When Is Predicting Training Loss Worth the Cost of Initialization Selection?

Anonymous authors
Paper under double-blind review

Abstract

Selecting among random initializations consumes computation that could instead train the selected network. We study this tradeoff under a fixed total compute budget using two predictors of future training loss: a separate frozen empirical neural tangent kernel (NTK) for each candidate, and a shared infinite-width limiting NTK used with each candidate's own initial residual. For two-hidden-layer ReLU networks trained by full-batch gradient descent, the expected squared discrepancy between frozen-NTK predictions and actual outputs vanishes with width, uniformly over a fixed number of candidates and updates. We derive the limiting expected training loss after selection. Assume there are at least two candidates and the limiting initial-output covariance is positive definite. We compare selection by predicted future loss with selection by initial loss at the same training horizon. Future-loss selection yields a strictly lower limiting expected training loss if and only if the limiting dynamics scale residual directions by different absolute factors. Under a compute budget, this gain must also offset the updates forgone in scoring; we give an explicit sufficient condition in a realizable class with two error-decay rates. Extensive simulations illustrate the theoretical results and assess the tradeoff between prediction accuracy and computational cost.

1 Introduction

Random initialization can materially affect the outcome of neural-network training. If several initializations are available, one may try to predict which candidate will train best before committing the full training budget. Computing candidate scores uses resources that could otherwise support additional training updates. We ask when predicting future training loss is worth this cost when the architecture, optimizer, initialization distributions, and total compute budget are fixed.

We compare two NTK-based predictors. The first holds a separate empirical neural tangent kernel (NTK) fixed at each candidate's initialization. It can capture candidate-specific training dynamics, but constructing several kernels is comparatively expensive. The second reuses one infinite-width limiting NTK across all candidates while preserving their different initial residuals. It is cheaper, but discards finite-width differences among their kernels. For each method, we subtract its scoring cost from the budget and determine how many training updates remain. We score each candidate by its predicted loss after that number of updates. We compare them with random selection, initial-loss selection, brief trial training that retains the winner's trained state, and successive halving (Jamieson & Talwalkar, 2016). Every method is charged for candidate generation, scoring, exploratory updates, and final training.

This setting differs from architecture and hyperparameter search. Training-free scores based on activations, linear regions, gradient kernels, or NTK quantities are commonly used to rank architectures (Mellor et al., 2021; Xu et al., 2021; Chen et al., 2021), while Successive Halving and Hyperband allocate partial training resources across configurations (Jamieson & Talwalkar, 2016; Li et al., 2018). Here the architecture and training method do not change, and the selected network is one of the candidates drawn independently from the same prescribed distributions. Selection changes the distribution of the winner, not the distribution used to generate candidates.

Our analysis builds on Gaussian-process (GP) and NTK descriptions of wide networks (Jacot et al., 2018; Matthews et al., 2018; Lee et al., 2019). For an explicit two-hidden-layer ReLU model trained by discrete full-batch gradient descent, we keep the data dimension, sample size, candidate count,

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.

and training horizon fixed as width grows. We first show that freezing each candidate's empirical NTK at initialization approximates its actual training trajectory. Together with GP and NTK limits, this yields the limiting expected training loss of the selected candidate.

The limiting comparison has a simple interpretation. Initial-loss selection ranks residuals by Euclidean norm. Future-loss selection instead accounts for how training scales different residual directions. With at least two candidates and full-support limiting Gaussian initial outputs, future-loss selection improves expected loss at a fixed horizon exactly when those directions are scaled by unequal magnitudes. Under a fixed total budget, however, this selection gain must also exceed the progress lost through fewer training updates. We separate these two effects and give an explicit sufficient condition for the overall comparison in a class with two error-decay rates that is realized by the stated ReLU model. This construction shows that the selection gain can outweigh the cost of fewer training updates.

The experiments distinguish prediction accuracy from end-to-end performance. Experiments on 200 independent data sets per model show a budget-dependent crossover at lower hidden width 128: initial-loss selection is better at the smaller budget, while empirical NTK is better at the larger budget. With four candidates at the larger budget, empirical NTK also outperforms brief trials and successive halving. In the four-candidate comparisons, empirical NTK predicts training loss more accurately at a common horizon; under equal budgets, limiting NTK can achieve lower final test loss with additional updates. Comparisons across three widths examine how these two effects vary with network size.

Our contributions are:

1. We prove that frozen empirical-NTK dynamics approximate actual ReLU training as width grows, with candidate count and update count held fixed. We also derive the limiting expected training loss after selection.
2. For at least two candidates with full-support limiting Gaussian initial outputs, we characterize exactly when future-loss selection improves on initial-loss selection at a common training horizon and give a computable sufficient condition for that gain to survive the cost of fewer updates.
3. We show empirically that more accurate training-loss predictions need not yield lower final loss under a fixed compute budget. The comparisons reveal how prediction accuracy and scoring cost jointly affect initialization selection.

We define the model, scores, and compute allocation in Section 2 and establish the training approximation and selection criteria in Sections 3–4. We report experiments in Section 5, discuss related work in Section 6, and conclude with future directions in Section 7.

2 Problem, scores, and compute allocation

Fix training data $(x_i, y_i)_{i=1}^n$, $x_i \in \mathbb{R}^d$, and write $Y = (y_i)_{i=1}^n$ and $v_i = (x_i, 1)$. Index the M candidates by $m = 1, \dots, M$. The hidden layers have widths w and $2w$. Each candidate independently initializes

$$W_{jr} \sim N\left(0, \frac{2}{d+1}\right), \qquad V_{uj}, a_u \sim N(0,1), \qquad c_u \sim N(0, \tau^2),$$

with all entries and arrays mutually independent. Here $W \in \mathbb{R}^{w \times (d+1)}$, $V \in \mathbb{R}^{2w \times w}$, and $c, a \in \mathbb{R}^{2w}$. ReLU acts componentwise. The hidden features $h_i \in \mathbb{R}^w$ and scalar outputs f_i are

$$h_i = w^{-1/2} \operatorname{ReLU}(W v_i), \qquad f_i = (2w)^{-1/2} a^\top \operatorname{ReLU}(V h_i + c), \qquad F = (f_i)_{i=1}^n. \tag{1}$$

The width factors are included explicitly in the network output; gradient descent updates the arrays W, V, c, a. We minimize $L(F) = \|Y - F\|^2/(2n)$ by simultaneous gradient descent on the raw arrays W, V, c, a with learning rate η. Write θ for the concatenated parameter arrays and $J_t = \partial F/\partial \theta$ for the output Jacobian at update t. With the ReLU derivative set to zero at zero, the exact output update is induced by

$$\theta_{t+1} = \theta_t + \alpha J_t^\top (Y - F_t), \qquad \alpha = \eta/n. \tag{2}$$

The empirical initial full NTK is $K_0 = J_0 J_0^\top$. We write I for the identity matrix, $\|\cdot\|$ for the Euclidean vector norm, and $\|\cdot\|_F$ and $\|\cdot\|_{\mathrm{op}}$ for the Frobenius and operator matrix norms, respectively.

During training, changes in weights and ReLU activation patterns generally change the Jacobian and NTK. To predict future loss without training every candidate, we replace the evolving NTK by its initial value. The prediction uses the observed training labels and initial network; the selected network is subsequently trained by updating all its parameters. For candidate m, the predicted output and loss after k updates are:

$$F_{k,m}^{\mathrm{fr}} = Y - (I - \alpha K_{0,m})^k (Y - F_{0,m}), \qquad S_{k,m} = \frac{1}{2n} \|(I - \alpha K_{0,m})^k (Y - F_{0,m})\|^2. \tag{3}$$

A less expensive score replaces $K_{0,m}$ by the deterministic limiting kernel $\overline{K}$ in (5) while preserving each candidate's initial residual:

$$\widetilde{S}_{k,m} = \frac{1}{2n} \|(I - \alpha \overline{K})^k (Y - F_{0,m})\|^2. \tag{4}$$

Initial-loss selection uses $S_{0,m}$. Selection by predicted future loss uses $\widehat{m}_{w,k} \in \arg\min_{1 \le m \le M} S_{k,m}$ and trains that original candidate for k actual updates. A one-step-trial baseline instead chooses the candidate with smallest actual loss after one update, retains that paid step for the winner, and performs $k - 1$ additional updates so that its endpoint is update k. All ties use the smallest index.

We calibrate the allocation before evaluating any candidate trajectory. The reference budget is the measured cost of drawing one random candidate, training it for b updates, and evaluating it. Let g be the cost of one update, i_M the cost of generating M candidates, and e the cost of evaluating its final training and test losses. Then

$$C_{\mathrm{random}}(b) = i_1 + bg + e,$$

Let a_M be the cost of choosing the minimum among M candidate scores. Let q_b be the measured per-candidate cost of constructing an empirical NTK and propagating its residual through at most b steps. Scoring all candidates and training the winner for k updates costs

$$C_{\mathrm{target}}(k) = i_M + Mq_b + a_M + kg + e.$$

For the method that reuses one limiting NTK, let q_K be the one-time cost of constructing $\overline{K}$ and let q_b^R be the per-candidate cost of a forward pass, initial-residual formation, and residual propagation. Its cost is

$$C_{\mathrm{limit}}(k) = i_M + q_K + Mq_b^R + a_M + kg + e.$$

Let $C_{\mathcal{M}}(j)$ be method $\mathcal{M}$'s total cost when the winner receives j updates, including candidate generation, selection, training, and final evaluation. We allocate

$$k_{\mathcal{M}} = \max\{j \in \{r_{\mathcal{M}}, \ldots, b\} : C_{\mathcal{M}}(j) \le C_{\mathrm{random}}(b)\}.$$

Here $r_{\mathcal{M}}$ is 0 for random, initial-loss, and NTK selection, 1 or 2 for the corresponding trial, and 4 for successive halving. An empty feasible set means that the required screening and final evaluation cannot fit within the budget. Each NTK method predicts loss at its allocated horizon k and trains only the winner for k actual updates. The scoring charges q_b and q_b^R are calibrated at the reference horizon b. The budget also includes every initialization, evaluation, exploratory update, and retained training path required by the method. Trajectories computed only for diagnostics or oracle comparisons do not affect selection and are excluded from the deployed cost.

Selection procedure. We first measure primitive costs and fix each method's allocation, then generate the candidates and select the candidate with the smallest score. For initial-loss and NTK-based methods, we train the winner from initialization. For training-based methods, we retain the winner's updated parameters and count its exploratory updates toward its final update total.

The implementation forms four analytic Jacobian-block Gram matrices rather than an $n \times P$ Jacobian, where P is the number of scalar parameters. With both widths of order w, forward computation, kernel construction, and k residual propagations together cost $O(nwd + nw^2 + n^2(w + d) + kn^2)$ arithmetic and $O(nw + n^2)$ working memory, besides $O(w^2 + wd)$ parameters.

3 From actual ReLU training to quadratic loss limits

We state the limiting objects explicitly. For a covariance matrix C with strictly positive diagonal entries, define $\vartheta_{ij} = \arccos(C_{ij}/\sqrt{C_{ii}C_{jj}})$ and

$$\mathcal{R}(C)_{ij} = \frac{\sqrt{C_{ii}C_{jj}}}{2\pi}[\sin\vartheta_{ij} + (\pi - \vartheta_{ij})\cos\vartheta_{ij}], \quad \mathcal{P}(C)_{ij} = \frac{\pi - \vartheta_{ij}}{2\pi}.$$

With $s_{ij} = v_i^\top v_j$, put

$$C_1 = \frac{2s}{d+1}, \quad T = \mathcal{R}(C_1), \quad D_1 = \mathcal{P}(C_1), \quad C_2 = T + \tau^2 \mathbf{1}\mathbf{1}^\top,$$

$$\Sigma = \mathcal{R}(C_2), \qquad D_2 = \mathcal{P}(C_2), \qquad \overline{K} = \Sigma + D_2 + T \odot D_2 + s \odot D_1 \odot D_2. \tag{5}$$

Here $s = (s_{ij})$, $\mathbf{1}$ is the all-ones vector, and $\odot$ denotes entrywise multiplication. The four terms in $\overline{K}$ are the limiting kernel contributions from derivatives with respect to a, c, V, W, respectively.

The next theorem connects three objects: the actual network trajectory, the trajectory obtained by freezing each candidate's empirical NTK at initialization, and the limiting quadratic-form scores obtained as width grows. The first connection justifies the finite-width predictor; the second makes expected loss after selection analytically tractable.

Theorem 1 (Initialization limit, approximation of training, and selected loss)**.** *Fix d, n, M, a maximum number of updates H, the data, $\tau > 0$, and finite $\eta \geq 0$ independently of w. Let $F_{t,m}^{\mathrm{act}}$ and $L_{t,m}^{\mathrm{act}} = \|Y - F_{t,m}^{\mathrm{act}}\|^2/(2n)$ denote actual training output and loss. Let $G_1, \ldots, G_M$ be independent $N(0, \Sigma)$ vectors, $A = I - (\eta/n)\overline{K}$, $B_t = (A^t)^\top A^t$, and*

$$Q_{t,m} = \frac{(Y - G_m)^\top B_t (Y - G_m)}{2n}.$$

Then, for every finite $p \geq 1$,

$$\mathbb{E}\|K_{0,m} - \overline{K}\|_F^p \to 0, \qquad (F_{0,m}, K_{0,m})_{m=1}^M \Rightarrow (G_m, \overline{K})_{m=1}^M,$$

and the finite score array $S = (S_{t,m})_{t \leq H, m \leq M}$ converges jointly to $Q = (Q_{t,m})_{t \leq H, m \leq M}$. Moreover, $\mathbb{E}\Psi(S) \to \mathbb{E}\Psi(Q)$ for every continuous Ψ satisfying $|\Psi(z)| \leq C(1 + \|z\|^q)$ for some finite $C, q > 0$. Moreover,

$$\mathbb{E} \max_{m \leq M, t \leq H} \|F_{t,m}^{\mathrm{act}} - F_{t,m}^{\mathrm{fr}}\|^2 \to 0, \qquad \mathbb{E} \max_{m \leq M, t \leq H} |L_{t,m}^{\mathrm{act}} - S_{t,m}| \to 0. \tag{6}$$

Consequently, for fixed $0 \leq k \leq b \leq H$,

$$\mathbb{E}L_{b,1}^{\mathrm{act}} - \mathbb{E}L_{k,\widehat{m}_{w,k}}^{\mathrm{act}} \longrightarrow \Delta_\infty(b,k) := \mathbb{E}Q_{b,1} - \mathbb{E} \min_{m \leq M} Q_{k,m}. \tag{7}$$

Proof roadmap. Conditional Gaussian heads give the GP limit after feature-covariance laws of large numbers. The head, bias, and upper-weight kernel blocks follow from conditional row averages. The lower block is more delicate because upper gates and weights are dependent. We first average the independent Gaussian head as a quadratic form, then condition on the lower features and use a rank-at-most-two Gaussian subspace calculation to show that the dependence correction is $O(w^{-1})$.

For (6), we control gate changes and all four Jacobian blocks uniformly on a parameter ball. Exact integration along update segments handles ReLU gate crossings, and an induction places each segment inside the ball before the local bounds are applied. Global moment bounds then yield convergence in expected squared output error. The supplement gives the complete proof, including singular covariance cases.

Equation (7) combines two distinct approximations. First, each frozen empirical-NTK trajectory approximates the finite network with the same initialization; this step does not require a deterministic kernel limit and can be assessed directly at finite width. Second, the empirical scores converge jointly to quadratic forms defined by the limiting kernel and GP initialization. Uniform integrability then transfers this joint limit to expected loss after selection.

Remark *Although G_m is Gaussian, $Q_{t,m}$ is a noncentral Gaussian quadratic form. Its first two moments are*

$$\mathbb{E}Q_{t,m} = \frac{Y^\top B_t Y + \operatorname{tr}(B_t \Sigma)}{2n}, \tag{8}$$

$$\operatorname{Var}(Q_{t,m}) = \frac{2\operatorname{tr}[(B_t\Sigma)^2] + 4Y^\top B_t \Sigma B_t Y}{4n^2}. \tag{9}$$

For each candidate, the scores at different steps are computed from a single Gaussian initial output G_m, and are therefore generally dependent across steps. Different candidates have independent initial outputs. A deterministic limiting kernel can still produce random candidate scores through their initial outputs.

4 WHEN SELECTION BY PREDICTED FUTURE LOSS IS WORTHWHILE

Predicting future loss may select a better candidate when both methods receive the same number of updates. To beat a cheaper method under the same total compute budget, that advantage must also compensate for the training updates forgone because of scoring costs. Let

$$\widehat{m}_0 \in \operatorname*{arg\,min}_{1\le m\le M} Q_{0,m}, \qquad D_k = \mathbb{E}\left[Q_{k,\widehat{m}_0} - \min_{m\le M} Q_{k,m}\right].$$

Ties are broken by the smallest candidate index. Here D_k is the expected selection gain at a common horizon k: it compares the candidate chosen by initial loss with the candidate chosen by predicted loss, after both have received k updates.

Let $\mathrm{range}(\Sigma) = \{\Sigma u : u \in \mathbb{R}^n\}$ be the column space of Σ. We write $\Sigma \succ 0$ when Σ is positive definite, that is, $v^\top \Sigma v > 0$ for every nonzero $v \in \mathbb{R}^n$. Recall that $B_k = (A^k)^\top A^k$, where $A = I - (\eta/n)\overline{K}$.

Proposition 1 (Candidate selection and the cost of fewer updates)**.** *For general n, $D_k \ge 0$, with strict inequality exactly when $\mathbb{P}(Q_{k,\widehat{m}_0} > \min_{m\le M} Q_{k,m}) > 0$. Moreover, for $M \ge 2$, a sufficient condition for $D_k > 0$ is the existence of residuals $r, s \in Y - \mathrm{range}(\Sigma)$ such that*

$$\|r\|^2 < \|s\|^2, \qquad r^\top B_k r > s^\top B_k s. \tag{10}$$

If initial-loss selection affords $k_0 \ge k$ updates, the difference between its limiting expected loss and that of selection by predicted future loss, under the same total compute budget, is

$$\mathbb{E}Q_{k_0,\widehat{m}_0} - \mathbb{E}\min_{m\le M} Q_{k,m} = D_k - \mathbb{E}[Q_{k,\widehat{m}_0} - Q_{k_0,\widehat{m}_0}]. \tag{11}$$

The corresponding difference in expected finite-width loss converges to (11)*. If $\|A\|_{\mathrm{op}} \le 1$, the second term is nonnegative: giving the candidate selected by initial loss the additional updates cannot increase its limiting expected loss.*

Proposition 2 (Exact condition under full support)**.** *If $M \ge 2$ and $\Sigma \succ 0$, then, for every integer $k \in \{0, \dots, H\}$,*

$$D_k > 0 \quad \Longleftrightarrow \quad B_k \notin \{\beta I : \beta \ge 0\}. \tag{12}$$

No contraction condition is required for this comparison after the same number of updates.

The Gaussian initial output is supported on the column space of Σ, so $Y - G$ is supported on $Y - \mathrm{range}(\Sigma)$. Condition (10) means two possible initial residuals whose loss ranking reverses after k updates. When $\Sigma \succ 0$, every residual neighborhood has positive probability, and Proposition 2 reduces this question to a simple geometric condition: training must scale residual directions by unequal magnitudes. If Σ is singular, the inequalities in (10) must hold for residuals in $Y - \mathrm{range}(\Sigma)$, the support of the initial residual distribution. Whether Σ is positive definite or singular, a selection advantage after the same number of updates does not by itself imply an advantage under the same total compute budget.

Under the assumptions of Proposition 2, $D_k > 0$ for $k \ge 1$ if and only if the magnitudes $|1 - (\eta/n)\lambda_j(\overline{K})|^k$ are not all equal. Unequal kernel eigenvalues alone are insufficient; what matters is their induced scaling after k updates. Since $\overline{K} \succeq 0$, contraction is equivalent to $0 \le \eta \le 2n/\lambda_{\max}(\overline{K})$. We use contraction only to ensure that giving a candidate more updates cannot increase its limiting loss. Neither the common-horizon comparison nor Theorem 1 requires it.

Equation (11) therefore yields three cases. If $D_k = 0$, prediction does not improve selection at the common horizon, so any positive scoring cost makes it weakly worse under contraction. If $D_k > 0$ but is smaller than the benefit of the cheaper method's additional updates, prediction chooses a better candidate at step k but still has higher final loss. Predicted-loss selection is preferable only when its selection gain exceeds the loss reduction from the additional updates.

The next result gives an explicit sufficient condition under which selection by predicted future loss achieves a strictly lower limiting expected training loss than initial-loss selection under the same total compute budget. It considers a class in which components of the training error decay at two different rates. Its constants depend only on fixed chi-square order statistics and can be evaluated before training. Let k be the updates available to predicted-loss selection and Δ the additional updates available to initial-loss selection under the same budget.

