m^2} F'_{\infty,T}(\eta_{\infty,T}) > 0$. □

Lemma D.2. *Since $\eta_{\infty,T} \in (\eta_\ell, \eta_u)$, there exists a deterministic $\delta > 0$ such that*

$$[\eta_{\infty,T} - \delta, \eta_{\infty,T} + \delta] \subset (\eta_\ell, \eta_u).$$

For such δ, we have

$$\sup_{|\eta - \eta_{\infty,T}| \leq \delta} |F_{n,T}(\eta) - F_{\infty,T}(\eta)| = O_p(n^{-1/2}),$$
$$\sup_{|\eta - \eta_{\infty,T}| \leq \delta} |F'_{n,T}(\eta) - F'_{\infty,T}(\eta)| = O_p(n^{-1/2}).$$

Proof. For fixed T, both $F_T(\eta; \chi, A)$ and $\partial_\eta F_T(\eta; \chi, A)$ are polynomials in the entries of (η, χ, A). Hence, on the compact interval $[\eta_{\infty,T} - \delta, \eta_{\infty,T} + \delta]$, their derivatives with respect to (χ, A) are uniformly bounded on a fixed neighborhood of $(-y, K)$. Define the product norm $\|(\chi, A)\|_\times = \left(\|\chi\|^2 + \|A\|_{\mathrm{F}}^2\right)^{1/2}$. Then Lemma 3.3 gives $\|(f_n^{(0)}(X), E_n)\|_\times = O_p(n^{-1/2})$, where we defined $E_n = A_n - K$. Thus, noting that $(\chi_n^{(0)}, A_n) \xrightarrow{p} (-y, K)$ holds, we have

$$\begin{aligned}
&\sup_{|\eta - \eta_{\infty,T}| \leq \delta} |F_{n,T}(\eta) - F_{\infty,T}(\eta)| \\
&= \sup_{|\eta - \eta_{\infty,T}| \leq \delta} \left| \int_0^1 D_{\chi,A} F_T(\eta; -y + s f_n^{(0)}(X), K + sE_n)[f_n^{(0)}(X), E_n] ds \right| \\
&\leq \sup_{|\eta - \eta_{\infty,T}| \leq \delta} \int_0^1 \|D_{\chi,A} F_T(\eta; -y + s f_n^{(0)}(X), K + sE_n)\|_{\mathrm{op}} \|(f_n^{(0)}(X), E_n)\|_\times ds \\
&= O_p(n^{-1/2}),
\end{aligned}$$

and, similarly,

$$\sup_{|\eta - \eta_{\infty,T}| \leq \delta} |F'_{n,T}(\eta) - F'_{\infty,T}(\eta)| = O_p(n^{-1/2}).$$

□

Lemma D.3. *Under the assumptions of Theorem 3.6, $\eta_{n,T}$ lies in (η_ℓ, η_u) with probability tending to one, and we have*

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

Proof. By Lemmas D.1, D.2, and continuity of $F'_{\infty,T}$, there exists a deterministic $\gamma > 0$ such that

$$\sup_{|\eta - \eta_{\infty,T}| \leq \delta} F'_{\infty,T}(\eta) \leq -2\gamma.$$

Hence we have

$$\mathrm{P}\left(\sup_{|\eta - \eta_{\infty,T}| \leq \delta} F'_{n,T}(\eta) \leq -\gamma \right) \to 1.$$

Since Corollary C.4 implies $F_{\infty,T}(\eta_{\infty,T}) = 0$, Lemma D.2 yields

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

Thus, for fixed $\epsilon > 0$, there exists $M_0 < \infty$ such that

$$\liminf_{n \to \infty} \mathrm{P}\left(|F_{n,T}(\eta_{\infty,T})| \leq \frac{M_0}{\sqrt{n}} \right) \geq 1 - \epsilon.$$

Choose $M > M_0/\gamma$. For all sufficiently large n, we have $M/\sqrt{n} < \delta$. On the intersection of the two preceding high-probability events, the mean value theorem gives

$$F_{n,T}\left(\eta_{\infty,T} - \frac{M}{\sqrt{n}}\right) \geq \frac{-M_0 + \gamma M}{\sqrt{n}} > 0,$$
$$F_{n,T}\left(\eta_{\infty,T} + \frac{M}{\sqrt{n}}\right) \leq \frac{M_0 - \gamma M}{\sqrt{n}} < 0.$$

Proposition C.5 and Corollary C.4 imply that, for all sufficiently large n, $\phi_{n,T}$ has almost surely a unique global minimizer characterized by the root of $F_{n,T}$. Therefore, on the same event, this root belongs to

$$\left(\eta_{\infty,T} - \frac{M}{\sqrt{n}}, \eta_{\infty,T} + \frac{M}{\sqrt{n}}\right) \subset (\eta_\ell, \eta_u).$$

Thus, with probability at least $1 - \epsilon + o(1)$, we have

$$|\eta_{n,T} - \eta_{\infty,T}| < \frac{M}{\sqrt{n}}.$$

Since $\epsilon > 0$ was arbitrary, we obtain

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

and the finite-width optimizer is interior with probability tending to one. □

Theorem D.4 (Limit distribution of the optimal learning rate at fixed T)**.** *Under the assumptions of Theorem 3.6, we have*

$$\sqrt{n}(\eta_{n,T} - \eta_{\infty,T}) \xrightarrow{d} \Omega_{\eta,T} := -\Omega_{F,T}/F'_{\infty,T}(\eta_{\infty,T}).$$

Proof. By Lemma C.6, applied at $(\eta_{\infty,T}, -y, K)$, we have

$$F_{n,T}(\eta_{\infty,T}) = g_{F,T}^\top f_n^{(0)}(X) + \langle Q_{F,T}, A_n - K\rangle_{\mathrm{F}} + o_p(n^{-1/2}).$$

Here we used $F_{\infty,T}(\eta_{\infty,T}) = 0$ by Corollary C.4. Combining this with Lemma 3.3 gives

$$\sqrt{n}F_{n,T}(\eta_{\infty,T}) \xrightarrow{d} \Omega_{F,T}. \tag{18}$$

By the mean value theorem, there exists a random $\tilde{\eta}_{n,T}$ between $\eta_{n,T}$ and $\eta_{\infty,T}$ such that

$$F_{n,T}(\eta_{n,T}) - F_{n,T}(\eta_{\infty,T}) = F'_{n,T}(\tilde{\eta}_{n,T})(\eta_{n,T} - \eta_{\infty,T}).$$

On the event $\{\eta_{n,T} \in (\eta_\ell, \eta_u)\}$, by Corollary C.4, we have $F_{n,T}(\eta_{n,T}) = 0$. Moreover, on the same event, Lemma D.2 implies $F'_{n,T}(\tilde{\eta}_{n,T}) = F'_{\infty,T}(\eta_{\infty,T}) + O_p(n^{-1/2})$. Thus by Lemma D.3, we have

$$F'_{n,T}(\tilde{\eta}_{n,T}) = F'_{\infty,T}(\eta_{\infty,T}) + o_p(1)$$

unconditionally. By Lemma D.1, this further implies that $F'_{n,T}(\tilde{\eta}_{n,T}) < 0$ holds with probability tending to one. Thus, on the intersection of this event and $\{\eta_{n,T} \in (\eta_\ell, \eta_u)\}$, both of which have probability tending to one, we have

$$\sqrt{n}(\eta_{n,T} - \eta_{\infty,T}) = -\frac{\sqrt{n}F_{n,T}(\eta_{\infty,T})}{F'_{n,T}(\tilde{\eta}_{n,T})}.$$

Finally, by equation 18, we obtain

$$\sqrt{n}(\eta_{n,T} - \eta_{\infty,T}) \xrightarrow{d} -\frac{\Omega_{F,T}}{F'_{\infty,T}(\eta_{\infty,T})} = \Omega_{\eta,T}.$$

□

Theorem D.5 (Limit distribution of the optimal loss at fixed T)**.** *Under the assumptions of Theorem 3.6, we have*

$$\sqrt{n}\left(\phi_{n,T}(\eta_{n,T}) - \phi_{\infty,T}(\eta_{\infty,T})\right) \xrightarrow{d} \Omega_{L,T} := g_{L,T}^\top Z + \langle Q_{L,T}, \Xi\rangle_{\mathrm{F}}.$$

Proof. By Lemma C.6, we have

$$\phi_{n,T}(\eta_{\infty,T}) - \phi_{\infty,T}(\eta_{\infty,T}) = g_{L,T}^\top f_n^{(0)}(X) + \langle Q_{L,T}, A_n - K\rangle_{\mathrm{F}} + o_p(n^{-1/2}).$$

Hence Lemma 3.3 yields

$$\sqrt{n}(\phi_{n,T}(\eta_{\infty,T}) - \phi_{\infty,T}(\eta_{\infty,T})) \xrightarrow{d} \Omega_{L,T}. \tag{19}$$

It remains to replace $\phi_{n,T}(\eta_{\infty,T})$ by $\phi_{n,T}(\eta_{n,T})$. By Taylor's theorem, there exists a random $\bar{\eta}_{n,T}$ between $\eta_{n,T}$ and $\eta_{\infty,T}$ such that

$$\phi_{n,T}(\eta_{\infty,T}) - \phi_{n,T}(\eta_{n,T}) = \phi'_{n,T}(\eta_{n,T})(\eta_{\infty,T} - \eta_{n,T}) + \frac{1}{2}\phi''_{n,T}(\bar{\eta}_{n,T})(\eta_{\infty,T} - \eta_{n,T})^2$$

On the event $\{\eta_{n,T} \in (\eta_\ell, \eta_u)\}$, Proposition C.3 gives $\phi'_{n,T}(\eta_{n,T}) = 0$. Thus, on the same event, we have

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

Since T is fixed, equation 10, Lemma 3.3, and Lemma D.3 imply $\phi''_{n,T}(\bar{\eta}_{n,T}) = O_p(1)$. Therefore, combining this with Lemma D.3, we have

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

with probability tending to one. Together with equation 19, this yields

$$\sqrt{n}(\phi_{n,T}(\eta_{n,T}) - \phi_{\infty,T}(\eta_{\infty,T})) \xrightarrow{d} \Omega_{L,T}.$$

□

Lemma D.6. *Under the assumptions of Theorem 3.6, we have*

$$c_{n,T} = \frac{\phi''_{\infty,T}(\eta_{\infty,T})}{2}(\eta_{n,T} - \eta_{\infty,T})^2(1 + o_p(1)).$$

Proof. We next expand the transfer gap. Since $\eta_{\infty,T}$ is the unique global minimizer, Proposition C.3 gives $\phi'_{\infty,T}(\eta_{\infty,T}) = 0$. By Taylor's theorem, there exists a random $\eta^\dagger_{n,T}$ between $\eta_{n,T}$ and $\eta_{\infty,T}$ such that

$$c_{n,T} = \frac{\phi''_{\infty,T}(\eta_{\infty,T})}{2}(\eta_{n,T} - \eta_{\infty,T})^2 + \frac{\phi'''_{\infty,T}(\eta^\dagger_{n,T})}{6}(\eta_{n,T} - \eta_{\infty,T})^3.$$

For fixed T, by Lemma D.3, with probability tending to one, we have

$$|\phi'''_{\infty,T}(\eta^\dagger_{n,T})| \le \sup_{\eta_\ell \le \eta \le \eta_u} |\phi'''_{\infty,T}(\eta)|.$$

Therefore, by Lemmas D.1 and D.3, we have

$$c_{n,T} = \frac{\phi''_{\infty,T}(\eta_{\infty,T})}{2}(\eta_{n,T} - \eta_{\infty,T})^2(1 + o_p(1)).$$

□

Proof of Theorem 3.6. We verify that the limiting Gaussian variables $\Omega_{L,T}$ and $\Omega_{\eta,T}$ of $a_{n,T}$ and $b_{n,T}$ are nondegenerate. We first consider $\Omega_{\eta,T}$. For $s > 0$, direct substitution into the definition of F_T gives

$$F_T(\eta_{\infty,T}; -y, sK) = sF_{\infty,T}(s\eta_{\infty,T}).$$

Differentiating both sides with respect to s at $s = 1$, applying Lemma C.6, and using $F_{\infty,T}(\eta_{\infty,T}) = 0$, we have

$$\langle Q_{F,T}, K\rangle_{\mathrm{F}} = F_{\infty,T}(\eta_{\infty,T}) + \eta_{\infty,T}F'_{\infty,T}(\eta_{\infty,T}) = \eta_{\infty,T}F'_{\infty,T}(\eta_{\infty,T}).$$

Recall by Lemma 3.3 that $\Xi = \alpha K + XGX^\top$ with $\alpha \sim \mathcal{N}(0, 2)$, where α is independent of G and Z. Noting that Corollary C.4 gives $\eta_{\infty,T} > 0$, the contribution of the αK component of Ξ to $\Omega_{\eta,T}$ is

$$-\frac{\alpha\langle Q_{F,T}, K\rangle_{\mathrm{F}}}{F'_{\infty,T}(\eta_{\infty,T})} = -\eta_{\infty,T}\alpha \sim \mathcal{N}(0, 2\eta^2_{\infty,T}).$$

Since this component is independent of the remaining Gaussian terms, $\Omega_{\eta,T}$ is nondegenerate. By Theorem D.4 and the continuous mapping theorem, this proves

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

Next, we consider $\Omega_{L,T}$. By the definition of $g_{L,T}$, we have

$$g_{L,T}^\top K g_{L,T} = \frac{1}{m^2} \sum_{j=1}^{r} c_j \lambda_j r_{j,T}^{4T},$$

where $r_{j,T}$ are defined in equation 17. Under $|\mathcal{J}_y| \geq 2$, at least one of the $r_{j,T}$ is nonzero. Hence we have $g_{L,T}^\top K g_{L,T} > 0$. Since $Z \sim \mathcal{N}(0, K)$ is independent of Ξ by Lemma 3.3, we have

$$\mathrm{Var}(\Omega_{L,T}) = \mathrm{Var}(g_{L,T}^\top Z) + \mathrm{Var}(\langle Q_{L,T}, \Xi \rangle_{\mathrm{F}}) = g_{L,T}^\top K g_{L,T} + \mathrm{Var}(\langle Q_{L,T}, \Xi \rangle_{\mathrm{F}}) > 0.$$

Therefore, $\Omega_{L,T}$ is also nondegenerate. By Theorem D.5 and the continuous mapping theorem, this proves

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

Finally, by Lemmas D.6 and D.1, we obtain

$$c_{n,T} = \Theta_p(b_{n,T}^2) = \Theta_p(n^{-1}).$$

□

Proof of Corollary 3.7. By Theorems D.5 and the continuous mapping theorem, we have

$$\sqrt{n} a_{n,T} \xrightarrow{d} |\Omega_{L,T}|,$$

where $\Omega_{L,T}$ is a nondegenerate Gaussian random variable. Since $\mathrm{P}(|\Omega_{L,T}| = 0) = 0$, another application of the continuous mapping theorem gives

$$\frac{1}{\sqrt{n} a_{n,T}} \xrightarrow{d} \frac{1}{|\Omega_{L,T}|}.$$

Therefore, we have

$$\frac{1}{a_{n,T}} = O_p(\sqrt{n}).$$

Combining this with $c_{n,T} = \Theta_p(n^{-1})$ in Theorem 3.6 yields the claim. □

