LEARNING DECISION-STUMP THRESHOLDS IN CONTEXT: DYNAMICS OF SOFTMAX ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating a decision threshold requires locating observations near an unknown boundary. We study how gradient-based pretraining learns this statistical rule in a two-parameter softmax-attention model with a fixed feature and inequality direction. Pretraining uses labeled contexts and their true thresholds; a fresh threshold must be inferred from context alone. Under a large-resolution initialization, constant-step gradient descent on m tasks with n examples each produces a frozen estimator with error $\widetilde{O}((m\wedge n)^{-1}+N^{-1})$ for each fixed interior threshold and every fresh-context size N. The two terms separate finite-pretraining accuracy from fresh-context localization. The mechanism is coordinated parameter divergence: population training calibrates the relative label and feature scores, then increases the attention scale as $t^{1/4}$, giving population threshold error $O(t^{-1/4})$. To transfer this mechanism to a fixed finite corpus, we control gradient errors relative to the shrinking directions of progress at successive parameter scales. This certifies a growing training interval without requiring long-time tracking of the population trajectory. We also identify the boundary limitation of the one-head model and explain statistically what a reflected symmetrization could achieve.

1 INTRODUCTION

In-context learning (ICL) is the ability of a pretrained model to perform a new task from examples supplied in its input, without updating its parameters (Brown et al., 2020). Garg et al. (2022) cast ICL as a statistical learning problem: pretraining across many datasets produces a reusable prediction *procedure*, which is applied to a new dataset through the context alone. In their formulation, as in much of the subsequent theory, each task is a supervised-prediction problem: the context contains labeled examples, and the model predicts the response at a new query point. A growing body of work explains how gradient-based pretraining acquires such procedures, characterizing convergence, phases of learning, and the evolution of the in-context predictor for linear regression and, more recently, induction and sequence tasks (Zhang et al., 2024; Chen et al., 2024a; Zhang et al., 2025c; Chen et al., 2024b; Musat et al., 2025).

The same template applies to estimating a parameter of the data-generating process. In *amortized estimation*, one repeatedly draws a parameter from a prior and a dataset given that parameter, trains a single network to map datasets to parameter estimates, and applies the frozen network to new datasets (Zammit-Mangion et al., 2025). Under squared loss, the population objective is the Bayes risk, so the network approximates the posterior mean. Neural Bayes estimators use permutation-invariant architectures for this purpose (Sainsbury-Dale et al., 2024); prior-fitted networks apply the same idea to predictive distributions and underlie tabular foundation models (Müller et al., 2025; Hollmann et al., 2025); and recent work estimates dataset-level quantities such as mutual information with permutation-invariant neural models (Peyrard & Cho, 2025), or uses TabPFN for semi-supervised parameter estimation (Zhang et al., 2025a). A trained amortized estimator is an in-context estimator: it infers a latent parameter from its input without parameter updates.

Recent theory for amortized estimators primarily studies the statistical properties of the trained estimator rather than its optimization dynamics. Rödder et al. (2025) decompose neural-estimator risk into Bayes, approximation, generalization, and algorithm-dependent terms, while leaving the algorithm-dependent term outside their analysis; related work studies the statistical accuracy of neural posterior and likelihood estimators (Frazier et al., 2024). The ICL training-dynamics works

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.

discussed above instead concern supervised prediction. *How gradient-based pretraining constructs an amortized estimator, and how the number and size of the simulated datasets limit its accuracy, have, to our knowledge, not been analyzed.*

We study this question for decision thresholds. Thresholds define the splits of decision trees, which are used both as interpretable models and as building blocks of accurate tabular ensembles (Tan et al., 2022; Erickson et al., 2025). Tabular foundation models (TFMs) are now competitive with tree ensembles on standard tabular benchmarks, particularly on smaller datasets (Erickson et al., 2025; Hollmann et al., 2025), yet their synthetic pretraining priors themselves contain decision trees and tree ensembles (Hollmann et al., 2025; Qu et al., 2025; Zhang et al., 2025b). This motivates isolating computations that locate boundaries induced by latent splits. We isolate one such computation: estimating a single threshold when the split feature and inequality direction are known.

Specifically, we consider the following model. Given a threshold $\alpha \in (0,1)$, a task supplies a context $D_{N,\alpha} = \{(X_i, Y_i)\}_{i=1}^N$ with N independent features $X_i \sim \mathrm{Unif}(0,1)$ and labels $Y_i = \mathbf{1}\{X_i \le \alpha\}$. Write $M_N^+ = \max\{X_i : Y_i = 1\}$, with value zero if no label is positive. For $0 \le \varepsilon < \alpha$, $\mathbb{P}(\alpha - M_N^+ > \varepsilon) = (1-\varepsilon)^N \le e^{-N\varepsilon}$, and hence, $N(\alpha - M_N^+) \Rightarrow \mathrm{Exp}(1)$. Thus, the estimator M_N^+ localizes the threshold at the nonregular N^{-1} scale, faster than the $N^{-1/2}$ scale typical of regular smooth parametric models (Hansen, 2000). Because the benchmark estimator is explicit, the setting lets us compare what gradient training learns against a known statistical rule. We approximate this estimator with a single softmax-attention head:

$$F_N(\lambda, \mu; D_{N,\alpha}) = \frac{\sum_{i=1}^N X_i e^{\lambda Y_i + \mu X_i}}{\sum_{i=1}^N e^{\lambda Y_i + \mu X_i}}. \tag{1}$$

The parameters λ and μ favor positive labels and larger feature values, respectively, and will be trained by gradient descent on simulated training tasks. Once the relative scale λ/μ is calibrated, increasing their common scale concentrates attention on the largest positive observation.

Implementing a sharp statistical rule with smooth attention requires increasing parameter scale. In the infinite-context limit, within-class concentration leaves $O(\mu^{-1})$ threshold bias; in finite contexts, the output must also approximate M_N^+. Thus, improving resolution requires parameter growth even as the gradient driving it weakens, creating a tension between statistical accuracy and optimization with finite pretraining data. In our model, it leads to two concrete questions: (1) Does gradient descent coordinate the relative scores while continuing to increase the attention scale? (2) As the useful gradient signal vanishes, how far can this sharpening process remain reliable when the population gradient is replaced by an empirical one computed from only m tasks with n observations each?

Our contributions answer both questions:

- **A learned amortized estimator with an explicit budget.** Constant-step gradient descent on m simulated tasks with n observations each learns a frozen estimator with error $\widetilde{O}((m \wedge n)^{-1} + N^{-1})$ for every fresh-context size N and each fixed threshold separated from the lower endpoint of the task range (Theorem 3.1). The first error term is the finite-pretraining error, and the second is the statistical error of the fresh context. The bound attains the N^{-1} benchmark, up to logarithmic factors, when $m \wedge n = \widetilde{\Omega}(N)$; in particular, the certified accuracy is limited by the simulated context size n, however large the fresh context.
- **Balance, then sharpen.** Population gradient flow first moves the ratio λ/μ to a boundary-aligned value, then increases μ as $t^{1/4}$. The estimator converges uniformly at rate $O(t^{-1/4})$ while both parameters λ and μ diverge (Theorem 5.2). The radial drift $\dot{\mu}(t)$ scales as the product of the μ^{-1} bias and the μ^{-2} tilted variance, which explains why sharpening never stops but steadily slows. This mechanism remains stable under discretization, and therefore transfers to population gradient descent with sufficiently small step size.
- **Finite-corpus control along a diverging trajectory.** Normalizing gradient errors by the shrinking progress margins at each parameter scale certifies continued sharpening for empirical gradient descent up to scale $(m \wedge n)/\log(mn)$ (Corollary 5.3). The argument follows the accuracy-relevant properties while avoiding full population-trajectory tracking over the growing horizon.

The paper is structured as follows. The model is precisely introduced in Section 2. Next, we describe our main results: the statistical results of the pretrained model are presented in Section 3, and Sections 4 and 5 establish how it is learned by gradient training dynamics.

Numerical experiments are provided in Section 6 to test the predictions of the theory. A further discussion of related works is given in Section 7, before we conclude with a discussion of extensions in Section 8. The technical details and proofs are deferred to the appendices.

2 Problem and attention estimator

Tasks and training data. Fix $\rho \in (0, 1/2)$, and write $\ell = 1 - \rho$ and $\Delta = 1 - 2\rho$. Each pretraining task draws a threshold $\alpha \sim \mathrm{Unif}([\rho, \ell])$, followed by n independent features $X_i \sim \mathrm{Unif}(0, 1)$ with corresponding labels $Y_i = \mathbf{1}\{X_i \le \alpha\}$. We observe m independent blocks $Z_j = (\alpha_j, D_{n,\alpha_j})$. A fresh task supplies an independent context of size N, but its threshold is withheld. Thus m counts pretraining tasks, n counts observations per pretraining task, and N is the fresh-context size.

Statistical benchmark. Let $M_N^+ = \max\{X_i : Y_i = 1\}$, with value zero if no label is positive. For $0 \le \varepsilon < \alpha$, $\mathbb{P}(\alpha - M_N^+ > \varepsilon) = (1-\varepsilon)^N \le e^{-N\varepsilon}$. The natural localization scale is therefore N^{-1}. This rule can be specified without pretraining; our question is whether the stated gradient procedure learns an attention estimator with comparable fresh-context accuracy. Section 8.1 compares the resulting rate with the Bayes risk, and also discusses a symmetrized estimator.

One-head architecture. Introduce context tokens $z_i = (X_i, Y_i, 0)^\top$ and a readout token $z_{\mathrm{r}} = (0, 0, 1)^\top$. A self-attention head with $Q = (0, 0, 1)$, $K_\theta = (\mu, \lambda, 0)$, $V = (1, 0, 0)$, and $O = 1$ computes equation 1 when the readout attends only to context tokens. Only $\theta = (\lambda, \mu)$ is trained. Equivalently, writing $\mathbf{D}_{n,\alpha} = [(X_i, Y_i)^\top]_{i=1}^n$,

$$F_n(\theta; D_{n,\alpha}) = \mathrm{softmax}([\mu\ \lambda]\mathbf{D}_{n,\alpha})\, \mathbf{D}_{n,\alpha}^\top (1, 0)^\top. \tag{2}$$

For $\mu > 0$, the ratio $k = \lambda/\mu$ measures the relative preference for labels and features, and μ measures concentration.

Objective and optimization. We use the empirical squared-error objective and its infinite-task, infinite-context counterpart, given by, respectively,

$$\widehat{\mathcal{R}}_{m,n}(\theta) = \frac{1}{2m}\sum_{j=1}^{m}(F_n(\theta; D_{n,\alpha_j}) - \alpha_j)^2, \qquad \mathcal{R}_\rho(\theta) = \frac{1}{2}\mathbb{E}_\alpha(F(\theta; \alpha) - \alpha)^2, \tag{3}$$

where F is the infinite-context limit of F_n. Empirical training is ordinary full-batch gradient descent with a constant step size $\gamma > 0$:

$$\widehat{\theta}_{s+1} = \widehat{\theta}_s - \gamma \nabla \widehat{\mathcal{R}}_{m,n}(\widehat{\theta}_s). \tag{4}$$

In our analysis, population gradient flow provides the reference dynamics.

Assumption 2.1 (Large-resolution initialization)**.** *Fix $\kappa > 0$. The deterministic initialization satisfies $\mu_0 \ge \mu_{\mathrm{init}}$ and $k_0 = \lambda_0/\mu_0 \ge \kappa$, where μ_{init} depends only on ρ, κ, and fixed model constants. It is independent of the pretraining corpus and is shared by population and empirical training.*

This assumption requires a large initial concentration while allowing the initial score ratio to be inaccurate. Training must still learn the appropriate balance of scores.

3 Learned estimation and transfer

We begin by stating the statistical outcome of the pretrained model. The next sections will explain which parameters give an accurate estimator and how gradient training reaches them. Define $A_{m,n} := (m \wedge n)/\log(C_{\mathrm{stat}} mn)$, where $C_{\mathrm{stat}} > 1$ is a fixed constant. All constants below may depend on the fixed initialization, model constants, and explicitly subscripted fixed quantities such as η, but not on m, n, N.

Theorem 3.1 (Learned threshold estimation)**.** *Suppose Assumption 2.1 holds, and fix $\eta \in (0, \Delta)$. For every sufficiently small fixed step size $\gamma > 0$ and all sufficiently large $A_{m,n}$, there exist a*

nonempty deterministic interval of gradient-descent iteration indices $\mathcal{I}_{m,n,\gamma}$ with both endpoints are of order $\Theta(A^4_{m,n}/\gamma)$, and a pretraining event $\mathcal{E}_{\mathrm{train}}$ with $\mathbb{P}_{\mathrm{pre}}(\mathcal{E}_{\mathrm{train}}) \geq 1 - Ce^{-cm} - C(mn)^{-c}$. Select any $\mathcal{F}_{\mathrm{pre}}$-measurable $\widehat{s} \in \mathcal{I}_{m,n,\gamma}$ and freeze $\widehat{\theta} := \widehat{\theta}^{\gamma}_{\widehat{s}}$. On $\mathcal{E}_{\mathrm{train}}$, for each fixed $\alpha^{\mathrm{fr}} \in [\rho+\eta, \ell]$, every $N \geq 1$, and every $u \geq 1$,

$$\mathbb{P}_{\mathrm{fr}}\left(|F_N(\widehat{\theta}; D_N^{\mathrm{fr}}) - \alpha^{\mathrm{fr}}| \leq C_\eta \left[\frac{1 + \log(1 + CA_{m,n}u)}{A_{m,n}} + \frac{u}{N} \right] \middle| \mathcal{F}_{\mathrm{pre}} \right) \geq 1 - 4e^{-u}, \tag{5}$$

where $\mathcal{F}_{\mathrm{pre}}$ is the σ-field generated by the pretraining corpus; thus $\widehat{s}$ is selected before the fresh threshold, context size, confidence level, or observations are specified.