Proposition 3. *Consider the training losses $Q_{t,m}$ in the infinite-width limit $w \to \infty$, as defined in Theorem 1. Suppose $n, M \geq 2$, $Y = 0$, and, in a common orthonormal basis with $p = n - 1$,*

$$\Sigma = \mathrm{diag}(s_s I_p, s_f), \qquad B_t = \mathrm{diag}(r_s^t I_p, r_f^t), \qquad s_s, s_f > 0, \quad 0 < r_f < r_s < 1.$$

Let $X_m \sim \chi_p^2$ and $Z_m \sim \chi_1^2$ be mutually independent, let $\widehat{m}_0 \in \arg\min_{1 \leq m \leq M}(s_s X_m + s_f Z_m)$, with smallest-index tie breaking, and define

$$c_{M,p} = \mathbb{E} \min_m X_m, \qquad v_s = s_s \mathbb{E} X_{\widehat{m}_0}, \qquad v_f = s_f \mathbb{E} Z_{\widehat{m}_0}.$$

For integers $k, \Delta \geq 0$ with $k + \Delta \leq H$, set $a_\Delta = v_s r_s^\Delta - s_s c_{M,p}$ and $b_\Delta = s_f - v_f r_f^\Delta$. Then

$$\mathbb{E} Q_{k+\Delta, \widehat{m}_0} - \mathbb{E} \min_{m \leq M} Q_{k,m} \geq \frac{a_\Delta r_s^k - b_\Delta r_f^k}{2n}. \tag{13}$$

Here $v_s > s_s c_{M,p}$ and $b_\Delta > 0$. Hence selection by predicted future loss has strictly lower limiting expected training loss than initial-loss selection whenever

$$a_\Delta > 0, \qquad k > \frac{\log(b_\Delta / a_\Delta)}{\log(r_s / r_f)}.$$

For every fixed $n, M \geq 2$ and integer $\Delta \geq 1$, these conditions are satisfied by some data and parameter settings of the model in (1)*, with a finite training horizon H independent of width.*

Here H is the maximum allowed number of training updates. The supplement gives data and parameter settings for the same ReLU model that satisfy the conditions of Proposition 3. The supplement also provides a proof that the strict loss advantage under the same compute budget holds for sufficiently wide finite networks.

5 Experiments

For each observation, the input $x = (x_1, x_2, x_3, x_4)^\top$ contains four independent features, each uniform on $[-1, 1]$. Training and test inputs are independent draws from this distribution. We use two regression functions:

$$f_{\mathrm{smooth}}(x) = 0.8 + 0.35x_1 - 0.20x_2 + 0.15x_3 x_4, \tag{14}$$

$$f_{\mathrm{mixed}}(x) = 0.45 \sin(2.2x_1) - 0.35x_2 x_3 + 0.25x_4^2 - 0.10. \tag{15}$$

We use a two-hidden-layer network with widths 128 and 256. Each data set contains 32 training observations and 256 test observations. We set the learning rate to $\eta = 0.05$ and the initialization standard deviation of the upper-layer biases to $\tau = 0.5$.

We generate each training label as $y = f(x) + \varepsilon$, where $\varepsilon \sim N(0, 0.1^2)$ is independent across observations and of the inputs. We evaluate test loss against the noiseless values $f(x)$ on an independent test sample.

For each data set, we generate groups of independently initialized candidate networks using the distributions in Section 2. We apply all selection methods to the same candidate groups and evaluate them on the same data. We use these paired results to calculate loss differences and confidence intervals. The supplement provides the experimental protocol, computation costs, and training-update allocations.

For each regression model, we independently generate 200 data sets, each containing training and test samples. For each data set, we generate 128 groups of four independently initialized networks. We compare all methods using the same groups. To study candidate counts $M = 2, 3, 4$, we use the first two, the first three, or all four networks in each group.

We repeat each comparison at budgets $b = 20$ and $b = 50$. Each method selects a candidate using its own budget-based training allocation. For each method comparison, we first average the paired loss differences over the 128 candidate groups sharing each data set. We then use the resulting 200 independent data-set means to estimate the overall mean difference and its confidence interval.

Table 1 reports, for each regression model and selection method with four candidates, the maximum number of training updates available within each compute budget.

Table 1: Training-update allocations at $w = 128, M = 4$ after charging measured initialization, selection, updates, and final training and test evaluation. Each model uses its own maximum feasible integer. SH denotes successive halving.

Model	b	Random	Initial	One-step	Two-step	SH	Empirical	Limiting
Eq. (14)	20	20	14	11	8	5	10	12
Eq. (14)	50	50	44	41	38	35	39	41
Eq. (15)	20	20	13	10	7	4	9	12
Eq. (15)	50	50	43	40	37	34	38	40

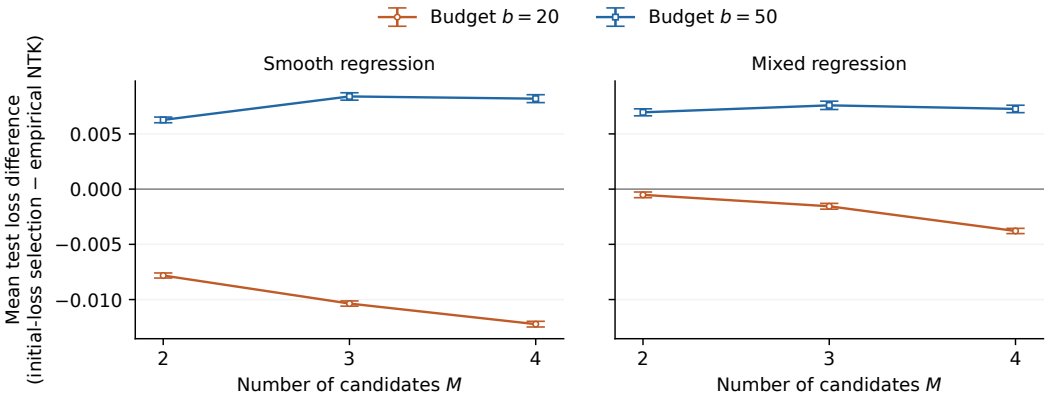

Figure 1: Difference in test half-MSE at $w = 128$: initial-loss selection minus empirical-NTK selection. Positive values favor empirical-NTK selection. Orange circles denote budget $b = 20$; blue squares denote $b = 50$. Error bars show pointwise 95% confidence intervals for the mean paired loss difference at each candidate count and budget, calculated from 200 independent data-set means, each averaging 128 paired candidate groups.

5.1 Budget dependence and comparison with training-based selection

We first compare initial-loss selection with empirical-NTK selection across candidate counts and budgets in Figure 1. We report the difference in final test loss between initial-loss selection and empirical-NTK selection (initial-loss selection minus empirical-NTK selection), averaged over data sets. Positive values favor empirical-NTK selection. We obtain mean differences of -0.01223 $[-0.01248, -0.01197]$ for the smooth model and -0.00379 $[-0.00403, -0.00355]$ for the mixed model at $M = 4, b = 20$. We obtain corresponding differences of 0.00819 $[0.00783, 0.00855]$ and 0.00726 $[0.00692, 0.00760]$ at $b = 50$. We therefore find lower mean test loss with initial-loss selection at the smaller budget and with empirical-NTK selection at the larger budget.

We next compare empirical-NTK selection with training-based selection methods. We use the one-step trial introduced in Section 2 and add a two-step trial that gives each candidate two updates before selection. In both cases, we select the candidate with the lowest observed training loss and continue training from its updated parameters. We also evaluate successive halving: we train four candidates for two updates, retain two and train them to update four, and then continue training the better survivor. We include the computation used to evaluate and train every candidate in the total budget, including candidates that are later discarded. We use the remaining budget to continue training the selected candidate. We use successive halving only for groups of four candidates.

We report all six comparisons with empirical-NTK selection at $M = 4, b = 50$ in Table 2. For both regression models, we find lower mean test loss with empirical NTK than with random selection, initial-loss selection, either trial method, or successive halving. Each corresponding pointwise confidence interval excludes zero. However, we obtain still lower mean test loss with limiting-NTK selection. In Figure 2, we show how these paired loss differences vary across the 200 independent data sets; each point averages the 128 candidate groups within one data set.

Table 2: Difference in test half-MSE at $w = 128, M = 4, b = 50$: the named method minus empirical-NTK selection. Positive values favor empirical-NTK selection. MCSE denotes Monte Carlo standard error. Pointwise 95% intervals use independent data-set means and a normal critical value.

Model	Comparator	Mean	MCSE	95% interval
Eq. (14)	Random	0.01431	0.00026	[0.01379, 0.01482]
Eq. (14)	Initial loss	0.00819	0.00018	[0.00783, 0.00855]
Eq. (14)	One-step trial	0.00946	0.00018	[0.00912, 0.00981]
Eq. (14)	Two-step trial	0.01096	0.00017	[0.01062, 0.01130]
Eq. (14)	Successive halving	0.01241	0.00017	[0.01208, 0.01273]
Eq. (14)	Limiting NTK	-0.00028	0.00006	[-0.00039, -0.00017]
Eq. (15)	Random	0.01965	0.00027	[0.01911, 0.02019]
Eq. (15)	Initial loss	0.00726	0.00017	[0.00692, 0.00760]
Eq. (15)	One-step trial	0.00817	0.00017	[0.00784, 0.00851]
Eq. (15)	Two-step trial	0.00918	0.00017	[0.00884, 0.00951]
Eq. (15)	Successive halving	0.00986	0.00016	[0.00955, 0.01017]
Eq. (15)	Limiting NTK	-0.00079	0.00005	[-0.00089, -0.00070]

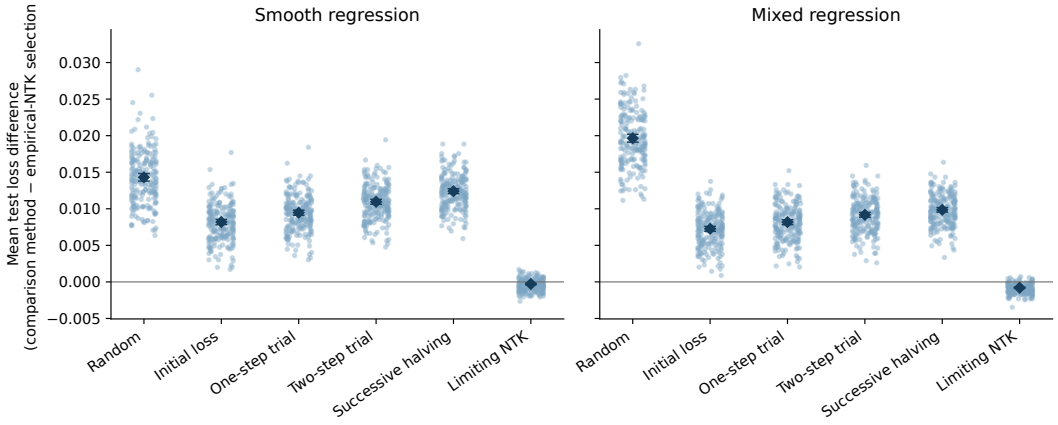

Figure 2: Paired test-loss differences at $w = 128, M = 4, b = 50$. Each dot is the mean of 128 paired groups on one of 200 independently generated data sets. Diamonds and bars show the overall mean and pointwise 95% normal-approximation interval. Positive values favor empirical NTK.

Prediction accuracy and computation cost. We first compare the accuracy of the two NTK predictors after the same number of training updates, using the update count allocated to empirical-NTK selection for both predictors. We measure the squared difference between each predicted and actual training loss, average over candidates and groups within each data set, and then average over the 200 data sets. We find lower mean squared prediction error with empirical NTK in all four model–budget combinations at $M = 4$ (Table 6 in the supplement). We then compare final test loss under the same total compute budget, allowing each method the training updates left after paying its selection costs. At $b = 50$, limiting-NTK selection receives two more updates in each model (Table 1) and achieves lower mean test loss (Table 2). Thus, we obtain more accurate predictions with empirical NTK at the same update count, but lower final test loss with limiting NTK under the same budget.

Width dependence. We compare the two NTK methods at $w = 64, 128, 256$, with four candidates and 10 or 40 updates on the same data sets. We find more accurate training-loss predictions and lower mean test loss with empirical NTK in all 12 comparisons across the two regression models; 11 pointwise test-loss intervals exclude zero. As width increases, we observe smaller differences between the kernels and their predictions, and a smaller test-loss advantage for empirical NTK (Figure 1 and Tables 9–11 in Section 7 of the supplement). Under the same compute budget, however, we often obtain lower test loss with limiting NTK because its lower scoring cost allows more training updates (Table 8 in Section 7 of the supplement).

Additional synthetic and real-data experiments. We also compare the methods in 24 additional synthetic settings and on the diabetes data set. We vary the regression function, network width, sample size, and compute budget. Alongside the two regression functions above, we include oscillatory and radial functions. We draw the four input coordinates independently from the uniform distribution on $[-1, 1]$, add independent Gaussian noise to the training labels, and evaluate test loss against the noiseless regression values. We use 20 data sets per synthetic setting or 20 random splits of the diabetes data, with 64 candidate groups per data set or split. For diabetes, the confidence intervals reflect variation from data splitting and network initialization with the source data held fixed. We report the design and results in Section 8 of the supplement, including how candidate selection and additional training updates contribute to the loss differences.

6 Related work

Wide-network dynamics and random initialization. We use GP limits to describe network outputs at initialization and NTK limits to describe their training dynamics (Jacot et al., 2018; Matthews et al., 2018; Lee et al., 2019; Novak et al., 2020). Related results cover nonzero initial functions and discrete NTK limits (Yang & Littwin, 2021), finite-time optimization (Arora et al., 2019), neural-network/NTK-regression approximation under gradient flow (Chen et al., 2024), and finite-width fluctuations (Bordelon & Pehlevan, 2023; Yaida, 2020). We study a different target: expected loss after selecting among random initializations under discrete updates.

Initialization design and selection among candidates. For rectified linear units, the Kaiming method scales the weight variance according to the number of inputs to each unit, while Layer-Sequential Unit-Variance (LSUV) initialization starts from orthogonal weights and rescales each layer to have unit output variance (He et al., 2015; Mishkin & Matas, 2016). MetaInit and GradInit adjust the initial network using optimization-related objectives (Dauphin & Schoenholz, 2019; Zhu et al., 2021); Laor constructs a new initialization from clusters of evaluated weight samples (Boongasame et al., 2025). These methods construct or adjust initial parameters. We study selection among parameter vectors drawn independently from the fixed normal distributions in Section 2.

Performance proxies and resource allocation. Training-free architecture search uses activations, gradient kernels, NTK conditioning, or linear regions (Mellor et al., 2021; Xu et al., 2021; Chen et al., 2021). Mok et al. (2022) examine limitations of initial-NTK scores and propose a measure that uses a small amount of training. We keep the architecture and optimizer fixed and vary only the parameter values sampled independently from the specified normal distributions. Our prediction depends on each candidate's initial residual and a specified number of future updates. Ahuja et al. (2020) learn a predictor of final objective values from previously solved instances, then use it to choose among independently sampled initializations for a new instance. We instead derive predictions from the NTK on the current data and analyze selection under a compute budget. Koren et al. (2026) screen random seeds using two training epochs per candidate and include this cost in their comparison.

Successive Halving allocates training in stages, and Hyperband runs several candidate-resource schedules (Jamieson & Talwalkar, 2016; Li et al., 2018). Our allocation is fixed in advance. We compare with Successive Halving on four initializations, charge every evaluation and update, and continue from surviving parameter states.

7 Conclusion and future work

We give conditions under which selecting an initialization by predicted training loss offsets the training updates lost to selection costs. We distinguish the benefit of choosing a better candidate from the benefit of training it longer. In our experiments, empirical NTK predicts training loss more accurately at the same update count, while the cheaper limiting NTK often achieves lower test loss under the same compute budget. We use each method's scoring cost to determine its prediction horizon and the number of training updates available to the selected network. We aim to extend the analysis to deeper networks and other optimizers, and develop a practical rule for choosing which NTK predictor to use under a given compute budget.

Reproducibility statement. We provide proofs and experimental details in the supplement. The code used for the simulations and analyses has been submitted as supplementary material.

AI use statement. Generative AI tools assisted with theoretical development, proof drafting and auditing, experimental design, implementation of simulations and synthetic-data generation, interpretation of results, manuscript editing, and consistency checks. The authors take responsibility for the final content.

REFERENCES

Kartik Ahuja, Amit Dhurandhar, and Kush R. Varshney. Learning to initialize gradient descent using gradient descent, 2020. arXiv:2012.12141.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 322–332. PMLR, 2019.

Laor Boongasame, Jirapond Muangprathub, and Karanrat Thammarak. Laor initialization: A new weight initialization method for the backpropagation of deep learning. *Big Data and Cognitive Computing*, 9(7):181, 2025. doi: 10.3390/bdcc9070181.

Blake Bordelon and Cengiz Pehlevan. Dynamics of Finite Width Kernel and Prediction Fluctuations in Mean Field Neural Networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0426.

Guhan Chen, Yicheng Li, and Qian Lin. On the impacts of the random initialization in the neural tangent kernel theory. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Wuyang Chen, Xinyu Gong, and Zhangyang Wang. Neural architecture search on ImageNet in four GPU hours: A theoretically inspired perspective. In *International Conference on Learning Representations*, 2021.

Yann N. Dauphin and Samuel S. Schoenholz. MetaInit: Initializing learning by learning to initialize. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 240–248. PMLR, 2016.

Oded Koren, Yoav Gvili, and Liron J. Friedman. Strategic approach for enhancing deep learning models. *Algorithms*, 19(4):311, 2026. doi: 10.3390/a19040311.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

Joe Mellor, Jack Turner, Amos Storkey, and Elliot J. Crowley. Neural architecture search without training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7588–7598. PMLR, 2021.

Dmytro Mishkin and Jiri Matas. All you need is a good init. In *International Conference on Learning Representations*, 2016.

Jisoo Mok, Byunggook Na, Ji-Hoon Kim, Dongyoon Han, and Sungroh Yoon. Demystifying the neural tangent kernel from a practical perspective: Can it be trusted for neural architecture search without training? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11861–11870, 2022.

Roman Novak, Lechao Xiao, Jiri Hron, Jaehoon Lee, Alexander A. Alemi, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. Neural tangents: Fast and easy infinite neural networks in Python. In *International Conference on Learning Representations*, 2020.

Jingjing Xu, Liang Zhao, Junyang Lin, Rundong Gao, Xu Sun, and Hongxia Yang. KNAS: Green neural architecture search. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11613–11625. PMLR, 2021.

Sho Yaida. Non-Gaussian processes and neural networks at finite widths. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 107, pp. 165–192, 2020.

Greg Yang and Etai Littwin. Tensor programs IIb: Architectural universality of neural tangent kernel training dynamics. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 11762–11772, 2021.

Chen Zhu, Renkun Ni, Zheng Xu, Kezhi Kong, W. Ronny Huang, and Tom Goldstein. GradInit: Learning to initialize neural networks for stable and efficient training. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

SUPPLEMENT: WHEN IS PREDICTING TRAINING LOSS WORTH THE COST OF INITIALIZATION SELECTION?

Anonymous authors
Paper under double-blind review

1 ORGANIZATION OF THE SUPPLEMENT

We first prove that holding the empirical neural tangent kernel (NTK) fixed approximates actual training, and then derive the limit of the initialization distribution. Section 3 first controls gate changes in both layers uniformly on a parameter ball, bounds all four Jacobian blocks, and shows that every update segment remains in that ball before applying segment calculus. A global polynomial bound and a prescribed probability for the complementary event then yield convergence in expected squared output error and uniform fourth-moment bounds. The proof establishes the training approximation directly from the finite-width empirical kernel.

Section 4 derives the Gaussian-process (GP) limit of the initial outputs and the limit of the full neural tangent kernel (NTK). Conditional Gaussian integration, empirical covariance limits, and a Gaussian projection calculation in a subspace of dimension at most two control the dependence between the upper-layer gates and upper weights in the bottom Jacobian block. Uniform integrability yields joint convergence of the scores and convergence of their expectations. Section 5 then applies these results to selection rules that depend on the initialization, exact compute allocations, an explicit example that improves on random initialization, and the comparison with initial-loss selection.

Throughout, d, n, M, H, the data, η, and $\tau > 0$ are fixed independently of width. The statements concern exact simultaneous gradient descent in the parameterization displayed in the paper and loss on the training labels. Repeated inputs and singular kernels are covered.

Section 6 gives the experimental designs, inference procedures, and cost definitions. Section 7 compares fixed-horizon and equal-budget performance across widths. Section 8 compares the two NTK scores across additional settings and decomposes their final loss difference.