E Infinite-width large-horizon analysis

E.1 Proof of Theorem 3.1

Note that we have

$$\begin{aligned}
\phi_{\infty,T}(\eta) &= \frac{1}{2m} \left\| \left(I_m - \frac{\eta}{m} K \right)^T y \right\|^2 = \frac{1}{2m} \left\| \left[P_0 + \sum_{j=1}^{r} \left(1 - \frac{\eta \lambda_j}{m} \right) P_j \right]^T y \right\|^2 \\
&= \frac{1}{2m} \left\| P_0 y + \sum_{j=1}^{r} \left(1 - \frac{\eta \lambda_j}{m} \right)^T P_j y \right\|^2 = \frac{c_0}{2m} + \frac{1}{2m} \sum_{j=1}^{r} c_j \left(1 - \frac{\eta \lambda_j}{m} \right)^{2T}.
\end{aligned}$$

Here, $c_0/(2m)$ is an irreducible training loss. No choice of learning rate or number of steps changes it. Define

$$R_T(\eta) := (2m \phi_{\infty,T}(\eta) - c_0)^{1/(2T)} = \left(\sum_{j=1}^{r} c_j \left| 1 - \frac{\eta \lambda_j}{m} \right|^{2T} \right)^{1/(2T)}.$$

Then we have

$$\arg\min_{\eta\geq 0}\phi_{\infty,T}(\eta)=\arg\min_{\eta\geq 0}R_T(\eta).$$

Define

$$R_\infty(\eta):=\max_{j\in\mathcal{J}_y}\left|1-\frac{\eta\lambda_j}{m}\right|=\left|1-\frac{\eta\lambda_-^y}{m}\right|\vee\left|1-\frac{\eta\lambda_+^y}{m}\right|.$$

Choose an index $j_*\in\mathcal{J}_y$ such that $|1-\eta\lambda_{j_*}/m|=R_\infty(\eta)$. Then we have

$$c_{j_*}R_\infty(\eta)^{2T}\leq\sum_{j\in\mathcal{J}_y}c_j\left|1-\frac{\eta\lambda_j}{m}\right|^{2T}\leq\left(\sum_{j=1}^r c_j\right)R_\infty(\eta)^{2T},$$

which implies

$$\min_{j\in\mathcal{J}_y}c_j^{1/(2T)}R_\infty(\eta)\leq R_T(\eta)\leq\left(\sum_{j=1}^r c_j\right)^{1/(2T)}R_\infty(\eta).$$

Since $\mathcal{J}_y$ is finite and all active coefficients are strictly positive, we have

$$\min_{j\in\mathcal{J}_y}c_j^{1/(2T)}\to 1,\qquad\left(\sum_{j=1}^r c_j\right)^{1/(2T)}\to 1.$$

Thus, on every compact interval $\mathcal{I}$, we have

$$\sup_{\eta\in\mathcal{I}}|R_T(\eta)-R_\infty(\eta)|\leq\left(\left|\min_{j\in\mathcal{J}_y}c_j^{1/(2T)}-1\right|\vee\left|\left(\sum_{j=1}^r c_j\right)^{1/(2T)}-1\right|\right)\sup_{\eta\in\mathcal{I}}|R_\infty(\eta)|\to 0.$$

It remains to verify that the global minimizers of R_T are contained in a fixed compact interval. Fix any $j_0\in\mathcal{J}_y$. Since $\eta_{\infty,T}$ is a global minimizer over $[0,\infty)$, we have

$$R_T(\eta_{\infty,T})\leq R_T(0)=\left(\sum_{j\in\mathcal{J}_y}c_j\right)^{1/(2T)}.$$

On the other hand, for any $\eta\geq 0$, we have

$$R_T(\eta)\geq c_{j_0}^{1/(2T)}\left|1-\frac{\eta\lambda_{j_0}}{m}\right|.$$

Therefore we have

$$\left|1-\frac{\eta_{\infty,T}\lambda_{j_0}}{m}\right|\leq c_{j_0}^{-1/(2T)}R_T(\eta_{\infty,T})\leq\left(\frac{\sum_{j\in\mathcal{J}_y}c_j}{c_{j_0}}\right)^{1/(2T)}\leq\left(\frac{\sum_{j\in\mathcal{J}_y}c_j}{c_{j_0}}\right)^{1/2}.$$

Hence there exists a deterministic constant $M<\infty$, independent of T, such that $\eta_{\infty,T}\in[0,M]$ for every $T\geq 1$. Enlarging M if necessary, we may also assume $\eta_*^y\in[0,M]$. Since $\sup_{\eta\in[0,M]}|R_T(\eta)-R_\infty(\eta)|\to 0$ holds and R_∞ has the unique minimizer η_*^y, we have

$$\eta_{\infty,T}\to\eta_*^y.$$

If $\lambda_+^y=\lambda_-^y=\lambda$, the loss is a positive multiple of $(1-\eta\lambda/m)^{2T}$ plus a constant, whose unique minimizer is m/λ. □

E.2 Expansion of the infinite-width optimizer and asymptotic loss

Proposition E.1 (First-order large-T expansion of the optimal learning rate)**.** *Suppose Assumption 3.2 holds and define*

$$\bar{\vartheta} := \frac{1}{q_*} \max_{2\le j\le r-1} \left|1 - \frac{\eta_* \lambda_j}{m}\right|,$$

where $\max \emptyset := 0$*. Then* $0 \le \bar{\vartheta} < 1$*. Fix any constant* ϑ *satisfying* $\bar{\vartheta} < \vartheta < 1$*. Then, as* $T \to \infty$*, we have*

$$\eta_{\infty,T} = \eta_* - \frac{m q_*}{(\lambda_1 + \lambda_r)(2T-1)} L_* + O(T^{-2}) + O\left(\frac{\vartheta^{2T}}{T}\right).$$

Proof. Near η_*, define the positive endpoint magnitudes

$$r_1(\eta) = \frac{\eta \lambda_1}{m} - 1, \qquad r_r(\eta) = 1 - \frac{\eta \lambda_r}{m}.$$

Then we have $r_1(\eta_*) = r_r(\eta_*) = q_* > 0$. By the definition of $\bar{\vartheta}$, we have

$$\max_{2\le j\le r-1} \frac{|1 - \eta_* \lambda_j / m|}{r_1(\eta_*)} = \bar{\vartheta} < \vartheta.$$

Since Theorem 3.1 gives $\eta_{\infty,T} \to \eta_*$, for all sufficiently large T, we have

$$\max_{2\le j\le r-1} \frac{|1 - \eta_{\infty,T} \lambda_j / m|}{r_1(\eta_{\infty,T})} \le \vartheta. \tag{20}$$

By Corollary C.4, the score equation is given by

$$\begin{aligned} 0 = F_{\infty,T}(\eta_{\infty,T}) &= y^\top K B_\infty(\eta_{\infty,T})^{2T-1} y = y^\top K \left(I_m - \frac{\eta_{\infty,T}}{m} K\right)^{2T-1} y \\ &= \sum_{j=1}^{r} c_j \lambda_j \left(1 - \frac{\lambda_j \eta_{\infty,T}}{m}\right)^{2T-1} \\ &= -c_1 \lambda_1 (r_1(\eta_{\infty,T}))^{2T-1} + c_r \lambda_r r_r(\eta_{\infty,T})^{2T-1} + \sum_{j=2}^{r-1} c_j \lambda_j \left(1 - \frac{\lambda_j \eta_{\infty,T}}{m}\right)^{2T-1}. \end{aligned}$$

Hence we have

$$|c_r \lambda_r r_r(\eta_{\infty,T})^{2T-1} - c_1 \lambda_1 r_1(\eta_{\infty,T})^{2T-1}| \le \sum_{j=2}^{r-1} c_j \lambda_j \left|1 - \frac{\lambda_j \eta_{\infty,T}}{m}\right|^{2T-1}.$$

Using equation 20, we obtain

$$\left| c_r \lambda_r \left(\frac{r_r(\eta_{\infty,T})}{r_1(\eta_{\infty,T})}\right)^{2T-1} - c_1 \lambda_1 \right| \le \left(\sum_{j=2}^{r-1} c_j \lambda_j\right) \vartheta^{2T-1} = O(\vartheta^{2T-1}).$$

Therefore, we have

$$\left(\frac{r_r(\eta_{\infty,T})}{r_1(\eta_{\infty,T})}\right)^{2T-1} = \frac{c_1 \lambda_1}{c_r \lambda_r} \left(1 + O(\vartheta^{2T-1})\right).$$

Taking logarithms gives

$$(2T-1)\left(\log r_r(\eta_{\infty,T}) - \log r_1(\eta_{\infty,T})\right) = L_* + O(\vartheta^{2T}).$$

Writing $\eta_{\infty,T} = \eta_* + \delta$, we have

$$r_1(\eta_{\infty,T}) = q_* + \frac{\lambda_1 \delta}{m}, \qquad r_r(\eta_{\infty,T}) = q_* - \frac{\lambda_r \delta}{m}. \tag{21}$$

Thus, using $\log(1+x) = x + O(x^2)$, we obtain

$$\log r_r(\eta_{\infty,T}) - \log r_1(\eta_{\infty,T}) = \log\left(\frac{1 - \frac{\lambda_r \delta}{m q_*}}{1 + \frac{\lambda_1 \delta}{m q_*}}\right) = -\frac{\lambda_r \delta}{m q_*} - \frac{\lambda_1 \delta}{m q_*} + O(\delta^2) = -\frac{\lambda_1 + \lambda_r}{m q_*} \delta + O(\delta^2).$$

Combining the results, we have

$$L_* + O(\vartheta^{2T}) = (2T-1)\left(-\frac{\lambda_1+\lambda_r}{mq_*}\delta + O(\delta^2)\right).$$

Since the left-hand side is $O(1)$, we have $\delta = O(T^{-1})$. Thus

$$L_* + O(\vartheta^{2T}) = -(2T-1)\frac{\lambda_1+\lambda_r}{mq_*}\delta + O(T^{-1}).$$

Solving for δ, we have

$$\eta_{\infty,T} - \eta_* = \delta = -\frac{L_* m q_*}{(\lambda_1+\lambda_r)(2T-1)} + O(T^{-2}) + O\left(\frac{\vartheta^{2T}}{T}\right).$$

□

Corollary E.2 (Loss and curvature at the large-horizon optimum)**.** *Under the assumptions of Proposition E.1, we have*

$$\phi_{\infty,T}(\eta_{\infty,T}) - \frac{c_0}{2m} = \Theta\left(q_*^{2T}\right),$$
$$\left|F'_{\infty,T}(\eta_{\infty,T})\right| = \Theta\left(Tq_*^{2T-2}\right), \tag{22}$$
$$\phi''_{\infty,T}(\eta_{\infty,T}) = \Theta\left(T^2 q_*^{2T-2}\right). \tag{23}$$

Proof. Define $r_1(\eta)$ and $r_r(\eta)$ as in the proof of Proposition E.1. By equation 21 and Proposition E.1, we have

$$r_r(\eta_{\infty,T}) = q_*\left[1 + \frac{\lambda_r}{\lambda_1+\lambda_r}\frac{L_*}{2T-1} + O(T^{-2}) + O\left(\frac{\vartheta^{2T}}{T}\right)\right],$$
$$r_1(\eta_{\infty,T}) = q_*\left[1 - \frac{\lambda_1}{\lambda_1+\lambda_r}\frac{L_*}{2T-1} + O(T^{-2}) + O\left(\frac{\vartheta^{2T}}{T}\right)\right].$$

In particular, both endpoint magnitudes are $q_*(1+O(T^{-1}))$. More precisely, for every fixed integer ℓ, the expansion

$$\left(1 + \frac{a}{2T-1} + O(T^{-2})\right)^{2T-\ell} = \exp(a)\left(1 + O(T^{-1})\right)$$

gives

$$r_r(\eta_{\infty,T})^{2T-\ell} = q_*^{2T-\ell} a_r^*\left(1 + O(T^{-1}) + O(\vartheta^{2T})\right), \tag{24}$$
$$r_1(\eta_{\infty,T})^{2T-\ell} = q_*^{2T-\ell} a_1^*\left(1 + O(T^{-1}) + O(\vartheta^{2T})\right). \tag{25}$$

We now control the remaining active modes. Since $\eta_{\infty,T} \to \eta_*$ by Theorem 3.1, we have

$$\frac{1}{q_*}\max_{2\le j\le r-1}\left|1 - \frac{\eta_{\infty,T}\lambda_j}{m}\right| \to \bar{\vartheta} < 1,$$

which implies, for sufficiently large T,

$$\max_{2\le j\le r-1}\left|1 - \frac{\eta_{\infty,T}\lambda_j}{m}\right| \le \vartheta q_*.$$

Hence we have

$$\sum_{j=2}^{r-1} c_j\lambda_j^2\left|1 - \frac{\eta_{\infty,T}\lambda_j}{m}\right|^{2T-2} = O\left((\vartheta q_*)^{2T-2}\right) = o\left(q_*^{2T-2}\right). \tag{26}$$

Using the spectral representation of the loss, we therefore have

$$\phi_{\infty,T}(\eta_{\infty,T}) - \frac{c_0}{2m} = \frac{1}{2m}\sum_{j=1}^{r} c_j\left(1 - \frac{\eta_{\infty,T}\lambda_j}{m}\right)^{2T} = \frac{q_*^{2T}}{2m}\left[c_r a_r^* + c_1 a_1^* + o(1)\right] = \Theta(q_*^{2T}),$$

where the last equality follows since $c_r a_r^* + c_1 a_1^* + o(1)$ converges to a strictly positive constant.

Next, recall that by Proposition C.3,

$$F'_{\infty,T}(\eta) = -\frac{2T-1}{m}\sum_{j=1}^{r} c_j \lambda_j^2 \left(1 - \frac{\eta\lambda_j}{m}\right)^{2T-2}.$$

Note that all summands in the preceding display are nonnegative before multiplication by the minus sign. Applying equation 24, equation 25, and equation 26 gives

$$F'_{\infty,T}(\eta_{\infty,T}) = -\frac{2T-1}{m} q_*^{2T-2}\left[c_r\lambda_r^2 a_r^* + c_1\lambda_1^2 a_1^* + o(1)\right] = -\Theta\left(T q_*^{2T-2}\right).$$

The last equality follows since the limiting coefficient in square brackets is again strictly positive.

Finally, noting that

$$\phi'_{\infty,T}(\eta) = -\frac{T}{m^2} F_{\infty,T}(\eta)$$

holds by Proposition C.3, the curvature at the optimizer satisfies

$$\phi''_{\infty,T}(\eta_{\infty,T}) = -\frac{T}{m^2} F'_{\infty,T}(\eta_{\infty,T}).$$

Substituting the previous result yields

$$\phi''_{\infty,T}(\eta_{\infty,T}) = \frac{T}{m^2}\Theta\left(T q_*^{2T-2}\right) = \Theta(T^2 q_*^{2T-2}).$$

□

E.3 Endpoint contractions and spectral separation

Lemma E.3 (Endpoint contraction asymptotics)**.** *For $r_{j,T}$, $(j \in [r])$ defined in equation 17, we have*

$$\frac{r_{1,T}^{2T-1}}{q_*^{2T-1}} \to -a_1^*, \qquad \frac{r_{r,T}^{2T-1}}{q_*^{2T-1}} \to a_r^*, \qquad \frac{r_{1,T}^{2T-2}}{q_*^{2T-2}} \to a_1^*, \qquad \frac{r_{r,T}^{2T-2}}{q_*^{2T-2}} \to a_r^*.$$

Moreover, if $r > 2$, there exists a constant $\vartheta_0 \in (0,1)$ such that

$$|r_{j,T}| \le \vartheta_0 q_* \qquad (2 \le j \le r-1)$$

for all sufficiently large T.