The first term measures the finite attention resolution certified by pretraining, and the second measures localization from a fresh context. Removing the conditioning gives total failure probability at most $Ce^{-cm} + C(mn)^{-c} + 4e^{-u}$. In particular, the error is $\widetilde{O}((m \wedge n)^{-1} + N^{-1})$, with high probability. Note that the same trained parameters can be applied to any fresh task, and the probability guarantee is pointwise in both α^{fr} and N.

Theorem 3.1 is the statistical consequence of the detailed continuation and transfer statements, Theorems A.1 and A.2, in Appendix A, where the iteration interval is defined precisely. Its proof is assembled in Appendix A.4. We remark that the constants that appear are theoretical; the result does not supply a calibrated practical stopping rule or assert that accuracy deteriorates beyond the certified interval.

Fresh-query classification. For an independent $X_q \sim \mathrm{Unif}(0,1)$, predict $\widehat{Y}_q = \mathbf{1}\{X_q \leq F_N\}$. Since $F_N \in [0,1]$, conditional misclassification probability is exactly $|F_N - \alpha^{\mathrm{fr}}|$. The estimation guarantee therefore gives the same rate for fresh-query classification; see Corollary A.3.

4 From parameters to statistical accuracy

Two parameter properties suffice for the accuracy of the estimator: the positive class must receive enough weight, and attention within that class must concentrate near the threshold. We first quantify these properties without assuming that the parameters were learned.

Class separation and concentration. Let $p_\theta(\alpha)$ be the population attention mass on $[0, \alpha]$, and let $q_\theta(\alpha) := 1 - p_\theta(\alpha)$. Exact integration gives

$$\frac{p_\theta(\alpha)}{q_\theta(\alpha)} = e^{\mu(k+\alpha-1)} \frac{1 - e^{-\mu\alpha}}{1 - e^{-\mu(1-\alpha)}}, \qquad F(\theta;\alpha) - \alpha = q_\theta(\alpha)(1-\alpha) - \mu^{-1} + O(e^{-\rho\mu}). \tag{6}$$

The error has two sources: mass assigned to the negative class and finite concentration within the positive class. The first is controlled by k and the second by μ. Across tasks, the two classes exchange dominance near $\alpha = 1 - k$; this is a transition in task space, not the unknown threshold within a particular context.

A sufficient progress region. For suitable fixed c_0, $C_{\mathrm{band}} > 1$, and sufficiently large μ_*, define

$$\mathcal{G} := \left\{ (\lambda, \mu) : \mu \geq \mu_*, \quad \ell + \frac{\log\mu + c_0}{\mu} \leq \frac{\lambda}{\mu} \leq \ell + C_{\mathrm{band}} \frac{\log\mu}{\mu} \right\}. \tag{7}$$

Throughout this region, negative-class mass is small enough that $\sup_{\alpha \in [\rho,\ell]} |F(\theta;\alpha) - \alpha| \leq C/\mu$. The logarithmic correction supplies sufficient class separation even at the lower endpoint ρ. The upper boundary facilitates the training analysis, but accurate prediction may also occur above it.

Proposition 4.1 (Accuracy of a fixed attention estimator)**.** *Fix $\eta \in (0, \Delta)$. For $\theta \in \mathcal{G}$ with sufficiently large μ, each fixed $\alpha \in [\rho + \eta, \ell]$, every $N \geq 1$, and every $u \geq 1$,*

$$\mathbb{P}_{\mathrm{fr}}\left(|F_N(\theta; D_{N,\alpha}) - \alpha| > C_\eta \left[\frac{1 + \log(1 + C_{\mathrm{fr}}\mu u)}{\mu} + \frac{u}{N} \right] \right) \leq 4e^{-u}. \tag{8}$$

Appendix G.1.2 proves Proposition 4.1. For shorter context sizes N, the proof compares softmax pooling with the largest positive observation. For N large enough to resolve the concentration scale,

it controls the weighted empirical average directly around its population limit. Combining the two arguments avoids a logarithmic dependence on N in the pretraining term.

Proposition 4.1 is the interface between statistics and optimization. Thus, to obtain Theorem 3.1, we show that training produces parameters in $\mathcal{G}$ with $\mu \asymp A_{m,n}$. Once a checkpoint is frozen, conditioning on its pretraining corpus makes the fixed-parameter bound in this result applicable.

5 From pretraining to parameter guarantees

Our proof separates deterministic optimization geometry from finite-corpus perturbations. We first use population gradient flow to identify a scale-dependent funnel in which the concentration parameter increases and the induced estimator becomes more accurate. Then, we transfer this geometry to population gradient descent and use scale-adapted empirical-gradient estimates to preserve the radial progress, inward boundary motion, and step-length control needed by empirical gradient descent. This avoids pointwise tracking of the empirical trajectory over a training horizon that grows with the pretraining sample size.

5.1 Population capture and increasing concentration

We study population gradient flow $\dot{\theta}(t) = -\nabla \mathcal{R}_\rho(\theta(t))$. For population gradient descent with fixed step size $\gamma > 0$, write $\theta_{s+1}^\gamma = \theta_s^\gamma - \gamma \nabla \mathcal{R}_\rho(\theta_s^\gamma)$ with initialization $\theta_0^\gamma = \theta_0$, and let μ_s^γ denote its second coordinate. For empirical gradient descent, write $\widehat{\theta}_{s+1}^\gamma = \widehat{\theta}_s^\gamma - \gamma \nabla \widehat{\mathcal{R}}_{m,n}(\widehat{\theta}_s^\gamma)$ with $\widehat{\theta}_0^\gamma = \theta_0$, and let $\widehat{\mu}_s^\gamma$ denote its second coordinate.

Population training must first learn the balance of scores and then increase their scale. The gradient identities make this coupling explicit. Here Cov_θ and Var_θ refer to the distribution tilted by the attention weights.

Lemma 5.1 (Gradient identities)**.** *For every $\theta \in \mathbb{R}^2$ and $\alpha \in [\rho, \ell]$, $\partial_\lambda F(\theta;\alpha) = \mathrm{Cov}_\theta(X, Y \mid \alpha)$ and $\partial_\mu F(\theta;\alpha) = \mathrm{Var}_\theta(X \mid \alpha)$. Consequently, $\nabla \mathcal{R}_\rho(\theta) = \frac{1}{\Delta}\int_\rho^\ell \{F(\theta;\alpha) - \alpha\}\big(\mathrm{Cov}_\theta(X, Y \mid \alpha), \mathrm{Var}_\theta(X \mid \alpha)\big)^\top d\alpha$.*

Theorem 5.2 (Population capture and sharpening)**.** *Under Assumption 2.1, population gradient flow enters $\mathcal{G}$ in finite time and remains there. Writing $\tau_{\mathcal{G}} := \inf\{t \geq 0 : \theta(t) \in \mathcal{G}\}$, there exists $c_{\mathrm{rad}} \in (0,1)$ such that $\inf_{0 \leq t \leq \tau_{\mathcal{G}}} \mu(t) \geq c_{\mathrm{rad}} \mu(0)$. Moreover, there exist constants $0 < c < C < \infty$ such that, for every $t \geq \tau_{\mathcal{G}}$,*

$$\frac{c}{\mu(t)^3} \leq \dot{\mu}(t) \leq \frac{C}{\mu(t)^3}, \qquad \sup_{\alpha \in [\rho,\ell]} |F(\theta(t);\alpha) - \alpha| \leq \frac{C}{\mu(t)}, \qquad \mathcal{R}_\rho(\theta(t)) \leq \frac{C}{\mu(t)^2}. \tag{9}$$

Writing $\mu_{\mathrm{ent}} := \mu(\tau_{\mathcal{G}})$, it follows that

$$\mu_{\mathrm{ent}}^4 + c(t - \tau_{\mathcal{G}}) \leq \mu(t)^4 \leq \mu_{\mathrm{ent}}^4 + C(t - \tau_{\mathcal{G}}). \tag{10}$$

Finally,

$$\frac{\lambda(t)}{\mu(t)} = \ell + O\left(\frac{\log \mu(t)}{\mu(t)}\right). \tag{11}$$

Hence, $\lambda(t), \mu(t) \to \infty$, while the induced estimator converges uniformly:

$$\sup_{\alpha \in [\rho,\ell]} |F(\theta(t);\alpha) - \alpha| = O(t^{-1/4}), \qquad \mathcal{R}_\rho(\theta(t)) = O(t^{-1/2}). \tag{12}$$

This theorem shows that population training reaches the progress region $\mathcal{G}$ required by Proposition 4.1. The concentration parameter may decrease before capture, but it can only shrink by at most a fixed multiplicative factor beforehand. Once this region is reached, the radial drift $\dot{\mu}(t)$ becomes positive and optimization acts primarily by sharpening the learned softmax rule. Appendix C gives the proof.

This mechanism is stable under discretization. For every sufficiently small fixed step size γ, population gradient descent reaches the interior of the same scale-dependent funnel at a deterministic

robust-entry index $s_0 = s_0(\gamma)$, and thereafter satisfies

$$(\mu^\gamma_{s_0})^4 + c\gamma(s - s_0) \le (\mu^\gamma_s)^4 \le (\mu^\gamma_{s_0})^4 + C\gamma(s - s_0),$$
$$\sup_{\alpha\in[\rho,\ell]} |F(\theta^\gamma_s;\alpha) - \alpha| \le \frac{C}{\{(\mu^\gamma_{s_0})^4 + c\gamma(s - s_0)\}^{1/4}}. \tag{13}$$

Before robust entry, the discrete trajectory shadows the flow for only a fixed physical time. After entry, a one-step barrier argument replaces pointwise trajectory tracking. This post-entry formulation is the one that survives finite pretraining. Proposition D.8 in Appendix D proves equation 13.

5.2 Finite pretraining and the shrinking gradient signal

Population training can increase μ indefinitely. A finite corpus perturbs the gradient, and the permissible perturbation becomes smaller as the population gradient weakens. In $\mathcal{G}$, radial progress has size μ^{-3} and the inward velocities at the two boundaries have size at least a constant times μ^{-4}. Determining whether these directions remain reliable therefore requires relative gradient control.

On a parameter range $\mu \asymp R$, we therefore normalize the radial error by μ^3 and the boundary-normal errors by μ^4. Let $\mathcal{E}(R)$ denote the maximum of these normalized errors and the normalized speed error (defined precisely in Appendix E, equation 120). The empirical-process argument gives

$$\mathcal{E}(R) \le C\left\{\sqrt{\frac{RL}{m}} + \frac{RL}{m} + \frac{R}{n}\right\}, \qquad L = \log\frac{C(1+R)mn}{\delta}, \tag{14}$$

with probability at least $1-\delta$, provided $n \ge CRL$. The terms distinguish sampling across tasks from finite-context bias. Their scale dependence comes from moment bounds adapted to concentrated attention; uniform bounds on individual task gradients would obscure this dependence.

Keeping equation 14 below the fixed normalized progress margins permits R up to a sufficiently small constant times $A_{m,n}$. This yields the following result:

Corollary 5.3 (Certified concentration from finite pretraining)**.** *Under Assumption 2.1, for sufficiently small fixed γ and sufficiently large $A_{m,n}$, with probability at least $1 - Ce^{-cm} - C(mn)^{-c}$, empirical gradient descent enters $\mathcal{G}$ and remains there throughout a deterministic training interval. Every checkpoint in its late subinterval satisfies*

$$c_A A_{m,n} \le \widehat{\mu} \le C_A A_{m,n}, \qquad \sup_{\alpha\in[\rho,\ell]} |F(\widehat{\theta};\alpha) - \alpha| \le C/A_{m,n}.$$

The endpoints of the late subinterval occur after $\Theta(A^4_{m,n}/\gamma)$ post-entry gradient steps.

This is a summary of Theorem A.1, with details given in Appendix A.5.2.

Why continuation does not require trajectory tracking. Before entry, a compact approximation argument compares empirical and population training over a fixed time interval. After entry, the proof uses overlapping parameter ranges. On each range, the normalized estimates preserve positive radial drift, inward boundary motion, and a bound on the step length. A one-step barrier argument prevents exit from the useful region, including on the step entering the next range. Intersecting the events for all deterministic ranges makes this argument valid for the data-dependent trajectory on the same reused corpus.

Thus the proof follows the properties sufficient for accurate prediction over the growing time horizon. The abstract formulation appears in Appendix A. The certified upper scale only marks the reach of the present estimates, but the achievable accuracy may improve with further empirical training.

6 Experiments

We examine the accuracy of a frozen estimator and the dynamics producing it. The transfer experiment uses one checkpoint per pretraining corpus across all fresh-context sizes. Population experiments examine capture and increasing concentration. Appendix H gives reproducibility details, initialization checks, and exploratory departures from the model assumptions.

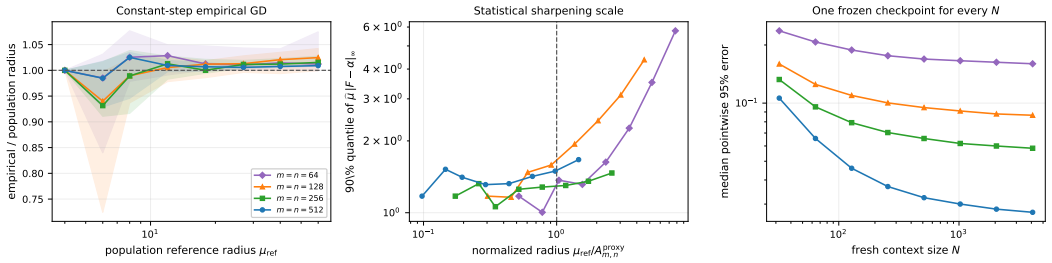

Figure 1: Finite pretraining followed by frozen transfer. The first two panels test the radius and prediction-error controls of Corollary 5.3; the right panel tests the two-term error prediction of Theorem 3.1 using one checkpoint per pretraining corpus, frozen and reused for every N. The proxy indicates the order of the theoretical stopping scale; its comparison constants are not calibrated.