2 EXACT MODEL AND NOTATION

Fix integers $d, n, M \geq 1$, $H \geq 0$, finite deterministic $x_i \in \mathbb{R}^d$, $Y = (y_i)_{i=1}^n \in \mathbb{R}^n$, $\tau > 0$, and $0 \leq \eta < \infty$. Write $[M] = \{1, \ldots, M\}$ and use I for the identity matrix, $\mathbf{1}$ for an all-ones vector, and $\mathbf{1}\{\cdot\}$ for an indicator. Put $v_i = (x_i, 1)$, $s_{ij} = v_i^\top v_j$, $\lambda = (2w)^{-1/2}$, and $\alpha = \eta/n$. Each of M independent candidates has raw arrays

$$W \in \mathbb{R}^{w \times (d+1)}, \qquad V \in \mathbb{R}^{2w \times w}, \qquad c, a \in \mathbb{R}^{2w},$$

whose initialization entries are mutually independent with

$$W_{jr} \sim N(0, 2/(d+1)), \qquad V_{uj}, a_u \sim N(0,1), \qquad c_u \sim N(0, \tau^2).$$

ReLU acts componentwise. At every parameter state define

$$z_{1i} = W v_i, \quad h_i = w^{-1/2} \operatorname{ReLU}(z_{1i}), \quad z_{2i} = V h_i + c,$$
$$\phi_i = \operatorname{ReLU}(z_{2i}), \quad f_i = \lambda a^\top \phi_i, \quad F = (f_i)_{i=1}^n.$$

Let $\mathsf{D}_{\ell i} = \operatorname{diag}(\mathbf{1}\{z_{\ell i} > 0\})$, with gate zero at zero. The parameter vector θ concatenates (W, V, c, a) with its sum-of-squares metric. Its chosen Jacobian, including at zero faces, has blocks

$$J_i^a = \lambda \phi_i^\top, \quad J_i^c = \lambda (\mathsf{D}_{2i} a)^\top, \quad \nabla_V f_i = \lambda (\mathsf{D}_{2i} a) h_i^\top, \quad \nabla_W f_i = b_i v_i^\top, \qquad b_i = \frac{\mathsf{D}_{1i} V^\top \mathsf{D}_{2i} a}{\sqrt{2} w}. \tag{1}$$

Matrix blocks are vectorized when placed in a Jacobian row. All arrays are updated simultaneously:

$$\theta_{t+1} = \theta_t + \alpha J(\theta_t)^\top r_t, \qquad r_t = Y - F(\theta_t), \qquad 0 \le t < H. \tag{2}$$

This is gradient descent on $\|Y - F\|^2/(2n)$. Write $K_{0,m} = J_{0,m} J_{0,m}^\top$ and

$$r_{t,m}^{\mathrm{fr}} = (I - \alpha K_{0,m})^t (Y - F_{0,m}), \quad F_{t,m}^{\mathrm{fr}} = Y - r_{t,m}^{\mathrm{fr}}, \quad S_{w,t,m} = \frac{\|r_{t,m}^{\mathrm{fr}}\|^2}{2n}. \tag{3}$$

The score $S_{w,t,m}$ is the main text's $S_{t,m}$, with width dependence explicit. Write $F_{t,m}^{\mathrm{act}} = F(\theta_{t,m})$ and $L_{w,t,m}^{\mathrm{act}} = \|Y - F_{t,m}^{\mathrm{act}}\|^2/(2n)$ for actual output and loss. The kernel normalization and initial output $F_{0,m}$ agree with the main text. Ties in candidate selection are resolved by the smallest index. Selection by predicted loss after k updates is $\widehat{m}_{w,k} \in \operatorname{argmin}_{m \in [M]} S_{w,k,m}$. The analysis defines hypothetical training paths for every candidate on one probability space, but an implementation needs to train only the selected candidate.

Vector norms are Euclidean and matrix norms are Frobenius unless explicitly marked op. Constants never depend on w but may depend on the fixed parameters and specified confidence or moment orders. Unqualified maxima over (m, t) mean $m \in [M]$ and $0 \le t \le H$.

For a positive-diagonal covariance matrix C, put $\vartheta_{ij} = \arccos(C_{ij}/\sqrt{C_{ii}C_{jj}})$ and

$$\mathcal{R}(C)_{ij} = \frac{\sqrt{C_{ii}C_{jj}}}{2\pi}[\sin\vartheta_{ij} + (\pi - \vartheta_{ij})\cos\vartheta_{ij}], \qquad \mathcal{P}(C)_{ij} = \frac{\pi - \vartheta_{ij}}{2\pi}.$$

Set

$$\begin{aligned} C_1 &= 2s/(d+1), & T &= \mathcal{R}(C_1), & D_1 &= \mathcal{P}(C_1), \\ C_2 &= T + \tau^2 \mathbf{1}\mathbf{1}^\top, & \Sigma &= \mathcal{R}(C_2), & D_2 &= \mathcal{P}(C_2). \end{aligned}$$

$$\overline{K} = \Sigma + D_2 + T \odot D_2 + s \odot D_1 \odot D_2, \qquad A = I - \alpha\overline{K}, \quad B_t = (A^t)^\top A^t. \tag{4}$$

Here $\odot$ is entrywise multiplication; D_1, D_2 are deterministic $n \times n$ arrays, distinct from state-dependent diagonal gates $\mathsf{D}_{\ell i}$. The four terms in (4) correspond to a, c, V, W. The displayed derivatives give the bias contribution D_2 and the lower-weight contribution $s \odot D_1 \odot D_2$.

Implementation correspondence. In the accompanying network code, the `kaiming` option configured to give hidden widths w and $2w$ draws parameters independently from the normal distributions above and implements this forward map and simultaneous update. Its empirical NTK is the sum of these four Gram matrices, and its predicted loss is computed by holding that kernel fixed and propagating the initial residual $Y - F_0$.

3 Approximation of actual training over a fixed number of updates

Theorem 3.1 (Unconditional approximation over fixed training steps)**.** *Under Section 2, as integer $w \to \infty$,*

$$\mathbb{E}\max_{m,t}\|F_{t,m}^{\mathrm{act}} - F_{t,m}^{\mathrm{fr}}\|^2 \longrightarrow 0, \qquad \sup_{w \ge 1}\mathbb{E}\max_{m,t}\big(\|F_{t,m}^{\mathrm{act}}\|^4 + \|F_{t,m}^{\mathrm{fr}}\|^4\big) < \infty.$$

The first expectation is $O(w^{-1/3}(1 + \log w)^2) + O(w^{-2})$. The result permits arbitrary fixed finite learning rates and singular kernels, and the proof retains the dependence induced by training.

3.1 Initialization event and moment bounds

Write $\ell_w = 1 + \log w$ and $\epsilon = w^{-1/3}$. For every fixed $a_* > 0$ there is an initialization event E_{a_*} with $\mathbb{P}(E_{a_*}^c) \le C_{a_*} w^{-a_*}$, simultaneously over candidates and inputs, on which

$$\begin{aligned} &\|W_0\|_F,\ \|V_0\|_{\mathrm{op}},\ \|a_0\|,\ \|c_0\| \le C\sqrt{w}, \\ &\|h_{i0}\| \le C, \quad \|\phi_{i0}\| \le C\sqrt{w}, \quad \|J_0\|_{\mathrm{op}} \le C_0, \end{aligned} \tag{5}$$

$$\max_u |a_{0u}| \le C_{a_*}\sqrt{\ell_w}, \quad \max_j |q_{i0,j}| \le C_{a_*}\sqrt{\ell_w/w},$$

$$q_{i0} = \frac{V_0^\top \mathsf{D}_{2i0} a_0}{\sqrt{2}w}, \quad \max_m \|F_{0,m}\| \le D_{a_*}\sqrt{\ell_w}, \tag{6}$$

$$\#\{j : |z_{1i0,j}| \le \epsilon\} \le Cw\epsilon + C_{a_*}\ell_w,$$
$$\#\{u : |z_{2i0,u}| \le \epsilon\} \le Cw\epsilon + C_{a_*}\ell_w. \tag{7}$$

The constants C, C_0 in (5) can be chosen independently of a_*. For every finite $p > 0$,

$$\sup_w \mathbb{E} \max_m \|F_{0,m}\|^p < \infty. \tag{8}$$

To prove these claims, completing the square gives $\mathbb{E}e^{tG} = e^{t^2/2}$ for a standard normal G, hence $\mathbb{P}(|G| > u) \le 2e^{-u^2/2}$. Gaussian integration also gives $\mathbb{E}e^{s\sum_{j=1}^k G_j^2} = (1-2s)^{-k/2}$ for $s < 1/2$. Markov at $s = 1/4$ bounds each required Gaussian vector norm by $C\sqrt{w}$ with exponentially small failure probability; d is fixed.

For the spectral bound, a maximal $1/4$-separated subset of the sphere in $\mathbb{R}^k$ is a $1/4$-net with at most 9^k points: disjoint radius-$1/8$ balls fit inside a radius-$9/8$ ball. Net approximations x', y' to unit x, y give $|y^\top V x| \le |y'^\top V x'| + \|V\|_{\mathrm{op}}/2$. Thus

$$\mathbb{P}(\|V_0\|_{\mathrm{op}} > 2u) \le 2\exp(3w\log 9 - u^2/2).$$

This proves the spectral claim and, by integration, all uniform moments of $\|V_0\|_{\mathrm{op}}/\sqrt{w}$. The vector bounds give the corresponding normalized moments for W, a, c. ReLU is Lipschitz, so $\|h_{i0}\| \le \|W_0\|_F\|v_i\|/\sqrt{w}$ and $\|\phi_{i0}\| \le \|V_0\|_{\mathrm{op}}\|h_{i0}\| + \|c_0\|$. Formula (1) bounds all four blocks by fixed constants, including $\|b_{i0}\| \le \|V_0\|_{\mathrm{op}}\|a_0\|/(\sqrt{2}w)$. Summing the fixed number of rows proves (5).

Conditionally on (W_0, V_0, c_0), the unmasked $q_{i0,j}$ is centered Gaussian of variance

$$\frac{\sum_u V_{0,uj}^2 \mathsf{D}_{2i0,uu}}{2w^2} \le \frac{\|V_0\|_{\mathrm{op}}^2}{2w^2} \le C/w$$

on the spectral event. Scalar Gaussian tails and a union bound over at most Mnw coordinates give its bound in (6); this bound applies to the jointly dependent coordinates and gates. A $2Mw$ union bound handles a. Conditionally on (W, V, c), f_{i0} is Gaussian of variance $\|\phi_{i0}\|^2/(2w) \le C$ on the norm event, giving the output bound. Enlarge D_{a_*} so that $\max_m \|r_{0,m}\| \le D_{a_*}\sqrt{\ell_w}$ as well.

Lower preactivations are independent across rows with variance $2\|v_i\|^2/(d+1) \ge 2/(d+1)$, so their small-ball probabilities are at most $C\epsilon$. Conditionally on W_0, upper preactivations are independent across rows with variance $\|h_{i0}\|^2 + \tau^2 \ge \tau^2$, giving the same bound uniformly in W_0. For a Bernoulli sum N_b with mean at most μ, $\mathbb{E}e^{N_b} \le e^{(e-1)\mu}$, hence $\mathbb{P}(N_b \ge e\mu + u) \le e^{-u}$. Take $\mu = Cw\epsilon$ and $u = (a_* + 2)\ell_w$ plus a fixed constant, and union-bound over the fixed inputs and candidates. This proves (7). Exponential failures are bounded by $C_{a_*}w^{-a_*}$ since $\sup_{w\ge 1} w^{a_*}e^{-cw} < \infty$. The Gaussian calculations condition on the specified parameter arrays before the bounds are combined into E_{a_*}.

Finally define the normalized quantities

$$T_W = \|W_0\|_F/\sqrt{w}, \qquad T_V = \|V_0\|_{\mathrm{op}}/\sqrt{w}, \qquad T_c = \|c_0\|/\sqrt{w}.$$

Deterministically $\|\phi_{i0}\|/\sqrt{2w} \le C(T_V T_W + T_c)$. Their normalized moments and Hölder's inequality bound every moment of this expression. Conditional Gaussian integration in the independent head yields $\mathbb{E}|f_{i0}|^p = c_p\mathbb{E}(\|\phi_{i0}\|^2/(2w))^{p/2} \le C_p$. Finite sums prove (8).

3.2 Gate changes and the four Jacobian blocks on a parameter ball

Fix $B < \infty$ and $R = B\sqrt{\ell_w}$. For sufficiently large w, $R \le \sqrt{w}$. On E_{a_*}, every $\|\theta - \theta_0\| \le R$ satisfies

$$\|z_{1i} - z_{1i0}\| \le CR, \quad \|h_i - h_{i0}\| \le CR/\sqrt{w}, \quad \|h_i\| \le C, \quad \|z_{2i} - z_{2i0}\| \le CR. \tag{9}$$

The first two follow by Lipschitz continuity. For the last use the exact difference $V_0(h_i - h_{i0}) + (V - V_0)h_i + (c - c_0)$, bounded by $C(R + R^2/\sqrt{w})$. A changed gate implies $|z_0| \le |z - z_0|$, including zero endpoints. Thus its flip count obeys, deterministically throughout the ball,

$$s_{\ell i} \le \#\{|z_{\ell i0}| \le \epsilon\} + \|z_{\ell i} - z_{\ell i0}\|^2/\epsilon^2 \le C_{a_*,B} w^{2/3} \ell_w. \tag{10}$$

This deterministic bound holds simultaneously throughout the parameter ball, including gate-change sets that depend on the parameters.

We claim

$$\sup_{\|\theta - \theta_0\| \le R} \|J(\theta) - J_0\|_F \le \delta_w := C_{a_*,B} w^{-1/6} \ell_w = o(1). \tag{11}$$

Suppress the input index and let $g = \mathsf{D}_2 a$, $g_0 = \mathsf{D}_{20} a_0$. Then

$$\|g - g_0\| \le R + C_{a_*}\sqrt{s_2 \ell_w}, \qquad \|g\| \le C\sqrt{w} + R.$$

The head difference is at most $CR/\sqrt{w}$, and the bias difference is at most $\lambda(R + C_{a_*}\sqrt{s_2\ell_w})$. The identity $gh^\top - g_0 h_0^\top = (g - g_0)h^\top + g_0(h - h_0)^\top$, with $\|uv^\top\|_F = \|u\|\|v\|$, bounds the V block by $C\{R/\sqrt{w} + \sqrt{s_2 \ell_w / w}\}$.

For the W block the exact decomposition is

$$\begin{aligned} b - b_0 = {} & (\mathsf{D}_1 - \mathsf{D}_{10}) q_0 \\ & + \frac{\mathsf{D}_1}{\sqrt{2}w}\big[(V - V_0)^\top \mathsf{D}_2 a + V_0^\top \mathsf{D}_2 (a - a_0) + V_0^\top (\mathsf{D}_2 - \mathsf{D}_{20}) a_0\big]. \end{aligned}$$

The first term is at most $C_{a_*}\sqrt{s_1 \ell_w / w}$. The bracket terms after scaling are at most $CR(\sqrt{w} + R)/w$, $CR/\sqrt{w}$, and $C_{a_*}\sqrt{s_2 \ell_w / w}$, respectively. Therefore

$$\|b - b_0\| \le C_{a_*}\{\sqrt{s_1 \ell_w / w} + R/\sqrt{w} + R^2/w + \sqrt{s_2 \ell_w / w}\}.$$

Multiply by $\|v_i\|$, substitute (10), and sum over the four blocks and fixed rows to obtain (11). Retaining unmasked q_0, rather than b_0, is essential at newly activated lower gates. Gate-dependent matrices were bounded by their full operator norms. For sufficiently large w, $\|J(\theta)\|_{\mathrm{op}} \le C_* := C_0 + 1$ throughout the ball. The value C_* is fixed before choosing B; only the width threshold depends on B.

3.3 Controlling each update segment

Along $\theta + su$, $0 \le s \le 1$, bottom preactivations are affine. Partition at their finitely many nonidentical zeros; on each interval bottom activations are affine and upper preactivations are degree-at-most-two polynomials. Further partition at their nonidentical roots. The output is a degree-at-most-three polynomial on every resulting interval and is continuous at endpoints, hence absolutely continuous. An identically zero preactivation has zero derivative, and the chosen zero gate gives its correct derivative. Elsewhere strict signs are locally fixed, so the product rule gives $J(\theta + su)u$ almost everywhere. Therefore

$$F(\theta + u) - F(\theta) = \int_0^1 J(\theta + su)u \, ds. \tag{12}$$

If the segment lies in the ball, subtracting $J(\theta)u$ yields a remainder of norm at most $2\delta_w\|u\|$. This identity includes all gate crossings and simultaneous-update product terms.

Put $g_* = 1 + \alpha C_*^2$, $T_H = \alpha C_* \sum_{j=0}^{H-1} g_*^j$, with empty sum zero. Choose $B = 2D_{a_*} T_H + 1$ and then a sufficiently large width for this fixed ball. Inductively, $\|r_t\| \le g_*^t \|r_0\|$ and the cumulative path length through step t is at most $\alpha C_* \|r_0\| \sum_{j=0}^{t-1} g_*^j$. These hold at zero. At step t, the endpoint Jacobian bound gives proposed step length at most $\alpha C_* g_*^t \|r_0\|$. Adding prior lengths gives at most $T_H D_{a_*}\sqrt{\ell_w} < R/2$. Thus the entire proposed segment is in the ball before using its interior. Equation (12) now gives $\|F_{t+1} - F_t\| \le \alpha C_*^2 \|r_t\|$ and closes the residual induction. In particular, maximum residual and total length are at most $C_H\|r_0\|$, and the step remainder ρ_t satisfies $\|\rho_t\| \le 2\alpha C_* \delta_w \|r_t\|$.

3.4 Comparison of the residual trajectories

The identity $K_t - K_0 = (J_t - J_0)J_t^\top + J_0(J_t - J_0)^\top$ gives $\|K_t - K_0\|_{\mathrm{op}} \le 2C_*\delta_w$. With $e_t = -\rho_t$ and $d_t = r_t - r_t^{\mathrm{fr}}$,

$$r_{t+1} = (I - \alpha K_t)r_t + e_t, \qquad d_{t+1} = (I - \alpha K_0)d_t - \alpha(K_t - K_0)r_t + e_t, \qquad d_0 = 0.$$

The last two terms have combined norm at most $4\alpha C_*\delta_w g_*^t\|r_0\|$ and $\|I - \alpha K_0\|_{\mathrm{op}} \le g_*$. Iteration gives $\|d_t\| \le 4\alpha C_*\delta_w t g_*^{t-1}\|r_0\|$ for $t \ge 1$. Consequently on E_{a_*},

$$\max_{m,t} \|F_{t,m}^{\mathrm{act}} - F_{t,m}^{\mathrm{fr}}\| \le C_H\delta_w \max_m \|r_{0,m}\|, \tag{13}$$

$$\mathcal{E}_w := \max_{m,t}(\|F_{t,m}^{\mathrm{act}}\| + \|F_{t,m}^{\mathrm{fr}}\|) \le C_H(1 + \max_m \|F_{0,m}\|).$$

The output bound follows by bounding the actual residual and the residual predicted by the fixed empirical NTK separately and then adding Y.

3.5 Global bounds and the contribution from the complementary event

On the entire probability space define

$$P_t = 1 + \sqrt{w} + \sum_m (\|W_{t,m}\|_F + \|V_{t,m}\|_F + \|c_{t,m}\| + \|a_{t,m}\|).$$

All width denominators can be bounded by one. Thus $\|h_i\| \le CP_t$, $\|\phi_i\| \le CP_t^2$, $\|F_{t,m}\| \le CP_t^3$, and each Jacobian block in (1) is at most CP_t^2. Hence $\|r_{t,m}\| \le CP_t^3$ and

$$P_{t+1} \le P_t + CP_t^5 \le C'P_t^5, \qquad P_t \le C_H P_0^{5^t}.$$

Actual outputs are bounded by $C_H P_0^{3\cdot 5^H}$. Since $\|K_0\|_{\mathrm{op}} \le CP_0^4$, outputs predicted by holding the empirical NTK fixed are bounded by $C_H P_0^{4H+3}$. The inequality $4H + 3 \le 3 \cdot 5^H$ holds at zero and persists since $3 \cdot 5^H + 4 \le 3 \cdot 5^{H+1}$. Therefore $\mathcal{E}_w \le C_H P_0^{3\cdot 5^H}$ globally.

For a Gaussian vector with k fixed-variance entries, Minkowski applied to its squared norm gives $\mathbb{E}\|G\|^{2s} \le C_s k^s$ for $s \ge 1$, with lower powers by Hölder. Apply this to $2w^2$ entries for V and $O(w)$ entries for other blocks. Finite sums give $\mathbb{E}P_0^r \le C_r w^r$ for fixed $r \ge 1$, and consequently

$$\mathbb{E}\mathcal{E}_w^p \le C_{p,H} w^{3p\cdot 5^H}, \qquad \mathbb{E}\mathcal{E}_w^8 \le C_H w^{24\cdot 5^H}.$$

These global bounds have polynomial width dependence because $\|V_0\|_F$ is order w.

Choose $a_* = 24 \cdot 5^H + 4$ before constructing E_{a_*} and choosing B. Cauchy–Schwarz yields

$$\mathbb{E}[\mathcal{E}_w^4 \mathbf{1}_{E_{a_*}^c}] \le (\mathbb{E}\mathcal{E}_w^8)^{1/2}\mathbb{P}(E_{a_*}^c)^{1/2} \le C_H w^{12\cdot 5^H - a_*/2} = C_H w^{-2}.$$

On the good event, (13) and (8) give a uniform fourth moment. Finitely many excluded widths each have finite moments by the global envelope, proving the supremum over every positive integer width.

Let $D_w = \max_{m,t} \|F_{t,m}^{\mathrm{act}} - F_{t,m}^{\mathrm{fr}}\|$. On the good event, unconditional initial moments give

$$\mathbb{E}[D_w^2 \mathbf{1}_{E_{a_*}}] \le C_H\delta_w^2 \mathbb{E}\max_m \|r_{0,m}\|^2 \le C_H w^{-1/3}\ell_w^2.$$

On its complement $D_w \le \mathcal{E}_w$ and $\mathcal{E}_w^2 \le 1 + \mathcal{E}_w^4$, so the expectation is at most $C_H w^{-2}$. This proves Theorem 3.1. If $H = 0$ or $\eta = 0$, the paths coincide directly.

3.6 From output approximation to loss after selection

Set $L_{w,t,m}^{\mathrm{act}} = \|Y - F_{t,m}^{\mathrm{act}}\|^2/(2n)$. Difference of squares gives

$$\max_{m,t} |L_{w,t,m}^{\mathrm{act}} - S_{w,t,m}| \le \frac{D_w}{2n} \max_{m,t}(\|Y - F_{t,m}^{\mathrm{act}}\| + \|Y - F_{t,m}^{\mathrm{fr}}\|).$$

Cauchy–Schwarz and the fourth moments prove

$$\varepsilon_w := \mathbb{E}\max_{m,t} |L_{w,t,m}^{\mathrm{act}} - S_{w,t,m}| \longrightarrow 0. \tag{14}$$

The same expected-error bound applies after any measurable choice of candidate and training step. Gate indicators, paths predicted by the fixed empirical NTK, and comparisons among finitely many candidates with smallest-index tie breaking are Borel measurable, including when a preactivation equals zero. For fixed k, put $\Delta_{w,k} = \max_m |L^{\mathrm{act}}_{w,k,m} - S_{w,k,m}|$. Pointwise,

$$L^{\mathrm{act}}_{w,k,\widehat{m}_{w,k}} \le S_{w,k,\widehat{m}_{w,k}} + \Delta_{w,k} = \min_m S_{w,k,m} + \Delta_{w,k} \le \min_m L^{\mathrm{act}}_{w,k,m} + 2\Delta_{w,k}.$$

Thus the expected difference from the candidate with the smallest actual loss is at most $2\varepsilon_w$. The same bound applies to the difference between the actual loss improvement from selection and the improvement predicted by the fixed empirical NTK. This argument does not require the predicted scores to be free of ties or the selected index to converge.