Proof. Proposition E.1 gives

$$\eta_{\infty,T} - \eta_* = -\frac{m q_* L_*}{(\lambda_1+\lambda_r)(2T-1)} + o(T^{-1}).$$

By $1 - \eta_*\lambda_1/m = -q_*$ and $1 - \eta_*\lambda_r/m = q_*$, we have

$$r_{1,T} = -q_*\left(1 - \frac{\lambda_1 L_*}{(\lambda_1+\lambda_r)(2T-1)} + o((2T-1)^{-1})\right),$$

$$r_{r,T} = q_*\left(1 + \frac{\lambda_r L_*}{(\lambda_1+\lambda_r)(2T-1)} + o((2T-1)^{-1})\right).$$

Thus, noting that $2T-1$ is odd, we obtain

$$\frac{r_{1,T}^{2T-1}}{q_*^{2T-1}} \to -a_1^*, \qquad \frac{r_{r,T}^{2T-1}}{q_*^{2T-1}} \to a_r^*, \qquad \frac{r_{1,T}^{2T-2}}{q_*^{2T-2}} \to a_1^*, \qquad \frac{r_{r,T}^{2T-2}}{q_*^{2T-2}} \to a_r^*.$$

For an interior positive eigenvalue $\lambda_j \in (\lambda_r, \lambda_1)$, the definition of η_* gives

$$-q_* < 1 - \frac{\eta_*\lambda_j}{m} < q_*.$$

By Proposition E.1, we have

$$\max_{2\le j\le r-1} \frac{|1-\eta_*\lambda_j/m|}{q_*} = \bar{\vartheta} < 1.$$

Since Theorem 3.1 gives $\eta_{\infty,T} \to \eta_*$, there exists $\vartheta_0 \in (\bar{\vartheta}, 1)$ such that, for all sufficiently large T,

$$\max_{2\le j\le r-1} \frac{|r_{j,T}|}{q_*} = \max_{2\le j\le r-1} \frac{|1-\eta_{\infty,T}\lambda_j/m|}{q_*} \le \vartheta_0.$$

□

F MATRIX-POWER PERTURBATION BOUNDS

F.1 FRÉCHET EXPANSION OF MATRIX POWERS

Lemma F.1 (First-order expansion of matrix powers)**.** *Let $k \geq 1$. Equip $\mathbb{R}^{p\times p}$ with any submultiplicative matrix norm $\|\cdot\|$, and define $\Psi_k(B) = B^k$ as a map from this normed space into itself. Then Ψ_k is Fréchet differentiable at every B, and for every $C \in \mathbb{R}^{p\times p}$, we have*

$$D\Psi_k(B)[C] = \sum_{s=0}^{k-1} B^s C B^{k-1-s}.$$

Consequently, for every fixed C, we have

$$(B+\epsilon C)^k = B^k + \epsilon \sum_{s=0}^{k-1} B^s C B^{k-1-s} + O(\epsilon^2)$$

as $\epsilon \to 0$. No commutativity between B and C is required.

Proof. We first recall that for arbitrary square matrices X and Y of the same size and every $k \geq 1$, we have

$$X^k - Y^k = \sum_{s=0}^{k-1} X^s (X-Y) Y^{k-1-s}.$$

Applying this identity with $X = B + E$ and $Y = B$, we obtain

$$\Psi_k(B+E) - \Psi_k(B) = (B+E)^k - B^k = \sum_{s=0}^{k-1} (B+E)^s E B^{k-1-s}.$$

Thus if we define

$$R_k(B,E) := \sum_{s=0}^{k-1} ((B+E)^s - B^s) E B^{k-1-s},$$

we can write

$$\Psi_k(B+E) - \Psi_k(B) = \sum_{s=0}^{k-1} B^s E B^{k-1-s} + R_k(B,E).$$

In $R_k(B,E)$, the summand corresponding to $s=0$ vanishes. For $s \geq 1$, applying the same identity once more gives

$$(B+E)^s - B^s = \sum_{j=0}^{s-1} (B+E)^j E B^{s-1-j}.$$

Therefore, we have

$$R_k(B,E) = \sum_{s=1}^{k-1} \sum_{j=0}^{s-1} (B+E)^j E B^{s-1-j} E B^{k-1-s}.$$

Let $\|\cdot\|$ be any submultiplicative matrix norm. Then for each $s \in \{1,\dots,k-1\}$ and $j \in \{0,\dots,s-1\}$, we have

$$\begin{aligned}\|(B+E)^j E B^{s-1-j} E B^{k-1-s}\| &\leq \|B+E\|^j \|E\| \|B\|^{s-1-j} \|E\| \|B\|^{k-1-s} \\ &\leq (\|B\| + \|E\|)^{k-2} \|E\|^2.\end{aligned}$$

Since there are $k(k-1)/2$ terms in the double sum, we have

$$\|R_k(B,E)\| \leq \frac{k(k-1)}{2} (\|B\| + \|E\|)^{k-2} \|E\|^2.$$

Thus, we have

$$\frac{\|\Psi_k(B+E) - \Psi_k(B) - \sum_{s=0}^{k-1} B^s E B^{k-1-s}\|}{\|E\|} = \frac{\|R_k(B,E)\|}{\|E\|}$$
$$\leq \frac{k(k-1)}{2}(\|B\| + \|E\|)^{k-2}\|E\| \to 0$$

as $\|E\| \to 0$. Since the map $E \mapsto \sum_{s=0}^{k-1} B^s E B^{k-1-s}$ is linear and bounded, it is the Fréchet derivative of Ψ_k at B. Therefore, we have

$$D\Psi_k(B)[C] = \sum_{s=0}^{k-1} B^s C B^{k-1-s}.$$

Finally, taking $E = \epsilon C$ for a fixed matrix C gives

$$(B+\epsilon C)^k = \Psi_k(B+\epsilon C) = \Psi_k(B) + D\Psi_k(B)[\epsilon C] + R_k(B, \epsilon C)$$
$$= B^k + \epsilon \sum_{s=0}^{k-1} B^s C B^{k-1-s} + O(\epsilon^2),$$

which proves the final assertion. □

F.2 Uniform noncommutative power bounds

Lemma F.2 (Uniform local power perturbation bounds)**.** *Let $K \succeq 0$ be a finite-dimensional symmetric positive semi-definite matrix and let P project onto $\mathrm{range}(K)$. Suppose there exists deterministic sequence $q_T > 0$, bounded away from zero, such that*

$$\sup_{\eta \in \mathcal{J}_T} \|P B_K(\eta) P\|_{\mathrm{op}} \leq q_T, \qquad q_0 := \inf_T q_T > 0.$$

Let $A = K + E \succeq 0$ be symmetric and have the same kernel as K. Then we have

$$PA = AP, \qquad PK = KP, \qquad PE = EP.$$

Suppose $T\|E\|_{\mathrm{op}} \leq c$ holds for a fixed c. Suppose $\sup_T \sup_{\eta \in \mathcal{J}_T} |\eta| < \infty$. Then, uniformly over $\eta \in \mathcal{J}_T$ and $k \leq 2T$, we have

$$\|P B_A(\eta)^k P\|_{\mathrm{op}} \lesssim q_T^k, \tag{27}$$
$$\|P(B_A(\eta)^k - B_K(\eta)^k)P\|_{\mathrm{op}} \lesssim k q_T^{k-1} \|E\|_{\mathrm{op}}. \tag{28}$$

Proof. Define $\mathcal{H} = \mathrm{range}(K) = \ker(K)^\perp$. Since A and K are symmetric and have the same kernel, we have

$$\mathrm{range}(A) = \ker(A)^\perp = \ker(K)^\perp = \mathrm{range}(K) = \mathcal{H}.$$

Consequently, $\mathcal{H}$ and $\mathcal{H}^\perp$ are invariant under A, K, and $E = A - K$. Equivalently, we have[4]

$$PA = AP, \qquad PK = KP, \qquad PE = EP,$$

and hence P also commutes with $B_A(\eta)$ and $B_K(\eta)$. In particular, we have

$$P B_A(\eta)^k P = (P B_A(\eta) P)^k, \qquad P B_K(\eta)^k P = (P B_K(\eta) P)^k.$$

These identities hold as operators on $\mathcal{H}$.

[4]This equivalence can be proved as follows. For any $x \in \mathbb{R}^m$, decompose $x = Px + (I_m - P)x$. Here, $Px \in \mathcal{H}$ and $(I_m - P)x \in \mathcal{H}^\perp$. If $\mathcal{H}$ and $\mathcal{H}^\perp$ are invariant under A, we have $APx \in \mathcal{H}$ and $A(I_m - P)x \in \mathcal{H}^\perp$. Thus,

$$PAx = P[APx + A(I_m - P)x] = P(APx) + P[A(I_m - P)x] = APx + 0.$$

Since x is arbitrary, we conclude $PA = AP$. On the other hand, if $PA = AP$, then $P(Ax) = A(Px) = Ax$ for $x \in \mathcal{H}$ and $P(Ax) = A(Px) = 0$ for $x \in \mathcal{H}^\perp$; thus, $A\mathcal{H} \subset \mathcal{H}$ and $A\mathcal{H}^\perp \subset \mathcal{H}^\perp$ hold. Similar arguments apply to K and E.

In the remainder of the proof, we regard these matrices as linear operators from $\mathcal{H}$ to itself. Fix $\eta \in \mathcal{J}_T$ and define

$$B_0 = PB_K(\eta)P\big|_{\mathcal{H}} = B_K(\eta)\big|_{\mathcal{H}}, \qquad \Delta = -\frac{\eta}{m}PEP\big|_{\mathcal{H}} = -\frac{\eta}{m}E\big|_{\mathcal{H}}.$$

Then

$$PB_A(\eta)P\big|_{\mathcal{H}} = B_0 + \Delta.$$

With respect to the orthogonal decomposition $\mathbb{R}^m = \mathcal{H} \oplus \mathcal{H}^\perp$, the operators A and K have the block representations

$$A = \begin{bmatrix} A|_{\mathcal{H}} & 0 \\ 0 & 0 \end{bmatrix}, \qquad K = \begin{bmatrix} K|_{\mathcal{H}} & 0 \\ 0 & 0 \end{bmatrix}. \tag{29}$$

Consequently, we have

$$B_A(\eta) = \begin{bmatrix} B_0 + \Delta & 0 \\ 0 & I_{\mathcal{H}^\perp} \end{bmatrix}, \qquad B_K(\eta) = \begin{bmatrix} B_0 & 0 \\ 0 & I_{\mathcal{H}^\perp} \end{bmatrix}. \tag{30}$$

By assumption, we have $\|B_0\|_{\mathrm{op}} \le q_T$. Set $M_\eta = \sup_{T\ge 1} \sup_{\eta\in\mathcal{J}_T} |\eta|/m < \infty$. Then it follows that $\|\Delta\|_{\mathrm{op}} \le M_\eta \|E\|_{\mathrm{op}}$. Since $T\|E\|_{\mathrm{op}} \le c$, for $k \le 2T$, we have

$$\frac{k\|\Delta\|_{\mathrm{op}}}{q_T} \le \frac{2M_\eta T\|E\|_{\mathrm{op}}}{q_T} \le \frac{2M_\eta c}{q_0}.$$

Thus $k\|\Delta\|_{\mathrm{op}}/q_T$ is bounded uniformly in η, k, and T.

We first prove equation 27. For $s \le k \le 2T$, we have

$$\begin{aligned}\|(B_0+\Delta)^s\|_{\mathrm{op}} &\le (\|B_0\|_{\mathrm{op}} + \|\Delta\|_{\mathrm{op}})^s \le (q_T + \|\Delta\|_{\mathrm{op}})^s = q_T^s \left(1 + \frac{\|\Delta\|_{\mathrm{op}}}{q_T}\right)^s \\ &\le q_T^s \exp\left(\frac{s\|\Delta\|_{\mathrm{op}}}{q_T}\right) \le q_T^s \exp\left(\frac{2M_\eta c}{q_0}\right) \lesssim q_T^s,\end{aligned}$$

which proves equation 27. By equation 30, for every $k \ge 1$,

$$B_A(\eta)^k - B_K(\eta)^k = \begin{bmatrix} (B_0+\Delta)^k - B_0^k & 0 \\ 0 & 0 \end{bmatrix}.$$

Therefore,

$$\left\|B_A(\eta)^k - B_K(\eta)^k\right\|_{\mathrm{op},\mathbb{R}^m} = \left\|(B_0+\Delta)^k - B_0^k\right\|_{\mathrm{op},\mathcal{H}}.$$

Thus it is sufficient to establish the desired power bounds on $\mathcal{H}$. The telescoping identity $\Gamma^k - \Lambda^k = \sum_{s=0}^{k-1} \Gamma^s(\Gamma-\Lambda)\Lambda^{k-1-s}$ implies

$$(B_0+\Delta)^k - B_0^k = \sum_{s=0}^{k-1} (B_0+\Delta)^s \Delta B_0^{k-1-s}.$$

Consequently, we have

$$\begin{aligned}\|(B_0+\Delta)^k - B_0^k\|_{\mathrm{op}} &\le \sum_{s=0}^{k-1} \|(B_0+\Delta)^s\|_{\mathrm{op}} \|\Delta\|_{\mathrm{op}} \|B_0^{k-1-s}\|_{\mathrm{op}} \\ &\lesssim \|\Delta\|_{\mathrm{op}} \sum_{s=0}^{k-1} q_T^s q_T^{k-1-s} = k q_T^{k-1} \|\Delta\|_{\mathrm{op}} \lesssim k q_T^{k-1} \|E\|_{\mathrm{op}},\end{aligned}$$

which implies equation 28. □

F.3 Application to the finite-width kernel

Lemma F.3 (Uniform finite-width contraction near the optimal learning rate)**.** *Suppose m and d are fixed and $T = T_n$ satisfies $T \to \infty$ and $T/\sqrt{n} \to 0$. Suppose Assumption 3.2 holds. Let $P = \sum_{j=1}^{r} P_j$ denote the orthogonal projector onto* $\mathrm{range}(K)$*. Then uniformly over $|\eta - \eta_{\infty,T}| \le C/T$ and $0 \le k \le 2T$, we have*

$$\|PB_\infty(\eta)^k P\|_{\mathrm{op}} \lesssim q_*^k.$$

For any $c > 0$, define the event

$$\mathcal{E}_n = \{\ker(A_n) = \ker(K),\ T\|A_n - K\|_{\mathrm{op}} \le c\}.$$

Then we have $\mathrm{P}(\mathcal{E}_n) \to 1$, and uniformly over $|\eta - \eta_{\infty,T}| \le C/T$ and $0 \le k \le 2T$, we have

$$\|PB_n(\eta)^k P\|_{\mathrm{op}} \lesssim q_*^k, \qquad \|P\left(B_n(\eta)^k - B_\infty(\eta)^k\right)P\|_{\mathrm{op}} \lesssim k q_*^{k-1}\|A_n - K\|_{\mathrm{op}}$$

on $\mathcal{E}_n$. On this event, it also holds that

$$PA_n = A_n P, \qquad PK = KP, \qquad P(A_n - K) = (A_n - K)P.$$

Proof. We confirm that the assumptions of Lemma F.2 are satisfied with probability tending to one. First, we have

$$\|PB_\infty(\eta)P\|_{\mathrm{op}} = \left\|\sum_{j=1}^{r}\left(1 - \frac{\eta\lambda_j}{m}\right)P_j\right\|_{\mathrm{op}} = \max_{j\in[r]}\left|1 - \frac{\eta\lambda_j}{m}\right| = \left|1 - \frac{\eta\lambda_1}{m}\right| \vee \left|1 - \frac{\eta\lambda_r}{m}\right|.$$