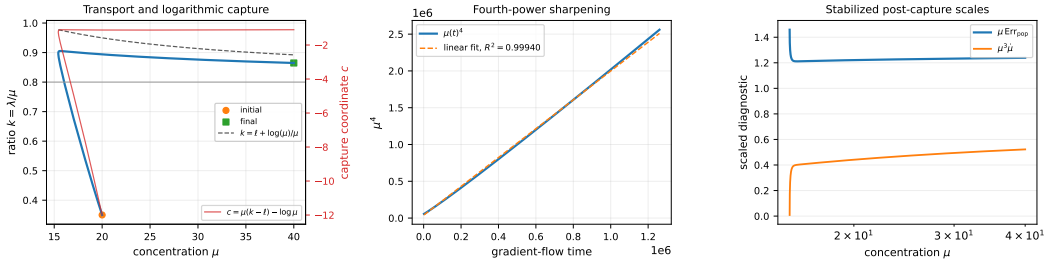

Figure 2: Exact one-head population dynamics. The ratio and rescaled endpoint coordinate $c = \mu(k - \ell) - \log\mu$ display capture into the logarithmic tube (left), μ^4 is nearly affine after entry (center), and both $\mu \sup_\alpha |F - \alpha|$ and $\mu^3\dot{\mu}$ stabilize at constant scale (right).

Finite pretraining and frozen transfer.

We use two complementary corpus sweeps to probe the predictions of Corollary 5.3 and Theorem 3.1. Let μ_s^{pop} be the deterministic population-GD reference path and set $A^{\mathrm{proxy}}_{m,n} := (m \wedge n)/\log(mn)$. The theory-guided stopping diagnostic fixes $s_q := \min\{s : \mu_s^{\mathrm{pop}} \geq qA^{\mathrm{proxy}}_{m,n}\}$ for $q \in \{0.5, 0.75, 1\}$ before any corpus is sampled. For each of seven balanced and axis-swapped configurations, we run full-batch empirical GD with $\gamma = 400$ on 30 corpora. From the shared post-entry start through s_1, all 210 empirical radii remain within $[0.7, 1.3]$ times their population references, and 209 of 210 runs have population prediction error at most $3/A^{\mathrm{proxy}}_{m,n}$ at every s_q; full results appear in Appendix H. These predetermined checkpoints provide a controlled diagnostic of the radius and prediction-error bounds in Corollary 5.3. The theoretical step-size threshold and comparison constants remain existential, so practical stopping-rule calibration and failure-boundary estimation lie outside the scope of this experiment.

For transfer, on the fixed reference-radius grid, we choose the largest $r \leq A^{\mathrm{proxy}}_{m,n}$ and freeze each empirical iterate at the precomputed iteration index $\min\{s : \mu_s^{\mathrm{pop}} \geq r\}$. The grid and population path are fixed in advance, so checkpoint selection precedes observation of the empirical corpus. We reuse the resulting parameters for every $N \in \{32, 64, \ldots, 4096\}$ and test the two additive terms in equation 5. Figure 1 reports, for each corpus, the maximum over the threshold grid of the 95th-percentile absolute error, followed by the median of this quantity across corpora. The curves decay to a pretraining-limited floor. At $N = 4096$, the error falls from 0.160 to 0.0272 as balanced pretraining grows from $m = n = 64$ to $m = n = 512$, with a log–log slope of -1.02 against $A^{\mathrm{proxy}}_{m,n}$. The supplementary axis swaps give nearly identical errors. These results are consistent with Theorem 3.1 at the prescribed checkpoints.

Population dynamics and finite-step robustness. Theorem 5.2 predicts capture into a logarithmic tube, $\mu(t)^4 = \Theta(t)$ after capture, and uniform error $O(\mu(t)^{-1})$. We set $\rho = 0.2$ and integrate the exact population vector field from $(k_0, \mu_0) = (0.35, 20)$. Figure 2 follows the complete transient: the radius first falls from 20 to 15.42 as the ratio crosses the task range, then grows in the endpoint

layer. At $\mu = 40$, the rescaled endpoint coordinate $\mu(k-\ell)-\log\mu$ is -1.10, while $\mu^3\dot{\mu} = 0.522$ and $\mu \sup_\alpha |F(\theta;\alpha) - \alpha| = 1.239$. After capture, a linear fit of μ^4 against flow time has $R^2 = 0.9994$, and the tested error is at most $1.51/\mu(t)$. For the discrete-time prediction equation 13, we initialize population GD in the post-entry regime and test $\gamma \in \{50, 100, 200, 400\}$. Over the plotted range $\mu \geq 14$, the relative radial increment is below 0.4%, and $(\mu_{s+1}^4 - \mu_s^4)/\gamma$ is nearly independent of γ. For additional details, see Figure 3 in Appendix H.

7 Related Work

Algorithmic and statistical accounts of ICL. Prior work connects transformer predictions to gradient-based and closed-form regression algorithms (Akyürek et al., 2023; Von Oswald et al., 2023), establishes algorithm selection and generalization for constrained pretraining ERM (Bai et al., 2023), interprets ICL as latent-concept inference (Xie et al., 2022), and constructs attention-based solvers for posterior predictive quantities (Kang et al., 2026). These characterize computations or statistical solutions; gradient training of transformer weights requires separate analysis.

Learned statistical estimators and tabular priors. Transformer-based statistical ICL covers posterior predictive distributions, density estimation, and parameter estimation (Kang et al., 2026; Hollmann et al., 2025; Zhang et al., 2025a). Meta-statistical learning similarly treats an entire dataset as input to a permutation-invariant network trained across datasets (Peyrard & Cho, 2025). We analyze the optimization and finite-corpus transfer of such an estimator for a threshold. This choice reflects tree-based generators in TFM priors: TabICL includes XGBoost tasks, while MITRA combines tree and ensemble priors with structural causal models (Qu et al., 2025; Zhang et al., 2025b). Full query-label objectives and multidimensional structure remain outside our scope.

Training dynamics and invariant structures. Convergence analyses use Polyak–Łojasiewicz (PL) or related gradient-dominance arguments for linear, looped, and representation-based models (Zhang et al., 2024; Gatmiry et al., 2024; Yang et al., 2024), and also give classification guarantees (Shen et al., 2025). Trajectory analyses track competing attention weights, spectral dynamics, and learning phases (Huang et al., 2024; Chen et al., 2024a; Zhang et al., 2025c; He et al., 2025); finite-step and mini-batch effects can create moving stability boundaries in linear self-attention (Balasubramanian, 2026). For induction, Chen et al. (2024b) study staged feature learning and attention-scale growth against finite-context errors. Related work analyzes dynamics in a symmetry-motivated low-dimensional reduction (Musat et al., 2025) and establishes invariant population-dynamics subspaces (Musat et al., 2026). Our framework combines regional capture, inward drift, increasing resolution, and scale-adapted estimates that preserve these properties on a finite corpus, thereby linking gradient training to threshold-estimation guarantees.

Boundary estimation and attention resolution. Nonlinear and nonparametric analyses study learned low-dimensional structure and local prediction rules (Oko et al., 2024; Collins et al., 2024; Li et al., 2024); experiments with discrete functions include thresholds (Bhattamishra et al., 2024). Directly observed threshold models exhibit inverse-sample-size localization (Hansen, 2000), whereas indirect or high-dimensional boundary-estimation rates depend on ill-posedness, smoothness, and sparsity (Goldenshluger et al., 2006; Feng et al., 2022). Parameter divergence with functional convergence appears in implicit bias and max-margin attention selection (Soudry et al., 2018; Tarzanagh et al., 2023; Vasudeva et al., 2025), and low-entropy attention is studied by Varre et al. (2026). Pretraining guarantees distinguish task and context counts (Wu et al., 2024; Oko et al., 2024); related work treats finite-token approximation (Bohbot et al., 2026) and finite-data softmax optimization with preconditioning and regularization (Goel et al., 2026). We instead connect ordinary constant-step descent to boundary estimation through relative gradient control along a diverging trajectory.

8 Discussion

The analysis connects a statistical rule to a training mechanism. Calibrating the relative label and feature scores suppresses the wrong label class, and increasing concentration localizes the boundary within the correct class. Population training acquires this calibration and continues to increase resolution. Relative gradient estimates show how much of that progress remains certified from a fixed finite pretraining corpus.

8.1 BAYES BENCHMARK AND SYMMETRIZATION

Let $M_N^- := \min\{X_i : Y_i = 0\}$, with value one if no label is negative, and define the posterior bracket

$$L_N := \max\{\rho, M_N^+\}, \qquad U_N := \min\{\ell, M_N^-\}.$$

Under the uniform prior on $[\rho, \ell]$, the posterior distribution of α given the context is $\mathrm{Unif}([L_N, U_N])$. Consequently, its midpoint $B_N := (L_N + U_N)/2$ is the exact Bayes estimator under both absolute and squared loss. The corresponding prior-averaged Bayes risks satisfy

$$R_{N,1}^{\mathrm{Bayes}} = \frac{1}{4}\mathbb{E}(U_N - L_N) = \frac{1 + o(1)}{2N}, \qquad R_{N,2}^{\mathrm{Bayes}} = \frac{1}{12}\mathbb{E}(U_N - L_N)^2 = \frac{1 + o(1)}{2N^2}.$$

These constants follow from the local spacing limit $N(\alpha - L_N, U_N - \alpha) \Rightarrow (E_L, E_R)$ away from the prior endpoints, where E_L, E_R are independent $\mathrm{Exp}(1)$ variables; the endpoint layers have vanishing prior mass. In its hard-attention limit, the trained one-head rule approaches the left endpoint M_N^+. For fixed interior α, $\mathbb{E}|M_N^+ - \alpha| = (1 + o(1))N^{-1}$ and $\mathbb{E}(M_N^+ - \alpha)^2 = (2 + o(1))N^{-2}$. Prior-averaging these fixed-interior asymptotics gives the same leading constants because the prior-endpoint layers contribute only lower-order terms. Thus, the one-head rule attains the Bayes-optimal rates, while its asymptotic absolute-risk and squared-risk constants are, respectively, two and four times the Bayes constants. The learned softmax estimator inherits these rates whenever its finite-pretraining resolution term is no larger than the fresh-context term.

For comparison, define two oppositely oriented heads with tied parameters (k, μ):

$$F_{\pm,N}(k, \mu; D_N) := \frac{\sum_{i=1}^N X_i \exp\{\pm\mu(kY_i + X_i)\}}{\sum_{i=1}^N \exp\{\pm\mu(kY_i + X_i)\}}, \qquad G_N := \frac{F_{+,N} + F_{-,N}}{2}.$$

When both label classes are resolved and the two softmax endpoint errors are $o(N^{-1})$, $F_{+,N}$ approaches M_N^+ and $F_{-,N}$ approaches M_N^- closely enough that G_N attains the leading Bayes constant. This statement concerns statistical representation. The optimization and finite-pretraining theorems concern the one-head objective, while training under the symmetrized loss is left for future work.

8.2 BOUNDARY COVERAGE

On the training interval $[\rho, \ell]$, the one-head population predictor is uniformly accurate all the way to both endpoints. The all-N transfer theorem above imposes the fixed margin $\alpha^{\mathrm{fr}} \geq \rho + \eta$ only because short fresh contexts may not resolve the weak score separation near the lower endpoint. Once the fresh context is large enough relative to the frozen concentration scale, the self-normalized transfer argument applies on the full closed interval $[\rho, \ell]$; see Corollary G.7.

A genuine architectural obstruction appears when the physical task family is extended from $[\rho, \ell]$ to $[0, 1]$.

Proposition 8.1 (Exact one-head boundary obstruction)**.** *For $S_{k,\mu}(x, y) := \mu(ky + x)$, let $F_+(k, \mu; \alpha)$ and $F_-(k, \mu; \alpha)$ be the population softmax averages of X under scores $S_{k,\mu}(X, Y)$ and $-S_{k,\mu}(X, Y)$, respectively. For fixed $k = \lambda/\mu$,*

$$F_+(k, \mu; 0) = 1 - \frac{1}{\mu} + \frac{1}{e^\mu - 1}, \qquad F_-(k, \mu; 1) = \frac{1}{\mu} - \frac{1}{e^\mu - 1}. \tag{15}$$

Consequently,

$$F_+(k, \mu; 0) \to 1, \qquad F_-(k, \mu; 1) \to 0 \qquad \text{as } \mu \to \infty. \tag{16}$$

Hence no single fixed orientation in this score family can be uniformly consistent on the full physical task interval $[0, 1]$.

Appendix A.5.3 gives the proof of Proposition 8.1.

Scope and further questions. The theory assumes a fixed split feature, noiseless labels, a uniform design, and large-resolution initialization; the two orientations are symmetric. Appendix H explores label noise and Beta designs, while Proposition C.25 and Lemma C.26 bound the pre-capture scale loss used in proving Theorem 5.2. Extending the theory to small initialization, noisy or nonuniform settings, learned split features, and sharp finite-pretraining limits remains open.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing and review, including improving readability, checking internal consistency, locating relevant literature, and formatting LaTeX. We also used generative AI to provide critical ingredients for proving mathematical claims, and assist in the writing of proofs. The authors independently reviewed all AI-assisted changes and verified the cited sources, mathematical statements, proofs, and computational results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The model, data-generating process, optimization procedure, and all assumptions are specified in Sections 2–5. Complete proofs of the theoretical claims appear in the supplementary material in the dependency order summarized by its proof guide. Appendix H provides the numerical procedures, parameter settings, initialization checks, and additional diagnostics used in the experiments.

REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*, 2023.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36:57125–57211, 2023.

Krishnakumar Balasubramanian. Unbiased gradients, moving stability boundaries: Exact mini-batch geometry in linear self-attention. *arXiv preprint arXiv:2605.21292*, 2026. URL https://arxiv.org/abs/2605.21292.

Satwik Bhattamishra, Arkil Patel, Phil Blunsom, and Varun Kanade. Understanding in-context learning in transformers and LLMs by learning to learn discrete functions. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f7e7fabd73b3df96c54a320862afcb78-Abstract-Conference.html.

Léa Bohbot, Cyril Letrouit, Gabriel Peyré, and François-Xavier Vialard. Token sample complexity of attention. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, Seoul, South Korea, 2026. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality (extended abstract). In *Proceedings of the Thirty-Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4573–4573. PMLR, 2024a.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Unveiling induction heads: Provable training dynamics and feature learning in transformers. *arXiv preprint arXiv:2409.10559*, 2024b.