4 Initial GP, full NTK, and score limits

Theorem 4.1 (Joint initialization and score limit)**.** *Under Section 2, for every finite $p \ge 1$,*

$$\mathbb{E}\|K_{0,m} - \overline{K}\|_F^p \to 0, \qquad (F_{0,m}, K_{0,m})_{m=1}^M \Rightarrow (G_m, \overline{K})_{m=1}^M,$$

where G_m are independent $N(0,\Sigma)$ vectors. Jointly over the finite candidate/time array,

$$(S_{w,t,m})_{t,m} \Rightarrow (Q_{t,m})_{t,m}, \qquad Q_{t,m} = \frac{(Y-G_m)^\top B_t (Y-G_m)}{2n}.$$

For every continuous function Ψ of polynomial growth on this array, $\mathbb{E}\Psi(S_w) \to \mathbb{E}\Psi(Q)$. For initial outputs, the conclusion is weak convergence; the kernel converges in every finite L^p. Scores at different times share the same G_m and are generally neither independent nor Gaussian.

4.1 Gaussian integration, moments, and uniform integrability

For variances $q_1, q_2 > 0$ and $\rho = \cos\vartheta$, $0 < \vartheta < \pi$, take independent standard normals X, Y and set $Z_1 = \sqrt{q_1}X$, $Z_2 = \sqrt{q_2}(\cos\vartheta X + \sin\vartheta Y)$. In polar coordinates $(X,Y) = r(\cos u, \sin u)$ the density times area is $(2\pi)^{-1}e^{-r^2/2} r\,dr\,du$. The strict-positive half-planes intersect on $\vartheta - \pi/2 < u < \pi/2$, giving probability $(\pi - \vartheta)/(2\pi)$. For the ReLU product,

$$\int_0^\infty r^3 e^{-r^2/2}\,dr = 2, \qquad \int_{\vartheta-\pi/2}^{\pi/2} \cos u \cos(u-\vartheta)\,du = \frac{(\pi-\vartheta)\cos\vartheta + \sin\vartheta}{2}.$$

The first follows by $t = r^2/2$; the second by $2\cos u\cos(u-\vartheta) = \cos\vartheta + \cos(2u-\vartheta)$. Multiplication gives $\mathcal{R}$. At $\vartheta = 0, \pi$, the same common-normal representation converges almost surely. Positive marginal variances ensure nonzero limiting coordinates almost surely, so gate products converge and are bounded by one. ReLU products are dominated by $\sqrt{q_1 q_2}(|X| + |Y|)^2$. Dominated convergence gives both endpoint formulas and continuity on positive-diagonal covariance matrices, including singular correlations and varying variances on a bounded neighborhood. No inverse covariance is used. Gaussian realization also makes $\mathcal{R}(C)$ positive semidefinite.

All scalar Gaussian moments are finite by integration. For nonnegative iid Z_j and $r \ge 1$, convexity gives $\mathbb{E}(w^{-1}\sum_j Z_j)^r \le \mathbb{E}Z_1^r$; for $0 < r < 1$ Jensen gives a bound by $(\mathbb{E}Z_1)^r$. Thus normalized Gaussian squared norms have all fixed uniform moments. Finite sums and Hölder preserve them without independence of factors. The spectral net argument in Section 3 further implies for sufficiently large fixed L and $t \ge L$,

$$\mathbb{P}(\|V\|_{\mathrm{op}}/\sqrt{w} > t) \le 2e^{-t^2/16},$$

uniformly in $w \ge 1$. Integrating $r\int t^{r-1}\mathbb{P}(X > t)\,dt$ gives every fixed spectral moment; in particular $\mathbb{E}\|V\|_{\mathrm{op}}^4 = O(w^2)$.

We use the following explicit uniform-integrability argument. If $Z_w \to 0$ in probability and $\sup_w \mathbb{E}|Z_w|^q < \infty$ for some $q > p > 0$, then

$$\mathbb{E}|Z_w|^p \le \delta^p + R^p\mathbb{P}(|Z_w| > \delta) + R^{p-q}\mathbb{E}|Z_w|^q.$$

Let w, then R, increase, and then δ decrease. This proves L^p convergence, including finite-dimensional versions by finite sums. For weak convergence, bounded continuous truncations and the same tail bound give expectation convergence for every continuous polynomial-growth function when sufficiently high uniform moments are available.

4.2 Bottom and upper feature covariances; the initial GP

At initialization write $l_{ij} = \mathbf{1}\{z_{1i,j} > 0\}$, $e_{iu} = \mathbf{1}\{z_{2i,u} > 0\}$, and

$$T_{w,i\ell} = h_i^\top h_\ell = w^{-1} \sum_j \mathrm{ReLU}(z_{1i,j}) \,\mathrm{ReLU}(z_{1\ell,j}), \qquad D_{1,w,i\ell} = w^{-1} \sum_j l_{ij} l_{\ell j}.$$

Rows of W are iid with covariance C_1 on the input list. The preceding integrals give expectations $T_{i\ell}$ and $D_{1,i\ell}$. Each feature summand has all moments, using $|AB| \le (A^2 + B^2)/2$; gates are bounded. Independence makes the centered-average variance the single-row variance divided by w. Thus both arrays converge entrywise in L^2. Convexity gives every uniform moment of T_w and $D_{1,w}$ is bounded, so the uniform-integrability argument upgrades convergence to every finite L^p and then to matrix norms.

Conditionally on W, the $2w$ upper row vectors on the inputs are iid centered Gaussian with covariance $C_{2,w} = T_w + \tau^2 \mathbf{1}\mathbf{1}^\top$, whose diagonal is at least τ^2. Let $\Sigma_{w,i\ell} = (2w)^{-1} \sum_u \phi_{i,u} \phi_{\ell,u}$. Its conditional mean is $\mathcal{R}(C_{2,w})_{i\ell}$. Gaussian fourth moments and Cauchy–Schwarz give

$$\mathbb{E}|\Sigma_{w,i\ell} - \mathcal{R}(C_{2,w})_{i\ell}|^2 \le \frac{3}{2w} \mathbb{E}[(T_{w,ii} + \tau^2)(T_{w,\ell\ell} + \tau^2)] = O(w^{-1}).$$

For $r \ge 1$, conditional Jensen bounds $\mathbb{E}|\Sigma_{w,i\ell}|^r$ by $C_r \mathbb{E}[(T_{w,ii} + \tau^2)^{r/2} (T_{w,\ell\ell} + \tau^2)^{r/2}]$, uniformly finite; conditional Jensen gives the same bound for its conditional mean. Continuity of $\mathcal{R}$ and the bottom limit imply $\mathcal{R}(C_{2,w}) \to \Sigma$ in probability. Uniform integrability proves $\Sigma_w \to \Sigma$ in every finite L^p.

Conditionally on (W, V, c) the independent head makes F_0 exactly $N(0, \Sigma_w)$. Its characteristic function at t is $\exp(-t^\top \Sigma_w t/2)$, a bounded variable converging in probability to $\exp(-t^\top \Sigma t/2)$. Expectations converge, proving $F_0 \Rightarrow N(0, \Sigma)$. Conditional scalar moments give $\mathbb{E}|F_{0,i}|^r = c_r \mathbb{E} \Sigma_{w,ii}^{r/2}$ and hence $\sup_w \mathbb{E}\|F_0\|^r < \infty$ for every finite $r > 0$. The characteristic-function argument also covers singular Σ.

4.3 Head, bias, and upper-weight Gram blocks

The disjoint raw parameter blocks in (1) have additive Gram matrices. The head block is $K_0^a = \Sigma_w$. Let

$$U_{w,i\ell} = (2w)^{-1} \sum_u a_u^2 e_{iu} e_{\ell u}.$$

Conditionally on W, the triples (V_u, c_u, a_u) are iid, and a_u is independent of the gates. Thus

$$\mathbb{E}(U_{w,i\ell} \mid W) = \mathcal{P}(C_{2,w})_{i\ell}, \qquad \mathbb{E}|U_{w,i\ell} - \mathcal{P}(C_{2,w})_{i\ell}|^2 \le 3/(2w).$$

Moreover $|U_{w,i\ell}| \le (2w)^{-1} \|a\|^2$ has all uniform moments. Continuity of $\mathcal{P}$ and uniform integrability prove $U_w \to D_2$ in every finite L^p. The other blocks are $K_0^c = U_w$ and $K_0^V = T_w \odot U_w$. For each entry,

$$\|T_{w,i\ell} U_{w,i\ell} - T_{i\ell} D_{2,i\ell}\|_{L^p} \le \|T_{w,i\ell} - T_{i\ell}\|_{L^{2p}} \|U_{w,i\ell}\|_{L^{2p}} + |T_{i\ell}| \|U_{w,i\ell} - D_{2,i\ell}\|_{L^p} \to 0.$$

4.4 Bottom block: independent head and gate/weight dependence

Fix input indices i, ℓ and let $\mathsf{B} = \mathsf{D}_{1i} \mathsf{D}_{1\ell}$, a diagonal $w \times w$ zero-one matrix. The bottom Gram entry is $s_{i\ell} Z_w$, where

$$Z_w = a^\top \mathsf{Q} a, \qquad \mathsf{Q} = \frac{\mathsf{D}_{2i} V \mathsf{B} V^\top \mathsf{D}_{2\ell}}{2w^2}.$$

This defines Z_w even when $s_{i\ell} = 0$. With $\mathsf{Q}_s = (\mathsf{Q} + \mathsf{Q}^\top)/2$, the quadratic form and trace are unchanged and $\|\mathsf{Q}_s\|_F \le \|\mathsf{Q}\|_F$. Given (W, V, c), a remains standard normal. Independence and one-dimensional moments imply

$$\mathbb{E}[a_r a_s a_t a_v] = \delta_{rs}\delta_{tv} + \delta_{rt}\delta_{sv} + \delta_{rv}\delta_{st}.$$

Indeed odd multiplicities give zero, two distinct pairs give one, and four identical indices give three. Expanding therefore yields conditional mean $\operatorname{tr} \mathsf{Q}$ and variance $2\|\mathsf{Q}_s\|_F^2$. Since $\|\mathsf{Q}\|_{\mathrm{op}} \le \|V\|_{\mathrm{op}}^2/(2w^2)$ and $\|\mathsf{Q}\|_F^2 \le 2w \|\mathsf{Q}\|_{\mathrm{op}}^2$,

$$\mathbb{E}|Z_w - \operatorname{tr} \mathsf{Q}|^2 \le \mathbb{E}\|V\|_{\mathrm{op}}^4 / w^3 = O(w^{-1}). \tag{15}$$

No independence between V and its gates was assumed.

The trace is $(2w)^{-1}\sum_u X_u$ with $X_u = e_{iu}e_{\ell u}V_u^\top \mathsf{B} V_u/w$. Given W these are iid, with conditional second moments bounded by $\mathbb{E}\|V_u\|^4/w^2 = 1 + 2/w \le 3$. Hence

$$\mathbb{E}|\operatorname{tr}\mathsf{Q} - \mathbb{E}[X_1 \mid W]|^2 \le 3/(2w). \tag{16}$$

The conditional mean retains the dependence between gates and weights. To evaluate it, condition on W and write $\xi = (V_u, c_u/\tau) \sim N(0, I_{w+1})$, $t_i = (h_i, \tau)$, $t_\ell = (h_\ell, \tau)$. Let Π project onto $\operatorname{span}(t_i, t_\ell)$, of rank at most two. Both vectors are nonzero. The event $E = \{\xi^\top t_i > 0, \xi^\top t_\ell > 0\}$ depends only on $\Pi\xi$. With $\widetilde{\mathsf{B}} = \operatorname{diag}(\mathsf{B}, 0)$, the row contribution is $X_u = w^{-1}\xi^\top \widetilde{\mathsf{B}}\xi \mathbf{1}_E$. In orthonormal coordinates the parallel and perpendicular Gaussian parts are independent; the latter has mean zero and second moment $I - \Pi$. Thus cross terms vanish and

$$\mathbb{E}[\xi\xi^\top \mathbf{1}_E \mid W] = p_w I + H_w, \quad p_w = \mathcal{P}(C_{2,w})_{i\ell}, \quad H_w = \Pi\mathbb{E}[\xi\xi^\top \mathbf{1}_E \mid W]\Pi - p_w\Pi = \Pi H_w \Pi.$$

For unit u, $\mathbb{E}[(u^\top\Pi\xi)^2\mathbf{1}_E] \le \|\Pi u\|^2 \le 1$. Hence $\|H_w\|_{\mathrm{op}} \le 1 + p_w \le 2$ and its sum of singular values is at most four. The matrix $\widetilde{\mathsf{B}}$ has operator norm at most one. Diagonalizing symmetric H_w gives $|\operatorname{tr}(\widetilde{\mathsf{B}}H_w)| \le \sum_j |\lambda_j(H_w)| \le 4$. Therefore

$$\mathbb{E}[X_1 \mid W] = p_w\frac{\operatorname{tr}\mathsf{B}}{w} + \frac{\operatorname{tr}(\widetilde{\mathsf{B}}H_w)}{w}, \qquad |\mathbb{E}[X_1 \mid W] - p_w D_{1,w,i\ell}| \le 4/w. \tag{17}$$

This holds uniformly in W, including coincident t vectors and singular covariances. Combining (15)–(17) with the bounded-factor limits proves $Z_w \to D_{1,i\ell}D_{2,i\ell}$ in probability. Also

$$|Z_w| \le \|V\|_{\mathrm{op}}^2\|a\|^2/(2w^2)$$

has every fixed uniform moment by Hölder and the spectral bounds. Uniform integrability upgrades convergence to every finite L^p. Thus $K_0^W \to s \odot D_1 \odot D_2$ in those norms.

4.5 Full kernel, joint limit, and score expectations

Finite sums of entries and blocks prove $\mathbb{E}\|K_0 - \overline{K}\|_F^p \to 0$ and $\sup_w \mathbb{E}\|K_0\|_F^p < \infty$. For each fixed u, $u^\top\overline{K}u$ is the probability limit of nonnegative $u^\top K_0 u$, so $\overline{K}$ is positive semidefinite. Since the kernel limit is deterministic, $(F_0, K_0) \Rightarrow (G, \overline{K})$: on a large compact set in F and a bounded neighborhood of $\overline{K}$, uniform continuity of a bounded continuous test function permits replacing K_0 by $\overline{K}$; tightness of F_0 controls the complement. Then use its weak limit. Independent candidate arrays have factorized characteristic functions, giving independent G_m.

Integer matrix powers and quadratic expressions are continuous. The finite score array therefore has the claimed weak limit. For every $r \ge 1$,

$$\max_{m,t} S_{w,t,m} \le C(1 + \max_m \|K_{0,m}\|_F)^{2H}(\|Y\| + \max_m \|F_{0,m}\|)^2$$

has uniformly bounded rth moment by Hölder. The truncation argument proves all polynomial-growth continuous expectation limits, including scores, powers, and finite minima, completing Theorem 4.1.

For completeness, the exact scalar moments of each limiting loss are

$$\mathbb{E}Q_t = \frac{Y^\top B_t Y + \operatorname{tr}(B_t\Sigma)}{2n}, \qquad \operatorname{Var}(Q_t) = \frac{2\operatorname{tr}(B_t\Sigma B_t\Sigma) + 4Y^\top B_t\Sigma B_t Y}{4n^2}. \tag{18}$$

Indeed write $G = \Sigma^{1/2}Z$. The centered numerator is $Z^\top CZ - \operatorname{tr}C - 2d^\top Z$, with $C = \Sigma^{1/2}B_t\Sigma^{1/2}$ and $d = \Sigma^{1/2}B_tY$. The fourth-moment identity above gives quadratic variance $2\operatorname{tr}C^2$, linear variance $4\|d\|^2$, and zero covariance by vanishing odd moments. Cyclic traces give (18).

The score is nonconstant exactly when $A^t\Sigma(A^t)^\top \ne 0$. If this covariance vanishes, $A^tG = 0$ almost surely, so the score is constant. Otherwise C is positive semidefinite and nonzero. Orthogonally diagonalize it and condition on all but one rotated normal coordinate with strictly positive eigenvalue. Each level set of the resulting scalar quadratic has at most two roots, each of conditional probability zero. Thus Q_t is atomless and nonconstant. Singular square roots and covariances are allowed throughout.

5 LIMITS FOR SELECTION AND COMPARISONS UNDER A FIXED COMPUTE BUDGET

5.1 EXPECTED ACTUAL LOSS AFTER SELECTION

For fixed $0 \le k \le b \le H$, the minimum identity and Theorem 4.1 give

$$\mathbb{E}S_{w,b,1} - \mathbb{E}S_{w,k,\widehat{m}_{w,k}} \longrightarrow \mathbb{E}Q_{b,1} - \mathbb{E}\min_{m \le M} Q_{k,m}.$$

The minimum is a continuous function of the finite score array, including arrays with ties, so its expectation converges directly. Equation (14) bounds the difference between the displayed score comparison and the corresponding actual-loss comparison by $2\varepsilon_w$. Consequently,

$$\Delta_w := \mathbb{E}L^{\mathrm{act}}_{w,b,1} - \mathbb{E}L^{\mathrm{act}}_{w,k,\widehat{m}_{w,k}} \longrightarrow \Delta_\infty(b,k) := \mathbb{E}Q_{b,1} - \mathbb{E}\min_{m \le M} Q_{k,m}. \tag{19}$$

The baseline may be candidate one in the pool or an independent identically distributed candidate; their expected losses are equal. The initialization limits are independent of the proof that the fixed empirical NTK approximates actual training; only the last transfer to actual training uses that approximation.

5.2 COMPARING THE TWO NTK SCORES AT A COMMON HORIZON

For the same candidate-specific $F_{0,m}$, define

$$\widetilde{S}_{w,t,m} = \frac{1}{2n}\left\|(I - \alpha\overline{K})^t(Y - F_{0,m})\right\|^2.$$

The existing kernel convergence and moment bounds imply

$$\delta^{\mathrm{score}}_w := \mathbb{E}\max_{m \le M,\, t \le H} |S_{w,t,m} - \widetilde{S}_{w,t,m}| \longrightarrow 0. \tag{20}$$

Indeed, the telescoping identity $U^t - V^t = \sum_{j=0}^{t-1} U^{t-1-j}(U - V)V^j$, which does not require commutation, and the difference-of-squared-norms inequality give

$$\max_{m,t} |S_{w,t,m} - \widetilde{S}_{w,t,m}| \le CD_{K,w}T^{2H-1}_{K,w}R^2_w,$$

where $D_{K,w} = \max_m \|K_{0,m} - \overline{K}\|_{\mathrm{op}}$, $T_{K,w} = 1 + \|\overline{K}\|_{\mathrm{op}} + \max_m \|K_{0,m}\|_{\mathrm{op}}$, and $R_w = \max_m \|Y - F_{0,m}\|$. Section 4 gives $\mathbb{E}D^2_{K,w} \to 0$ and the required uniform moments of $T_{K,w}$ and R_w; Hölder's inequality proves (20). No independence between $K_{0,m}$ and $F_{0,m}$ is used.

For fixed $k \le H$, let $\widehat{m}_{w,k}$ and $\widetilde{m}_{w,k}$ be the smallest-index minimizers of $S_{w,k,m}$ and $\widetilde{S}_{w,k,m}$. Combining their minimizing properties with (14) yields

$$\mathbb{E}\left|L^{\mathrm{act}}_{w,k,\widehat{m}_{w,k}} - L^{\mathrm{act}}_{w,k,\widetilde{m}_{w,k}}\right| \le 2\varepsilon_w + 2\delta^{\mathrm{score}}_w \longrightarrow 0.$$

5.3 EXACT COMPUTE ALLOCATION AND DEPENDENCE ON WIDTH

Fix $1 \le b \le H$. Let $g > 0$ be the cost of one simultaneous update, $i_M, i_1 \ge 0$ initialization costs, $e \ge 0$ final evaluation cost, and $q_b \ge 0$ per-candidate score cost calibrated at baseline budget b. Let $a_M \ge 0$ be the cost of choosing the minimum score, as in the main text. Define

$$C_{\mathrm{random}}(b) = i_1 + bg + e, \qquad C_{\mathrm{target}}(k) = i_M + Mq_b + a_M + kg + e,$$
$$\Omega = \frac{i_M - i_1 + Mq_b + a_M}{g}.$$

The score predicts loss after the allocated k updates, while q_b is the conservative score cost measured for the full baseline budget. In exact arithmetic, k is feasible when $k \le b - \Omega$. For the positive-update analysis, the largest feasible integer in $\{1, \dots, b\}$, with zero used as an empty-set marker, is

$$k^* = \begin{cases} 0, & b - \Omega < 1, \\ \min\{b, \lfloor b - \Omega \rfloor\}, & b - \Omega \ge 1. \end{cases}$$

For $1 \le k < b$, the allocation equals k exactly when $b-k-1 < \Omega \le b-k$. The allocation is capped at $k = b$ when $\Omega \le 0$, and it is zero when $\Omega > b-1$. This marker distinguishes the absence of a feasible positive update from feasibility of selection itself. A zero-update selection is feasible exactly when $C_{\mathrm{target}}(0) \le C_{\mathrm{random}}(b)$, or $\Omega \le b$. Thus $b-1 < \Omega \le b$ permits zero updates, whereas $\Omega > b$ makes the entire method infeasible. The experimental allocation includes zero in its feasible set and records an empty set separately. If both inequalities are strict, a sufficiently small perturbation of the measured costs leaves the allocation unchanged. The implementation uses a feasibility tolerance of 10^{-12}, which replaces Ω by $\Omega - 10^{-12}/g$ in these conditions. Away from that tolerance at a boundary, the exact and numerical allocations agree.