Since $\|PB_\infty(\eta_*)P\|_{\mathrm{op}} = \left|1 - \frac{\eta_*\lambda_1}{m}\right| = \left|1 - \frac{\eta_*\lambda_r}{m}\right| = q_*$ holds, we have

$$|\|PB_\infty(\eta)P\|_{\mathrm{op}} - q_*| \le \left|\left|1 - \frac{\eta\lambda_1}{m}\right| - q_*\right| \vee \left|\left|1 - \frac{\eta\lambda_r}{m}\right| - q_*\right| \lesssim |\eta - \eta_*|.$$

Thus, there exists a constant C_1 such that, uniformly over $|\eta - \eta_{\infty,T}| \le C/T$,

$$\|PB_\infty(\eta)P\|_{\mathrm{op}} \le q_* + \frac{C_1}{T}. \tag{31}$$

This implies

$$\|PB_\infty(\eta)^k P\|_{\mathrm{op}} \le \|PB_\infty(\eta)P\|_{\mathrm{op}}^k \le \left(q_* + \frac{C_1}{T}\right)^k \le q_*^k \exp\left(\frac{kC_1}{q_*T}\right) \lesssim q_*^k \qquad (0 \le k \le 2T).$$

By Lemma 3.3, we have

$$\|A_n - K\|_{\mathrm{op}} = O_p(n^{-1/2}), \qquad \|\chi_n^{(0)} + y\| = O_p(n^{-1/2}).$$

The assumption $T/\sqrt{n} \to 0$ therefore implies $T\|A_n - K\|_{\mathrm{op}} = o_p(1)$. Consequently, we have $\mathrm{P}(\mathcal{E}_n) \to 1$.

Note that $\{q_* + \frac{C_1}{T}\}_T$ is bounded away from zero. Thus, by Lemma F.2, on the event $\mathcal{E}_n$, uniformly over $|\eta - \eta_{\infty,T}| \le C/T$ and $0 \le k \le 2T$, we have

$$\|PB_n(\eta)^k P\|_{\mathrm{op}} \lesssim \left(q_* + \frac{C_1}{T}\right)^k = q_*^k\left(1 + \frac{C_1}{q_*T}\right)^k \le q_*^k \exp\left(\frac{C_1 k}{q_*T}\right) \le q_*^k \exp\left(\frac{2C_1}{q_*}\right) \lesssim q_*^k,$$

and similarly,

$$\|P\left(B_n(\eta)^k - B_\infty(\eta)^k\right)P\|_{\mathrm{op}} \lesssim k\left(q_* + \frac{C_1}{T}\right)^{k-1}\|A_n - K\|_{\mathrm{op}} \lesssim k q_*^{k-1}\|A_n - K\|_{\mathrm{op}}.$$

We also note that on this event, it holds that

$$PA_n = A_n P, \qquad PK = KP, \qquad P(A_n - K) = (A_n - K)P$$

by Lemma F.2. □

G PROOF OF THE JOINT WIDTH-HORIZON TRANSFER THEOREM

Throughout this section, assume the conditions in Theorem 3.4.

By assumption, we have $\eta_\ell < \eta_* < \eta_u < \frac{2m}{\lambda_1}$. Since Proposition E.1 implies $\eta_{\infty,T} = \eta_* + O(T^{-1})$, there exists a constant $C > 0$ such that, for all sufficiently large T,

$$\{\eta : |\eta - \eta_{\infty,T}| \le C/T\} \subset (\eta_\ell, \eta_u).$$

Let $P = \sum_{j=1}^{r} P_j$ be the orthogonal projector onto $\mathrm{range}(K)$, and set $E_n = A_n - K$. By Lemma C.2, for all sufficiently large n, we have $\ker(A_n) \overset{\text{a.s.}}{=} \ker(K)$ and $f_n^{(0)}(X) \in \mathrm{range}(H_n) \overset{\text{a.s.}}{=} \mathrm{range}(K)$. Consequently, we have

$$P f_n^{(0)}(X) \overset{\text{a.s.}}{=} f_n^{(0)}(X), \qquad P E_n \overset{\text{a.s.}}{=} E_n P \overset{\text{a.s.}}{=} E_n. \tag{32}$$

G.1 UNIFORM SCORE APPROXIMATION

Lemma G.1 (Uniform approximation of the score and its derivative)**.** *Note that Proposition E.1 places $\eta_{\infty,T}$ in a C/T-neighborhood of η_* for some C. For this neighborhood, we have*

$$\sup_{|\eta-\eta_{\infty,T}|\le C/T} |F_{n,T}(\eta) - F_{\infty,T}(\eta)| = O_p\left(\frac{T q_*^{2T-2}}{\sqrt{n}}\right),$$
$$\sup_{|\eta-\eta_{\infty,T}|\le C/T} |F'_{n,T}(\eta) - F'_{\infty,T}(\eta)| = O_p\left(\frac{T^2 q_*^{2T-3}}{\sqrt{n}}\right).$$

Proof. Note that if the $O_p(\cdot)$ bound holds on an event whose probability tends to one, the same $O_p(\cdot)$ bound holds unconditionally. Thus, it suffices to show upper bounds for $\sup_{|\eta-\eta_{\infty,T}|\le C/T} |F_{n,T}(\eta) - F_{\infty,T}(\eta)|$ and $\sup_{|\eta-\eta_{\infty,T}|\le C/T} |F'_{n,T}(\eta) - F'_{\infty,T}(\eta)|$ on the event $\mathcal{E}_n$ defined in Lemma F.3. We first consider the score. Set

$$M_{n,2T-1}(\eta) = A_n B_n(\eta)^{2T-1}, \qquad M_{\infty,2T-1}(\eta) = K B_\infty(\eta)^{2T-1}.$$

On the event $\mathcal{E}_n$, we have

$$\begin{aligned}
&\sup_{|\eta-\eta_{\infty,T}|\le C/T} \|P\left(M_{n,2T-1}(\eta) - M_{\infty,2T-1}(\eta)\right)P\|_{\mathrm{op}} \\
&= \sup_{|\eta-\eta_{\infty,T}|\le C/T} \|P(A_n-K)B_n(\eta)^{2T-1}P + PK(B_n(\eta)^{2T-1} - B_\infty(\eta)^{2T-1})P\|_{\mathrm{op}} \\
&\le \|A_n - K\|_{\mathrm{op}} \sup_{|\eta-\eta_{\infty,T}|\le C/T} \|PB_n(\eta)^{2T-1}P\|_{\mathrm{op}} \\
&\quad + \|K\|_{\mathrm{op}} \sup_{|\eta-\eta_{\infty,T}|\le C/T} \left\|P\left(B_n(\eta)^{2T-1} - B_\infty(\eta)^{2T-1}\right)P\right\|_{\mathrm{op}} \\
&= O_p\left(\frac{q_*^{2T-1}}{\sqrt{n}} + \frac{T q_*^{2T-2}}{\sqrt{n}}\right) \\
&= O_p\left(\frac{T q_*^{2T-2}}{\sqrt{n}}\right).
\end{aligned}$$

By Lemma F.3, on the same event, we have

$$A_n = PA_nP, \qquad B_n(\eta)P = PB_n(\eta). \tag{33}$$

Therefore, for every integer $k \ge 0$, we have

$$A_n B_n(\eta)^k = PA_nPB_n(\eta)^k = PA_nPB_n(\eta)^k P = A_nPB_n(\eta)^k P.$$

Consequently, it holds that

$$\begin{aligned}
\sup_{|\eta-\eta_{\infty,T}|\le C/T} \|M_{n,2T-1}(\eta)\|_{\mathrm{op}} &\le \sup_{|\eta-\eta_{\infty,T}|\le C/T} \|A_n\|_{\mathrm{op}} \|PB_n(\eta)^{2T-1}P\|_{\mathrm{op}} \\
&\lesssim q_*^{2T-1}(\|K\|_{\mathrm{op}} + \|A_n - K\|_{\mathrm{op}}) \\
&= O_p(q_*^{2T-1}).
\end{aligned}$$

Since A_n commutes with $B_n(\eta)$, the matrix $M_{n,2T-1}(\eta)$ is symmetric. Therefore,

$$F_{n,T}(\eta) - F_{\infty,T}(\eta) = y^\top (M_{n,2T-1}(\eta) - M_{\infty,2T-1}(\eta))y - 2(\chi_n^{(0)} + y)^\top M_{n,2T-1}(\eta)y + (\chi_n^{(0)} + y)^\top M_{n,2T-1}(\eta)(\chi_n^{(0)} + y).$$

On the event $\mathcal{E}_n$, since both $M_{n,2T-1}(\eta)$ and $M_{\infty,2T-1}(\eta)$ vanish on $\ker(K)$, we may insert the projector P in the first term:

$$y^\top (M_{n,2T-1}(\eta) - M_{\infty,2T-1}(\eta))y = y^\top P(M_{n,2T-1}(\eta) - M_{\infty,2T-1}(\eta))Py.$$

Thus, on the event $\mathcal{E}_n$, we obtain

$$\begin{aligned}
\sup_{|\eta-\eta_{\infty,T}|\le C/T} |F_{n,T}(\eta) - F_{\infty,T}(\eta)| &\le \|y\|^2 \sup_{|\eta-\eta_{\infty,T}|\le C/T} \|P(M_{n,2T-1}(\eta) - M_{\infty,2T-1}(\eta))P\|_{\mathrm{op}} \\
&\quad + 2\|\chi_n^{(0)} + y\|\|y\| \sup_{|\eta-\eta_{\infty,T}|\le C/T} \|M_{n,2T-1}(\eta)\|_{\mathrm{op}} \\
&\quad + \|\chi_n^{(0)} + y\|^2 \sup_{|\eta-\eta_{\infty,T}|\le C/T} \|M_{n,2T-1}(\eta)\|_{\mathrm{op}} \\
&= O_p\left(\frac{Tq_*^{2T-2}}{\sqrt{n}}\right) + O_p\left(\frac{q_*^{2T-1}}{\sqrt{n}}\right) + O_p\left(\frac{q_*^{2T-1}}{n}\right) \\
&= O_p\left(\frac{Tq_*^{2T-2}}{\sqrt{n}}\right).
\end{aligned}$$

We next consider the derivative of the score. Define

$$R_{n,2T-2}(\eta) := A_n^2 B_n(\eta)^{2T-2}, \qquad R_{\infty,2T-2}(\eta) := K^2 B_\infty(\eta)^{2T-2}.$$

Then on the event $\mathcal{E}_n$, we have by Lemma F.3 that

$$\begin{aligned}
&\sup_{|\eta-\eta_{\infty,T}|\le C/T} \|P\left(R_{n,2T-2}(\eta) - R_{\infty,2T-2}(\eta)\right)P\|_{\mathrm{op}} \\
&= \sup_{|\eta-\eta_{\infty,T}|\le C/T} \|P(A_n^2 - K^2)B_n(\eta)^{2T-2}P + PK^2(B_n(\eta)^{2T-2} - B_\infty(\eta)^{2T-2})P\|_{\mathrm{op}} \\
&\le \|A_n^2 - K^2\|_{\mathrm{op}} \sup_{|\eta-\eta_{\infty,T}|\le C/T} \|PB_n(\eta)^{2T-2}P\|_{\mathrm{op}} \\
&\quad + \|K\|_{\mathrm{op}}^2 \sup_{|\eta-\eta_{\infty,T}|\le C/T} \left\|P\left(B_n(\eta)^{2T-2} - B_\infty(\eta)^{2T-2}\right)P\right\|_{\mathrm{op}} \\
&\lesssim \|(A_n - K)A_n + K(A_n - K)\|_{\mathrm{op}} q_*^{2T-2} + Tq_*^{2T-3}\|A_n - K\|_{\mathrm{op}} \\
&\le \|A_n - K\|_{\mathrm{op}}(\|A_n\|_{\mathrm{op}} + \|K\|_{\mathrm{op}})q_*^{2T-2} + Tq_*^{2T-3}\|A_n - K\|_{\mathrm{op}} \\
&= O_p\left(\frac{q_*^{2T-2}}{\sqrt{n}} + \frac{Tq_*^{2T-3}}{\sqrt{n}}\right) \\
&= O_p\left(\frac{Tq_*^{2T-3}}{\sqrt{n}}\right).
\end{aligned}$$

Moreover, by equation 33, it holds for every integer $k \ge 0$ that

$$A_n^2 B_n(\eta)^k = PA_nPA_nPB_n(\eta)^k = PA_n^2PB_n(\eta)^k = PA_n^2PB_n(\eta)^kP = A_n^2PB_n(\eta)^kP.$$

Thus, we have

$$\sup_{|\eta-\eta_{\infty,T}|\le C/T} \|R_{n,2T-2}(\eta)\|_{\mathrm{op}} \le \sup_{|\eta-\eta_{\infty,T}|\le C/T} \|A_n^2\|_{\mathrm{op}}\|PB_n(\eta)^{2T-2}P\|_{\mathrm{op}} = O_p(q_*^{2T-2}).$$

Note that by Proposition C.3, we have

$$\begin{aligned}
&F'_{n,T}(\eta) - F'_{\infty,T}(\eta) \\
&= -\frac{2T-1}{m}[(\chi_n^{(0)})^\top R_{n,2T-2}(\eta)\chi_n^{(0)} - y^\top R_{\infty,2T-2}(\eta)y] \\
&= -\frac{2T-1}{m}[y^\top (R_{n,2T-2}(\eta) - R_{\infty,2T-2}(\eta))y - 2(\chi_n^{(0)} + y)^\top R_{n,2T-2}(\eta)y] \\
&\quad - \frac{2T-1}{m}(\chi_n^{(0)} + y)^\top R_{n,2T-2}(\eta)(\chi_n^{(0)} + y).
\end{aligned}$$

On the event $\mathcal{E}_n$, since both $R_{n,2T-2}(\eta)$ and $R_{\infty,2T-2}(\eta)$ vanish on $\ker(K)$, we may insert the projector P in the first term:

$$y^\top (R_{n,2T-2}(\eta) - R_{\infty,2T-2}(\eta))y = y^\top P(R_{n,2T-2}(\eta) - R_{\infty,2T-2}(\eta))Py.$$