Liam Collins, Advait Parulekar, Aryan Mokhtari, Sujay Sanghavi, and Sanjay Shakkottai. In-context learning with transformers: Softmax attention adapts to function lipschitzness. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a8633d27d782f66fe660c2fb4bae446e-Abstract-Conference.html.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. TabArena: A living benchmark for

machine learning on tabular data. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025. doi: 10.52202/085713-0519. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1697e3fb412da11dc9488249f9e7bbc9-Abstract-Datasets_and_Benchmarks_Track.html.

Huijie Feng, Yang Ning, and Jiwei Zhao. Nonregular and minimax estimation of individualized thresholds in high dimension with binary responses. *The Annals of Statistics*, 50(4):2284–2305, 2022.

David T. Frazier, Ryan Kelly, Christopher Drovandi, and David J. Warne. The statistical accuracy of neural posterior and likelihood estimation. *arXiv preprint arXiv:2411.12068*, 2024. URL https://arxiv.org/abs/2411.12068.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598, 2022.

Khashayar Gatmiry, Nikunj Saunshi, Sashank J. Reddi, Stefanie Jegelka, and Sanjiv Kumar. Can looped transformers learn to implement multi-step gradient descent for in-context learning? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15130–15152. PMLR, 2024.

Gautam Goel, Mahdi Soltanolkotabi, and Peter Bartlett. Training dynamics of softmax self-attention: Fast global convergence via preconditioning. *arXiv preprint arXiv:2603.01514*, 2026.

A. Goldenshluger, A. Tsybakov, and A. Zeevi. Optimal change-point estimation from indirect observations. *The Annals of Statistics*, 34(1):350 – 372, 2006. doi: 10.1214/009053605000000750. URL https://doi.org/10.1214/009053605000000750.

Bruce E. Hansen. Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603, 2000. doi: 10.1111/1468-0262.00124. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00124.

Jianliang He, Xintian Pan, Siyu Chen, and Zhuoran Yang. In-context linear regression demystified: Training dynamics and mechanistic interpretability of multi-head softmax attention. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22686–22742. PMLR, 2025.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19660–19722. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huang24d.html.

Gyeonghun Kang, Changwoo J. Lee, and Xiang Cheng. Transformers can learn posterior predictive distributions in-context. *arXiv preprint arXiv:2605.26713*, 2026.

Zihao Li, Yuan Cao, Cheng Gao, Yihan He, Han Liu, Jason M Klusowski, Jianqing Fan, and Mengdi Wang. One-layer transformer provably learns one-nearest neighbor in context. *Advances in Neural Information Processing Systems*, 37:82166–82204, 2024.

Samuel Müller, Arik Reuter, Noah Hollmann, David Rügamer, and Frank Hutter. Position: The future of bayesian prediction is prior-fitted. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 81861–81875, 2025. URL https://proceedings.mlr.press/v267/muller25d.html.

Tiberiu Musat, Tiago Pimentel, Lorenzo Noci, Alessandro Stolfo, Mrinmaya Sachan, and Thomas Hofmann. On the emergence of induction heads for in-context learning. *arXiv preprint arXiv:2511.01033*, 2025.

Tiberiu Musat, Tiago Pimentel, Nicolas Zucchet, and Thomas Hofmann. Invariant learning dynamics of transformers in inductive reasoning tasks. *arXiv preprint arXiv:2607.11875*, 2026.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Pretrained transformer efficiently learns low-dimensional target functions in-context. *Advances in Neural Information Processing Systems*, 37:77316–77365, 2024.

Maxime Peyrard and Kyunghyun Cho. Meta-statistical learning: Supervised learning of statistical estimators. *arXiv preprint arXiv:2502.12088*, 2025. URL https://arxiv.org/abs/2502.12088.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50817–50847. PMLR, 2025.

Almut Rödder, Manuel Hentschel, and Sebastian Engelke. Theoretical guarantees for neural estimators in parametric statistics. *arXiv preprint arXiv:2506.18508*, 2025. URL https://arxiv.org/abs/2506.18508.

Matthew Sainsbury-Dale, Andrew Zammit-Mangion, and Raphaël Huser. Likelihood-free parameter estimation with neural bayes estimators. *The American Statistician*, 78(1):1–14, 2024. doi: 10.1080/00031305.2023.2249522.

Wei Shen, Ruida Zhou, Jing Yang, and Cong Shen. On the training convergence of transformers for in-context classification of Gaussian mixtures. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54732–54771. PMLR, 2025.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. URL https://jmlr.org/papers/v19/18-188.html.

Yan Shuo Tan, Abhineet Agarwal, and Bin Yu. A cautionary tale on fitting decision trees to data from additive models: generalization lower bounds. In *International Conference on Artificial Intelligence and Statistics*, pp. 9663–9685. PMLR, 2022.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/970f59b22f4c72aec75174aae63c7459-Abstract-Conference.html.

Aditya Varre, Mark Rofin, and Nicolas Flammarion. Gradient flow polarizes softmax outputs towards low-entropy solutions. *arXiv preprint arXiv:2603.06248*, 2026. URL https://arxiv.org/abs/2603.06248.

Bhavya Vasudeva, Puneesh Deora, and Christos Thrampoulidis. Implicit bias and fast convergence rates for self-attention. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=pKilnjQsb0.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fa3c139cf8084de7bfd944f1c90c8695-Abstract-Conference.html.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. In-context learning with representations: Contextual generalization of trained transformers. *Advances in Neural Information Processing Systems*, 37, 2024.

Andrew Zammit-Mangion, Matthew Sainsbury-Dale, and Raphaël Huser. Neural methods for amortized inference. *Annual Review of Statistics and Its Application*, 12:311–335, 2025. doi: 10.1146/annurev-statistics-112723-034123.

Qiong Zhang, Yan Shuo Tan, Qinglong Tian, and Pengfei Li. TabPFN: One model to rule them all? *arXiv preprint arXiv:2505.20003*, 2025a. URL https://arxiv.org/abs/2505.20003.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

Xiyuan Zhang, Danielle C. Maddix, Junming Yin, Nick Erickson, Abdul Fatir Ansari, Boran Han, Shuai Zhang, Leman Akoglu, Christos Faloutsos, Michael W. Mahoney, Cuixiong Hu, Huzefa Rangwala, George Karypis, and Bernie Wang. MITRA: Mixed synthetic priors for enhancing tabular foundation models. In *Advances in Neural Information Processing Systems*, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/177d68f4adef163b7b123b5c5adb3c60-Abstract-Conference.html.

Yedi Zhang, Aaditya K. Singh, Peter E. Latham, and Andrew M. Saxe. Training dynamics of in-context learning in linear attention. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76047–76087. PMLR, 2025c.

SUPPLEMENTARY MATERIAL

Roadmap. The appendix is organized in the dependency order of the arguments:

- Appendix A gives the proof framework, the exact deterministic training-time window, and the detailed finite-pretraining and freeze-then-transfer theorems. It also assembles Theorem 3.1 and records the short proofs of the gradient identities, the finite-pretraining summary, and the boundary obstruction.
- Appendix B derives the two-component mixture representation, the bias and ratio identities, and global well-posedness of the population flow.
- Appendix C proves population capture and sharpening (Theorem 5.2) through outer ratio correction, interior transport, logarithmic endpoint capture, and radial persistence.
- Appendix D transfers the population-flow geometry to constant-step population gradient descent using finite-horizon Euler shadowing and discrete barrier arguments.
- Appendix E establishes the compact and scale-adapted empirical-process bounds used to compare finite-corpus and population gradients.
- Appendix F proves the finite-pretraining theorem by combining compact pre-entry tracking with continuation across overlapping statistical shells; it then derives the certified radius and deterministic training-time window.
- Appendix G proves the fixed-parameter fresh-context bound (Proposition 4.1), the freeze-then-transfer theorem, and the large-context extension to the full training interval.
- Appendix H provides the reproducibility protocol, detailed finite-step and transfer diagnostics, initialization and matched-design scope checks, and exploratory label-noise experiments.

A PROOF FRAMEWORK AND TECHNICAL STATEMENTS

This section records the exact training-time window construction and the two technical theorems used to prove the streamlined learned-estimation theorem in the main manuscript. Main-text results appear here only through their assembly and short proofs. The subsequent appendices provide the population, empirical-process, and transfer arguments in dependency order.

A.1 CONTINUATION GEOMETRY

The continuation argument uses the concrete region $\mathcal{G}$ in equation 7. Its resolution coordinate is μ and its shape coordinate is $k = \lambda/\mu$. Within this region, the population radial drift is $\Theta(\mu^{-3})$, the vector-field speed is $O(\mu^{-3})$, the inward margins at the side barriers h_- and h_+ are $\Omega(\mu^{-4})$, and the functional error is $O(\mu^{-1})$; hence the intrinsic training clock is μ^4. The population analysis establishes finite-time entry and subsequent invariance. The empirical continuation argument then preserves the radial and boundary signs on overlapping shells by normalizing their errors by μ^3 and μ^4, respectively.

A.2 DETERMINISTIC TRAINING-TIME WINDOW

Fix $C_{\mathrm{stat}} > 1$ as in Appendix F.4. The certified radius is $M_{\mathrm{stat}}(m,n) \asymp A_{m,n}$. Let $T_0 > \tau_{\mathcal{G}}$ be the deterministic post-capture population-flow time chosen in Appendix D, and let $s_0(\gamma) := \lceil T_0/\gamma \rceil$ be the corresponding population-GD robust-entry index. Define $\mathbb{T}_\gamma := \{T_0 + \gamma j : j = 0, 1, 2, \ldots\}$. For $T = T_0 + \gamma(s - s_0)$, write $\widehat{\theta}^\gamma(T) := \widehat{\theta}^\gamma_s$ and $\widehat{\mu}^\gamma(T) := \widehat{\mu}^\gamma_s$. The empirical continuation argument provides deterministic entry radii $r^{\mathrm{gd}}_- \le r^{\mathrm{gd}}_+$ and constants $0 < c^{\mathrm{gd}}_- \le c^{\mathrm{gd}}_+$, which define the radial envelopes

$$M^{\mathrm{gd}}_\pm(T) := \left\{ (r^{\mathrm{gd}}_\pm)^4 + c^{\mathrm{gd}}_\pm (T - T_0) \right\}^{1/4}. \tag{17}$$

Choose $c_{\mathrm{rate}} > 0$ sufficiently small, depending only on the deterministic envelope constants. Define

$$T^{\mathrm{gd}}_{\mathrm{stat}}(m,n,\gamma) := \max\left\{ T \in \mathbb{T}_\gamma : M^{\mathrm{gd}}_+(T) \le M_{\mathrm{stat}}(m,n) \right\},$$
$$T^{\mathrm{gd}}_{\mathrm{rate}}(m,n,\gamma) := \min\left\{ T \in \mathbb{T}_\gamma : T \le T^{\mathrm{gd}}_{\mathrm{stat}},\ M^{\mathrm{gd}}_-(T) \ge c_{\mathrm{rate}} M_{\mathrm{stat}}(m,n) \right\}.$$

The resulting deterministic statistical-scale training-time window is

$$\mathcal{T}_{\mathrm{stat}}(m,n,\gamma) := [T^{\mathrm{gd}}_{\mathrm{rate}}(m,n,\gamma), T^{\mathrm{gd}}_{\mathrm{stat}}(m,n,\gamma)] \cap \mathbb{T}_\gamma. \tag{18}$$

Both endpoints depend only on (m,n,γ) and fixed model constants; they are determined before the empirical training trajectory is observed. The theorem below shows that, once the pretraining sizes are sufficiently large, these definitions are well posed and the resulting window is nonempty.

Theorem A.1 (Finite-pretraining sharpening horizon)**.** *Under Assumption 2.1, there exist constants $\gamma_0 > 0$ and $A_0 < \infty$ such that, for every fixed $0 < \gamma \le \gamma_0$ and every m, n satisfying $A_{m,n} \ge A_0$, there exists a pretraining event $\mathcal{E}_{\mathrm{train}}$ with $\mathbb{P}(\mathcal{E}_{\mathrm{train}}) \ge 1 - Ce^{-cm} - C(mn)^{-c}$ on which, simultaneously for every $T \in [T_0, T^{\mathrm{gd}}_{\mathrm{stat}}] \cap \mathbb{T}_\gamma$,*

$$\widehat{\theta}^\gamma(T) \in \mathcal{G}, \qquad M^{\mathrm{gd}}_-(T) \le \widehat{\mu}^\gamma(T) \le M^{\mathrm{gd}}_+(T). \tag{19}$$

For the same A_0, the training-time window $\mathcal{T}_{\mathrm{stat}}(m,n,\gamma)$ is nonempty. Moreover, there exist constants $0 < c_A < C_A < \infty$ such that every $T \in \mathcal{T}_{\mathrm{stat}}(m,n,\gamma)$ satisfies

$$c_A A_{m,n} \le \widehat{\mu}^\gamma(T) \le C_A A_{m,n}, \qquad \sup_{\alpha \in [\rho,\ell]} \left| F(\widehat{\theta}^\gamma(T); \alpha) - \alpha \right| \le \frac{C}{A_{m,n}}. \tag{20}$$

Finally, $T^{\mathrm{gd}}_{\mathrm{rate}} - T_0 \asymp T^{\mathrm{gd}}_{\mathrm{stat}} - T_0 \asymp A^4_{m,n}$. Equivalently, finite pretraining certifies order $A^4_{m,n}/\gamma$ post-entry gradient steps.

Appendix F proves the theorem and the well-posedness of the late window $\mathcal{T}_{\mathrm{stat}}$ used below for transfer.