For $b = 20$, $k = 16$, choose declared normalized costs $g = i_1 = e = 1$, $i_2 = 2$, $a_2 = 0$, and $q_b = 5/4$. Then $\Omega = 7/2$, $C_{\mathrm{random}} = 22$, $C_{\mathrm{target}}(16) = 43/2$, and $C_{\mathrm{target}}(17) = 45/2$. The update and scoring costs are positive; 16 updates fit the budget and 17 exceed it. This analytic example uses declared normalized costs and sets the minimum-selection cost to zero.

To combine this allocation with a limit as width grows, one must use fixed normalized costs, a deterministic sequence of width-dependent costs that eventually gives the same allocation, or a fixed k whose feasibility has been proved. For allocations obtained from random timing measurements, applying the width limit requires eventual stabilization of the allocated step count. If the score cost depends on the predicted step, say $q(k)$, feasibility must instead be checked over the full set $\{k : i_M + Mq(k) + a_M + kg + e \le C_{\mathrm{random}}(b)\}$; the full feasible-set calculation determines the allocation.

5.4 An open class of ReLU problems where selection improves on a random initialization

For $n = 1$, fix $d \ge 1$, $x \in \mathbb{R}^d$ and put $s = \|x\|^2 + 1$, $q = s/(d+1)$. The diagonal Gaussian formulas give $T = q$, $D_1 = D_2 = 1/2$, and scalar GP variance $\nu^2 = (q+\tau^2)/2 > 0$. The four blocks give

$$\kappa = \frac{q+\tau^2}{2} + \frac{1}{2} + \frac{q}{2} + \frac{s}{4} = q + \frac{\tau^2}{2} + \frac{1}{2} + \frac{s}{4} > 0.$$

For $0 < \eta < 1/\kappa$, let $\gamma = 1 - \eta\kappa \in (0,1)$, $G_m = \nu Z_m$ for iid standard normals, $u = y/\nu$, and $c_M(u) = \mathbb{E}\min_{m \le M}(u - Z_m)^2$. For $M \ge 2$ and $1 \le k < b \le H$,

$$Q_{t,m} = \frac{\nu^2}{2}\gamma^{2t}(u - Z_m)^2, \qquad \Delta_\infty(b,k) = \frac{\nu^2}{2}\gamma^{2k}\big[\gamma^{2(b-k)}(1+u^2) - c_M(u)\big]. \tag{21}$$

The expected loss improvement is positive exactly when $c_M(u)/(1+u^2) < \gamma^{2(b-k)}$. The factor on the right accounts for the $b-k$ updates that are unavailable after paying for selection.

The finite minimum of squared standard normals is positive almost surely, so $c_M(0) > 0$. It is bounded by Z_1^2 and is strictly smaller on the positive-probability event $|Z_2| < |Z_1|$. The nonnegative integrable difference therefore has positive expectation, proving $c_M(0) < 1$. Every

$$0 < \eta < \frac{1 - c_M(0)^{1/(2(b-k))}}{\kappa}$$

gives a positive expected loss improvement and lies strictly inside $(0, 1/\kappa)$. Thus the limiting dynamics contract at every step.

The function c_M is continuous: if $u_j \to u$, the finite minimum converges pointwise, and for $|u_j| \le R$ it is bounded by the integrable variable $(R + |Z_1|)^2$. The quantities ν^2 and κ are continuous in (x, τ) for $\tau > 0$. Formula (21) is therefore continuous when $0 < \eta\kappa < 1$. Every point at which the expected improvement is strictly positive has an open neighborhood with the same property, including nonzero labels around a point with $y = 0$. Combining this neighborhood with compute costs for which the allocation is unchanged gives a nonempty open set in $(x, y, \tau, \eta, \text{cost ratios})$, with $d, n = 1, M, b, k$ fixed. By (19), actual expected loss is also improved at every point in this set when width is sufficiently large. The required width may depend on the point in this set.

Explicit example with two candidates. For independent standard normals X, Y, $\min(X^2, Y^2) = (X^2 + Y^2 - |X^2 - Y^2|)/2$. The orthogonal transform $U = (X+Y)/\sqrt{2}$, $V = (X-Y)/\sqrt{2}$ produces

independent standard normals, with $X^2 - Y^2 = 2UV$. Direct integration gives $\mathbb{E}|U| = \sqrt{2/\pi}$, so $c_2(0) = 1 - 2/\pi$. At $d = 1$, $x = y = 0$, $\tau = 1/2$, the four kernel blocks are $3/8, 1/2, 1/4, 1/4$. Thus

$$\nu^2 = 3/8, \quad \kappa = 11/8, \quad \eta = 1/20, \quad \gamma = 149/160,$$

and for $b = 20, k = 16$,

$$\Delta_\infty = \frac{3}{16}(149/160)^{32}\big[(149/160)^8 - (1 - 2/\pi)\big] > 0. \tag{22}$$

This is exact. To verify its sign without decimal integration, $\pi/4 = \int_0^1 (1 + t^2)^{-1} dt < 1$, so $1 - 2/\pi < 1/2$. Bernoulli's inequality (by induction, or direct expansion at power four) gives $(149/160)^4 > 1 - 4(11/160) = 29/40$. Squaring gives $(149/160)^8 > 841/1600$, and the bracket in (22) exceeds $41/1600$. Hence $\Delta_\infty > (123/25600)(149/160)^{32} > 0$. The costs above allocate exactly 16 updates, and (19) gives $\Delta_w > \Delta_\infty/2$ for all sufficiently large w. The guarantee is asymptotic: an explicit width threshold, including a guarantee at $w = 128$, remains unquantified.

5.5 Predicted future loss versus initial loss

Define $\widehat{m}_0 \in \operatorname{argmin}_m Q_{0,m}$ and $\widehat{m}_k \in \operatorname{argmin}_m Q_{k,m}$, using the smallest index to break ties. When both selected candidates receive k updates,

$$D_k = \mathbb{E}[Q_{k,\widehat{m}_0} - \min_{m \le M} Q_{k,m}] \ge 0. \tag{23}$$

The integrable difference is nonnegative, so $D_k > 0$ if and only if $\mathbb{P}(Q_{k,\widehat{m}_0} > \min_{m \le M} Q_{k,m}) > 0$. Nonconstant scores or positive score correlation alone do not replace this necessary and sufficient criterion.

For $M \ge 2$ a sufficient condition is a strict rank reversal on the affine residual support $\mathcal{S} = Y - \operatorname{range}(\Sigma)$: there exist $r, s \in \mathcal{S}$ such that

$$\|r\|^2 < \|s\|^2, \qquad r^\top B_k r > s^\top B_k s. \tag{24}$$

The induced Gaussian measure has positive density relative to Lebesgue measure on its affine support: choose an orthonormal basis of the covariance range and diagonalize its strictly positive restricted covariance. Continuity supplies small relative-open neighborhoods U, V of r, s for which both inequalities persist for every pair. By independence, there is positive probability that candidate one lies in V and every other candidate lies in U. Initial-loss selection then chooses a residual in U, while selection by predicted loss after k updates chooses candidate one in V. This proves strictness for any finite number of candidates.

Proposition 5.1 (Exact condition under full support)**.** *Suppose $\Sigma \succ 0$, $2 \le M < \infty$, and $k \in \{0, \dots, H\}$. Then*

$$D_k > 0 \quad \Longleftrightarrow \quad B_k \notin \{\beta I : \beta \ge 0\}. \tag{25}$$

The equivalence holds for every fixed label vector $Y \in \mathbb{R}^n$ and requires no contraction assumption.

Proof. Write $R_m = Y - G_m$ and define

$$q_0(r) = \frac{\|r\|^2}{2n}, \qquad q_k(r) = \frac{r^\top B_k r}{2n}.$$

The residuals are independent with law $N(Y, \Sigma)$, whose density is strictly positive on $\mathbb{R}^n$. Moreover, $B_k = (A^k)^\top A^k$ is symmetric positive semidefinite. Thus the nonnegative gap

$$Z = Q_{k,\widehat{m}_0} - \min_{m \le M} Q_{k,m}$$

is integrable, since

$$0 \le Z \le \frac{\|B_k\|_{\mathrm{op}}}{2n} \sum_{m=1}^{M} \|R_m\|^2.$$

If $B_k = \beta I$ for some $\beta \ge 0$, then $Q_{k,m} = \beta Q_{0,m}$ for every candidate. Every candidate minimizing initial loss therefore also minimizes predicted loss after k updates, including when $\beta = 0$, so $Z = 0$ almost surely and $D_k = 0$.

Conversely, suppose $B_k \notin \{\beta I : \beta \geq 0\}$. The spectral theorem gives unit eigenvectors $e_{\mathrm{hi}}, e_{\mathrm{lo}}$ with eigenvalues $\lambda_{\mathrm{hi}} > \lambda_{\mathrm{lo}} \geq 0$. Choose $a > 1$ such that $\lambda_{\mathrm{hi}} > a^2\lambda_{\mathrm{lo}}$; if $\lambda_{\mathrm{lo}} = 0$, any $a > 1$ works, while otherwise one may take $1 < a < \sqrt{\lambda_{\mathrm{hi}}/\lambda_{\mathrm{lo}}}$. Put $r = e_{\mathrm{hi}}$ and $s = ae_{\mathrm{lo}}$. Then

$$q_0(r) < q_0(s), \qquad q_k(r) > q_k(s).$$

By continuity, there are open neighborhoods U of r and V of s, and $\delta > 0$, such that for every $u \in U$ and $v \in V$,

$$q_0(u) < q_0(v), \qquad q_k(u) - q_k(v) \geq \delta.$$

Let $p_U = \mathbb{P}(R_1 \in U)$ and $p_V = \mathbb{P}(R_1 \in V)$; full support gives $p_U, p_V > 0$. Independence gives positive probability $p_V p_U^{M-1}$ to the event

$$R_1 \in V, \qquad R_2, \ldots, R_M \in U.$$

On this event every candidate minimizing initial loss lies among candidates $2, \ldots, M$, whereas candidate one uniquely minimizes predicted loss after k updates, and hence $Z \geq \delta$. Therefore

$$D_k = \mathbb{E}Z \geq \delta p_V p_U^{M-1} > 0.$$

This proves (25). □

Since $\overline{K}$ is symmetric and $A = I - (\eta/n)\overline{K}$, the eigenvalues of B_k are $|1 - (\eta/n)\lambda_j(\overline{K})|^{2k}$. Thus, for $k \geq 1$, Proposition 5.1 requires these absolute factors to differ across eigendirections after k updates. Unequal eigenvalues of $\overline{K}$ are not sufficient when the corresponding update factors have equal absolute value.

When the covariance is singular, it is not sufficient that B_k be nonscalar on all of $\mathbb{R}^n$. For example, $\Sigma = \mathrm{diag}(1, 0)$, $Y = 0$, and $B_k = \mathrm{diag}(1, 2)$ give equal scores on the support of the residual distribution. More generally, there is no improvement after the same number of updates if $Q_k = cQ_0 + d$ with $c > 0$ on that support, or if Q_k is constant there. These examples are sufficient conditions for no improvement, not a complete classification for singular covariance.

To transfer the result for initial-loss selection, let $\widehat{m}_{w,0} \in \mathrm{argmin}_m S_{w,0,m}$. For $M = 1$, selecting the only candidate is simply a coordinate projection. For $M \geq 2$, the present model has $\Sigma_{ii} = (T_{ii} + \tau^2)/2 > 0$, so $\Sigma \neq 0$ and Q_0 has no point masses by Section 4. Independent candidates therefore tie with probability zero. Whenever there is no tie, a positive gap between the smallest and second-smallest initial scores preserves the selected index under a sufficiently small perturbation. Joint weak convergence and this almost-sure continuity imply convergence of the loss coordinate corresponding to the selected candidate. More explicitly, exclude events on which the initial-score gap is at most δ or the score array is too large. As $\delta \downarrow 0$ and the truncation level tends to infinity, the probability of these excluded events tends to zero, while selection is continuous on the remaining compact set. The bound $|S_{w,k,\widehat{m}_{w,0}}| \leq \max_m S_{w,k,m}$ gives uniform integrability. Hence

$$\mathbb{E}S_{w,k,\widehat{m}_{w,0}} \to \mathbb{E}Q_{k,\widehat{m}_0}.$$

Equation (14) transfers this convergence to actual losses. Selection by predicted loss after k updates enters through the minimum directly and does not require the predicted scores to have a continuous distribution. A hypothetical model with $\Sigma = 0$ would not support the no-tie argument for initial loss, but the present assumption $\tau > 0$ excludes that case.

Now suppose the two methods have different computation costs. Initial-loss selection receives k_0 updates, while selection by predicted future loss receives k, where $0 \leq k \leq k_0 \leq H$. Adding and subtracting $\mathbb{E}Q_{k,\widehat{m}_0}$ gives

$$\mathbb{E}Q_{k_0,\widehat{m}_0} - \mathbb{E}\min_{m \leq M} Q_{k,m} = D_k - \mathbb{E}[Q_{k,\widehat{m}_0} - Q_{k_0,\widehat{m}_0}]. \tag{26}$$

Selection by predicted future loss has lower limiting expected loss exactly when the right side is positive. The corresponding difference in expected finite-width actual loss converges to it by applying the preceding selection argument at both update counts and using the approximation in Theorem 3.1. If $\|A\|_{\mathrm{op}} \leq 1$, its symmetric eigenvalues satisfy $|\lambda| \leq 1$, so $\lambda^{2k} - \lambda^{2k_0} \geq 0$. Thus $B_k - B_{k_0}$ is positive semidefinite and the second expectation in (26) is nonnegative even though the residual was selected using initial loss. Without contraction, its sign is not determined. A positive D_k alone does not guarantee lower loss under the same total compute budget.

5.6 A SUFFICIENT CONDITION FOR OUTPERFORMING INITIAL-LOSS SELECTION

We now give a computable sufficient condition for the right side of (26) to be positive. Unlike the full-support sign criterion, the condition compares the reduction in loss obtained by choosing a better candidate with the reduction that initial-loss selection obtains from a fixed number of additional updates.

Proposition 5.2 (Sufficient condition under a fixed compute budget)**.** *Suppose $n, M \geq 2$, $Y = 0$, and, in a common orthonormal basis with $p = n - 1$,*

$$\Sigma = \mathrm{diag}(s_s I_p, s_f), \qquad B_t = \mathrm{diag}(r_s^t I_p, r_f^t), \qquad s_s, s_f > 0, \quad 0 < r_f < r_s < 1. \tag{27}$$

Let $X_m \sim \chi_p^2$ and $Z_m \sim \chi_1^2$ be mutually independent, let $\widehat{m}_0 \in \mathrm{argmin}_{1 \leq m \leq M}(s_s X_m + s_f Z_m)$, with smallest-index tie breaking, and define

$$c_{M,p} = \mathbb{E} \min_{m \leq M} X_m, \qquad v_s = s_s \mathbb{E} X_{\widehat{m}_0}, \qquad v_f = s_f \mathbb{E} Z_{\widehat{m}_0}. \tag{28}$$

Then

$$v_s > s_s c_{M,p}, \qquad 0 < v_f < s_f. \tag{29}$$

For integers $k, \Delta \geq 0$ with $k + \Delta \leq H$, put

$$a_\Delta = v_s r_s^\Delta - s_s c_{M,p}, \qquad b_\Delta = s_f - v_f r_f^\Delta.$$

The difference between the limiting expected loss of initial-loss selection with $k + \Delta$ updates and that of selection by predicted future loss with k updates satisfies

$$\mathbb{E} Q_{k+\Delta, \widehat{m}_0} - \mathbb{E} \min_{m \leq M} Q_{k,m} \geq \frac{a_\Delta r_s^k - b_\Delta r_f^k}{2n}. \tag{30}$$

In particular, it is strictly positive if

$$a_\Delta > 0, \qquad k > \frac{\log(b_\Delta / a_\Delta)}{\log(r_s / r_f)}. \tag{31}$$

For every fixed $n, M \geq 2$ and positive integer Δ, there is a sufficiently large H, fixed independently of width, for which the network in Section 2 realizes a nonempty parameter class satisfying (31)*. There are open intervals of normalized compute costs for which the two methods receive k and $k + \Delta$ updates.*

Proof. In the basis of (27), the initial loss of candidate m is $Q_{0,m} = (s_s X_m + s_f Z_m)/(2n)$, with independent components $s_s X_m$ and $s_f Z_m$. Thus $\widehat{m}_0$ is the same initial-loss selector as above. Under the training dynamics, the first decays more slowly than the second. Consequently,

$$\mathbb{E} Q_{k+\Delta, \widehat{m}_0} = \frac{r_s^{k+\Delta} v_s + r_f^{k+\Delta} v_f}{2n}. \tag{32}$$

Let $\widehat{m}_s \in \mathrm{argmin}_{1 \leq m \leq M} X_m$, using the smallest index to break ties. Evaluating the candidate chosen by this auxiliary rule after k updates gives an upper bound on the minimum predicted loss. Since $\widehat{m}_s$ is measurable with respect to $(X_m)_m$ and the fast variables are independent, $\mathbb{E} Z_{\widehat{m}_s} = 1$. Therefore

$$\mathbb{E} \min_{m \leq M} Q_{k,m} \leq \frac{s_s c_{M,p} r_s^k + s_f r_f^k}{2n}. \tag{33}$$

Subtracting (33) from (32) proves (30); rearranging its strict positivity gives (31) once $a_\Delta > 0$.

Pathwise, $X_{\widehat{m}_0} \geq \min_m X_m$. The inequality is strict with positive probability: one candidate can have the smallest X_m while a second has slightly larger X_m and sufficiently smaller Z_m to win the initial score, with all other candidates above both. The product density is positive on this open event, so $v_s > s_s c_{M,p}$.

For the fast coordinate, write $W = s_s X + s_f Z$ and let $\bar{F}_W$ denote its survival function. Symmetry and absence of ties give

$$\mathbb{E} Z_{\widehat{m}_0} = M \mathbb{E}\{Z \bar{F}_W(W)^{M-1}\}.$$

After averaging over X, the factor multiplying the density of Z is a strictly decreasing function of Z. Its covariance with Z is strictly negative, while $M\mathbb{E}\bar{F}_W(W)^{M-1} = 1$ normalizes the selected density. Thus $0 < \mathbb{E}Z_{\widehat{m}_0} < \mathbb{E}Z = 1$, proving (29); in particular $b_\Delta > 0$.

For realizability, take $d = n$, $x_i = e_i \in \mathbb{R}^n$, $Y = 0$, and any $\tau > 0$. Then $s_{ii} = 2$ and $s_{ij} = 1$ for $i \neq j$. Put

$$t_0 = \frac{2}{n+1}, \qquad t_1 = \frac{\sqrt{3} + 2\pi/3}{\pi(n+1)}, \qquad q = t_0 + \tau^2, \qquad \rho = \frac{t_1 + \tau^2}{q},$$

$$\sigma = \frac{q}{2}, \qquad u = \frac{q}{2\pi}\{\sqrt{1-\rho^2} + (\pi - \arccos\rho)\rho\}, \qquad h = \frac{\pi - \arccos\rho}{2\pi},$$

and

$$\kappa = \sigma + 1 + \frac{t_0}{2}, \qquad z = u + (4/3 + t_1)h.$$

Substitution in (4) gives equicorrelation matrices with diagonal/off-diagonal pairs (σ, u) and (κ, z) for Σ and $\overline{K}$. Their contrast and mean eigenvalues are

$$s_s = \sigma - u, \qquad s_f = \sigma + (n-1)u,$$

$$\lambda_s = \kappa - z, \qquad \lambda_f = \kappa + (n-1)z.$$

Here $0 < s_s < s_f$ and $0 < \lambda_s < \lambda_f$. In particular, $\lambda_s > 0$ follows from the strictly positive contrast eigenvalue of the D_2 block. Setting $r_j = (1 - \eta\lambda_j/n)^2$, every

$$0 < \eta < \min\left\{\frac{n}{\lambda_f},\ \frac{n}{\lambda_s}\left[1 - \left(\frac{s_s c_{M,p}}{v_s}\right)^{1/(2\Delta)}\right]\right\} \tag{34}$$

gives $0 < r_f < r_s < 1$ and $a_\Delta > 0$. A finite k satisfying (31) then exists; choose any fixed $H \geq k + \Delta$.

The constants are directly computable. With f_p, f_1 the chi-square densities,

$$c_{M,p} = \int_0^\infty \mathbb{P}(\chi_p^2 > x)^M \, dx,$$

$$v_s = M s_s \int_0^\infty \int_0^\infty x f_p(x) f_1(y) \bar{F}_W(s_s x + s_f y)^{M-1} \, dx \, dy,$$

$$v_f = M s_f \int_0^\infty \int_0^\infty y f_p(x) f_1(y) \bar{F}_W(s_s x + s_f y)^{M-1} \, dx \, dy.$$

Their integration dimension does not grow with n.