Thus, on the event $\mathcal{E}_n$, we have

$$\begin{aligned}
&\sup_{|\eta-\eta_{\infty,T}|\le C/T} |F'_{n,T}(\eta) - F'_{\infty,T}(\eta)| \\
&\le \frac{2T-1}{m}\|y\|^2 \sup_{|\eta-\eta_{\infty,T}|\le C/T} \|P(R_{n,2T-2}(\eta) - R_{\infty,2T-2}(\eta))P\|_{\mathrm{op}} \\
&\quad + \frac{2T-1}{m} 2\|\chi_n^{(0)} + y\|\|y\| \sup_{|\eta-\eta_{\infty,T}|\le C/T} \|R_{n,2T-2}(\eta)\|_{\mathrm{op}} \\
&\quad + \frac{2T-1}{m}\|\chi_n^{(0)} + y\|^2 \sup_{|\eta-\eta_{\infty,T}|\le C/T} \|R_{n,2T-2}(\eta)\|_{\mathrm{op}} \\
&= O_p\left(\frac{T^2 q_*^{2T-3}}{\sqrt{n}}\right) + O_p\left(\frac{T q_*^{2T-2}}{\sqrt{n}}\right) + O_p\left(\frac{T q_*^{2T-2}}{n}\right) \\
&= O_p\left(\frac{T^2 q_*^{2T-3}}{\sqrt{n}}\right).
\end{aligned}$$

□

G.2 Localization of the finite-width optimizer

Lemma G.2 ($n^{-1/2}$-consistency of the finite-width optimizer)**.** *Under the assumptions of Theorem 3.4, with probability tending to one, we have $\eta_{n,T} \in (\eta_\ell, \eta_u)$ and $F_{n,T}(\eta_{n,T}) = 0$. Moreover, we have*

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

which implies

$$|\eta_{n,T} - \eta_{\infty,T}| = o_p(T^{-1}).$$

Proof. Define $s_T = T q_*^{2T-2}$. By equation 22, there exist constants $0 < \underline{C} < \overline{C} < \infty$ such that

$$\underline{C} s_T \le -F'_{\infty,T}(\eta_{\infty,T}) \le \overline{C} s_T \tag{34}$$

for all sufficiently large T. Differentiating equation 11 once more gives

$$F''_{\infty,T}(\eta) = \frac{(2T-1)(2T-2)}{m^2} y^\top K^3 B_\infty(\eta)^{2T-3} y.$$

By Lemma F.3, uniformly over $|\eta - \eta_{\infty,T}| \le C/T$, we have

$$\|PB_\infty(\eta)^k P\|_{\mathrm{op}} \lesssim q_*^k \qquad (0 \le k \le 2T),$$

which implies

$$\sup_{|\eta-\eta_{\infty,T}|\le C/T} |F''_{\infty,T}(\eta)| \lesssim T^2 \|Py\|^2 \|K\|_{\mathrm{op}}^3 \|PB_\infty(\eta)^{2T-3}P\|_{\mathrm{op}} \lesssim T^2 q_*^{2T-3}.$$

For every fixed $M > 0$, the assumption $T/\sqrt{n} \to 0$ implies $M/\sqrt{n} \le C/T$ for all sufficiently large n. Hence we have

$$\sup_{|\eta-\eta_{\infty,T}|\le M/\sqrt{n}} |F'_{\infty,T}(\eta) - F'_{\infty,T}(\eta_{\infty,T})| = O\left(\frac{MT^2 q_*^{2T-3}}{\sqrt{n}}\right) = o(s_T). \tag{35}$$

Since Lemma G.1 gives

$$\sup_{|\eta-\eta_{\infty,T}|\le M/\sqrt{n}} |F'_{n,T}(\eta) - F'_{\infty,T}(\eta)| = O_p\left(\frac{T^2 q_*^{2T-3}}{\sqrt{n}}\right) = o_p(s_T), \tag{36}$$

we have $\sup_{|\eta-\eta_{\infty,T}|\le M/\sqrt{n}} F'_{n,T}(\eta) \le F'_{\infty,T}(\eta_{\infty,T}) + o_p(s_T)$. Combining this with equation 34, we obtain

$$\mathrm{P}\left(\sup_{|\eta-\eta_{\infty,T}|\le M/\sqrt{n}} F'_{n,T}(\eta) \le -\frac{\underline{C}}{2} s_T\right) \to 1 \tag{37}$$

for every fixed $M > 0$.

By Corollary C.4 and Theorem 3.1, we have $F_{\infty,T}(\eta_{\infty,T}) = 0$. Thus the first bound in Lemma G.1 gives

$$F_{n,T}(\eta_{\infty,T}) = O_p\left(\frac{s_T}{\sqrt{n}}\right).$$

Therefore, for any fixed ϵ, there exists $M_0 < \infty$ such that

$$\liminf_{n\to\infty} \mathrm{P}\left(|F_{n,T}(\eta_{\infty,T})| \le \frac{M_0 s_T}{\sqrt{n}}\right) \ge 1 - \epsilon. \tag{38}$$

Choose $M > 2M_0/\underline{C}$. On the intersection of this event and the event in equation 37, the mean value theorem gives

$$F_{n,T}\left(\eta_{\infty,T} - \frac{M}{\sqrt{n}}\right) \ge -\frac{M_0 s_T}{\sqrt{n}} + \frac{\underline{C} M s_T}{2\sqrt{n}} > 0,$$
$$F_{n,T}\left(\eta_{\infty,T} + \frac{M}{\sqrt{n}}\right) \le \frac{M_0 s_T}{\sqrt{n}} - \frac{\underline{C} M s_T}{2\sqrt{n}} < 0.$$

By Proposition C.3, $F_{n,T}$ is continuous. On the event in equation 37, it is strictly decreasing throughout the interval $|\eta - \eta_{\infty,T}| \le M/\sqrt{n}$. Hence the preceding sign inequalities and the intermediate value theorem imply that on the intersection of events in equation 37 and equation 38, $F_{n,T}$ has a unique root in

$$\left(\eta_{\infty,T} - \frac{M}{\sqrt{n}}, \eta_{\infty,T} + \frac{M}{\sqrt{n}}\right) \subset (\eta_\ell, \eta_u).$$

For all sufficiently large n, Proposition C.5 and Corollary C.4 imply that $\phi_{n,T}$ is almost surely strictly convex. Therefore this root is the unique minimizer of $\phi_{n,T}$ over $\mathcal{I}$, namely $\eta_{n,T}$. Consequently, with probability at least $1 - \epsilon + o(1)$, we have

$$|\eta_{n,T} - \eta_{\infty,T}| < \frac{M}{\sqrt{n}}.$$

Since $\epsilon > 0$ was arbitrary, this proves both the asserted interiority and $\eta_{n,T} - \eta_{\infty,T} = O_p(n^{-1/2})$. □

G.3 First-order perturbation expansion

We next derive the first-order triangular-array expansion of the score. Set

$$B_T = B_\infty(\eta_{\infty,T}), \qquad \Delta_n = -\frac{\eta_{\infty,T}}{m} E_n.$$

Then we have

$$B_n(\eta_{\infty,T}) = B_T + \Delta_n.$$

For $1 \le k \le 2T$, define

$$\mathcal{L}_{k,n} = \sum_{s=0}^{k-1} B_T^s \Delta_n B_T^{k-1-s}, \qquad \mathcal{R}_{k,n} = (B_T + \Delta_n)^k - B_T^k - \mathcal{L}_{k,n}.$$

Then by Lemma F.1, we have

$$\mathcal{L}_{k,n} = D\Psi_k(B_T)[\Delta_n], \qquad \Psi_k(B) = B^k.$$

Lemma G.3 (Linear and quadratic matrix-power perturbations)**.** *Let P denote the orthogonal projector onto* $\mathrm{range}(K)$*. Then we have*

$$\|P\mathcal{L}_{k,n}P\|_{\mathrm{op}} = O_p(k q_*^{k-1} \|E_n\|_{\mathrm{op}}), \qquad \|P\mathcal{R}_{k,n}P\|_{\mathrm{op}} = O_p(k^2 q_*^{k-2} \|E_n\|_{\mathrm{op}}^2).$$

Proof. Define the event $\mathcal{E}_n$ as in Lemma F.3. By Lemma F.3, we have $\mathrm{P}(\mathcal{E}_n) \to 1$. Note that if the $O_p(\cdot)$ bound holds on an event whose probability tends to one, the same $O_p(\cdot)$ bound holds unconditionally. Thus, it suffices to show that

$$\|P\mathcal{L}_{k,n}P\|_{\mathrm{op}} \lesssim kq_*^{k-1}\|E_n\|_{\mathrm{op}}, \qquad \|P\mathcal{R}_{k,n}P\|_{\mathrm{op}} \lesssim k^2 q_*^{k-2}\|E_n\|_{\mathrm{op}}^2$$

holds on $\mathcal{E}_n$.

First, note that Lemma F.3 gives

$$\|PB_T^j P\|_{\mathrm{op}} \lesssim q_*^j \qquad (0 \le j \le 2T).$$

Since Proposition E.1 implies $\eta_{\infty,T} = \eta_* + O(T^{-1})$, we have $\sup_T \eta_{\infty,T} < \infty$. Hence, by the definition of Δ_n, we have

$$\|\Delta_n\|_{\mathrm{op}} \lesssim \|E_n\|_{\mathrm{op}}. \tag{39}$$

Therefore, we obtain

$$\begin{aligned}\|P\mathcal{L}_{k,n}P\|_{\mathrm{op}} &\le \sum_{s=0}^{k-1}\|P(B_T^s\Delta_n B_T^{k-1-s})P\|_{\mathrm{op}} \le \sum_{s=0}^{k-1}\|PB_T^sP\|_{\mathrm{op}}\|\Delta_n\|_{\mathrm{op}}\|PB_T^{k-1-s}P\|_{\mathrm{op}} \\ &\lesssim \sum_{s=0}^{k-1} q_*^s\|\Delta_n\|_{\mathrm{op}}q_*^{k-1-s} = kq_*^{k-1}\|\Delta_n\|_{\mathrm{op}} \lesssim kq_*^{k-1}\|E_n\|_{\mathrm{op}}.\end{aligned}$$

Moreover, since $B_T + \Delta_n = B_n(\eta_{\infty,T})$, Lemma F.3 gives

$$\|P(B_T+\Delta_n)^jP\|_{\mathrm{op}} \lesssim q_*^j$$

uniformly over $0 \le j \le 2T$ on $\mathcal{E}_n$.

For the remainder, the proof of Lemma F.1 gives the identity

$$\mathcal{R}_{k,n} = \sum_{s=1}^{k-1}\sum_{j=0}^{s-1}(B_T+\Delta_n)^j\Delta_n B_T^{s-1-j}\Delta_n B_T^{k-1-s}$$

for $k \ge 2$, while $\mathcal{R}_{1,n} = 0$. Hence, using the preceding projected power bounds and equation 39, we obtain

$$\begin{aligned}\|P\mathcal{R}_{k,n}P\|_{\mathrm{op}} &\le \sum_{s=1}^{k-1}\sum_{j=0}^{s-1}\|P(B_T+\Delta_n)^jP\|_{\mathrm{op}}\|\Delta_n\|_{\mathrm{op}}\|PB_T^{s-1-j}P\|_{\mathrm{op}}\|\Delta_n\|_{\mathrm{op}}\|PB_T^{k-1-s}P\|_{\mathrm{op}} \\ &\lesssim \sum_{s=1}^{k-1}\sum_{j=0}^{s-1} q_*^j\|\Delta_n\|_{\mathrm{op}}^2 q_*^{s-1-j}q_*^{k-1-s} = \sum_{s=1}^{k-1} sq_*^{k-2}\|\Delta_n\|_{\mathrm{op}}^2 = \frac{k(k-1)}{2}q_*^{k-2}\|\Delta_n\|_{\mathrm{op}}^2 \\ &\lesssim k^2q_*^{k-2}\|E_n\|_{\mathrm{op}}^2.\end{aligned}$$

□

G.4 Limits of the first-order sensitivities of the loss and score

Lemma G.4 (Asymptotics of the local Fréchet coefficients)**.** *We have*

$$\frac{PQ_{F,T}P}{Tq_*^{2T-2}} \to Q_{F,*} := -\frac{2\eta_* h_*}{m}(u_1u_1^\top + u_ru_r^\top), \qquad \frac{PQ_{L,T}P}{Tq_*^{2T-1}} \to Q_{L,*}$$

in Frobenius norm, where $Q_{L,*}$ is defined by $Q_{L,*} = m^{-2}\eta_*(c_1a_1^*u_1u_1^\top - c_ra_r^*u_ru_r^\top)$. In addition, we have

$$\frac{\|Pg_{F,T}\|}{Tq_*^{2T-2}} \to 0, \qquad \frac{\|Pg_{L,T}\|}{Tq_*^{2T-1}} \to 0.$$

Proof. Set

$$\Gamma_{ij,T}^{(k)} := \sum_{s=0}^{k-1} r_{i,T}^{k-1-s}r_{j,T}^s = \begin{cases} kr_{i,T}^{k-1} & (i=j), \\ (r_{i,T}^k - r_{j,T}^k)/(r_{i,T}-r_{j,T}) & (i\ne j). \end{cases}.$$

Since $\lambda_i \neq \lambda_j$ for $i \neq j$ and $\eta_{\infty,T} \to \eta_* > 0$, the denominator $|r_{i,T} - r_{j,T}|$ is bounded away from zero for all sufficiently large T. Directly projecting $Q_{F,T}$ onto the two eigenspaces gives

$$P_i Q_{F,T} P_j = \left(\frac{r_{i,T}^{2T-1} + r_{j,T}^{2T-1}}{2} - \frac{\eta_{\infty,T}(\lambda_i + \lambda_j)}{2m} \Gamma_{ij,T}^{(2T-1)} \right) (P_i y)(P_j y)^\top.$$

In particular, we have

$$P_j Q_{F,T} P_j = \left(r_{j,T}^{2T-1} - \frac{\eta_{\infty,T}(2T-1)\lambda_j}{m} r_{j,T}^{2T-2} \right) (P_j y)(P_j y)^\top.$$

By Lemma E.3, for every $j = 2, \ldots, r-1$, we have $P_j Q_{F,T} P_j = o(T q_*^{2T-2})$. For $i \neq j$, noting that $|\Gamma_{ij,T}^{(2T-1)}| \lesssim |r_{i,T}|^{2T-1} + |r_{j,T}|^{2T-1}$, we have $P_i Q_{F,T} P_j = O(q_*^{2T-1}) = o(T q_*^{2T-2})$ by Lemma E.3. By Assumption 3.2, we can write $(P_1 y)(P_1 y)^\top = c_1 u_1 u_1^\top$ and $(P_r y)(P_r y)^\top = c_r u_r u_r^\top$. Using Lemma E.3 and $\eta_{\infty,T} \to \eta_*$, we have

$$\begin{aligned}
\frac{P_1 Q_{F,T} P_1}{T q_*^{2T-2}} &= \left(r_{1,T}^{2T-1} - \frac{\eta_{\infty,T}(2T-1)\lambda_1}{m} r_{1,T}^{2T-2} \right) \frac{(P_1 y)(P_1 y)^\top}{T q_*^{2T-2}} \\
&\to -\frac{\eta_* 2\lambda_1}{m} a_1^* (P_1 y)(P_1 y)^\top = -\frac{2\eta_* h_*}{m} u_1 u_1^\top, \\
\frac{P_r Q_{F,T} P_r}{T q_*^{2T-2}} &= \left(r_{r,T}^{2T-1} - \frac{\eta_{\infty,T}(2T-1)\lambda_r}{m} r_{r,T}^{2T-2} \right) \frac{(P_r y)(P_r y)^\top}{T q_*^{2T-2}} \\
&\to -\frac{\eta_* 2\lambda_r}{m} a_r^* (P_r y)(P_r y)^\top = -\frac{2\eta_* h_*}{m} u_r u_r^\top.
\end{aligned}$$