A.3 CHECKPOINT-CONDITIONED FRESH-TASK GUARANTEE

Theorem A.2 (Checkpoint-conditioned transfer)**.** *Fix $\eta \in (0, \Delta)$. There exists a constant $q_\eta^{\mathrm{fr}} < \infty$ such that the following holds whenever the conditions of Theorem A.1 are satisfied and $A_{m,n} \geq q_\eta^{\mathrm{fr}}$. Let $\mathcal{F}_{\mathrm{pre}}$ denote the σ-field generated by the pretraining corpus, and let $\widehat{T}$ be any deterministic or $\mathcal{F}_{\mathrm{pre}}$-measurable training time satisfying $\widehat{T} \in \mathcal{T}_{\mathrm{stat}}(m, n, \gamma)$ on $\mathcal{E}_{\mathrm{train}}$. Freeze $\widehat{\theta} := \widehat{\theta}^\gamma(\widehat{T})$ before the fresh threshold, fresh-context size, confidence parameter, or fresh observations are revealed.*

Then, on $\mathcal{E}_{\mathrm{train}}$, for every fixed $\alpha^{\mathrm{fr}} \in [\rho + \eta, \ell]$, every integer $N \geq 1$, and every $u \geq 1$,

$$\mathbb{P}_{\mathrm{fr}}\left(\left| F_N\left(\widehat{\theta}; D_N^{\mathrm{fr}}\right) - \alpha^{\mathrm{fr}} \right| \leq C_\eta \left[\frac{1 + \log\left(1 + C_{\mathrm{fr}} \widehat{\mu}^\gamma(\widehat{T}) u\right)}{\widehat{\mu}^\gamma(\widehat{T})} + \frac{u}{N} \right] \,\middle|\, \mathcal{F}_{\mathrm{pre}} \right) \geq 1 - 4e^{-u}. \tag{21}$$

The checkpoint is selected entirely from pretraining information and is independent of the fresh-context size N. The first term in equation 21 depends on the realized frozen concentration $\widehat{\mu}$, whereas the second is the fresh-context localization error. Substituting the scale bounds from Theorem A.1 yields the $A_{m,n}$-level statement in Theorem 3.1; Appendix A.4 performs this assembly.

Appendix G.2 proves the conditional statement from Proposition 4.1 and a conditioning argument.

Corollary A.3 (Fresh-query classification)**.** *Under the setting of Theorem 3.1, let $X_q^{\mathrm{fr}} \sim \mathrm{Unif}(0, 1)$ be an independent fresh query and define $Y_q^{\mathrm{fr}} := \mathbf{1}\{X_q^{\mathrm{fr}} \leq \alpha^{\mathrm{fr}}\}$ and $\widehat{Y}_q^{\mathrm{fr}} := \mathbf{1}\{X_q^{\mathrm{fr}} \leq F_N(\widehat{\theta}; D_N^{\mathrm{fr}})\}$. Then*

$$\mathbb{P}\left(\widehat{Y}_q^{\mathrm{fr}} \neq Y_q^{\mathrm{fr}} \,\middle|\, \widehat{\theta}, D_N^{\mathrm{fr}}, \alpha^{\mathrm{fr}}\right) = \left| F_N(\widehat{\theta}; D_N^{\mathrm{fr}}) - \alpha^{\mathrm{fr}} \right|. \tag{22}$$

Hence Theorem 3.1 yields the same $\widetilde{O}((m \wedge n)^{-1} + N^{-1})$ rate for fresh-query classification.

Proof of Corollary A.3. Condition on $\widehat{\theta}$, D_N^{fr}, and α^{fr}, so the two thresholds $F_N(\widehat{\theta}; D_N^{\mathrm{fr}})$ and α^{fr} are fixed points in $[0, 1]$. The indicators $\mathbf{1}\{X_q^{\mathrm{fr}} \leq F_N\}$ and $\mathbf{1}\{X_q^{\mathrm{fr}} \leq \alpha^{\mathrm{fr}}\}$ disagree exactly when X_q^{fr} lies between these two points. Since $X_q^{\mathrm{fr}} \sim \mathrm{Unif}(0, 1)$, the conditional probability of this interval is their absolute difference, proving equation 22. The rate follows from Theorem 3.1. □

A.4 ASSEMBLY OF THE MAIN LEARNED-ESTIMATION THEOREM

Proof of Theorem 3.1. Let $\mathcal{I}_{m,n,\gamma}$ contain the gradient-descent iteration indices corresponding to $\mathcal{T}_{\mathrm{stat}}(m, n, \gamma)$ under $T = T_0 + \gamma(s - s_0)$. Theorem A.1 makes this interval deterministic and nonempty and gives the stated scale and step counts. On $\mathcal{E}_{\mathrm{train}}$, substitute $c_A A_{m,n} \leq \widehat{\mu} \leq C_A A_{m,n}$ into equation 21, absorbing fixed constants. This gives equation 5 with its conditional probability unchanged. Adding the failure probability of $\mathcal{E}_{\mathrm{train}}$ gives the joint guarantee. □

A.5 SHORT PROOFS OF MAIN-TEXT RESULTS

A.5.1 GRADIENT IDENTITIES

Proof of Lemma 5.1. For fixed α, let $\mathbb{E}_\theta[\,\cdot \mid \alpha]$ denote expectation under the density on $[0, 1]$ proportional to $\exp\{\lambda Y(x) + \mu x\}$, where $Y(x) = \mathbf{1}\{x \leq \alpha\}$. Then $F(\theta; \alpha) = \mathbb{E}_\theta[X \mid \alpha]$. Differentiating this bounded exponential-family expectation gives

$$\partial_\lambda F(\theta; \alpha) = \mathrm{Cov}_\theta(X, Y \mid \alpha), \qquad \partial_\mu F(\theta; \alpha) = \mathrm{Cov}_\theta(X, X \mid \alpha) = \mathrm{Var}_\theta(X \mid \alpha).$$

Because $X, Y \in [0, 1]$, differentiation may be passed through the integral defining the population risk. Since $\alpha \sim \mathrm{Unif}([\rho, \ell])$ and $\Delta = \ell - \rho$, differentiating $\mathcal{R}_\rho(\theta) = \frac{1}{2\Delta} \int_\rho^\ell (F(\theta; \alpha) - \alpha)^2 \, d\alpha$ and substituting the two identities yields the claimed gradient formula. □

A.5.2 FINITE-PRETRAINING SUMMARY

Proof of Corollary 5.3. Apply Theorem A.1 on its event $\mathcal{E}_{\mathrm{train}}$. Its interval $[T_0, T_{\mathrm{stat}}^{\mathrm{gd}}] \cap \mathbb{T}_\gamma$ gives entry into and persistence within $\mathcal{G}$. Choose the late subinterval to be $\mathcal{T}_{\mathrm{stat}}(m, n, \gamma)$. Equation equa-

tion 20 gives the two asserted bounds at every checkpoint in this subinterval, and the theorem gives both the stated event probability and the $\Theta(A_{m,n}^4/\gamma)$ post-entry step count. □

A.5.3 One-head boundary obstruction

Proof of Proposition 8.1. At $\alpha = 0$, one has $Y = 0$ almost surely, so the factor involving k disappears and

$$F_+(k, \mu; 0) = \frac{\int_0^1 x e^{\mu x}\, dx}{\int_0^1 e^{\mu x}\, dx} = 1 - \frac{1}{\mu} + \frac{1}{e^\mu - 1}.$$

At $\alpha = 1$, one has $Y = 1$ almost surely. Under the reflected score, the constant factor $e^{-\mu k}$ cancels from the normalized average, giving

$$F_-(k, \mu; 1) = \frac{\int_0^1 x e^{-\mu x}\, dx}{\int_0^1 e^{-\mu x}\, dx} = \frac{1}{\mu} - \frac{1}{e^\mu - 1}.$$

Taking $\mu \to \infty$ proves equation 16. Thus each fixed orientation incurs order-one error at one physical endpoint, which rules out uniform consistency on $[0, 1]$. □

B Structural Reduction and Basic Identities

Throughout, $c, C > 0$ denote generic constants that may change from line to line, while named constants remain fixed after they are introduced. Dependence on fixed parameters is indicated by subscripts when needed.

B.1 Two-component structural reduction

For $\mu > 0$, define the component masses

$$A_{1,\mu}(\alpha) := \int_0^\alpha e^{\mu x}\, dx = \frac{e^{\mu\alpha} - 1}{\mu}, \qquad A_{2,\mu}(\alpha) := \int_\alpha^1 e^{\mu x}\, dx = \frac{e^\mu - e^{\mu\alpha}}{\mu}, \tag{23}$$

and the corresponding mixture weights

$$p_\theta(\alpha) := \frac{e^\lambda A_{1,\mu}(\alpha)}{e^\lambda A_{1,\mu}(\alpha) + A_{2,\mu}(\alpha)}, \qquad q_\theta(\alpha) := 1 - p_\theta(\alpha). \tag{24}$$

Let $m_{j,\mu}(\alpha)$ and $s_{j,\mu}^2(\alpha)$ denote the mean and variance of the density proportional to $e^{\mu x}$ on $I_1 = [0, \alpha]$ and $I_2 = (\alpha, 1]$, respectively. Then

$$m_{1,\mu}(\alpha) = \alpha - \frac{1}{\mu} + \frac{\alpha}{e^{\mu\alpha} - 1}, \qquad m_{2,\mu}(\alpha) = 1 - \frac{1}{\mu} + \frac{1 - \alpha}{e^{\mu(1-\alpha)} - 1},$$
$$s_{1,\mu}^2(\alpha) = \frac{1}{\mu^2} - \frac{\alpha^2 e^{-\mu\alpha}}{(1 - e^{-\mu\alpha})^2}, \qquad s_{2,\mu}^2(\alpha) = \frac{1}{\mu^2} - \frac{(1-\alpha)^2 e^{-\mu(1-\alpha)}}{(1 - e^{-\mu(1-\alpha)})^2}.$$

Finally, define the component mean separation $d_\mu(\alpha) := m_{2,\mu}(\alpha) - m_{1,\mu}(\alpha)$. The exact representation and its large-μ behavior are summarized in the following structural lemma.

Lemma B.1 (Two-component representation and asymptotics)**.** *For every $\mu > 0$ and $\alpha \in (0, 1)$,*

$$F(\theta; \alpha) = p_\theta(\alpha) m_{1,\mu}(\alpha) + q_\theta(\alpha) m_{2,\mu}(\alpha), \tag{25}$$
$$\mathrm{Cov}_\theta(X, Y \mid \alpha) = -p_\theta(\alpha) q_\theta(\alpha) d_\mu(\alpha), \tag{26}$$

and

$$\mathrm{Var}_\theta(X \mid \alpha) = p_\theta(\alpha) s_{1,\mu}^2(\alpha) + q_\theta(\alpha) s_{2,\mu}^2(\alpha) + p_\theta(\alpha) q_\theta(\alpha) d_\mu(\alpha)^2. \tag{27}$$

Moreover, whenever $k = \lambda/\mu$,

$$\frac{p_\theta(\alpha)}{q_\theta(\alpha)} = e^{\mu(k+\alpha-1)} \frac{1 - e^{-\mu\alpha}}{1 - e^{-\mu(1-\alpha)}}. \tag{28}$$

Uniformly over $\alpha \in [\rho, \ell]$, as $\mu \to \infty$,

$$m_{1,\mu}(\alpha) = \alpha - \frac{1}{\mu} + O(e^{-\rho\mu}), \qquad m_{2,\mu}(\alpha) = 1 - \frac{1}{\mu} + O(e^{-\rho\mu}), \tag{29}$$

$$s^2_{1,\mu}(\alpha) = \frac{1}{\mu^2} + O(e^{-\rho\mu}), \qquad s^2_{2,\mu}(\alpha) = \frac{1}{\mu^2} + O(e^{-\rho\mu}), \tag{30}$$

and

$$d_\mu(\alpha) = 1 - \alpha + O(e^{-\rho\mu}). \tag{31}$$

In particular, there exist constants $0 < c_s < C_s < \infty$ and $\mu_s > 0$, depending only on ρ, such that

$$\frac{c_s}{\mu^2} \le s^2_{j,\mu}(\alpha) \le \frac{C_s}{\mu^2}, \qquad j \in \{1, 2\}, \tag{32}$$

for every $\mu \ge \mu_s$, uniformly over $\alpha \in [\rho, \ell]$.

Proof. The total tilted mass on $[0, \alpha]$ is $e^\lambda A_{1,\mu}(\alpha)$, while the total tilted mass on $[\alpha, 1]$ is $A_{2,\mu}(\alpha)$. Hence $p_\theta(\alpha)$ and $q_\theta(\alpha)$ are the probabilities of the left and right components, respectively, which proves equation 25.

For the covariance identity, suppress the common dependence on α. Since $Y = 1$ on the left component and $Y = 0$ on the right component,

$$\begin{aligned}\mathrm{Cov}_\theta(X, Y \mid \alpha) &= \mathbb{E}_\theta[XY \mid \alpha] - \mathbb{E}_\theta[X \mid \alpha]\mathbb{E}_\theta[Y \mid \alpha] \\ &= p_\theta m_{1,\mu} - p_\theta \left(p_\theta m_{1,\mu} + q_\theta m_{2,\mu}\right) \\ &= p_\theta q_\theta (m_{1,\mu} - m_{2,\mu}) \\ &= -p_\theta q_\theta d_\mu.\end{aligned}$$

This proves equation 26. The variance formula equation 27 follows from the law of total variance. Finally,

$$\frac{p_\theta(\alpha)}{q_\theta(\alpha)} = e^\lambda \frac{A_{1,\mu}(\alpha)}{A_{2,\mu}(\alpha)} = e^{\lambda + \mu\alpha - \mu} \frac{1 - e^{-\mu\alpha}}{1 - e^{-\mu(1-\alpha)}}.$$

Using $\lambda = k\mu$ gives equation 28.

Since $\alpha \ge \rho$ and $1 - \alpha \ge \rho$ on $[\rho, \ell]$, the correction terms in $m_{1,\mu}(\alpha)$ and $m_{2,\mu}(\alpha)$ are uniformly $O(e^{-\rho\mu})$. This proves the mean expansions. The variance expansions follow similarly from $s^2_{1,\mu}(\alpha)$ and $s^2_{2,\mu}(\alpha)$. Subtracting the two mean expansions gives the expansion for d_μ. The two-sided variance bounds follow after increasing μ_s. □

The representation immediately yields the bias expansion used throughout the subsequent dynamics.