There are open intervals of normalized costs giving the required allocations for every Δ. For example take $g = i_1 = e = 1$, $i_M = M$, $a_M = 0$, $q_0 = 1/(2M)$, and $q_{\text{target}} = q_0 + \Delta/M$. Here q_0 and q_{target} are the per-candidate initial-loss and predicted-loss scoring costs. The normalized overheads are $M - 1/2$ and $M - 1/2 + \Delta$. After choosing k above, take the integer baseline budget $b = k + \Delta + M$ and fix $H \geq b$. Then b minus these two overheads equals $k + \Delta + 1/2$ and $k + 1/2$, respectively, so the allocations differ by exactly Δ and remain unchanged under sufficiently small perturbations of the costs. The selected-coordinate convergence and (14) transfer every strict limiting advantage to expected actual training loss at all sufficiently large widths. □

A fully explicit example with two data points and two candidates. For $n = M = 2$, take $d = 1$, $x_1 = 1$, $x_2 = -1$, $Y = 0$, $\tau = 1/2$, and $\eta = 1/20$. Define

$$\rho = \frac{1 + 4/\pi}{5}, \qquad u = \frac{5}{8\pi}\{\sqrt{1-\rho^2} + (\pi - \arccos\rho)\rho\},$$

$$p_2 = \frac{\pi - \arccos\rho}{2\pi}, \qquad z = u + (1 + 1/\pi)p_2.$$

Then

$$\Sigma = \begin{pmatrix} 5/8 & u \\ u & 5/8 \end{pmatrix}, \qquad \overline{K} = \begin{pmatrix} 17/8 & z \\ z & 17/8 \end{pmatrix}.$$

The covariance and kernel eigenvalues in the slow and fast directions are

$$s_s = \frac{5}{8} - u, \quad s_f = \frac{5}{8} + u, \qquad \lambda_s = \frac{17}{8} - z, \quad \lambda_f = \frac{17}{8} + z,$$

with $r_j = (1 - \lambda_j/40)^2$ for $j \in \{s, f\}$. To state a directly certifiable lower bound, let $c = 1 - 2/\pi$ and $h_s = 1/(1 + \sqrt{s_s/s_f})$. For independent standard normals, define

$$F(a, b) = \mathbb{E} \min_{m \le 2}(aZ_{m1}^2 + bZ_{m2}^2) = a + b - \frac{2}{\pi} \int_0^{\pi/2} \sqrt{a^2 \cos^2\theta + b^2 \sin^2\theta}\, d\theta.$$

Indeed, the difference of the two candidate energies becomes $2(aUV + bST)$ after independent orthogonal transforms, where U, V, S, T are standard normal. Conditional absolute-normal integration followed by polar coordinates gives its expected absolute value as $4/\pi$ times the displayed integral; the identity then follows from $\min(A, B) = (A + B - |A - B|)/2$. Almost-sure uniqueness of the winner permits differentiation under the expectation: the coordinate difference quotients are dominated by the integrable variable $\max_{m \le 2} Z_{mj}^2$. Writing the integral term as $J(a, b)$ gives $v_s = s_s\{1 - \partial_1 J(s_s, s_f)\}$ and $v_f = s_f\{1 - \partial_2 J(s_s, s_f)\}$. The bounds

$$\partial_1 J(a, b) \le \min\left\{\frac{2}{\pi}, \frac{\sqrt{a}}{\sqrt{a} + \sqrt{b}}\right\}, \qquad \partial_2 J(a, b) \le \frac{2}{\pi}$$

follow by lower-bounding the square-root denominator first by $a \cos\theta$ and then by $a\cos^2\theta + b\sin^2\theta$; the latter integral is elementary after $t = \tan\theta$. Hence $v_s \ge s_s h_s$ and $v_f \ge s_f c$. Also $J(a, b) \ge 2a/\pi$, so $F(a, b) \le ca + b$. Therefore, for $\Delta = 4$ and $k = 40$, the right side of (30) is at least

$$L = \frac{1}{4}\left[h_s s_s r_s^{44} + c s_f r_f^{44} - c s_s r_s^{40} - s_f r_f^{40}\right]. \tag{35}$$

Exact interval evaluation gives

$$\mathbb{E}Q_{44,\widehat{m}_0} - \mathbb{E} \min_{m \le 2} Q_{40,m} \ge L > 0.0001428271244. \tag{36}$$

An exact-rational computation encloses the conservative bound L in $[0.0001428271244696, 0.0001428271248857]$; no Monte Carlo estimate is used for the sign. The declared costs

$$g = i_1 = e = 1, \quad i_2 = 2, \quad a_2 = 0, \quad q_0 = 1/4, \quad q_{\text{target}} = 9/4, \quad b = 46$$

give total budget 48; take $H \ge 46$. Initial-loss selection and selection by predicted future loss cost 47.5 at 44 and 40 updates, respectively, while adding one further update to either method raises its cost to 48.5. These are analytic normalized costs, not runtime measurements.

The strict expected improvement and the two update allocations persist on an open neighborhood of this example, including nonzero labels and asymmetric inputs. Couple nearby residuals by a common standard Gaussian vector. Positive definiteness persists, initial-score ties have probability zero at the center, and both endpoint losses have a common integrable quadratic envelope. Dominated convergence therefore gives continuity of the expected loss difference. The explicit threshold above applies to the class in Proposition 5.2 with $Y = 0$.

For $n = 1$ and $\gamma \neq 0$, all later scores are positive scalar multiples of Q_0, so selection by initial loss and by predicted future loss gives the same ordering and $D_k = 0$. For $\gamma = 0, k \ge 1$, every predicted loss after k updates is zero, again giving no improvement from selection when both methods receive the same number of updates. The scalar example proves an improvement only over a random initialization. If initial-loss selection is less expensive and $|\gamma| \le 1$, its additional updates can only improve its comparison in (26).

5.7 Special cases and assumptions

With $M = 1$, there is no benefit from selecting among candidates. Formula (21) shows that selection has higher expected loss than the random baseline whenever $\gamma^{2(b-k)}(1 + u^2) < c_M(u)$, even though the initialization and training approximations hold. At $\eta\kappa = 1$ and $k \ge 1$, all limiting predicted losses are zero; this point lies outside the open scalar class for which a strict improvement was

proved. The limit and approximation theorems permit any fixed finite learning rate. Contraction is used only for the stated conclusions about the sign of additional training. The results allow repeated inputs, zero inputs before appending the constant coordinate, zero labels, and singular kernels. The appended constant ensures positive bottom-layer marginal variance, and $\tau > 0$ supplies the upper-layer Gaussian small-ball bounds. Extending the results to $\tau = 0$ would require a separate argument.

6 Experimental design, computation, and inference

6.1 Independent data sets and paired candidate pools

The $w = 128$ reference experiment uses inputs uniform on $[-1, 1]^4$, $n_{\text{train}} = 32$, $n_{\text{test}} = 256$, lower and upper widths 128 and 256, learning rate 0.05, upper-bias initialization standard deviation 0.5, and independent training noise with standard deviation 0.1. Each input $x = (x_1, x_2, x_3, x_4)^\top$ has independent $\text{Unif}[-1, 1]$ coordinates. Inputs are independent across observations and between the training and test samples; training noise is independent across observations and of all inputs. The regression functions are

$$f_{\text{smooth}}(x) = 0.8 + 0.35x_1 - 0.20x_2 + 0.15x_3x_4, \tag{37}$$

$$f_{\text{mixed}}(x) = 0.45\sin(2.2x_1) - 0.35x_2x_3 + 0.25x_4^2 - 0.10. \tag{38}$$

Training loss uses noisy labels; test loss uses the noiseless regression function at an independent finite test sample. Selection uses training data only.

For each regression model, we generate $D = 200$ independent data sets and $R = 128$ independently initialized four-candidate groups per data set. We compare all methods on the same groups. For $M = 2, 3, 4$, we use the first M candidates, and we select separately at each budget $b = 20, 50$. We average within each data set and use its mean as one independent observation for inference.

Random selection uses the first candidate. Initial-loss selection minimizes observed training loss at initialization. One-step and two-step trials apply the stated number of updates to every candidate, choose the lowest training loss, and continue the winner from its current parameters. Successive halving uses four candidates at update 2, two survivors at update 4, and one survivor thereafter. The two NTK selectors predict training loss at their own allocated horizons from initial outputs and either the candidate-specific empirical kernel or the limiting kernel. They then train the winner from initialization. Ties use the smallest original candidate index.

6.2 Measured primitive costs and feasible allocations

Primitive operation costs are measured before the main outcomes are generated. Allocations use the median measured cost of each operation.

Let i_M denote initialization of M candidates, g the cost of one gradient calculation and simultaneous update, f a training-loss evaluation, and $e = f + f_{\text{test}}$ the final training and test evaluations. Let a_M denote an M-candidate minimum selection, and let a_{SH} denote the two successive-halving ranking operations. The reference budget is $i_1 + bg + e$, conditional on the previously measured cost profile. Offline calibration is outside the per-selection budget. The other costs, with k the winner's total updates, are

$$\begin{aligned}
C_{\text{initial}}(k) &= i_M + Mf + a_M + kg + e,\\
C_{r\text{-step}}(k) &= i_M + Mf + a_M + [k + (M-1)r]g + e, \quad k \geq r,\\
C_{\text{SH}}(k) &= i_4 + 6f + a_{\text{SH}} + (k+8)g + e, \quad k \geq 4,\\
C_{\text{target}}(k) &= i_M + Mq_b + a_M + kg + e,\\
C_{\text{limit}}(k) &= i_M + q_K + Mq_b^R + a_M + kg + e.
\end{aligned}$$

Here q_b includes initial-output evaluation, empirical-kernel construction, b residual updates, and the endpoint score. The limiting method pays q_K once for kernel construction and q_b^R per candidate for initial-output evaluation, residual updates, and the endpoint score. We predict loss after the allocated k updates, while including the cost of propagating each score through the reference b updates.

For each model, method, candidate count, and budget, we use the largest integer $k \leq b$ whose total computation fits within the budget. Random, initial-loss, and NTK methods allow $k = 0$; one-step trials, two-step trials, and successive halving require at least 1, 2, and 4 updates, respectively. We mark a method as infeasible if no update count satisfies these requirements. We use successive halving only for $M = 4$. Each model has its own allocation, and all methods in this $w = 128$ study were feasible where defined.

Table 1: Training-update allocations at $w = 128$ after charging measured initialization, selection, updates, and final training and test evaluation. Each model uses the maximum number of updates available within its budget. SH denotes successive halving.

Model	M	b	Random	Initial	One-step	Two-step	SH	Empirical	Limiting
Eq. (37)	2	20	20	17	16	15	–	15	16
Eq. (37)	2	50	50	47	46	45	–	45	46
Eq. (37)	3	20	20	15	13	11	–	12	14
Eq. (37)	3	50	50	45	43	41	–	42	43
Eq. (37)	4	20	20	14	11	8	5	10	12
Eq. (37)	4	50	50	44	41	38	35	39	41
Eq. (38)	2	20	20	17	16	15	–	15	16
Eq. (38)	2	50	50	47	46	45	–	45	45
Eq. (38)	3	20	20	15	13	11	–	12	14
Eq. (38)	3	50	50	45	43	41	–	41	43
Eq. (38)	4	20	20	13	10	7	4	9	12
Eq. (38)	4	50	50	43	40	37	34	38	40

6.3 Paired normal-approximation intervals

For a comparator and the empirical NTK, let D_{dr} be comparator loss minus empirical-NTK loss in candidate group r of data set d. Positive values favor empirical NTK. We compute

$$Z_d = \frac{1}{R}\sum_{r=1}^{R} D_{dr}, \qquad \overline{Z} = \frac{1}{D}\sum_{d=1}^{D} Z_d, \qquad \mathrm{MCSE} = \sqrt{\frac{\sum_{d=1}^{D}(Z_d - \overline{Z})^2}{D(D-1)}}.$$

Each reported interval is $\overline{Z} \pm 1.959963984540054\,\mathrm{MCSE}$. These are pointwise 95% normal-approximation intervals. The sample variance of the data-set means already includes the variability from averaging finitely many candidate groups. The independent units for the reported intervals are therefore the D data-set means. The same procedure reports means and MCSEs for each individual method and for every available pairwise comparison.

6.4 Training and test loss comparisons

Tables 2–5 complement the main-text results. They report paired comparisons with empirical NTK, individual method means, and the budget comparison for all three candidate counts. The sign convention and pointwise interval construction are the same as in the main text.

Table 2: All four-candidate training half-MSE comparisons at $w = 128$ with empirical NTK. Positive differences favor empirical NTK; intervals are pointwise normal approximations.

Model	b	Comparator	Mean	MCSE	95% interval
Eq. (37)	20	Random	0.03321	0.00054	[0.03215, 0.03427]
Eq. (37)	20	Initial loss	-0.01264	0.00016	[-0.01295, -0.01234]
Eq. (37)	20	One-step trial	-0.00252	0.00006	[-0.00263, -0.00240]
Eq. (37)	20	Two-step trial	0.01067	0.00012	[0.01044, 0.01090]
Eq. (37)	20	Successive halving	0.02800	0.00031	[0.02740, 0.02861]
Eq. (37)	20	Limiting NTK	-0.00782	0.00008	[-0.00798, -0.00765]
Eq. (37)	50	Random	0.01264	0.00018	[0.01230, 0.01298]
Eq. (37)	50	Initial loss	0.00769	0.00011	[0.00747, 0.00791]
Eq. (37)	50	One-step trial	0.00906	0.00012	[0.00883, 0.00929]
Eq. (37)	50	Two-step trial	0.01063	0.00012	[0.01039, 0.01087]
Eq. (37)	50	Successive halving	0.01214	0.00013	[0.01190, 0.01239]
Eq. (37)	50	Limiting NTK	-0.00018	0.00003	[-0.00023, -0.00012]
Eq. (38)	20	Random	0.02600	0.00034	[0.02534, 0.02667]
Eq. (38)	20	Initial loss	-0.00426	0.00006	[-0.00437, -0.00415]
Eq. (38)	20	One-step trial	-0.00016	0.00004	[-0.00023, -0.00009]
Eq. (38)	20	Two-step trial	0.00506	0.00005	[0.00496, 0.00516]
Eq. (38)	20	Successive halving	0.01176	0.00010	[0.01156, 0.01196]
Eq. (38)	20	Limiting NTK	-0.00561	0.00005	[-0.00571, -0.00551]
Eq. (38)	50	Random	0.01902	0.00022	[0.01859, 0.01945]
Eq. (38)	50	Initial loss	0.00657	0.00010	[0.00637, 0.00676]
Eq. (38)	50	One-step trial	0.00765	0.00010	[0.00745, 0.00785]
Eq. (38)	50	Two-step trial	0.00885	0.00011	[0.00865, 0.00906]
Eq. (38)	50	Successive halving	0.00972	0.00011	[0.00951, 0.00993]
Eq. (38)	50	Limiting NTK	-0.00065	0.00002	[-0.00068, -0.00062]

Table 3: All four-candidate test half-MSE comparisons at $w = 128$ with empirical NTK. Positive differences favor empirical NTK; intervals are pointwise normal approximations.

Model	b	Comparator	Mean	MCSE	95% interval
Eq. (37)	20	Random	0.03651	0.00073	[0.03508, 0.03795]
Eq. (37)	20	Initial loss	-0.01223	0.00013	[-0.01248, -0.01197]
Eq. (37)	20	One-step trial	-0.00251	0.00009	[-0.00268, -0.00234]
Eq. (37)	20	Two-step trial	0.01016	0.00011	[0.00995, 0.01037]
Eq. (37)	20	Successive halving	0.02682	0.00022	[0.02640, 0.02724]
Eq. (37)	20	Limiting NTK	-0.00758	0.00008	[-0.00774, -0.00741]
Eq. (37)	50	Random	0.01431	0.00026	[0.01379, 0.01482]
Eq. (37)	50	Initial loss	0.00819	0.00018	[0.00783, 0.00855]
Eq. (37)	50	One-step trial	0.00946	0.00018	[0.00912, 0.00981]
Eq. (37)	50	Two-step trial	0.01096	0.00017	[0.01062, 0.01130]
Eq. (37)	50	Successive halving	0.01241	0.00017	[0.01208, 0.01273]
Eq. (37)	50	Limiting NTK	-0.00028	0.00006	[-0.00039, -0.00017]
Eq. (38)	20	Random	0.02509	0.00038	[0.02435, 0.02584]
Eq. (38)	20	Initial loss	-0.00379	0.00012	[-0.00403, -0.00355]
Eq. (38)	20	One-step trial	-0.00009	0.00011	[-0.00030, 0.00011]
Eq. (38)	20	Two-step trial	0.00460	0.00010	[0.00441, 0.00479]
Eq. (38)	20	Successive halving	0.01073	0.00010	[0.01054, 0.01093]
Eq. (38)	20	Limiting NTK	-0.00528	0.00007	[-0.00541, -0.00516]
Eq. (38)	50	Random	0.01965	0.00027	[0.01911, 0.02019]
Eq. (38)	50	Initial loss	0.00726	0.00017	[0.00692, 0.00760]
Eq. (38)	50	One-step trial	0.00817	0.00017	[0.00784, 0.00851]
Eq. (38)	50	Two-step trial	0.00918	0.00017	[0.00884, 0.00951]
Eq. (38)	50	Successive halving	0.00986	0.00016	[0.00955, 0.01017]
Eq. (38)	50	Limiting NTK	-0.00079	0.00005	[-0.00089, -0.00070]

Table 4: Individual method half-MSE at $w = 128, M = 4, b = 50$. MCSEs use independent data-set means.

Model	Endpoint	Method	Mean	MCSE
Eq. (37)	train	Random	0.04797	0.00033
Eq. (37)	train	Initial loss	0.04302	0.00029
Eq. (37)	train	One-step trial	0.04439	0.00029
Eq. (37)	train	Two-step trial	0.04596	0.00030
Eq. (37)	train	Successive halving	0.04747	0.00030
Eq. (37)	train	Empirical NTK	0.03533	0.00022
Eq. (37)	train	Limiting NTK	0.03515	0.00022
Eq. (37)	test	Random	0.05420	0.00046
Eq. (37)	test	Initial loss	0.04809	0.00030
Eq. (37)	test	One-step trial	0.04936	0.00031
Eq. (37)	test	Two-step trial	0.05085	0.00032
Eq. (37)	test	Successive halving	0.05230	0.00032
Eq. (37)	test	Empirical NTK	0.03989	0.00029
Eq. (37)	test	Limiting NTK	0.03962	0.00029
Eq. (38)	train	Random	0.07112	0.00058
Eq. (38)	train	Initial loss	0.05866	0.00047
Eq. (38)	train	One-step trial	0.05975	0.00048
Eq. (38)	train	Two-step trial	0.06095	0.00049
Eq. (38)	train	Successive halving	0.06181	0.00050
Eq. (38)	train	Empirical NTK	0.05210	0.00045
Eq. (38)	train	Limiting NTK	0.05144	0.00044
Eq. (38)	test	Random	0.07868	0.00042
Eq. (38)	test	Initial loss	0.06629	0.00037
Eq. (38)	test	One-step trial	0.06720	0.00037
Eq. (38)	test	Two-step trial	0.06820	0.00037
Eq. (38)	test	Successive halving	0.06889	0.00036
Eq. (38)	test	Empirical NTK	0.05903	0.00028
Eq. (38)	test	Limiting NTK	0.05823	0.00028

Table 5: Test half-MSE budget comparison at $w = 128$ for all candidate counts: initial loss minus empirical NTK.

Model	M	b	Mean	MCSE	95% interval
Eq. (37)	2	20	-0.00782	0.00012	[-0.00805, -0.00759]
Eq. (37)	2	50	0.00627	0.00013	[0.00601, 0.00652]
Eq. (37)	3	20	-0.01036	0.00012	[-0.01060, -0.01013]
Eq. (37)	3	50	0.00839	0.00017	[0.00806, 0.00873]
Eq. (37)	4	20	-0.01223	0.00013	[-0.01248, -0.01197]
Eq. (37)	4	50	0.00819	0.00018	[0.00783, 0.00855]
Eq. (38)	2	20	-0.00052	0.00013	[-0.00078, -0.00026]
Eq. (38)	2	50	0.00695	0.00016	[0.00664, 0.00727]
Eq. (38)	3	20	-0.00155	0.00013	[-0.00181, -0.00129]
Eq. (38)	3	50	0.00758	0.00019	[0.00721, 0.00796]
Eq. (38)	4	20	-0.00379	0.00012	[-0.00403, -0.00355]
Eq. (38)	4	50	0.00726	0.00017	[0.00692, 0.00760]

6.5 Prediction comparison at a common number of updates

For $M = 4$, we evaluate both predictors after the number of updates allocated to empirical-NTK selection. We measure each candidate's squared prediction error for training half-MSE and average over candidates and groups within each data set. We report the difference between limiting-NTK and empirical-NTK prediction errors across the 200 independent data-set means. This comparison holds the number of updates fixed; the budget comparison allows each method its own allocated number of updates.

Table 6: Common-horizon mean squared error for predicting training half-MSE at $w = 128, M = 4$: limiting NTK minus empirical NTK. Positive values favor empirical NTK. The horizons at $b = 20, 50$ are $k_E = 10, 39$ for Eq. (37) and $9, 38$ for Eq. (38); the units are squared half-MSE.