Thus we obtain

$$\frac{P Q_{F,T} P}{T q_*^{2T-2}} = \sum_{i=1}^r \sum_{j=1}^r \frac{P_i Q_{F,T} P_j}{T q_*^{2T-2}} = \frac{P_1 Q_{F,T} P_1}{T q_*^{2T-2}} + \frac{P_r Q_{F,T} P_r}{T q_*^{2T-2}} + o(1) \to Q_{F,*}.$$

The loss coefficient is handled in the same way. Projecting $Q_{L,T}$ gives

$$P_i Q_{L,T} P_j = -\frac{\eta_{\infty,T}}{2m^2} \Gamma_{ij,T}^{(2T)} (P_i y)(P_j y)^\top.$$

Thus all off-diagonal blocks and all interior diagonal blocks are $o(T q_*^{2T-1})$. At the two endpoints, we have by Lemma E.3 and $\eta_{\infty,T} \to \eta_*$ that

$$\begin{aligned}
\frac{P_1 Q_{L,T} P_1}{T q_*^{2T-1}} &= -\frac{\eta_{\infty,T} 2T r_{1,T}^{2T-1} (P_1 y)(P_1 y)^\top}{2m^2 T q_*^{2T-1}} \to -\frac{\eta_*}{m^2}(-a_1^*)(P_1 y)(P_1 y)^\top = \frac{\eta_*}{m^2} c_1 a_1^* u_1 u_1^\top, \\
\frac{P_r Q_{L,T} P_r}{T q_*^{2T-1}} &= -\frac{\eta_{\infty,T} 2T r_{r,T}^{2T-1} (P_r y)(P_r y)^\top}{2m^2 T q_*^{2T-1}} \to -\frac{\eta_*}{m^2} a_r^* (P_r y)(P_r y)^\top = -\frac{\eta_*}{m^2} c_r a_r^* u_r u_r^\top.
\end{aligned}$$

Thus we obtain

$$\frac{P Q_{L,T} P}{T q_*^{2T-1}} = \sum_{i=1}^r \sum_{j=1}^r \frac{P_i Q_{L,T} P_j}{T q_*^{2T-1}} = \frac{P_1 Q_{L,T} P_1}{T q_*^{2T-1}} + \frac{P_r Q_{L,T} P_r}{T q_*^{2T-1}} + o(1) \to Q_{L,*}.$$

Finally, on $\mathrm{range}(K)$, Lemma E.3 gives

$$\|P B_T^{2T-1} P\|_{\mathrm{op}} = \left\| \sum_{i=1}^r r_{i,T}^{2T-1} P_i \right\| \leq \sum_{i=1}^r |r_{i,T}|^{2T-1} \lesssim q_*^{2T-1},$$

and similarly, $\|P B_T^{2T} P\|_{\mathrm{op}} \lesssim q_*^{2T}$. Thus we have

$$\frac{\|P g_{F,T}\|}{T q_*^{2T-2}} \to 0, \qquad \frac{\|P g_{L,T}\|}{T q_*^{2T-1}} \to 0.$$

□

G.5 Score linearization and weak convergence of the optimal learning rate

Lemma G.5 (Score linearization with an explicit remainder)**.** *Define*

$$S_{0,n} := KB_T^{2T-1}, \qquad S_{1,n} := E_n B_T^{2T-1} + K\mathcal{L}_{2T-1,n},$$
$$S_{2,n} := K\mathcal{R}_{2T-1,n} + E_n\mathcal{L}_{2T-1,n} + E_n\mathcal{R}_{2T-1,n},$$

and

$$R_{F,n,T} := -2f_n^{(0)}(X)^\top(S_{1,n}+S_{2,n})y + f_n^{(0)}(X)^\top(S_{0,n}+S_{1,n}+S_{2,n})f_n^{(0)}(X) + y^\top S_{2,n}y.$$

Then we have

$$F_{n,T}(\eta_{\infty,T}) - F_{\infty,T}(\eta_{\infty,T}) = (g_{F,T})^\top f_n^{(0)}(X) + \langle Q_{F,T}, E_n\rangle_{\mathrm{F}} + R_{F,n,T}$$

with

$$|R_{F,n,T}| = O_p\left(\frac{T^2 q_*^{2T-3}}{n}\right),$$

which consequently implies

$$\frac{\sqrt{n}R_{F,n,T}}{Tq_*^{2T-2}} = o_p(1).$$

Proof. By Lemma C.2, we have $PE_n = E_nP = E_n$ almost surely for all sufficiently large n. Hence P also commutes with Δ_n, B_T, $\mathcal{L}_{k,n}$, and $\mathcal{R}_{k,n}$ almost surely, and all three matrices $S_{0,n}, S_{1,n}, S_{2,n}$ vanish on $\ker(K)$. In addition, $A_nB_n(\eta_{\infty,T})^{2T-1}$ and $S_{0,n}$ are symmetric. $S_{1,n}$ is the Fréchet derivative at K, in the symmetric direction E_n, of the map

$$\mathcal{F}(A) := A\left(I_m - \frac{\eta_{\infty,T}}{m}A\right)^{2T-1}.$$

Since $\mathcal{F}(A)$ is symmetric whenever A is symmetric, and $K + tE_n$ is symmetric for every $t \in \mathbb{R}$, the difference quotient $(\mathcal{F}(K+tE_n) - \mathcal{F}(K))/t$ is symmetric. Passing to the limit $t \to 0$ therefore shows that $S_{1,n} = D\mathcal{F}(K)[E_n]$ is symmetric. Consequently, $S_{2,n}$ is symmetric as well.

Note that we have

$$A_nB_n(\eta_{\infty,T})^{2T-1} = S_{0,n} + S_{1,n} + S_{2,n}.$$

Thus, using $\chi_n^{(0)} = -y + f_n^{(0)}(X)$ and $F_{\infty,T}(\eta_{\infty,T}) = y^\top S_{0,n}y$, we obtain

$$\begin{aligned}
&F_{n,T}(\eta_{\infty,T}) - F_{\infty,T}(\eta_{\infty,T}) \\
&= -2f_n^{(0)}(X)^\top S_{0,n}y + y^\top S_{1,n}y - 2f_n^{(0)}(X)^\top(S_{1,n}+S_{2,n})y \\
&\quad + f_n^{(0)}(X)^\top(S_{0,n}+S_{1,n}+S_{2,n})f_n^{(0)}(X) + y^\top S_{2,n}y \\
&= (g_{F,T})^\top f_n^{(0)}(X) + y^\top S_{1,n}y + R_{F,n,T}.
\end{aligned}$$

Since the definitions of $\mathcal{L}_{2T-1,n}$ and Δ_n give

$$y^\top S_{1,n}y = y^\top E_n B_T^{2T-1}y - \frac{\eta_{\infty,T}}{m}\sum_{s=0}^{2T-2} y^\top KB_T^s E_n B_T^{2T-2-s}y = \langle Q_{F,T}, E_n\rangle_{\mathrm{F}},$$

we have

$$F_{n,T}(\eta_{\infty,T}) - F_{\infty,T}(\eta_{\infty,T}) = (g_{F,T})^\top f_n^{(0)}(X) + \langle Q_{F,T}, E_n\rangle_{\mathrm{F}} + R_{F,n,T}.$$

We next bound the remainder. Lemma F.3 implies

$$\|PS_{0,n}P\|_{\mathrm{op}} = O_p(q_*^{2T-1}).$$

Lemmas F.3 and G.3 imply

$$\|PS_{1,n}P\|_{\mathrm{op}} = O_p(Tq_*^{2T-2}\|E_n\|_{\mathrm{op}}).$$

Lemma G.3 and $T\|E_n\|_{\mathrm{op}} = o_p(1)$ imply

$$\|PS_{2,n}P\|_{\mathrm{op}} = O_p(T^2 q_*^{2T-3}\|E_n\|_{\mathrm{op}}^2).$$

Thus we have

$$\|PS_{1,n}P\|_{\mathrm{op}} + \|PS_{2,n}P\|_{\mathrm{op}} = O_p(Tq_*^{2T-2}\|E_n\|_{\mathrm{op}})$$

and

$$\|PS_{0,n}P\|_{\mathrm{op}} + \|PS_{1,n}P\|_{\mathrm{op}} + \|PS_{2,n}P\|_{\mathrm{op}} = O_p(q_*^{2T-1}).$$

By equation 32, we have $f_n^{(0)}(X) \overset{\text{a.s.}}{=} Pf_n^{(0)}(X)$ and $S_{j,n} \overset{\text{a.s.}}{=} PS_{j,n}P$ for $j = 0, 1, 2$. Hence, the definition of $R_{F,n,T}$ and the preceding bounds give

$$\begin{aligned}
&|R_{F,n,T}| \\
&\le |2f_n^{(0)}(X)^\top (S_{1,n} + S_{2,n})y| + |f_n^{(0)}(X)^\top (S_{0,n} + S_{1,n} + S_{2,n}) f_n^{(0)}(X)| + |y^\top S_{2,n} y| \\
&= O_p\left(\|f_n^{(0)}(X)\|(\|PS_{1,n}P\|_{\mathrm{op}} + \|PS_{2,n}P\|_{\mathrm{op}}) + \|f_n^{(0)}(X)\|^2 \sum_{j=0}^{2} \|PS_{j,n}P\|_{\mathrm{op}} + \|PS_{2,n}P\|_{\mathrm{op}} \right) \\
&= O_p(\|f_n^{(0)}(X)\| Tq_*^{2T-2}\|E_n\|_{\mathrm{op}} + \|f_n^{(0)}(X)\|^2 q_*^{2T-1} + T^2 q_*^{2T-3}\|E_n\|_{\mathrm{op}}^2) \\
&= O_p\left(\frac{T^2 q_*^{2T-3}}{n} \right),
\end{aligned}$$

where the last equality follows from $\|f_n^{(0)}(X)\| = O_p(n^{-1/2})$ and $\|E_n\|_{\mathrm{op}} = O_p(n^{-1/2})$ by Lemma 3.3. Consequently, we have

$$\frac{\sqrt{n}|R_{F,n,T}|}{Tq_*^{2T-2}} = O_p\left(\frac{T}{q_*\sqrt{n}} \right) = o_p(1).$$

□

Theorem G.6 (Limit distribution of the optimal learning rate at growing T)**.** *Under the assumptions of Theorem 3.4, the finite-width optimizer is interior with probability tending to one and we have*

$$\sqrt{n}(\eta_{n,T} - \eta_{\infty,T}) \xrightarrow{d} \Omega_{\eta,*} := -\frac{2m}{(\lambda_1 + \lambda_r)^2}(u_1^\top \Xi u_1 + u_r^\top \Xi u_r),$$

where Ξ is defined in Lemma 3.3.

Proof. Note that by Corollary C.4, we have $F_{\infty,T}(\eta_{\infty,T}) = 0$. Thus, Lemma G.5 implies

$$\frac{\sqrt{n}F_{n,T}(\eta_{\infty,T})}{Tq_*^{2T-2}} = \frac{\sqrt{n}[(g_{F,T})^\top f_n^{(0)}(X) + \langle Q_{F,T}, E_n\rangle_{\mathrm{F}}]}{Tq_*^{2T-2}} + o_p(1).$$

By equation 32 and Lemmas G.4 and 3.3, we have

$$\left| \frac{\sqrt{n}(g_{F,T})^\top f_n^{(0)}(X)}{Tq_*^{2T-2}} \right| \overset{\text{a.s.}}{=} \left| \frac{\sqrt{n}(Pg_{F,T})^\top f_n^{(0)}(X)}{Tq_*^{2T-2}} \right| \le \frac{\|Pg_{F,T}\|}{Tq_*^{2T-2}} \|\sqrt{n} f_n^{(0)}(X)\| = o_p(1).$$

Moreover, since $E_n = PE_nP$ by equation 32, we have by Lemmas G.4 and 3.3 that

$$\frac{\sqrt{n}\langle Q_{F,T}, E_n\rangle_{\mathrm{F}}}{Tq_*^{2T-2}} = \left\langle \frac{PQ_{F,T}P}{Tq_*^{2T-2}}, \sqrt{n}E_n \right\rangle_{\mathrm{F}} \xrightarrow{d} \langle Q_{F,*}, \Xi\rangle_{\mathrm{F}} = -\frac{2\eta_* h_*}{m}(u_1^\top \Xi u_1 + u_r^\top \Xi u_r).$$

Therefore we have

$$\frac{\sqrt{n}F_{n,T}(\eta_{\infty,T})}{Tq_*^{2T-2}} = \frac{\sqrt{n}\langle Q_{F,T}, E_n\rangle_{\mathrm{F}}}{Tq_*^{2T-2}} + o_p(1) \xrightarrow{d} -\frac{2\eta_* h_*}{m}(u_1^\top \Xi u_1 + u_r^\top \Xi u_r). \tag{40}$$

By Proposition C.3, we can write

$$F'_{\infty,T}(\eta_{\infty,T}) = -\frac{2T-1}{m}\sum_{j=1}^{r} \lambda_j^2 \|P_j y\|^2 r_{j,T}^{2T-2}.$$

Thus, by Lemma E.3, we have

$$\frac{F'_{\infty,T}(\eta_{\infty,T})}{Tq_*^{2T-2}} = -\frac{2T-1}{mT}\sum_{j=1}^{r} \lambda_j^2 \|P_j y\|^2 \frac{r_{j,T}^{2T-2}}{q_*^{2T-2}} \to -\frac{2}{m}(\lambda_1^2 \|P_1 y\|^2 a_1^* + \lambda_r^2 \|P_r y\|^2 a_r^*).$$

By Assumption 3.2, we have $\|P_1 y\|^2 = c_1$ and $\|P_r y\|^2 = c_r$. Therefore we have

$$\frac{F'_{\infty,T}(\eta_{\infty,T})}{T q_*^{2T-2}} \to -\frac{2}{m}(c_1\lambda_1^2 a_1^* + c_r\lambda_r^2 a_r^*) = -\frac{2h_*}{m}(\lambda_1+\lambda_r) < 0. \tag{41}$$

By the mean value theorem, there exists a random $\tilde{\eta}_{n,T}$ between $\eta_{n,T}$ and $\eta_{\infty,T}$ such that

$$F_{n,T}(\eta_{n,T}) - F_{n,T}(\eta_{\infty,T}) = F'_{n,T}(\tilde{\eta}_{n,T})(\eta_{n,T} - \eta_{\infty,T}).$$

Since $F_{n,T}(\eta_{n,T}) = 0$ holds on an event with probability tending to one, we have on the same event that

$$-F_{n,T}(\eta_{\infty,T}) = F'_{n,T}(\tilde{\eta}_{n,T})(\eta_{n,T} - \eta_{\infty,T}). \tag{42}$$