Corollary B.2 (Uniform bias expansion)**.** *Uniformly over $\alpha \in [\rho, \ell]$, as $\mu \to \infty$,*

$$F(\theta; \alpha) - \alpha = q_\theta(\alpha)(1 - \alpha) - \frac{1}{\mu} + O(e^{-\rho\mu}). \tag{33}$$

The remainder is uniform over $\lambda \in \mathbb{R}$.

Proof. Using equation 25 and equation 29,

$$\begin{aligned}F(\theta; \alpha) &= p_\theta(\alpha)\left(\alpha - \frac{1}{\mu}\right) + q_\theta(\alpha)\left(1 - \frac{1}{\mu}\right) + O(e^{-\rho\mu}) \\ &= p_\theta(\alpha)\alpha + q_\theta(\alpha) - \frac{1}{\mu} + O(e^{-\rho\mu}).\end{aligned}$$

Since $p_\theta(\alpha) + q_\theta(\alpha) = 1$, subtracting α gives the claim. Uniformity in λ follows from $0 \le p_\theta(\alpha), q_\theta(\alpha) \le 1$. □

The odds identity equation 28 gives, uniformly over $\alpha \in [\rho, \ell]$,

$$\frac{p_\theta(\alpha)}{q_\theta(\alpha)} = e^{\mu(k + \alpha - 1)} \left\{1 + O(e^{-\rho\mu})\right\}. \tag{34}$$

Hence the two components exchange dominance near $\alpha_c := 1 - k$. Thus k determines the location of the attention transition across the task family, while μ determines its sharpness. This is the structural reduction underlying the population dynamics.

B.2 Effective coordinates and ratio dynamics

The preceding representation shows that (k, μ) are the natural coordinates for the population flow: k moves the transition location α_c, while μ controls concentration.

Lemma B.3 (Ratio dynamics and sign structure)**.** *Along the population gradient flow, whenever* $\mu(t) > 0$*,*

$$\dot{k}(t) = -\frac{S(\lambda(t), \mu(t))}{\mu(t)^2}, \tag{35}$$

where $S(\lambda, \mu) := \mu \partial_\lambda \mathcal{R}_\rho(\lambda, \mu) - \lambda \partial_\mu \mathcal{R}_\rho(\lambda, \mu)$*. Equivalently,*

$$S(\lambda, \mu) = \frac{1}{\Delta} \int_\rho^\ell \big(F(\theta; \alpha) - \alpha\big) H_\theta(\alpha)\, d\alpha, \tag{36}$$

where $H_\theta(\alpha) := \mu \mathrm{Cov}_\theta(X, Y \mid \alpha) - \lambda \mathrm{Var}_\theta(X \mid \alpha)$*. For every* $\lambda, \mu \geq 0$ *and* $\alpha \in (0, 1)$*,* $H_\theta(\alpha) \leq 0$*.*

Proof. Since $k = \lambda/\mu$, $\dot{k} = (\mu\dot{\lambda} - \lambda\dot{\mu})/\mu^2$. Using $\dot{\lambda} = -\partial_\lambda \mathcal{R}_\rho$ and $\dot{\mu} = -\partial_\mu \mathcal{R}_\rho$, gives equation 35 and $S(\lambda, \mu)$. Formula equation 36 follows directly from Lemma 5.1.

Suppose first that $\mu > 0$. By equation 26, $\mathrm{Cov}_\theta(X, Y \mid \alpha) = -p_\theta(\alpha) q_\theta(\alpha) d_\mu(\alpha) \leq 0$ because $d_\mu(\alpha) \geq 0$. Since $\mathrm{Var}_\theta(X \mid \alpha) \geq 0, \quad \lambda, \mu \geq 0$, it follows that $H_\theta(\alpha) \leq 0$. If $\mu = 0$, then $H_\theta(\alpha) = -\lambda \mathrm{Var}_\theta(X \mid \alpha) \leq 0$. □

B.3 Well-posedness of the population flow

The population dynamics used below are globally well posed. We record the standard ODE fact and its application here for completeness. The derivative identities needed below are those of Lemma 5.1 in the main text.

Lemma B.4 (Bounded locally Lipschitz gradient flows)**.** *Let* $G : \mathbb{R}^2 \to \mathbb{R}$ *be continuously differentiable, and suppose that* ∇G *is locally Lipschitz. If there exists* $B < \infty$ *such that* $\|\nabla G(\theta)\| \leq B$ *for every* $\theta \in \mathbb{R}^2$*, then, for every initial condition* $\theta(0) \in \mathbb{R}^2$*, the gradient flow* $\dot{\theta}(t) = -\nabla G(\theta(t))$ *admits a unique solution defined for all* $t \geq 0$*.*

Proof. Since $-\nabla G$ is locally Lipschitz, the Picard–Lindelöf theorem gives a unique maximal solution $\theta : [0, T_{\mathrm{life}}) \to \mathbb{R}^2$. The uniform gradient bound gives $\|\dot{\theta}(t)\| = \|\nabla G(\theta(t))\| \leq B$. Hence, for $t < T_{\mathrm{life}}$, $\|\theta(t)\| \leq \|\theta(0)\| + Bt$. Thus the trajectory is bounded on every finite time interval. By the continuation theorem for locally Lipschitz ODEs, a finite maximal time would force $\|\theta(t)\| \to \infty$ as $t \uparrow T_{\mathrm{life}}$, which is impossible. Therefore $T_{\mathrm{life}} = \infty$. □

Lemma B.5 (Population global well-posedness)**.** *For every initial condition* $\theta(0) \in \mathbb{R}^2$*, the population gradient flow* $\dot{\theta}(t) = -\nabla \mathcal{R}_\rho(\theta(t))$ *admits a unique solution defined for all* $t \geq 0$*.*

Proof. The map $\theta \mapsto \mathcal{R}_\rho(\theta)$ is smooth. Indeed, on every compact subset of $\mathbb{R}^2$, the integrands defining $F(\theta; \alpha)$ and all their θ-derivatives are uniformly bounded over $(x, \alpha) \in [0, 1] \times [\rho, \ell]$, so differentiation under the integral sign is valid.

Moreover, $|F(\theta; \alpha) - \alpha| \leq 1$. Since $X, Y \in [0, 1]$, $\mathrm{Var}_\theta(X \mid \alpha) \leq \frac{1}{4}$, $|\mathrm{Cov}_\theta(X, Y \mid \alpha)| \leq \frac{1}{4}$. Using Lemma 5.1, we obtain $|\partial_\lambda \mathcal{R}_\rho(\theta)| \leq \frac{1}{4}$ and $|\partial_\mu \mathcal{R}_\rho(\theta)| \leq \frac{1}{4}$. Thus $\|\nabla \mathcal{R}_\rho(\theta)\| \leq \frac{\sqrt{2}}{4}$ for every θ. The claim follows from Lemma B.4. □

The identities above reduce the population training problem to the interaction between the prediction bias $F(\theta; \alpha) - \alpha$ and the moving attention transition α_c. Section C shows how this interaction first corrects the ratio, then transports the transition across the task interval, and finally produces logarithmic capture near the endpoint.

C Proof of Population Capture and Sharpening

This section proves Theorem 5.2. The proof has four modules: outer ratio correction, transport through the task interior, logarithmic endpoint capture, and radial persistence. The first three produce finite-time entry into the logarithmic tube; the last prevents the concentration scale from leaving the large-μ regime before capture.

Fix $C_{\mathrm{band}} \geq 2$. For $\mu > 1$, set $L_{\mathrm{band}}(\mu) := \rho - C_{\mathrm{band}} \log \mu/\mu$, $U_{\mathrm{band}}(\mu) := \ell + C_{\mathrm{band}} \log \mu/\mu$, and $b_c(\mu) := \ell + (\log \mu + c)/\mu$. The endpoint drift profiles are

$$d_k(c) := \ell + \frac{\ell(1+\ell^2)}{\Delta} e^{-c} - \frac{\ell^2(1+\ell^2)}{2\Delta} e^{-2c}, \qquad d_{\mathrm{rad}}(c) := 1 + \frac{\ell^2}{\Delta} e^{-c} - \frac{\ell^3}{2\Delta} e^{-2c}. \tag{37}$$

Lemma C.16 selects c_0 with $d_k(c_0) < 0 < d_{\mathrm{rad}}(c_0)$. Define $\mathcal{I}_{\log}(\mu) := [b_{c_0}(\mu), U_{\mathrm{band}}(\mu)]$ and use the main-text progress region $\mathcal{G} = \{\mu \geq \mu_*,\ k \in \mathcal{I}_{\log}(\mu)\}$. Its side barriers are $h_-(\lambda, \mu) := \lambda/\mu - b_{c_0}(\mu)$ and $h_+(\lambda, \mu) := U_{\mathrm{band}}(\mu) - \lambda/\mu$.

Threshold convention. All constants and large-μ thresholds in this section depend only on the fixed structural parameters (and on κ or η when displayed). Because only finitely many such thresholds are used in the final assembly, we increase the base radius μ_* once so that every cited estimate holds whenever $\mu \geq \mu_*$.

C.1 Barrier tools and regional drift estimates

For compactness, write $B_\theta(\alpha) := F(\theta; \alpha) - \alpha$, $\sigma_\theta^2(\alpha) := \mathrm{Var}_\theta(X \mid \alpha)$, and $J_\theta(\alpha) := -H_\theta(\alpha)$. Since $J_\theta = -H_\theta$, Lemma B.3 gives

$$S(\lambda, \mu) = -\frac{1}{\Delta} \int_\rho^\ell B_\theta(\alpha) J_\theta(\alpha)\, d\alpha, \qquad \dot{k} = \frac{1}{\Delta \mu^2} \int_\rho^\ell B_\theta(\alpha) J_\theta(\alpha)\, d\alpha. \tag{38}$$

For $0 < \eta < \Delta/4$, define the lower layer, fixed interior strip, and upper rail tube by

$$\mathcal{L}_\eta(\mu) := [L_{\mathrm{band}}(\mu), \rho + \eta], \qquad \mathcal{I}_\eta := [\rho + \eta, \ell - \eta], \qquad \mathcal{T}_\eta(\mu) := [\ell - \eta, U_{\mathrm{band}}(\mu)].$$

Lemma C.1 (Finite-time entry from a one-sided error inequality)**.** *Let $T \in (t_0, \infty]$, and let $e : [t_0, T) \to [0, \infty)$, $\mu : [t_0, T) \to (0, \infty)$ be continuously differentiable. Define the entry time $\tau_e := \inf\{t \geq t_0 : e(t) = 0\}$. Suppose that, whenever $e(t) > 0$,*

$$\dot{e}(t) \leq -\frac{a}{\mu(t)^p} \tag{39}$$

for some constants $a > 0$ and $p > 0$.

(i) *If $\mu(t) \leq \mu(t_0)$ whenever $e(t) > 0$, then, for every $t < T \wedge \tau_e$,*

$$e(t) \leq e(t_0) - \frac{a}{\mu(t_0)^p}(t - t_0). \tag{40}$$

Consequently, if $\tau_e \geq T$, then $T - t_0 \leq \dfrac{\mu(t_0)^p e(t_0)}{a}$. In particular, if $T = \infty$, then $\tau_e < \infty$.

(ii) *If $p = 4$ and $\dot{\mu}(t) \leq \dfrac{A}{\mu(t)^3}$ whenever $e(t) > 0$, for some $A > 0$, then, for every $t < T \wedge \tau_e$,*

$$e(t) \leq e(t_0) - \frac{a}{4A} \log\left(1 + \frac{4A(t - t_0)}{\mu(t_0)^4}\right). \tag{41}$$

In particular, if $T = \infty$, then $\tau_e < \infty$.

Proof. If $e(t_0) = 0$, then $\tau_e = t_0$ and there is nothing to prove. Hence assume $e(t_0) > 0$. On $[t_0, T \wedge \tau_e)$, one has $e(t) > 0$, so the hypotheses are valid.

For (i), on $[t_0, T \wedge \tau_e)$, $\dot{e}(t) \leq -\frac{a}{\mu(t_0)^p}$, and integration gives equation 40. If $\tau_e \geq T$, then $e(t) \geq 0$ for all $t < T$, so letting $t \uparrow T$ gives $T - t_0 \leq \frac{\mu(t_0)^p e(t_0)}{a}$. If $T = \infty$, this is impossible; hence $\tau_e < \infty$.

For (ii), the radial bound gives $\frac{d}{dt}\mu(t)^4 = 4\mu(t)^3\dot{\mu}(t) \le 4A$. Hence, on $[t_0, T \wedge \tau_e)$, $\mu(t)^4 \le \mu(t_0)^4 + 4A(t-t_0)$. Using equation 39 with $p=4$ and integrating,

$$\begin{aligned} e(t) &\le e(t_0) - a\int_{t_0}^{t} \frac{ds}{\mu(s)^4} \\ &\le e(t_0) - a\int_{t_0}^{t} \frac{ds}{\mu(t_0)^4 + 4A(s-t_0)} \\ &= e(t_0) - \frac{a}{4A}\log\left(1 + \frac{4A(t-t_0)}{\mu(t_0)^4}\right). \end{aligned}$$

This proves equation 41. If $T=\infty$ and $\tau_e=\infty$, the logarithm diverges, forcing the right-hand side eventually to become negative, contradicting $e(t)\ge 0$. Hence $\tau_e < \infty$. □

Lemma C.2 (Moving-interval invariance). *Let $\mu_{\text{stop}} > 0$, and let $a, b : (\mu_{\text{stop}}, \infty) \to \mathbb{R}$ be C^1 functions satisfying $a(\mu) < b(\mu)$ for $\mu > \mu_{\text{stop}}$. Let k, μ be continuously differentiable on $[t_0, T)$, with $\mu(t_0) > \mu_{\text{stop}}$, and set $\tau_{\text{stop}} := \inf\{t \ge t_0 : \mu(t) \le \mu_{\text{stop}}\}$. Define $h_-(t) := k(t) - a(\mu(t))$, $h_+(t) := b(\mu(t)) - k(t)$. Assume that, whenever $t < T \wedge \tau_{\text{stop}}$,*

$$h_-(t) = 0 \Longrightarrow \dot{h}_-(t) > 0, \qquad h_+(t) = 0 \Longrightarrow \dot{h}_+(t) > 0.$$

If $h_-(t_0) \ge 0$ and $h_+(t_0) \ge 0$, then $k(t) \in [a(\mu(t)), b(\mu(t))]$ for every $t \in [t_0, T \wedge \tau_{\text{stop}})$.