Model	b	Mean difference	MCSE	95% interval
Eq. (37)	20	0.0010288	0.0000177	[0.0009942, 0.0010634]
Eq. (37)	50	0.0001311	0.0000018	[0.0001277, 0.0001346]
Eq. (38)	20	0.0001739	0.0000034	[0.0001672, 0.0001805]
Eq. (38)	50	0.0000605	0.0000011	[0.0000584, 0.0000626]

7 Width dependence of the two NTK selectors

7.1 Design and fixed-horizon comparisons

We extend the two-model experiment to lower hidden widths $w = 64, 128, 256$, with upper widths $2w$. Each width uses the same 200 independent training and test data sets per model and 128 candidate groups per data set. Candidate counts, budgets, learning rate, noise, and selection methods are those in Section 6. Data sets are paired across widths, while network parameter arrays are not nested across widths. We use the original $w = 128$ costs and results. We fixed the design and the two comparison horizons before generating results for the additional widths.

Widths 64 and 256 have separately measured operation costs for each model. The budget remains the cost of training one random candidate for $b = 20$ or 50 updates at the relevant width and model. Thus methods share a budget within each configuration, while the absolute time budget varies across widths. The scoring charge uses b residual updates and the selector predicts at its allocated k, as in Section 6. At $w = 256$, $M = 4$, and $b = 20$, successive halving has an empty feasible set for the smooth model; it is reported as infeasible.

To compare the predictors separately from their cost allocations, we evaluate both at fixed horizons $h = 10$ and 40 for every width. For each candidate, we compute the squared prediction error for actual training half-MSE. At the group level, we also record which candidate each score selects, the selected network's actual training and test half-MSE after h updates, and an indicator that the selectors choose the same candidate. Agreement fractions are obtained by averaging these indicators within each data set. Two further summaries directly measure the discrepancy between the kernels and scores: the candidate-average relative Frobenius norm $\|K_{0,m} - \overline{K}\|_F / \|\overline{K}\|_F$ and the groupwise maximum $\max_{m \le M} |S_{h,m} - \widetilde{S}_{h,m}|$. Means and intervals use the same independent-data-set construction as Section 6; cross-width contrasts are paired by data-set index. All intervals are pointwise.

7.2 Selection and update-count components

For the equal-budget comparison, let $\widehat{m}_E, \widehat{m}_L$ be the empirical- and limiting-NTK choices, with allocated update counts k_E, k_L. The choices are made at their respective allocated horizons. In the two-letter loss notation below, the first subscript identifies the selector and the second the update count. For either actual training or test half-MSE, write $L_{EE} = L_{k_E,\widehat{m}_E}$, $L_{EL} = L_{k_L,\widehat{m}_E}$, $L_{LE} = L_{k_E,\widehat{m}_L}$, and $L_{LL} = L_{k_L,\widehat{m}_L}$. The limiting-minus-empirical difference has the exact symmetric decomposition

$$\begin{aligned} L_{LL} - L_{EE} = {} & \tfrac{1}{2}\{(L_{LE} - L_{EE}) + (L_{LL} - L_{EL})\} \\ & + \tfrac{1}{2}\{(L_{EL} - L_{EE}) + (L_{LL} - L_{LE})\}. \end{aligned}$$

The first term compares the selected networks at common update counts; the second compares update counts for each selected network. Their sum equals the final loss difference for every group. Both terms can have either sign. This is a descriptive decomposition of the implemented selection rules and training trajectories; the first term reflects the kernels and the horizons used for selection together. We average the terms within each data set before computing their MCSEs and intervals.

7.3 RESULTS

Figure 1 and Tables 7–15 report $M = 4$. At either fixed horizon, empirical NTK has lower mean prediction squared error and lower mean selected-network training and test losses in every model–width configuration. All prediction-error and training-loss intervals exclude zero. Eleven of the 12 test-loss intervals exclude zero; at $w = 256, h = 10$ in the mixed model, limiting minus empirical test half-MSE is 0.000063 with interval $[-0.000019, 0.000145]$.

The observed mean kernel and score discrepancies decrease across the three widths. The mean relative Frobenius discrepancy is approximately 0.209, 0.148, and 0.105 for the smooth model and 0.208, 0.148, and 0.104 for the mixed model. At $h = 40$, selection agreement rises from 75.0% to 86.7% in the smooth model and from 83.7% to 91.9% in the mixed model. The empirical-NTK mean test-loss advantage at that horizon decreases from 0.00238 to 0.00064 and from 0.00060 to 0.00015, respectively. These kernel, score, and training-loss patterns agree with the common-horizon result in Section 5.2. We report the test-loss and selection-agreement trends as empirical findings.

Under equal budgets, limiting NTK has lower mean test loss in 11 of the 12 model–width–budget configurations, with the corresponding intervals excluding zero. The remaining configuration is the smooth model at $w = 64, b = 50$, where empirical NTK has lower loss. At $b = 50$, the symmetric selection component favors empirical NTK in both models at all three widths, while the update-count component favors limiting NTK. For example, at $w = 64$ in the smooth model, the limiting-minus-empirical components are 0.00257 and -0.00186, giving a total of 0.00071. At $w = 256$, they are 0.00070 and -0.00160, giving -0.00090. Thus the smaller fixed-horizon difference at greater width can coexist with a larger equal-budget advantage for the less expensive selector.

Table 7: Equal-budget train half-MSE: limiting NTK minus empirical NTK. Positive differences favor empirical NTK. $M = 4$; means, MCSEs, and pointwise 95% normal-z intervals use 200 independent data-set means, each averaging 128 candidate pools. Each budget is the random-reference cost at its own width and model, not a common number of seconds.

Model	w	b	Mean	MCSE	95% interval
Smooth	64	20	-0.011856	0.000119	$[-0.012089, -0.011623]$
Smooth	64	50	0.001021	5.732×10^{-5}	$[0.000908, 0.001133]$
Smooth	128	20	-0.007817	8.478×10^{-5}	$[-0.007983, -0.007651]$
Smooth	128	50	-0.000176	2.934×10^{-5}	$[-0.000233, -0.000118]$
Smooth	256	20	-0.012839	0.000146	$[-0.013124, -0.012554]$
Smooth	256	50	-0.000877	1.805×10^{-5}	$[-0.000913, -0.000842]$
Mixed	64	20	-0.005622	5.534×10^{-5}	$[-0.005730, -0.005514]$
Mixed	64	50	-0.000807	2.599×10^{-5}	$[-0.000858, -0.000757]$
Mixed	128	20	-0.005610	4.952×10^{-5}	$[-0.005707, -0.005513]$
Mixed	128	50	-0.000654	1.562×10^{-5}	$[-0.000685, -0.000624]$
Mixed	256	20	-0.005812	5.333×10^{-5}	$[-0.005917, -0.005708]$
Mixed	256	50	-0.000920	1.149×10^{-5}	$[-0.000943, -0.000898]$

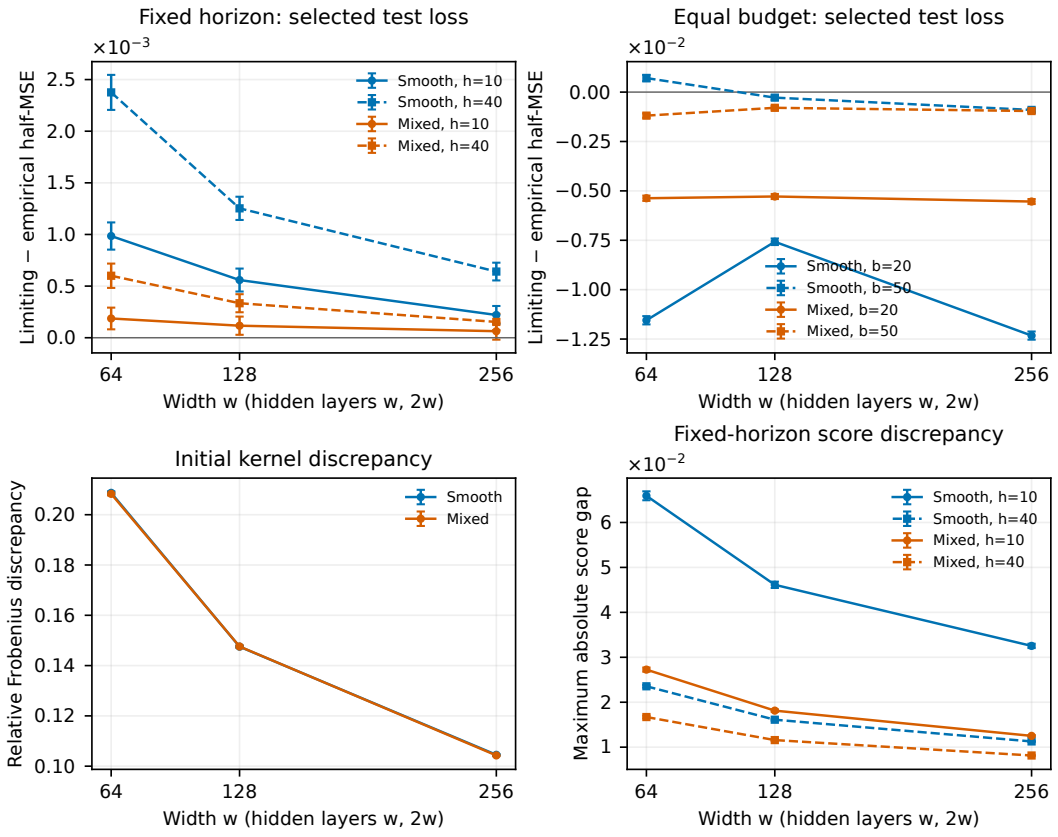

Figure 1: Width comparisons at $M = 4$. Top: limiting minus empirical selected-network test half-MSE at fixed update counts (left) and each method's equal-budget allocation (right); positive values favor empirical NTK. Bottom: the candidate-average relative kernel discrepancy (left) and the mean of each pool's maximum absolute score difference (right). Error bars are pointwise 95% normal-z intervals over 200 independent data-set means. Widths share data sets; equal budgets are defined separately at each width.

Table 8: Equal-budget test half-MSE: limiting NTK minus empirical NTK. Positive differences favor empirical NTK. $M = 4$; means, MCSEs, and pointwise 95% normal-z intervals use 200 independent data-set means, each averaging 128 candidate pools. Each budget is the random-reference cost at its own width and model, not a common number of seconds.

Model	w	b	Mean	MCSE	95% interval
Smooth	64	20	-0.011549	0.000107	$[-0.011759, -0.011340]$
Smooth	64	50	0.000710	8.266×10^{-5}	$[0.000548, 0.000872]$
Smooth	128	20	-0.007576	8.455×10^{-5}	$[-0.007741, -0.007410]$
Smooth	128	50	-0.000280	5.743×10^{-5}	$[-0.000392, -0.000167]$
Smooth	256	20	-0.012325	0.000105	$[-0.012531, -0.012118]$
Smooth	256	50	-0.000902	4.341×10^{-5}	$[-0.000987, -0.000817]$
Mixed	64	20	-0.005372	6.840×10^{-5}	$[-0.005506, -0.005238]$
Mixed	64	50	-0.001191	5.705×10^{-5}	$[-0.001303, -0.001079]$
Mixed	128	20	-0.005284	6.562×10^{-5}	$[-0.005413, -0.005155]$
Mixed	128	50	-0.000795	4.715×10^{-5}	$[-0.000887, -0.000702]$
Mixed	256	20	-0.005539	6.487×10^{-5}	$[-0.005667, -0.005412]$
Mixed	256	50	-0.000957	4.081×10^{-5}	$[-0.001037, -0.000877]$

Table 9: Fixed-horizon prediction squared-error difference. Limiting minus empirical predictor squared error against actual training half-MSE. $M = 4$; means, MCSEs, and pointwise 95% normal-z intervals use 200 independent data-set means, each averaging 128 candidate pools.

Model	w	h	Mean	MCSE	95% interval
Smooth	64	10	0.002013	3.317×10^{-5}	$[0.001948, 0.002078]$
Smooth	64	40	0.000263	3.430×10^{-6}	$[0.000256, 0.000269]$
Smooth	128	10	0.001029	1.765×10^{-5}	$[0.000994, 0.001063]$
Smooth	128	40	0.000122	1.629×10^{-6}	$[0.000119, 0.000125]$
Smooth	256	10	0.000523	7.992×10^{-6}	$[0.000508, 0.000539]$
Smooth	256	40	6.021×10^{-5}	7.654×10^{-7}	$[5.871\times10^{-5}, 6.171\times10^{-5}]$
Mixed	64	10	0.000416	1.123×10^{-5}	$[0.000394, 0.000438]$
Mixed	64	40	0.000114	1.953×10^{-6}	$[0.000111, 0.000118]$
Mixed	128	10	0.000177	3.370×10^{-6}	$[0.000170, 0.000183]$
Mixed	128	40	5.792×10^{-5}	1.048×10^{-6}	$[5.586\times10^{-5}, 5.997\times10^{-5}]$
Mixed	256	10	8.654×10^{-5}	1.565×10^{-6}	$[8.347\times10^{-5}, 8.960\times10^{-5}]$
Mixed	256	40	2.987×10^{-5}	5.255×10^{-7}	$[2.884\times10^{-5}, 3.090\times10^{-5}]$

Table 10: Fixed-horizon selected training loss. Limiting minus empirical selected-network training half-MSE. $M = 4$; means, MCSEs, and pointwise 95% normal-z intervals use 200 independent data-set means, each averaging 128 candidate pools.

Model	w	h	Mean	MCSE	95% interval
Smooth	64	10	0.001141	3.968×10^{-5}	$[0.001063, 0.001219]$
Smooth	64	40	0.002680	5.962×10^{-5}	$[0.002563, 0.002797]$
Smooth	128	10	0.000583	2.655×10^{-5}	$[0.000531, 0.000635]$
Smooth	128	40	0.001358	3.358×10^{-5}	$[0.001293, 0.001424]$
Smooth	256	10	0.000262	1.452×10^{-5}	$[0.000233, 0.000290]$
Smooth	256	40	0.000683	1.815×10^{-5}	$[0.000647, 0.000718]$
Mixed	64	10	0.000379	1.606×10^{-5}	$[0.000347, 0.000410]$
Mixed	64	40	0.001112	2.749×10^{-5}	$[0.001058, 0.001166]$
Mixed	128	10	0.000195	8.700×10^{-6}	$[0.000178, 0.000212]$
Mixed	128	40	0.000581	1.525×10^{-5}	$[0.000551, 0.000611]$
Mixed	256	10	0.000103	5.640×10^{-6}	$[9.181\times10^{-5}, 0.000114]$
Mixed	256	40	0.000284	8.965×10^{-6}	$[0.000266, 0.000302]$

Table 11: Fixed-horizon selected test loss. Limiting minus empirical selected-network test half-MSE. $M = 4$; means, MCSEs, and pointwise 95% normal-z intervals use 200 independent data-set means, each averaging 128 candidate pools.

Model	w	h	Mean	MCSE	95% interval
Smooth	64	10	0.000985	6.720×10^{-5}	[0.000853, 0.001117]
Smooth	64	40	0.002376	8.651×10^{-5}	[0.002206, 0.002545]
Smooth	128	10	0.000559	5.670×10^{-5}	[0.000447, 0.000670]
Smooth	128	40	0.001253	5.739×10^{-5}	[0.001140, 0.001365]
Smooth	256	10	0.000220	4.445×10^{-5}	[0.000133, 0.000307]
Smooth	256	40	0.000641	4.383×10^{-5}	[0.000555, 0.000727]
Mixed	64	10	0.000186	5.361×10^{-5}	[8.134×10^{-5}, 0.000291]
Mixed	64	40	0.000600	6.008×10^{-5}	[0.000482, 0.000718]
Mixed	128	10	0.000117	4.494×10^{-5}	[2.842×10^{-5}, 0.000205]
Mixed	128	40	0.000334	4.461×10^{-5}	[0.000247, 0.000422]
Mixed	256	10	6.321×10^{-5}	4.182×10^{-5}	[-1.875×10^{-5}, 0.000145]
Mixed	256	40	0.000153	4.019×10^{-5}	[7.416×10^{-5}, 0.000232]

Table 12: Fixed-horizon selection agreement. Fraction of pools in which the empirical and limiting predictors select the same candidate. $M = 4$; means, MCSEs, and pointwise 95% normal-z intervals use 200 independent data-set means, each averaging 128 candidate pools.

Model	w	h	Mean	MCSE	95% interval
Smooth	64	10	0.929609	0.001559	[0.926555, 0.932664]
Smooth	64	40	0.750117	0.003031	[0.744177, 0.756057]
Smooth	128	10	0.950508	0.001425	[0.947714, 0.953301]
Smooth	128	40	0.814805	0.002538	[0.809830, 0.819780]
Smooth	256	10	0.966250	0.001117	[0.964061, 0.968439]
Smooth	256	40	0.867227	0.002226	[0.862864, 0.871589]
Mixed	64	10	0.942461	0.001491	[0.939538, 0.945384]
Mixed	64	40	0.836680	0.002287	[0.832197, 0.841162]
Mixed	128	10	0.958047	0.001204	[0.955688, 0.960406]
Mixed	128	40	0.884180	0.002058	[0.880147, 0.888213]
Mixed	256	10	0.970000	0.001053	[0.967936, 0.972064]
Mixed	256	40	0.919453	0.001701	[0.916119, 0.922787]

Table 13: Maximum feasible integer update allocations for $M = 4$. Initial, 1-step, 2-step, SH, Emp., and Lim. denote initial loss, one-step, two-step, successive halving, empirical NTK, and limiting NTK. *infeas.* means no admissible allocation; no loss is imputed. Prediction costs retain the predeclared q_b/q_b^R convention.

Model	w	b	Random	Initial	1-step	2-step	SH	Emp.	Lim.
Smooth	64	20	20	14	11	8	5	9	12
Smooth	64	50	50	44	41	38	35	37	39
Smooth	128	20	20	14	11	8	5	10	12
Smooth	128	50	50	44	41	38	35	39	41
Smooth	256	20	20	12	9	6	*infeas.*	9	12
Smooth	256	50	50	42	39	36	33	39	41
Mixed	64	20	20	14	11	8	5	8	11
Mixed	64	50	50	44	41	38	35	36	39
Mixed	128	20	20	13	10	7	4	9	12
Mixed	128	50	50	43	40	37	34	38	40
Mixed	256	20	20	12	9	6	4	9	12
Mixed	256	50	50	42	39	36	34	39	41

Table 14: Symmetric equal-budget decomposition of train half-MSE at $b = 50$: limiting minus empirical. The selection and horizon components sum to the total difference. $M = 4$; means, MCSEs, and pointwise 95% normal-z intervals use 200 independent data-set means, each averaging 128 candidate pools.

Model	w	Component	Mean	MCSE	95% interval
Smooth	64	Total	0.001021	5.732×10^{-5}	[0.000908, 0.001133]
Smooth	64	Selection	0.002888	6.422×10^{-5}	[0.002762, 0.003013]
Smooth	64	Horizon	−0.001867	1.388×10^{-5}	[−0.001894, −0.001840]
Smooth	128	Total	−0.000176	2.934×10^{-5}	[−0.000233, −0.000118]
Smooth	128	Selection	0.001455	3.455×10^{-5}	[0.001387, 0.001523]
Smooth	128	Horizon	−0.001631	1.153×10^{-5}	[−0.001653, −0.001608]
Smooth	256	Total	−0.000877	1.805×10^{-5}	[−0.000913, −0.000842]
Smooth	256	Selection	0.000719	1.881×10^{-5}	[0.000682, 0.000756]
Smooth	256	Horizon	−0.001596	1.232×10^{-5}	[−0.001620, −0.001572]
Mixed	64	Total	−0.000807	2.599×10^{-5}	[−0.000858, −0.000757]
Mixed	64	Selection	0.001171	2.841×10^{-5}	[0.001115, 0.001226]
Mixed	64	Horizon	−0.001978	1.817×10^{-5}	[−0.002014, −0.001943]
Mixed	128	Total	−0.000654	1.562×10^{-5}	[−0.000685, −0.000624]
Mixed	128	Selection	0.000599	1.549×10^{-5}	[0.000568, 0.000629]
Mixed	128	Horizon	−0.001253	1.109×10^{-5}	[−0.001275, −0.001231]
Mixed	256	Total	−0.000920	1.149×10^{-5}	[−0.000943, −0.000898]
Mixed	256	Selection	0.000287	8.814×10^{-6}	[0.000270, 0.000305]
Mixed	256	Horizon	−0.001208	1.070×10^{-5}	[−0.001229, −0.001187]

Table 15: Symmetric equal-budget decomposition of test half-MSE at $b = 50$: limiting minus empirical. The selection and horizon components sum to the total difference. $M = 4$; means, MCSEs, and pointwise 95% normal-z intervals use 200 independent data-set means, each averaging 128 candidate pools.

Model	w	Component	Mean	MCSE	95% interval
Smooth	64	Total	0.000710	8.266×10^{-5}	[0.000548, 0.000872]
Smooth	64	Selection	0.002573	9.012×10^{-5}	[0.002396, 0.002749]
Smooth	64	Horizon	−0.001863	1.745×10^{-5}	[−0.001897, −0.001829]
Smooth	128	Total	−0.000280	5.743×10^{-5}	[−0.000392, −0.000167]
Smooth	128	Selection	0.001356	6.204×10^{-5}	[0.001234, 0.001477]
Smooth	128	Horizon	−0.001635	1.587×10^{-5}	[−0.001667, −0.001604]
Smooth	256	Total	−0.000902	4.341×10^{-5}	[−0.000987, −0.000817]
Smooth	256	Selection	0.000698	4.591×10^{-5}	[0.000608, 0.000788]
Smooth	256	Horizon	−0.001601	1.557×10^{-5}	[−0.001631, −0.001570]
Mixed	64	Total	−0.001191	5.705×10^{-5}	[−0.001303, −0.001079]
Mixed	64	Selection	0.000596	5.962×10^{-5}	[0.000479, 0.000713]
Mixed	64	Horizon	−0.001787	1.144×10^{-5}	[−0.001809, −0.001765]
Mixed	128	Total	−0.000795	4.715×10^{-5}	[−0.000887, −0.000702]
Mixed	128	Selection	0.000336	4.766×10^{-5}	[0.000243, 0.000430]
Mixed	128	Horizon	−0.001131	7.375×10^{-6}	[−0.001145, −0.001116]
Mixed	256	Total	−0.000957	4.081×10^{-5}	[−0.001037, −0.000877]
Mixed	256	Selection	0.000129	4.023×10^{-5}	[5.016×10^{-5}, 0.000208]
Mixed	256	Horizon	−0.001086	7.186×10^{-6}	[−0.001100, −0.001072]

8 Additional experiments across regression functions and network sizes

8.1 Design and methods

This study compares the two NTK scores across additional regression functions, network widths, sample sizes, and budgets.