By Lemma G.2, we have $\tilde{\eta}_{n,T} - \eta_{\infty,T} = O_p(n^{-1/2})$. Hence equation 35 and equation 36 in the proof of Lemma G.2 imply

$$F'_{n,T}(\tilde{\eta}_{n,T}) - F'_{\infty,T}(\eta_{\infty,T}) = o_p(T q_*^{2T-2}). \tag{43}$$

By equation 41, this further yields

$$\frac{F'_{n,T}(\tilde{\eta}_{n,T})}{T q_*^{2T-2}} = -\frac{2h_*}{m}(\lambda_1+\lambda_r) + o_p(1),$$

thus $F'_{n,T}(\tilde{\eta}_{n,T})$ is negative with probability tending to one. Combining this with equation 42, on an event with probability tending to one, we have

$$\sqrt{n}(\eta_{n,T} - \eta_{\infty,T}) = -\frac{\sqrt{n}F_{n,T}(\eta_{\infty,T})/T q_*^{2T-2}}{F'_{n,T}(\tilde{\eta}_{n,T})/T q_*^{2T-2}}.$$

Therefore, equation 40, equation 41, and equation 43 yield

$$\sqrt{n}(\eta_{n,T} - \eta_{\infty,T}) \xrightarrow{d} -\frac{\eta_*}{\lambda_1+\lambda_r}(u_1^\top \Xi u_1 + u_r^\top \Xi u_r) = -\frac{2m}{(\lambda_1+\lambda_r)^2}(u_1^\top \Xi u_1 + u_r^\top \Xi u_r).$$

□

G.6 Fluctuations and weak convergence of the optimal loss

Lemma G.7 (Loss linearization with a quadratic remainder)**.** *Define*

$$R_{L,n,T} := \frac{1}{2m}[y^\top \mathcal{R}_{2T,n} y - 2 f_n^{(0)}(X)^\top (\mathcal{L}_{2T,n} + \mathcal{R}_{2T,n}) y + f_n^{(0)}(X)^\top B_n(\eta_{\infty,T})^{2T} f_n^{(0)}(X)].$$

Then we have

$$\phi_{n,T}(\eta_{\infty,T}) - \phi_{\infty,T}(\eta_{\infty,T}) = (g_{L,T})^\top f_n^{(0)}(X) + \langle Q_{L,T}, E_n\rangle_{\mathrm{F}} + R_{L,n,T}.$$

Thus

$$|R_{L,n,T}| = O_p\left(\frac{T^2 q_*^{2T-2}}{n}\right),$$

which implies

$$\frac{\sqrt{n}R_{L,n,T}}{T q_*^{2T-1}} = o_p(1).$$

Proof. Since $B_n(\eta_{\infty,T})^{2T} = B_T^{2T} + \mathcal{L}_{2T,n} + \mathcal{R}_{2T,n}$, expanding $\chi_n^{(0)} = -y + f_n^{(0)}(X)$ in the quadratic loss gives

$$\begin{aligned}
\phi_{n,T}(\eta_{\infty,T}) - \phi_{\infty,T}(\eta_{\infty,T}) &= \frac{1}{2m}(\chi_n^{(0)})^\top B_n(\eta_{\infty,T})^{2T}\chi_n^{(0)} - \frac{1}{2m} y^\top B_T^{2T} y \\
&= -\frac{1}{m} f_n^{(0)}(X)^\top B_T^{2T} y + \frac{1}{2m} y^\top \mathcal{L}_{2T,n} y + R_{L,n,T} \\
&= (g_{L,T})^\top f_n^{(0)}(X) + \frac{1}{2m} y^\top \mathcal{L}_{2T,n} y + R_{L,n,T}.
\end{aligned}$$

By the definition of $\mathcal{L}_{2T,n}$, we have

$$\frac{1}{2m}y^\top \mathcal{L}_{2T,n} y = -\frac{\eta_{\infty,T}}{2m^2}\sum_{s=0}^{2T-1} y^\top B_T^s E_n B_T^{2T-1-s} y = \langle Q_{L,T}, E_n\rangle_{\mathrm{F}},$$

which proves

$$\phi_{n,T}(\eta_{\infty,T}) - \phi_{\infty,T}(\eta_{\infty,T}) = (g_{L,T})^\top f_n^{(0)}(X) + \langle Q_{L,T}, E_n\rangle_{\mathrm{F}} + R_{L,n,T}.$$

By Lemma G.3, we have

$$\|P\mathcal{L}_{2T,n}P\|_{\mathrm{op}} = O_p(Tq_*^{2T-1}\|E_n\|_{\mathrm{op}}), \qquad \|P\mathcal{R}_{2T,n}P\|_{\mathrm{op}} = O_p(T^2 q_*^{2T-2}\|E_n\|_{\mathrm{op}}^2).$$

By Lemma F.3, we also have

$$\|PB_n(\eta_{\infty,T})^{2T}P\|_{\mathrm{op}} = O_p(q_*^{2T}).$$

By equation 32, we have

$$f_n^{(0)}(X) \overset{\text{a.s.}}{=} Pf_n^{(0)}(X), \qquad \mathcal{L}_{2T,n} \overset{\text{a.s.}}{=} P\mathcal{L}_{2T,n}P, \qquad \mathcal{R}_{2T,n} \overset{\text{a.s.}}{=} P\mathcal{R}_{2T,n}P.$$

Lemma 3.3 then gives

$$\begin{aligned}
|R_{L,n,T}| &\lesssim |y^\top \mathcal{R}_{2T,n} y| + |2f_n^{(0)}(X)^\top(\mathcal{L}_{2T,n}+\mathcal{R}_{2T,n})y| + |f_n^{(0)}(X)^\top B_n(\eta_{\infty,T})^{2T} f_n^{(0)}(X)| \\
&= O_p(T^2 q_*^{2T-2}\|E_n\|_{\mathrm{op}}^2 + Tq_*^{2T-1}\|E_n\|_{\mathrm{op}}\|f_n^{(0)}(X)\| + q_*^{2T}\|f_n^{(0)}(X)\|^2) \\
&= O_p\left(\frac{T^2 q_*^{2T-2}}{n}\right),
\end{aligned}$$

and

$$\frac{\sqrt{n}|R_{L,n,T}|}{Tq_*^{2T-1}} = O_p\left(\frac{T}{q_*\sqrt{n}}\right) = o_p(1).$$

□

Theorem G.8 (Limit distribution of the optimal loss at growing T)**.** *Under the assumptions of Theorem 3.4, we have*

$$\frac{\sqrt{n}}{Tq_*^{2T-1}}(\phi_{n,T}(\eta_{n,T}) - \phi_{\infty,T}(\eta_{\infty,T})) \xrightarrow{d} \Omega_{L,*} = \langle Q_{L,*}, \Xi\rangle_{\mathrm{F}}.$$

Proof. Lemma G.7 implies

$$\frac{\sqrt{n}}{Tq_*^{2T-1}}(\phi_{n,T}(\eta_{\infty,T}) - \phi_{\infty,T}(\eta_{\infty,T})) = \frac{\sqrt{n}}{Tq_*^{2T-1}}[(g_{L,T})^\top f_n^{(0)}(X) + \langle Q_{L,T}, E_n\rangle_{\mathrm{F}}] + o_p(1).$$

By equation 32 and Lemmas G.4 and 3.3, we have

$$\left|\frac{\sqrt{n}(g_{L,T})^\top f_n^{(0)}(X)}{Tq_*^{2T-1}}\right| \overset{\text{a.s.}}{=} \left|\frac{\sqrt{n}(Pg_{L,T})^\top f_n^{(0)}(X)}{Tq_*^{2T-1}}\right| \le \frac{\|Pg_{L,T}\|}{Tq_*^{2T-1}}\|\sqrt{n}f_n^{(0)}(X)\| = o_p(1).$$

Moreover, since $E_n = PE_nP$ by equation 32, we have by Lemma G.4 that

$$\frac{\sqrt{n}\langle Q_{L,T}, E_n\rangle_{\mathrm{F}}}{Tq_*^{2T-1}} = \left\langle \frac{PQ_{L,T}P}{Tq_*^{2T-1}}, \sqrt{n}E_n\right\rangle_{\mathrm{F}} \xrightarrow{d} \langle Q_{L,*}, \Xi\rangle_{\mathrm{F}}.$$

Thus we have

$$\frac{\sqrt{n}}{Tq_*^{2T-1}}(\phi_{n,T}(\eta_{\infty,T}) - \phi_{\infty,T}(\eta_{\infty,T})) \xrightarrow{d} \langle Q_{L,*}, \Xi\rangle_{\mathrm{F}}. \tag{44}$$

It remains to replace $\eta_{\infty,T}$ by $\eta_{n,T}$ in the finite-width loss. By Taylor's theorem, there exists a random $\bar{\eta}_{n,T}$ between $\eta_{\infty,T}$ and $\eta_{n,T}$ such that

$$\phi_{n,T}(\eta_{\infty,T}) - \phi_{n,T}(\eta_{n,T}) = \phi'_{n,T}(\eta_{n,T})(\eta_{\infty,T} - \eta_{n,T}) + \frac{1}{2}\phi''_{n,T}(\bar{\eta}_{n,T})(\eta_{\infty,T} - \eta_{n,T})^2.$$

By Lemma G.2, with probability tending to one, we have $\eta_{n,T} \in (\eta_\ell, \eta_u)$ and $F_{n,T}(\eta_{n,T}) = 0$. Hence, by Proposition C.3, we have on the same event that $\phi'_{n,T}(\eta_{n,T}) = 0$. Therefore, with probability tending to one, we have

$$\phi_{n,T}(\eta_{\infty,T}) - \phi_{n,T}(\eta_{n,T}) = \frac{1}{2}\phi''_{n,T}(\bar{\eta}_{n,T})(\eta_{\infty,T} - \eta_{n,T})^2. \tag{45}$$

By Lemma G.2, we have $|\bar{\eta}_{n,T} - \eta_{\infty,T}| = o_p(T^{-1})$. Moreover, equation 10 and Lemma C.2 yield

$$\begin{aligned}
&\sup_{|\eta-\eta_{\infty,T}|\le C/T} |\phi''_{n,T}(\eta)| = \frac{T(2T-1)}{m^3} \sup_{|\eta-\eta_{\infty,T}|\le C/T} |(\chi_n^{(0)})^\top A_n^2 B_n(\eta)^{2T-2}\chi_n^{(0)}| \\
&\overset{\text{a.s.}}{=} \frac{T(2T-1)}{m^3} \sup_{|\eta-\eta_{\infty,T}|\le C/T} |(P\chi_n^{(0)})^\top (PA_n^2P)(PB_n(\eta)^{2T-2}P)(P\chi_n^{(0)})| \\
&\le \frac{T(2T-1)}{m^3} \|P\chi_n^{(0)}\|^2 \|A_n\|_{\mathrm{op}}^2 \sup_{|\eta-\eta_{\infty,T}|\le C/T} \|PB_n(\eta)^{2T-2}P\|_{\mathrm{op}}.
\end{aligned}$$

Here, $\|P\chi_n^{(0)}\| = O_p(1)$ and $\|A_n\|_{\mathrm{op}} = O_p(1)$ hold by Lemma 3.3, while Lemma F.3 implies $\sup_{|\eta-\eta_{\infty,T}|\le C/T} \|PB_n(\eta)^{2T-2}P\|_{\mathrm{op}} = O_p(q_*^{2T-2})$. Thus we have

$$\sup_{|\eta-\eta_{\infty,T}|\le C/T} |\phi''_{n,T}(\eta)| = O_p(T^2 q_*^{2T-2}),$$

which implies $\phi''_{n,T}(\bar{\eta}_{n,T}) = O_p(T^2 q_*^{2T-2})$. We have by equation 45 and Lemma G.2 that

$$\frac{\sqrt{n}}{Tq_*^{2T-1}}(\phi_{n,T}(\eta_{n,T}) - \phi_{n,T}(\eta_{\infty,T})) = O_p\left(\frac{T}{q_*\sqrt{n}}\right) = o_p(1).$$

Therefore, by equation 44 and Slutsky's theorem, we obtain

$$\frac{\sqrt{n}}{Tq_*^{2T-1}}(\phi_{n,T}(\eta_{n,T}) - \phi_{\infty,T}(\eta_{\infty,T})) \xrightarrow{d} \langle Q_{L,*}, \Xi\rangle_{\mathrm{F}}.$$

□

G.7 Convergence rates of the optimal loss and the optimal learning rate

Lemma G.9 (Quadratic expansion of the transfer gap)**.** *Under the assumptions of Theorem 3.4, we have*

$$c_{n,T} = \frac{\phi''_{\infty,T}(\eta_{\infty,T})}{2}(\eta_{n,T} - \eta_{\infty,T})^2(1 + o_p(1)).$$

Proof. Recall that $\eta_{\infty,T} \in (\eta_\ell, \eta_u)$ for sufficiently large T. Thus Corollary C.4 implies $\phi'_{\infty,T}(\eta_{\infty,T}) = 0$. By Taylor's theorem, there exists a random $\eta^\dagger_{n,T}$ between $\eta_{n,T}$ and $\eta_{\infty,T}$ such that

$$c_{n,T} = \phi_{\infty,T}(\eta_{n,T}) - \phi_{\infty,T}(\eta_{\infty,T}) = \frac{\phi''_{\infty,T}(\eta_{\infty,T})}{2}(\eta_{n,T}-\eta_{\infty,T})^2 + \frac{1}{6}\phi'''_{\infty,T}(\eta^\dagger_{n,T})(\eta_{n,T}-\eta_{\infty,T})^3.$$

Differentiating equation 10 gives

$$\phi'''_{\infty,T}(\eta) = -\frac{T(2T-1)(2T-2)}{m^4} y^\top K^3 B_\infty(\eta)^{2T-3} y.$$

Lemma G.2 gives $|\eta^\dagger_{n,T} - \eta_{\infty,T}| = O_p(n^{-1/2}) = o_p(T^{-1})$. Therefore, with probability tending to one, $\eta^\dagger_{n,T}$ belongs to the C/T-neighborhood of $\eta_{\infty,T}$ in Lemma F.3. Thus, Lemma F.3 yields

$$|\phi'''_{\infty,T}(\eta^\dagger_{n,T})| = O_p(T^3 q_*^{2T-3}).$$

Together with equation 23 and Lemma G.2, we have

$$\left|\frac{\phi'''_{\infty,T}(\eta^\dagger_{n,T})(\eta_{n,T} - \eta_{\infty,T})/6}{\phi''_{\infty,T}(\eta_{\infty,T})/2}\right| = O_p\left(\frac{T^3 q_*^{2T-3}}{T^2 q_*^{2T-2}\sqrt{n}}\right) = O_p\left(\frac{T}{q_*\sqrt{n}}\right) = o_p(1),$$

which implies

$$c_{n,T} = \frac{\phi''_{\infty,T}(\eta_{\infty,T})}{2}(\eta_{n,T} - \eta_{\infty,T})^2(1 + o_p(1)).$$