Proof. Suppose not, and define the first violation time

$$\tau := \inf\{t \in [t_0, T \wedge \tau_{\text{stop}}) : h_-(t) < 0 \text{ or } h_+(t) < 0\}.$$

By continuity, $h_-(t), h_+(t) \ge 0$ on $[t_0, \tau]$, and at least one of $h_-(\tau)$ and $h_+(\tau)$ is zero. Every active barrier has strictly positive derivative at τ, while every inactive barrier is positive at τ. Hence, by differentiability and continuity, both barriers remain nonnegative on $(\tau, \tau+\varepsilon)$ for some $\varepsilon > 0$. This contradicts the definition of τ. Therefore no violation occurs before $T \wedge \tau_{\text{stop}}$. □

C.1.1 Outer correction and dynamic-band entry

The sign identity from Section B.2 converts a uniform bias sign outside the support rails into a directional correction of the ratio. The next proposition quantifies this correction and controls the accompanying radial motion.

Proposition C.3 (Quantitative outer correction). *Assume $C_{\text{band}} \ge 2$. There exist constants $c_U, C_U > 0$ and $\mu_U > 0$, depending only on ρ and C_{band}, such that, for all (λ, μ) with $\mu \ge \mu_U$,*

$$k \ge U_{\text{band}}(\mu) \quad \Longrightarrow \quad \dot{k} \le -\frac{c_U}{\mu^4}, \qquad 0 < \dot{\mu} \le \frac{C_U}{\mu^3}.$$

Moreover, for every fixed $\kappa > 0$, there exist constants $c_L(\kappa) > 0$ and $\mu_L(\kappa) > 0$ such that, for all (λ, μ) with $\mu \ge \mu_L(\kappa)$,

$$\kappa \le k < L_{\text{band}}(\mu) \quad \Longrightarrow \quad \dot{k} \ge \frac{c_L(\kappa)}{\mu^3}, \qquad \dot{\mu} < 0.$$

Proof. We use the odds identity $p_\theta(\alpha)/q_\theta(\alpha) = e^{\mu(k+\alpha-1)}(1-e^{-\mu\alpha})/(1-e^{-\mu(1-\alpha)})$, where the second factor is uniformly bounded above and below on $\alpha \in [\rho, \ell]$ for all sufficiently large μ.

First assume $k \ge U_{\text{band}}(\mu)$. Then $\mu(k+\alpha-1) \ge C_{\text{band}} \log \mu$ and hence

$$q_\theta(\alpha) \lesssim \mu^{-C_{\text{band}}} \tag{42}$$

uniformly in α. By the bias expansion, $B_\theta(\alpha) = q_\theta(\alpha)(1-\alpha) - \frac{1}{\mu} + O(e^{-\rho\mu})$, so, since $C_{\text{band}} \ge 2$,

$$B_\theta(\alpha) \le -\frac{c}{\mu}, \qquad |B_\theta(\alpha)| \lesssim \frac{1}{\mu}. \tag{43}$$

Also $p_\theta \geq 1/2$ for large μ, and the component-variance bounds imply $\sigma_\theta^2(\alpha) \geq p_\theta(\alpha) s_{1,\mu}^2(\alpha) \gtrsim \mu^{-2}$. Since $\lambda = k\mu \geq \ell\mu$, we have

$$J_\theta(\alpha) = \mu p_\theta q_\theta d_\mu + \lambda \sigma_\theta^2(\alpha) \geq \lambda \sigma_\theta^2(\alpha) \gtrsim \frac{1}{\mu}. \tag{44}$$

Combining equation 43 and equation 44 gives $\int_\rho^\ell B_\theta(\alpha) J_\theta(\alpha)\, d\alpha \leq -c/\mu^2$. Therefore, by equation 38, $\dot{k} \leq -\frac{c_U}{\mu^4}$ after naming the constant.

Furthermore, equation 43 gives $\partial_\mu \mathcal{R}_\rho < 0$, and hence $\dot{\mu} > 0$. For the upper bound, using equation 42 and the mixture variance identity, $\sigma_\theta^2(\alpha) \lesssim \mu^{-2} + q_\theta(\alpha) \lesssim \mu^{-2}$. Thus $0 < \dot{\mu} \lesssim \int_\rho^\ell |B_\theta(\alpha)| \sigma_\theta^2(\alpha)\, d\alpha \lesssim \mu^{-3}$. This proves the upper outer estimates.

Now fix $\kappa > 0$ and assume $\kappa \leq k < L_{\mathrm{band}}(\mu)$. Then $\mu(k+\alpha-1) \leq -C_{\mathrm{band}} \log \mu,\ \alpha \in [\rho, \ell]$, and hence $p_\theta(\alpha) \lesssim \mu^{-C_{\mathrm{band}}}$ and $q_\theta(\alpha) = 1 + O(\mu^{-C_{\mathrm{band}}})$ uniformly in α. The bias expansion gives $B_\theta(\alpha) = (1-\alpha) - \frac{1}{\mu} + O(\mu^{-C_{\mathrm{band}}})$, so, since $1 - \alpha \geq \rho$,

$$B_\theta(\alpha) \geq c_\rho > 0 \tag{45}$$

for all sufficiently large μ.

Also $q_\theta \geq 1/2$ for large μ, and therefore $\sigma_\theta^2(\alpha) \geq q_\theta(\alpha) s_{2,\mu}^2(\alpha) \gtrsim \mu^{-2}$. Since $\lambda = k\mu \geq \kappa\mu$, $J_\theta(\alpha) \geq \lambda \sigma_\theta^2(\alpha) \gtrsim_\kappa \frac{1}{\mu}$. Thus $\int_\rho^\ell B_\theta(\alpha) J_\theta(\alpha)\, d\alpha \gtrsim_\kappa \mu^{-1}$. Using equation 38, we obtain $\dot{k} \geq \frac{c_L(\kappa)}{\mu^3}$ after naming the constant. Finally, equation 45 gives $\partial_\mu \mathcal{R}_\rho > 0$, and hence $\dot{\mu} = -\partial_\mu \mathcal{R}_\rho < 0$. This proves the lower outer estimates. □

Corollary C.4 (Upper moving-boundary correction)**.** *Assume $C_{\mathrm{band}} \geq 2$, and let c_U, C_U, μ_U be the constants from Proposition C.3. There exist constants $a_U, \mu_{U,\partial} > 0$, depending only on ρ and C_{band}, such that, whenever $\mu \geq \mu_{U,\partial},\ k \geq U_{\mathrm{band}}(\mu)$, one has $\dot{k} - U'_{\mathrm{band}}(\mu)\dot{\mu} \leq -\frac{a_U}{\mu^4}$, and $0 < \dot{\mu} \leq \frac{C_U}{\mu^3}$.*

Proof. By Proposition C.3, in the region $k \geq U_{\mathrm{band}}(\mu),\ \dot{k} \leq -\frac{c_U}{\mu^4},\ 0 < \dot{\mu} \leq \frac{C_U}{\mu^3}$. Since $U'_{\mathrm{band}}(\mu) = C_{\mathrm{band}} \frac{1 - \log \mu}{\mu^2}$, we have $|U'_{\mathrm{band}}(\mu)| \lesssim \frac{\log \mu}{\mu^2}$. Therefore $\dot{k} - U'_{\mathrm{band}}(\mu)\dot{\mu} \leq -c_U/\mu^4 + C(\log \mu)/\mu^5 \leq -c_U/(2\mu^4)$ for all sufficiently large μ. Taking $a_U := c_U/2$ and increasing the threshold to $\mu_{U,\partial}$ proves the claim. □

Proposition C.5 (Stopped finite-time entry into the dynamic band)**.** *Assume $C_{\mathrm{band}} \geq 2$.*

(i) *There exists $\mu_+ > 1$, depending only on ρ and C_{band}, such that if $k(0) > U_{\mathrm{band}}(\mu(0)),\ \mu(0) \geq \mu_+$, then*

$$\tau_{\mathrm{band}}^+ := \inf\{t \geq 0 : k(t) \leq U_{\mathrm{band}}(\mu(t))\} < \infty.$$

(ii) *For every fixed $\kappa > 0$, there exists $\mu_-(\kappa) > 1$, depending only on ρ, C_{band}, and κ, such that, if $\kappa \leq k(0) < L_{\mathrm{band}}(\mu(0)),\ \mu(0) > \mu_-(\kappa)$, and*

$$\tau_{\mathrm{band}}^- := \inf\{t \geq 0 : k(t) \geq L_{\mathrm{band}}(\mu(t))\}, \qquad \tau_{\mathrm{band},-}^{\mathrm{exit}} := \inf\{t \geq 0 : \mu(t) \leq \mu_-(\kappa)\},$$

then

$$\tau_{\mathrm{band}}^- \wedge \tau_{\mathrm{band},-}^{\mathrm{exit}} \leq \frac{\mu(0)^3}{c_L(\kappa)} \{L_{\mathrm{band}}(\mu(0)) - k(0)\}. \tag{46}$$

In particular, if $\tau_{\mathrm{band},-}^{\mathrm{exit}} > \tau_{\mathrm{band}}^-$, then $\tau_{\mathrm{band}}^- < \infty$ and the same bound holds with $\tau_{\mathrm{band}}^- \wedge \tau_{\mathrm{band},-}^{\mathrm{exit}}$ replaced by τ_{band}^-.

Proof. **Upper outer region.** Choose μ_+ larger than the threshold in Corollary C.4. Set $g_U(t) := k(t) - U_{\mathrm{band}}(\mu(t))$. Before τ^+_{band}, one has $g_U(t) > 0$, hence $k(t) > U_{\mathrm{band}}(\mu(t))$. Corollary C.4 gives

$$\dot{g}_U(t) = \dot{k}(t) - U'_{\mathrm{band}}(\mu(t))\dot{\mu}(t) \le -\frac{a_U}{\mu(t)^4}, \qquad 0 < \dot{\mu}(t) \le \frac{C_U}{\mu(t)^3}.$$

In particular, $\mu(t) \ge \mu(0) \ge \mu_+$ before τ^+_{band}, so the estimates remain valid throughout the pre-entry interval. Lemma C.1(ii) therefore implies that g_U reaches zero in finite time. Hence $\tau^+_{\mathrm{band}} < \infty$. **Lower outer region.** Fix $\kappa > 0$. Choose $\mu_-(\kappa) > e$ sufficiently large that the lower estimates in Proposition C.3 hold whenever $\mu \ge \mu_-(\kappa),\ \kappa \le k < L_{\mathrm{band}}(\mu)$. Set $g_L(t) := L_{\mathrm{band}}(\mu(t)) - k(t),\ \tau_\kappa := \inf\{t \ge 0 : k(t) < \kappa\}$. Before $\tau^-_{\mathrm{band}} \wedge \tau^{\mathrm{exit}}_{\mathrm{band},-} \wedge \tau_\kappa$, one has $\kappa \le k(t) < L_{\mathrm{band}}(\mu(t)),\ \mu(t) > \mu_-(\kappa)$. Therefore Proposition C.3 gives $\dot{k}(t) \ge \frac{c_L(\kappa)}{\mu(t)^3},\ \dot{\mu}(t) < 0$. In particular, k is strictly increasing on this stopped interval, so τ_κ cannot be the first stopping time.

Since $L'_{\mathrm{band}}(\mu) = C_{\mathrm{band}}\dfrac{\log\mu - 1}{\mu^2} > 0$ for $\mu > \mu_-(\kappa)$, we have $\dot{g}_L(t) = L'_{\mathrm{band}}(\mu(t))\dot{\mu}(t) - \dot{k}(t) \le -c_L(\kappa)/\mu(t)^3$. Moreover, $\dot{\mu}(t) < 0$, so $\mu(t) \le \mu(0)$ on the stopped interval. Lemma C.1(i) gives $\tau^-_{\mathrm{band}} \wedge \tau^{\mathrm{exit}}_{\mathrm{band},-} \wedge \tau_\kappa \le \mu(0)^3 g_L(0)/c_L(\kappa)$. Since τ_κ cannot occur before $\tau^-_{\mathrm{band}} \wedge \tau^{\mathrm{exit}}_{\mathrm{band},-}$, this implies $\tau^-_{\mathrm{band}} \wedge \tau^{\mathrm{exit}}_{\mathrm{band},-} \le \mu(0)^3 g_L(0)/c_L(\kappa)$. Finally, $g_L(0) = L_{\mathrm{band}}(\mu(0)) - k(0)$, so this is equation 46. The final assertion is immediate. □

C.1.2 Transport through the dynamic band

This subsection proves that a trajectory in the lower portion of the dynamic band is transported successively through the lower boundary layer and the fixed interior strip, after which it is captured by a forward-invariant upper rail tube. **Lower boundary-layer escape** The finite-time entry result places the trajectory inside the dynamic band. Entry from below gives only $k(t) \ge L_{\mathrm{band}}(\mu(t))$, so the trajectory may first lie in the lower layer $\mathcal{L}_\eta(\mu)$. The goal of this subsection is to prove finite-time entrance into the fixed interior strip $\mathcal{I}_\eta$.