This separate study uses its own computation cost profile and $M = 4$, $\eta = 0.05$, $\tau = 0.5$, 20 data draws or prespecified splits, and 64 paired candidate groups per draw or split. The four synthetic targets comprise the regression functions in Eqs. (37) and (38) and

$$f_{\mathrm{osc}}(x) = 0.35\sin(4x_1) + 0.30\cos(3x_2) + 0.20x_3x_4 + 0.15x_4,$$
$$f_{\mathrm{radial}}(x) = 0.65\exp(-1.5\|x\|^2) - 0.25 + 0.20x_1x_2.$$

Synthetic inputs are uniform on $[-1, 1]^4$, training noise has standard deviation 0.1, and each test set contains 512 independently drawn inputs with noiseless labels. At $n_{\mathrm{train}} = 32$, the study uses lower/upper widths $64/128$, $128/256$, and $256/512$; it also uses width $128/256$ at $n_{\mathrm{train}} = 64$.

These settings use a budget equal to the cost of training one random initialization for 50 updates. The reference width and sample size additionally use budgets equal to 30 and 80 such updates. This gives 16 synthetic settings at the 50-update budget and eight additional synthetic settings at the other budgets.

The real-data setting uses the scikit-learn diabetes regression data, 128 training observations, and the remaining 314 observations for testing, bringing the total to 25 settings. Inputs and targets are standardized using each training split only. Its 20 test splits overlap. Inference averages random splitting and initialization conditional on the fixed source data.

The study compares random selection, initial-loss selection, selection after training every candidate for two steps, successive halving that continues from the parameter states already reached, and the candidate-specific empirical-NTK method. It also evaluates a less expensive score that uses the same deterministic limiting NTK for every candidate:

$$\widetilde{S}_{k,m} = \frac{1}{2n}\left\|\left(I - \frac{\eta}{n}\overline{K}\right)^k (Y - F_{0,m})\right\|^2.$$

The same $\overline{K}$ is used for all candidates on one data set, while each candidate retains its own initial residual. The score accounts for how the limiting NTK weights each candidate's initial errors at a later training step, with a single kernel construction shared across candidates.

Section 5.2 establishes the asymptotic comparison of the two scores at a common horizon.

We measure computation costs separately for each array shape and use them to determine the budget and number of available updates. Training one random initialization for b updates defines the total budget. Here the final evaluation cost covers test loss; Section 6 includes both training and test losses. We include the cost of predicting through b updates for each empirical kernel. For the limiting-NTK method, we include one kernel construction per selection problem and propagation of all four residuals through b updates. Table 16 reports the maximum number of training updates available to each method within the budget.

Table 16: Numbers of updates under the same measured compute budget in the additional descriptive study. Synthetic regression functions with the same array shapes use the same row. The reported width is the lower hidden width; the upper hidden width is twice this value.

Profile	b	Random	Initial	Two-step	SH	Empirical NTK	Limiting NTK
$w = 64, n = 32$	50	50	45	39	37	38	41
$w = 128, n = 32$	30	30	23	17	15	19	22
$w = 128, n = 32$	50	50	43	37	35	39	41
$w = 128, n = 32$	80	80	73	67	65	67	70
$w = 256, n = 32$	50	50	45	39	37	44	45
$w = 128, n = 64$	50	50	45	39	36	39	43
Diabetes $w = 128, n = 128$	50	50	46	40	38	40	44

Test-loss differences are paired within each candidate group and averaged within each data set or split. Intervals use the normal critical value 1.959964 and the sample standard deviation of the 20 means divided by $\sqrt{20}$. For synthetic settings the independent unit is the data set. For diabetes, the intervals describe variation across random splits and network initializations, conditional on the fixed source data set. All intervals are pointwise 95% normal approximations.

8.2 Results across regression functions and network sizes

Empirical NTK has lower test loss than random, two-step, and successive-halving selection in all 25 settings. It has lower loss than initial-loss selection in 22 settings, with 21 intervals excluding zero in this direction. Limiting NTK has lower loss than empirical NTK in 24 settings, with 14 intervals excluding zero in this direction. Tables 17–21 use the same comparator-minus-empirical convention and columns as the main text. They report the mean, MCSE, and pointwise interval for every comparator within each setting.

Table 17: Additional study: test half-MSE for Eq. (37). Named comparator minus empirical NTK; positive values favor empirical NTK. Intervals are pointwise 95% normal approximations over 20 data-set or conditional split means.

Profile	b	Comparator	Mean	MCSE	95% interval
$w = 64, n = 32$	50	Random	0.01222	0.00109	[0.01008, 0.01436]
$w = 64, n = 32$	50	Initial loss	0.00574	0.00067	[0.00444, 0.00705]
$w = 64, n = 32$	50	Two-step trial	0.00834	0.00064	[0.00709, 0.00960]
$w = 64, n = 32$	50	Successive halving	0.00886	0.00060	[0.00767, 0.01004]
$w = 64, n = 32$	50	Limiting NTK	-0.00035	0.00037	[-0.00108, 0.00038]
$w = 128, n = 32$	30	Random	0.01608	0.00206	[0.01205, 0.02011]
$w = 128, n = 32$	30	Initial loss	-0.00274	0.00064	[-0.00400, -0.00148]
$w = 128, n = 32$	30	Two-step trial	0.00806	0.00062	[0.00686, 0.00927]
$w = 128, n = 32$	30	Successive halving	0.01236	0.00061	[0.01117, 0.01355]
$w = 128, n = 32$	30	Limiting NTK	-0.00549	0.00027	[-0.00601, -0.00497]
$w = 128, n = 32$	50	Random	0.01288	0.00108	[0.01076, 0.01499]
$w = 128, n = 32$	50	Initial loss	0.00828	0.00074	[0.00684, 0.00973]
$w = 128, n = 32$	50	Two-step trial	0.01137	0.00075	[0.00990, 0.01283]
$w = 128, n = 32$	50	Successive halving	0.01196	0.00068	[0.01064, 0.01329]
$w = 128, n = 32$	50	Limiting NTK	0.00004	0.00020	[-0.00035, 0.00043]
$w = 128, n = 32$	80	Random	0.01118	0.00070	[0.00981, 0.01256]
$w = 128, n = 32$	80	Initial loss	0.00994	0.00056	[0.00884, 0.01105]
$w = 128, n = 32$	80	Two-step trial	0.01122	0.00058	[0.01008, 0.01235]
$w = 128, n = 32$	80	Successive halving	0.01120	0.00054	[0.01014, 0.01226]
$w = 128, n = 32$	80	Limiting NTK	-0.00034	0.00019	[-0.00070, 0.00002]
$w = 256, n = 32$	50	Random	0.01525	0.00069	[0.01389, 0.01661]
$w = 256, n = 32$	50	Initial loss	0.00913	0.00060	[0.00794, 0.01031]
$w = 256, n = 32$	50	Two-step trial	0.01182	0.00062	[0.01061, 0.01303]
$w = 256, n = 32$	50	Successive halving	0.01252	0.00068	[0.01118, 0.01385]
$w = 256, n = 32$	50	Limiting NTK	-0.00005	0.00019	[-0.00043, 0.00032]
$w = 128, n = 64$	50	Random	0.01296	0.00068	[0.01163, 0.01430]
$w = 128, n = 64$	50	Initial loss	0.00728	0.00057	[0.00617, 0.00839]
$w = 128, n = 64$	50	Two-step trial	0.01005	0.00056	[0.00895, 0.01114]
$w = 128, n = 64$	50	Successive halving	0.01142	0.00059	[0.01026, 0.01257]
$w = 128, n = 64$	50	Limiting NTK	-0.00201	0.00014	[-0.00229, -0.00173]

Table 18: Additional study: test half-MSE for Eq. (38). Named comparator minus empirical NTK; positive values favor empirical NTK. Intervals are pointwise 95% normal approximations over 20 data-set or conditional split means.

Profile	b	Comparator	Mean	MCSE	95% interval
$w = 64, n = 32$	50	Random	0.01825	0.00085	[0.01658, 0.01992]
$w = 64, n = 32$	50	Initial loss	0.00657	0.00070	[0.00519, 0.00796]
$w = 64, n = 32$	50	Two-step trial	0.00844	0.00071	[0.00705, 0.00983]
$w = 64, n = 32$	50	Successive halving	0.00857	0.00069	[0.00721, 0.00993]
$w = 64, n = 32$	50	Limiting NTK	-0.00074	0.00023	[-0.00120, -0.00029]
$w = 128, n = 32$	30	Random	0.02333	0.00131	[0.02077, 0.02590]
$w = 128, n = 32$	30	Initial loss	0.00185	0.00041	[0.00104, 0.00266]
$w = 128, n = 32$	30	Two-step trial	0.00617	0.00040	[0.00539, 0.00695]
$w = 128, n = 32$	30	Successive halving	0.00756	0.00040	[0.00678, 0.00833]
$w = 128, n = 32$	30	Limiting NTK	-0.00264	0.00022	[-0.00308, -0.00221]
$w = 128, n = 32$	50	Random	0.02119	0.00075	[0.01973, 0.02266]
$w = 128, n = 32$	50	Initial loss	0.00743	0.00062	[0.00621, 0.00865]
$w = 128, n = 32$	50	Two-step trial	0.00924	0.00059	[0.00809, 0.01040]
$w = 128, n = 32$	50	Successive halving	0.00942	0.00059	[0.00826, 0.01058]
$w = 128, n = 32$	50	Limiting NTK	-0.00059	0.00024	[-0.00105, -0.00012]
$w = 128, n = 32$	80	Random	0.01719	0.00065	[0.01591, 0.01847]
$w = 128, n = 32$	80	Initial loss	0.00777	0.00050	[0.00679, 0.00875]
$w = 128, n = 32$	80	Two-step trial	0.00876	0.00046	[0.00785, 0.00967]
$w = 128, n = 32$	80	Successive halving	0.00871	0.00046	[0.00781, 0.00961]
$w = 128, n = 32$	80	Limiting NTK	-0.00021	0.00014	[-0.00049, 0.00006]
$w = 256, n = 32$	50	Random	0.02501	0.00105	[0.02296, 0.02706]
$w = 256, n = 32$	50	Initial loss	0.01083	0.00081	[0.00924, 0.01242]
$w = 256, n = 32$	50	Two-step trial	0.01266	0.00074	[0.01120, 0.01412]
$w = 256, n = 32$	50	Successive halving	0.01237	0.00061	[0.01118, 0.01357]
$w = 256, n = 32$	50	Limiting NTK	-0.00024	0.00014	[-0.00051, 0.00003]
$w = 128, n = 64$	50	Random	0.02017	0.00086	[0.01849, 0.02185]
$w = 128, n = 64$	50	Initial loss	0.00695	0.00053	[0.00591, 0.00798]
$w = 128, n = 64$	50	Two-step trial	0.00914	0.00048	[0.00819, 0.01009]
$w = 128, n = 64$	50	Successive halving	0.00991	0.00053	[0.00887, 0.01095]
$w = 128, n = 64$	50	Limiting NTK	-0.00184	0.00020	[-0.00223, -0.00145]

Table 19: Additional study: test half-MSE for Oscillatory. Named comparator minus empirical NTK; positive values favor empirical NTK. Intervals are pointwise 95% normal approximations over 20 data-set or conditional split means.

Profile	b	Comparator	Mean	MCSE	95% interval
$w = 64, n = 32$	50	Random	0.00881	0.00061	[0.00762, 0.01001]
$w = 64, n = 32$	50	Initial loss	0.00267	0.00048	[0.00173, 0.00362]
$w = 64, n = 32$	50	Two-step trial	0.00363	0.00050	[0.00264, 0.00461]
$w = 64, n = 32$	50	Successive halving	0.00369	0.00046	[0.00279, 0.00459]
$w = 64, n = 32$	50	Limiting NTK	-0.00071	0.00026	[-0.00123, -0.00020]
$w = 128, n = 32$	30	Random	0.01185	0.00124	[0.00941, 0.01428]
$w = 128, n = 32$	30	Initial loss	-0.00027	0.00077	[-0.00178, 0.00124]
$w = 128, n = 32$	30	Two-step trial	0.00264	0.00060	[0.00146, 0.00381]
$w = 128, n = 32$	30	Successive halving	0.00363	0.00060	[0.00245, 0.00482]
$w = 128, n = 32$	30	Limiting NTK	-0.00209	0.00032	[-0.00273, -0.00146]
$w = 128, n = 32$	50	Random	0.00961	0.00084	[0.00796, 0.01126]
$w = 128, n = 32$	50	Initial loss	0.00390	0.00051	[0.00289, 0.00490]
$w = 128, n = 32$	50	Two-step trial	0.00491	0.00045	[0.00403, 0.00578]
$w = 128, n = 32$	50	Successive halving	0.00494	0.00044	[0.00408, 0.00579]
$w = 128, n = 32$	50	Limiting NTK	-0.00036	0.00028	[-0.00091, 0.00020]
$w = 128, n = 32$	80	Random	0.00757	0.00069	[0.00623, 0.00892]
$w = 128, n = 32$	80	Initial loss	0.00429	0.00045	[0.00341, 0.00517]
$w = 128, n = 32$	80	Two-step trial	0.00480	0.00038	[0.00405, 0.00555]
$w = 128, n = 32$	80	Successive halving	0.00469	0.00035	[0.00401, 0.00537]
$w = 128, n = 32$	80	Limiting NTK	-0.00052	0.00019	[-0.00090, -0.00015]
$w = 256, n = 32$	50	Random	0.01076	0.00075	[0.00930, 0.01222]
$w = 256, n = 32$	50	Initial loss	0.00477	0.00051	[0.00376, 0.00578]
$w = 256, n = 32$	50	Two-step trial	0.00566	0.00055	[0.00459, 0.00674]
$w = 256, n = 32$	50	Successive halving	0.00589	0.00054	[0.00483, 0.00695]
$w = 256, n = 32$	50	Limiting NTK	-0.00031	0.00017	[-0.00065, 0.00002]
$w = 128, n = 64$	50	Random	0.01018	0.00060	[0.00901, 0.01135]
$w = 128, n = 64$	50	Initial loss	0.00378	0.00045	[0.00289, 0.00466]
$w = 128, n = 64$	50	Two-step trial	0.00463	0.00037	[0.00390, 0.00536]
$w = 128, n = 64$	50	Successive halving	0.00503	0.00037	[0.00430, 0.00575]
$w = 128, n = 64$	50	Limiting NTK	-0.00109	0.00017	[-0.00142, -0.00076]

Table 20: Additional study: test half-MSE for Radial. Named comparator minus empirical NTK; positive values favor empirical NTK. Intervals are pointwise 95% normal approximations over 20 data-set or conditional split means.

Profile	b	Comparator	Mean	MCSE	95% interval
$w = 64, n = 32$	50	Random	0.00946	0.00056	[0.00837, 0.01055]
$w = 64, n = 32$	50	Initial loss	0.00299	0.00039	[0.00222, 0.00377]
$w = 64, n = 32$	50	Two-step trial	0.00395	0.00042	[0.00313, 0.00477]
$w = 64, n = 32$	50	Successive halving	0.00399	0.00037	[0.00326, 0.00471]
$w = 64, n = 32$	50	Limiting NTK	-0.00062	0.00015	[-0.00091, -0.00032]
$w = 128, n = 32$	30	Random	0.01099	0.00057	[0.00988, 0.01210]
$w = 128, n = 32$	30	Initial loss	0.00009	0.00028	[-0.00046, 0.00065]
$w = 128, n = 32$	30	Two-step trial	0.00292	0.00030	[0.00234, 0.00351]
$w = 128, n = 32$	30	Successive halving	0.00400	0.00030	[0.00341, 0.00458]
$w = 128, n = 32$	30	Limiting NTK	-0.00186	0.00011	[-0.00207, -0.00165]
$w = 128, n = 32$	50	Random	0.00879	0.00043	[0.00794, 0.00964]
$w = 128, n = 32$	50	Initial loss	0.00360	0.00034	[0.00293, 0.00428]
$w = 128, n = 32$	50	Two-step trial	0.00460	0.00035	[0.00392, 0.00527]
$w = 128, n = 32$	50	Successive halving	0.00471	0.00031	[0.00411, 0.00532]
$w = 128, n = 32$	50	Limiting NTK	-0.00022	0.00013	[-0.00047, 0.00003]
$w = 128, n = 32$	80	Random	0.00741	0.00039	[0.00665, 0.00816]
$w = 128, n = 32$	80	Initial loss	0.00415	0.00033	[0.00351, 0.00479]
$w = 128, n = 32$	80	Two-step trial	0.00463	0.00033	[0.00399, 0.00527]
$w = 128, n = 32$	80	Successive halving	0.00456	0.00031	[0.00396, 0.00516]
$w = 128, n = 32$	80	Limiting NTK	-0.00009	0.00009	[-0.00026, 0.00008]
$w = 256, n = 32$	50	Random	0.01034	0.00043	[0.00949, 0.01119]
$w = 256, n = 32$	50	Initial loss	0.00463	0.00035	[0.00395, 0.00531]
$w = 256, n = 32$	50	Two-step trial	0.00553	0.00033	[0.00488, 0.00618]
$w = 256, n = 32$	50	Successive halving	0.00559	0.00033	[0.00494, 0.00623]
$w = 256, n = 32$	50	Limiting NTK	-0.00003	0.00011	[-0.00025, 0.00020]
$w = 128, n = 64$	50	Random	0.00907	0.00040	[0.00829, 0.00985]
$w = 128, n = 64$	50	Initial loss	0.00350	0.00037	[0.00278, 0.00423]
$w = 128, n = 64$	50	Two-step trial	0.00442	0.00037	[0.00368, 0.00515]
$w = 128, n = 64$	50	Successive halving	0.00480	0.00039	[0.00403, 0.00556]
$w = 128, n = 64$	50	Limiting NTK	-0.00087	0.00012	[-0.00111, -0.00063]

Table 21: Additional study: test half-MSE for Diabetes. Named comparator minus empirical NTK; positive values favor empirical NTK. Intervals are pointwise 95% normal approximations over 20 data-set or conditional split means.

Profile	b	Comparator	Mean	MCSE	95% interval
$w = 128, n = 128$	50	Random	0.01168	0.00110	[0.00953, 0.01384]
$w = 128, n = 128$	50	Initial loss	-0.00011	0.00107	[-0.00221, 0.00200]
$w = 128, n = 128$	50	Two-step trial	0.00807	0.00117	[0.00577, 0.01037]
$w = 128, n = 128$	50	Successive halving	0.01110	0.00116	[0.00884, 0.01337]
$w = 128, n = 128$	50	Limiting NTK	-0.00419	0.00064	[-0.00544, -0.00293]

8.3 DECOMPOSITION OF THE FINAL LOSS DIFFERENCE

We use the following exact identity to describe how candidate selection and subsequent updates contribute to the final loss difference. Let $\widehat{m}_E$ and $\widehat{m}_L$ be the candidates selected by the candidate-specific empirical NTK and the limiting NTK, respectively, and let $k_E < k_L$ be their assigned numbers of updates. The candidates are selected at their respective prediction horizons k_E and k_L. For actual training or test half-MSE $L_{k,m}$ at update k, define Δ as the final loss of the limiting-NTK method minus that of the candidate-specific method. Let A_{sel} compare the two selected candidates after k_E updates, and let P be the signed reduction in the limiting-NTK method's selected candidate between updates k_E and k_L. For every candidate group,

$$\begin{aligned}\Delta &:= L_{k_L,\widehat{m}_L} - L_{k_E,\widehat{m}_E} \\ &= \underbrace{L_{k_E,\widehat{m}_L} - L_{k_E,\widehat{m}_E}}_{A_{\text{sel}}\text{: difference after } k_E \text{ updates}} - \underbrace{\left(L_{k_E,\widehat{m}_L} - L_{k_L,\widehat{m}_L}\right)}_{P\text{: change during the additional updates}} .\end{aligned}$$

Positive Δ favors the candidate-specific empirical NTK, and negative Δ favors the limiting NTK. The empirical-NTK method therefore has lower mean loss exactly when $\mathbb{E}A_{\text{sel}} > \mathbb{E}P$. For actual finite-width training or test losses, A_{sel} and P can each have either sign. Thus A_{sel} compares the complete selection rules, jointly reflecting their kernels and prediction horizons. We formed this descriptive decomposition after examining the results.

The following summaries use test half-MSE. The two methods select the same candidate in 72.9% to 92.5% of groups across settings. The common-k_E term A_{sel} is positive in 23 of 25 settings, with 15 normal-approximation intervals excluding zero in that direction. Its unweighted mean is 0.00073. The term $-P$ is negative in 25 settings, with 25 intervals excluding zero, and its unweighted mean is -0.00182. At the common horizon k_L, the candidate chosen by empirical NTK has lower test loss in 23 settings, with 12 intervals excluding zero. These descriptive components use the same trajectories and data-set or conditional split units as the preceding tables.