□

Theorem G.10 (Rates of $a_{n,T}, b_{n,T}$, and $c_{n,T}$)**.** *Under the assumptions of Theorem 3.4, we have*

$$a_{n,T} = \Theta_p\left(\frac{Tq_*^{2T-1}}{\sqrt{n}}\right), \qquad b_{n,T} = \Theta_p(n^{-1/2}), \qquad c_{n,T} = \Theta_p\left(\frac{T^2q_*^{2T-2}}{n}\right).$$

Proof. We first show that the limit distribution of Theorem G.6 is nondegenerate. Recall from Lemma 3.3 that $\Xi = \alpha K + XGX^\top$, where $\alpha \sim \mathcal{N}(0, 2)$ is independent of G. The contribution of αK to $\Omega_{\eta,*}$ is

$$-\frac{2m}{(\lambda_1 + \lambda_r)^2}\alpha(\lambda_1 + \lambda_r) = -\frac{2m}{\lambda_1 + \lambda_r}\alpha = -\eta_*\alpha \sim \mathcal{N}(0, 2\eta_*^2),$$

which has strictly positive variance and is independent of the remaining Gaussian term. Hence $\Omega_{\eta,*}$ is nondegenerate, and we have

$$b_{n,T} = \Theta_p(n^{-1/2}). \tag{46}$$

We next show the non-degeneracy of the limit distribution of Theorem G.8. Since $c_r\lambda_r a_r^*/(c_1\lambda_1 a_1^*) = c_r\lambda_r e^{L_*}/(c_1\lambda_1) = 1$, we have

$$h_* := c_1\lambda_1 a_1^* = c_r\lambda_r a_r^* > 0. \tag{47}$$

This implies

$$\langle Q_{L,*}, K\rangle_{\mathrm{F}} = \frac{\eta_*}{m^2}(c_1 a_1^*\lambda_1 - c_r a_r^*\lambda_r) = 0.$$

Thus we can write

$$\Omega_{L,*} = \langle Q_{L,*}, XGX^\top\rangle_{\mathrm{F}} = \left\langle X^\top Q_{L,*}X, G\right\rangle_{\mathrm{F}}.$$

Recall from Lemma 3.3 that $\mathbb{E}(G_{rs}G_{tu}) = d^{-2}(\delta_{rt}\delta_{su} + \delta_{ru}\delta_{st})$. Thus for a deterministic symmetric matrix S, we have

$$\mathrm{Var}(\langle S, G\rangle_{\mathrm{F}}) = \sum_{r,s,t,u} S_{rs}S_{tu}\mathbb{E}(G_{rs}G_{tu}) = 2d^{-2}\|S\|_{\mathrm{F}}^2.$$

Therefore we have

$$\begin{aligned}
\mathrm{Var}(\Omega_{L,*}) &= 2d^{-2}\|X^\top Q_{L,*}X\|_{\mathrm{F}}^2 = 2\,\mathrm{tr}(Q_{L,*}KQ_{L,*}K) \\
&= 2\frac{\eta_*}{m^2}(c_1 a_1^* u_1^\top KQ_{L,*}Ku_1 - c_r a_r^* u_r^\top KQ_{L,*}Ku_r) \\
&= 2\frac{\eta_*}{m^2}(c_1 a_1^*\lambda_1^2 u_1^\top Q_{L,*}u_1 - c_r a_r^*\lambda_r^2 u_r^\top Q_{L,*}u_r) \\
&= 2\frac{\eta_*^2}{m^4}[(c_1 a_1^*\lambda_1)^2 + (c_r a_r^*\lambda_r)^2] \\
&> 0.
\end{aligned}$$

Hence $\Omega_{L,*}$ is a nondegenerate centered Gaussian random variable. Theorem G.8 therefore yields

$$a_{n,T} = \Theta_p\left(\frac{Tq_*^{2T-1}}{\sqrt{n}}\right).$$

Finally, Lemma G.9, equation 23 and equation 46 imply

$$c_{n,T} = \Theta_p\left(\frac{T^2q_*^{2T-2}}{n}\right).$$

□

Proof of Corollary 3.5. By Theorems G.8, G.10, and the continuous mapping theorem, we have

$$\frac{\sqrt{n}}{Tq_*^{2T-1}} a_{n,T} \xrightarrow{d} |\Omega_{L,*}|,$$

where $\Omega_{L,*}$ is a nondegenerate Gaussian random variable. Since $\mathrm{P}(|\Omega_{L,*}| = 0) = 0$, applying the continuous mapping theorem once more gives

$$\frac{Tq_*^{2T-1}}{\sqrt{n} a_{n,T}} \xrightarrow{d} \frac{1}{|\Omega_{L,*}|},$$

and consequently,

$$\frac{1}{a_{n,T}} = O_p\left(\frac{\sqrt{n}}{Tq_*^{2T-1}}\right).$$

Together with $c_{n,T} = \Theta_p\left(T^2 q_*^{2T-2}/n\right)$, this yields

$$\frac{c_{n,T}}{a_{n,T}} = \Theta_p\left(\frac{T}{q_*\sqrt{n}}\right) \xrightarrow{p} 0,$$

where the last convergence follows from $T/\sqrt{n} \to 0$ and the fact that $q_* > 0$ is fixed. □

H COUNTEREXAMPLE TO FAST TRANSFER

Proof of Proposition 3.8. Since $K = d^{-1} I_m$, we have by Corollary C.1 that

$$\phi_{\infty,T}(\eta) = \frac{\|y\|^2}{2m}\left(1 - \frac{\eta}{\eta_0}\right)^{2T}. \tag{48}$$

Then we have

$$\eta_{\infty,T} = \eta_0, \qquad \phi_{\infty,T}(\eta_{\infty,T}) = 0.$$

Define $\epsilon_n := n^{-1/2}$ and $M_n := \sqrt{n}(dA_n - I_m)$. Recall the Gaussian random variable and matrix α and G in Lemma 3.3, and define $M = \alpha I_m + dXGX^\top$. Here, $\mathcal{G} := dXGX^\top$ is a symmetric Gaussian matrix whose upper-triangular entries are independent and satisfy

$$\mathcal{G}_{ii} \sim \mathcal{N}(0, 2), \qquad \mathcal{G}_{ij} \sim \mathcal{N}(0, 1) \quad (i < j).$$

Then Lemma 3.3 and $XX^\top = I_m$ imply

$$(\chi_n^{(0)}, M_n) \xrightarrow{d} (-y, M). \tag{49}$$

We first localize the finite-width optimizer without using positive curvature at the infinite-width optimum. Since $K \neq 0$ and $A_n \succeq 0$, by Lemma C.2, we have for any fixed vector $u \in \mathrm{range}(K) \setminus \{0\}$ that $u^\top A_n u > 0$ almost surely, which implies that $A_n \succ 0$ almost surely. By Proposition C.5, we have $A_n \chi_n^{(0)} \neq 0$ almost surely, and consequently, $\chi_n^{(0)} \neq 0$ almost surely. Suppose $A_n = \sum_j \lambda_j(A_n) v_{n,j} v_{n,j}^\top$ is the orthonormal eigendecomposition of A_n. Then the score is expressed as

$$F_n(\eta) = \sum_j \lambda_j(A_n)\left(1 - \frac{\eta}{m}\lambda_j(A_n)\right)^{2T-1} (v_{n,j}^\top \chi_n^{(0)})^2.$$

Substituting $\eta = m/\lambda_{\max}(A_n)$ and $\eta = m/\lambda_{\min}(A_n)$, we have

$$F_n\left(\frac{m}{\lambda_{\max}(A_n)}\right) \geq 0, \qquad F_n\left(\frac{m}{\lambda_{\min}(A_n)}\right) \leq 0.$$

Thus the score characterization in Corollary C.4 implies that the unconstrained optimizer $\hat{\eta}_{n,T}$ satisfies

$$\frac{m}{\lambda_{\max}(A_n)} \leq \hat{\eta}_{n,T} \leq \frac{m}{\lambda_{\min}(A_n)}.$$

Since $A_n = d^{-1}(I_m + \epsilon_n M_n)$, we have $\lambda_{\max}(A_n) = d^{-1}[1 + \epsilon_n \lambda_{\max}(M_n)]$ and $\lambda_{\min}(A_n) = d^{-1}[1 + \epsilon_n \lambda_{\min}(M_n)]$. Noting that $M_n = O_p(1)$, we have $m/\lambda_{\max}(A_n) = \eta_0 + O_p(\epsilon_n)$ and

$m/\lambda_{\min}(A_n) = \eta_0 + O_p(\epsilon_n)$. Therefore $\hat{\eta}_{n,T}$ lies in (η_ℓ, η_u) with probability tending to one, and agrees there with $\eta_{n,T}$. In particular, we have

$$s_{n,T} := \epsilon_n^{-1}\left(\frac{\eta_{n,T}}{\eta_0} - 1\right) = O_p(1). \tag{50}$$

For $s \in \mathbb{R}$, the rescaled iteration matrix satisfies the exact identity

$$B_n(\eta_0(1+\epsilon_n s)) = -\epsilon_n(M_n + sI_m + \epsilon_n s M_n).$$

Consequently, the rescaled finite-width objective is

$$n^T \phi_{n,T}(\eta_0(1+\epsilon_n s)) = \frac{1}{2m}(\chi_n^{(0)})^\top (M_n + sI_m + \epsilon_n s M_n)^{2T} \chi_n^{(0)}. \tag{51}$$

Define $\mathcal{H}_T(s; D) := (2m)^{-1} y^\top (D + sI_m)^{2T} y$ for symmetric matrix D. Since T is fixed, the right-hand side of equation 51 is a polynomial in s whose coefficients converge jointly in distribution to those of $\mathcal{H}_T(s; M)$. Combining this with equation 49, for every $R < \infty$, we have

$$\left(M_n, \chi_n^{(0)}, n^T \phi_{n,T}(\eta_0(1+\epsilon_n s))\big|_{[-R,R]}\right) \xrightarrow{d} \left(M, -y, \mathcal{H}_T(\cdot; M)\big|_{[-R,R]}\right) \tag{52}$$

in $\mathbb{R}^{m\times m}_{\text{sym}} \times \mathbb{R}^m \times C([-R,R])$, where $C([-R,R])$ is equipped with the uniform norm.

For any symmetric D, suppose $\mu_1(D), \ldots, \mu_m(D)$ are eigenvalues of D, including zero and negative ones. Then a spectral decomposition writes $\mathcal{H}_T(s; D)$ as a nonnegative weighted sum of $(s+\mu_j(D))^{2T}$, with at least one strictly positive weight because $y \neq 0$. Moreover, for every $T \geq 1$, the map $s \mapsto (s+\lambda)^{2T}$ is strictly convex and coercive. Hence $\mathcal{H}_T(\cdot; D)$ is strictly convex and coercive, and therefore admits a unique minimizer $s_T(D)$. Since $\hat{\eta}_{n,T}$ agrees with $\eta_{n,T}$ with probability tending to one, the global minimizer of $s \mapsto n^T \phi_{n,T}(\eta_0(1+\epsilon_n s))$ agrees with $s_{n,T}$ on the same event. Thus the continuous mapping theorem and equation 52 yield

$$(M_n, \chi_n^{(0)}, s_{n,T}) \xrightarrow{d} (M, -y, s_T(M)).$$

Since $\phi_{\infty,T}(\eta_0) = 0$, we have $a_{n,T} = \phi_{n,T}(\eta_{n,T})$. Evaluating equation 51 at $s = s_{n,T}$ therefore gives

$$n^T a_{n,T} \xrightarrow{d} L_T := \mathcal{H}_T(s_T(M); M).$$

By equation 48, we further obtain

$$n^T c_{n,T} = \frac{\|y\|^2}{2m}|s_{n,T}|^{2T} \xrightarrow{d} C_T := \frac{\|y\|^2}{2m}|s_T(M)|^{2T}.$$

These convergences are joint; hence,

$$\left(\sqrt{n}\left(\frac{\eta_{n,T}}{\eta_0} - 1\right), n^T a_{n,T}, n^T c_{n,T}\right) \xrightarrow{d} (s_T(M), L_T, C_T).$$

Observe that the original random initial prediction is retained in equation 51. Since $\chi_n^{(0)} = -y + O_p(n^{-1/2})$, replacing $\chi_n^{(0)}$ by $-y$ changes the n^T-rescaled local objective by only $O_p(n^{-1/2})$, uniformly on compact sets of s. Hence the initialization fluctuation in $\chi_n^{(0)}$ does not contribute to the leading $O_p(1)$ local limit.

It remains to establish nondegeneracy. If $L_T = 0$, then $(M + s_T(M)I_m)^T y = 0$. By symmetry of M, this implies $(M + s_T(M)I_m)y = 0$, so y must be an eigenvector of M, and hence of $\mathcal{G}$. For a fixed nonzero y this event has probability zero when $m \geq 2$: writing $e = y/\|y\|$, the vector $(I_m - ee^\top)\mathcal{G}e = P_{e^\perp}\mathcal{G}e$ is a nondegenerate Gaussian vector on $e^\perp$. Thus $L_T > 0$ almost surely.

We next show that $C_T > 0$ almost surely. Since $M = \alpha I_m + \mathcal{G}$, for every $s \in \mathbb{R}$ we have $\mathcal{H}_T(s; M) = \mathcal{H}_T(s+\alpha; \mathcal{G})$. Therefore, adding the scalar perturbation αI_m only translates the minimizing argument, and hence we have

$$s_T(M) = s_T(\mathcal{G}) - \alpha.$$

Moreover, evaluating the objective at the minimizer gives

$$L_T = \mathcal{H}_T(s_T(M); M) = \mathcal{H}_T(s_T(\mathcal{G}); \mathcal{G}).$$

Thus L_T depends only on $\mathcal{G}$, whereas

$$C_T = \frac{\|y\|^2}{2m} |s_T(\mathcal{G}) - \alpha|^{2T}.$$

Conditional on $\mathcal{G}$, the quantity $s_T(\mathcal{G})$ is fixed and $\alpha \sim \mathcal{N}(0, 2)$ is independent of $\mathcal{G}$. Hence we have $\mathrm{P}(s_T(M) = 0 \mid \mathcal{G}) = 0$, and therefore $C_T > 0$ almost surely. Together with the preceding argument showing $L_T > 0$ almost surely, this implies

$$\mathrm{P}\left(0 < \frac{C_T}{L_T} < \infty\right) = 1.$$

Finally, the continuous mapping theorem proves

$$\frac{c_{n,T}}{a_{n,T}} \xrightarrow{d} \frac{C_T}{L_T} = \Theta_p(1).$$

□

Remark H.1 (Why the spectral condition matters)**.** In this example, we have $B_\infty(\eta_0) = 0$. Thus by Lemma C.6, the first-order loss sensitivity $D_{(\chi,A)}\Phi_T(\eta_0; -y, K)$ vanishes, and the $n^{-1/2}$ optimal-loss fluctuation used in Corollary 3.7 is absent. Instead, both loss gaps are of order n^{-T}. This is not a failure of learning rate consistency or of absolute transfer accuracy: $b_{n,T} \to 0$ and $c_{n,T} \to 0$ in probability still hold. What fails is the relative criterion $c_{n,T} = o_p(a_{n,T})$. We also note that $D_{(\chi,A)}F_T(\eta_0; -y, K)$ also vanishes when $T \geq 2$ while $D_{(\chi,A)}F_1(\eta_0; -y, K)[h, E] = -y^\top Ey$.