Proposition C.6 (Lower boundary-layer drift estimates)**.** *Assume $0 < \eta < \dfrac{\Delta}{4}$. There exist constants $c_{\mathrm{low},\eta}, C_{\mathrm{low},\eta}, \mu_{\mathrm{low},\eta} > 0$, depending only on $\rho, \eta, C_{\mathrm{band}}$, such that, whenever $\mu \ge \mu_{\mathrm{low},\eta},\ k \in [L_{\mathrm{band}}(\mu), \rho + \eta]$, one has $\dot{k} \ge \dfrac{c_{\mathrm{low},\eta}}{\mu^3}$, and*

$$\dot{\mu} < 0, \qquad |\dot{\mu}| \le \frac{C_{\mathrm{low},\eta}}{\mu}. \tag{47}$$

Moreover,

$$\frac{(-\dot{\mu})_+}{\mu} \le C_{\mathrm{low},\eta}\dot{k}. \tag{48}$$

Proof. All implicit constants may depend on $\rho, \eta, C_{\mathrm{band}}$. Increase $\mu_{\mathrm{low},\eta}$ so that, for $\mu \ge \mu_{\mathrm{low},\eta}$, $L_{\mathrm{band}}(\mu) \ge \frac{\rho}{2},\ \mu_{\mathrm{low},\eta} > e$. Thus

$$\frac{\rho}{2} \le k \le \rho + \eta. \tag{49}$$

Set $z := \mu(k + \alpha - 1)$. The odds identity gives

$$q_\theta(\alpha) \asymp \frac{1}{1+e^z}, \qquad p_\theta(\alpha)q_\theta(\alpha) \lesssim \frac{e^z}{(1+e^z)^2}, \tag{50}$$

uniformly for $\alpha \in [\rho, \ell]$. Hence, by the change of variables $z = \mu(k + \alpha - 1)$,

$$\int_\rho^\ell p_\theta(\alpha)q_\theta(\alpha)\,d\alpha \lesssim \frac{1}{\mu}. \tag{51}$$

The uniform component asymptotics imply

$$\begin{aligned} B_\theta(\alpha) &= q_\theta(\alpha)(1-\alpha) - \frac{1}{\mu} + O(e^{-\rho\mu}), \\ |B_\theta(\alpha)| &\lesssim 1, \\ \sigma_\theta^2(\alpha) &\lesssim \frac{1}{\mu^2} + p_\theta(\alpha)q_\theta(\alpha), \\ 0 \le J_\theta(\alpha) &\lesssim \mu p_\theta(\alpha)q_\theta(\alpha) + \frac{1}{\mu}. \end{aligned} \tag{52}$$

Here the last bound follows from $J_\theta = \mu p_\theta q_\theta d_\mu + k\mu\sigma_\theta^2$ and equation 49.

Let $J_\eta := [\rho, \ell - 2\eta]$. Since $\eta < \Delta/4$, this interval has positive length. For $\alpha \in J_\eta$, $k + \alpha - 1 \le (\rho + \eta) + (\ell - 2\eta) - 1 = -\eta$. Thus $p_\theta(\alpha) \lesssim e^{-\eta\mu}$ and $q_\theta(\alpha) = 1 + O(e^{-\eta\mu})$ on J_η. Since $1 - \alpha \ge \rho + 2\eta$, the bias expansion gives $B_\theta(\alpha) \gtrsim 1$. Also $\sigma_\theta^2(\alpha) \gtrsim \mu^{-2}$ on J_η. Using $\lambda = k\mu \ge \rho\mu/2$ and $\mathrm{Cov}_\theta(X, Y \mid \alpha) \le 0$, $J_\theta(\alpha) = -H_\theta(\alpha) \ge \lambda\sigma_\theta^2(\alpha) \gtrsim \mu^{-1}$. Consequently,

$$\int_{J_\eta} B_\theta(\alpha)J_\theta(\alpha)\, d\alpha \gtrsim \frac{1}{\mu}. \tag{53}$$

Now define $E_+ := \{\alpha : B_\theta(\alpha) \ge 0\}$, $E_- := \{\alpha : B_\theta(\alpha) < 0\}$. If $\alpha \in E_-$, then the bias expansion and $1 - \alpha \ge \rho$ imply

$$q_\theta(\alpha) \le \frac{C}{\mu}, \qquad |B_\theta(\alpha)| \lesssim \frac{1}{\mu}. \tag{54}$$

By equation 50, the set $\{q_\theta \le C/\mu\}$ is contained in $\{z \ge \log\mu - O(1)\}$, and therefore

$$\int_{\{q_\theta \le C/\mu\}} q_\theta(\alpha)\, d\alpha \lesssim \frac{1}{\mu^2}. \tag{55}$$

Combining equation 52, equation 54, and equation 55 gives

$$\int_{E_-} |B_\theta(\alpha)|J_\theta(\alpha)\, d\alpha \lesssim \frac{1}{\mu^2}, \tag{56}$$

and

$$\int_{E_-} |B_\theta(\alpha)|\sigma_\theta^2(\alpha)\, d\alpha \lesssim \frac{1}{\mu^3}. \tag{57}$$

Since $J_\theta \ge 0$, the nonnegative-bias region outside J_η contributes nonnegatively to $\int B_\theta J_\theta$. Hence equation 53 and equation 56 imply $\int_\rho^\ell B_\theta(\alpha)J_\theta(\alpha)\, d\alpha \ge c/\mu - C/\mu^2 \ge c/\mu$ after increasing $\mu_{\mathrm{low},\eta}$. By equation 38, $\dot{k} \ge \frac{c_{\mathrm{low},\eta}}{\mu^3}$. It remains to control $\dot{\mu}$ and the relative radial loss. Set

$$\begin{aligned} A_+ &:= \int_{E_+} B_\theta(\alpha)\sigma_\theta^2(\alpha)\, d\alpha, \qquad A_- := \int_{E_-} |B_\theta(\alpha)|\sigma_\theta^2(\alpha)\, d\alpha, \\ K_+ &:= \int_{E_+} B_\theta(\alpha)J_\theta(\alpha)\, d\alpha, \qquad K_- := \int_{E_-} |B_\theta(\alpha)|J_\theta(\alpha)\, d\alpha. \end{aligned}$$

The estimates above give $K_+ \gtrsim \mu^{-1}$, $K_- \lesssim \mu^{-2}$, and $K_+ - K_- \ge K_+/2$ for sufficiently large μ. Moreover, on E_+, $J_\theta \ge k\mu\sigma_\theta^2 \ge (\rho\mu/2)\sigma_\theta^2$, so $K_+ \ge (\rho\mu/2)A_+$. Since $-\dot{\mu} = (A_+ - A_-)/\Delta$ and $\dot{k} = (K_+ - K_-)/(\Delta\mu^2)$, we get $(-\dot{\mu})_+/\mu \le A_+/(\Delta\mu) \le 2K_+/(\Delta\rho\mu^2) \le (4/\rho)\dot{k}$. After increasing $C_{\mathrm{low},\eta}$, this proves equation 48.

Finally, $\int_{J_\eta} B_\theta(\alpha)\sigma_\theta^2(\alpha)\, d\alpha \gtrsim \mu^{-2}$, while the negative radial tail is $O(\mu^{-3})$ by equation 57. Hence $\int_\rho^\ell B_\theta(\alpha)\sigma_\theta^2(\alpha)\, d\alpha > 0$ for sufficiently large μ, and therefore $\dot{\mu} = -\Delta^{-1}\int_\rho^\ell B_\theta(\alpha)\sigma_\theta^2(\alpha)\, d\alpha < 0$. The size bound follows from equation 52 and equation 51:

$$|\dot{\mu}| \lesssim \int_\rho^\ell |B_\theta(\alpha)| \left(\frac{1}{\mu^2} + p_\theta(\alpha)q_\theta(\alpha)\right) d\alpha \lesssim \frac{1}{\mu}.$$

This proves equation 47. □

Corollary C.7 (Finite-time entrance into the interior strip)**.** *Assume* $0 < \eta < \dfrac{\Delta}{4}$. *Let* t_0 *satisfy* $\mu(t_0) > \mu_{\mathrm{low},\eta},\ k(t_0) \in [L_{\mathrm{band}}(\mu(t_0)), \rho + \eta]$. *Define*

$$\tau_{\mathrm{int}} := \inf\{t \geq t_0 : k(t) \geq \rho + \eta\}, \qquad \tau_{\mathrm{low}} := \inf\{t \geq t_0 : \mu(t) \leq \mu_{\mathrm{low},\eta}\}.$$

Then

$$\tau_{\mathrm{int}} \wedge \tau_{\mathrm{low}} - t_0 \leq \frac{\mu(t_0)^3}{c_{\mathrm{low},\eta}}\big(\rho + \eta - k(t_0)\big). \tag{58}$$

In particular, if $\tau_{\mathrm{low}} > \tau_{\mathrm{int}}$*, then* $\tau_{\mathrm{int}} < \infty$ *and*

$$\tau_{\mathrm{int}} - t_0 \leq \frac{\mu(t_0)^3}{c_{\mathrm{low},\eta}}\big(\rho + \eta - k(t_0)\big). \tag{59}$$

Proof. If $k(t_0) = \rho + \eta$, then $\tau_{\mathrm{int}} = t_0$ and the claim is immediate. Assume $k(t_0) < \rho + \eta$. First we show that the trajectory cannot exit below $L_{\mathrm{band}}(\mu)$ before $\tau_{\mathrm{int}} \wedge \tau_{\mathrm{low}}$. Set $h(t) := k(t) - L_{\mathrm{band}}(\mu(t))$.

At any point with $h(t) = 0,\ t < \tau_{\mathrm{int}} \wedge \tau_{\mathrm{low}}$, one has $k(t) = L_{\mathrm{band}}(\mu(t)) \leq \rho + \eta,\ \mu(t) > \mu_{\mathrm{low},\eta}$. Thus Proposition C.6 gives $\dot{k}(t) \geq \frac{c_{\mathrm{low},\eta}}{\mu(t)^3},\ \dot{\mu}(t) < 0$. Since $L'_{\mathrm{band}}(\mu) > 0$ for $\mu > \mu_{\mathrm{low},\eta} > e$, it follows that $\dot{h}(t) = \dot{k}(t) - L'_{\mathrm{band}}(\mu(t))\dot{\mu}(t) > 0$. Hence h cannot cross from nonnegative to negative before $\tau_{\mathrm{int}} \wedge \tau_{\mathrm{low}}$. Therefore $k(t) \in [L_{\mathrm{band}}(\mu(t)), \rho + \eta)$ for every $t < \tau_{\mathrm{int}} \wedge \tau_{\mathrm{low}}$.

Now set $g_{\mathrm{low}}(t) := \rho + \eta - k(t)$. On $[t_0, \tau_{\mathrm{int}} \wedge \tau_{\mathrm{low}})$, Proposition C.6 gives $\dot{g}_{\mathrm{low}}(t) = -\dot{k}(t) \leq -c_{\mathrm{low},\eta}/\mu(t)^3$ and $\dot{\mu}(t) < 0$. Therefore $\mu(t) \leq \mu(t_0)$ on this interval. Lemma C.1(i) gives $\tau_{\mathrm{int}} \wedge \tau_{\mathrm{low}} - t_0 \leq \mu(t_0)^3 g_{\mathrm{low}}(t_0)/c_{\mathrm{low},\eta}$. Since $g_{\mathrm{low}}(t_0) = \rho + \eta - k(t_0)$, this proves equation 58. If $\tau_{\mathrm{low}} > \tau_{\mathrm{int}}$, then the stopped time equals τ_{int}, giving equation 59. □

The large-parameter preservation estimate later rules out $\tau_{\mathrm{low}} \leq \tau_{\mathrm{int}}$ under the initialization assumptions of the global dynamics theorem. **Interior Laplace asymptotics and the slingshot mechanism** Once the trajectory reaches the fixed interior strip $\mathcal{I}_\eta$, the transition point $\alpha_c = 1 - k$ remains uniformly separated from the endpoints of the task interval. The gradient then admits a uniform Laplace expansion, which shows that k and λ increase while μ decreases. **Uniform interior Laplace expansion.**

Lemma C.8 (Uniform interior Laplace expansion)**.** *Fix* $0 < \eta < \dfrac{\Delta}{4}$. *Uniformly for* $k \in [\rho+\eta, \ell-\eta]$, *as* $\mu \to \infty$,

$$\dot{\mu} = -\frac{k^3}{2\Delta\mu} + O_\eta(\mu^{-2}), \qquad \dot{\lambda} = \frac{k^2}{2\Delta\mu} + O_\eta(\mu^{-2}). \tag{60}$$

Consequently,

$$\dot{k} = \frac{k^2 + k^4}{2\Delta\mu^2} + O_\eta(\mu^{-3}). \tag{61}$$

Proof. Set $\alpha_c := 1 - k$. For $k \in [\rho + \eta, \ell - \eta]$, one has $\alpha_c \in [\rho + \eta, \ell - \eta]$, so α_c remains at least distance η from both endpoints. Introduce $z := \mu(\alpha - \alpha_c),\ \alpha = \alpha_c + \frac{z}{\mu}$, and define $Q(z) := \frac{1}{1+e^z},\ \mathfrak{l}(z) := \frac{e^z}{(1+e^z)^2}$. The range of z is $I_{\mu,k} := [\mu(\rho - \alpha_c), \mu(\ell - \alpha_c)]$, with $[-\eta\mu, \eta\mu] \subseteq I_{\mu,k}$. Consequently, for every fixed $m \geq 0$,

$$\sup_{k \in [\rho+\eta, \ell-\eta]} \int_{\mathbb{R}\setminus I_{\mu,k}} (1 + |z|)^m \mathfrak{l}(z)\, dz = O_{\eta,m}(e^{-c_\eta \mu}) \tag{62}$$

for some $c_\eta > 0$.

By equation 28,

$$\frac{p_\theta(\alpha)}{q_\theta(\alpha)} = e^z h_\mu(\alpha), \qquad h_\mu(\alpha) := \frac{1 - e^{-\mu\alpha}}{1 - e^{-\mu(1-\alpha)}}.$$

Uniformly for $\alpha \in [\rho, \ell]$,

$$h_\mu(\alpha) = 1 + O(e^{-\rho\mu})$$
$$q_\theta(\alpha) = Q(z) + O(e^{-\rho\mu}\mathfrak{l}(z))$$
$$p_\theta(\alpha)q_\theta(\alpha) = \mathfrak{l}(z) + O(e^{-\rho\mu}\mathfrak{l}(z)).$$

By Lemma B.1, $d_\mu(\alpha) = 1 - \alpha + O(e^{-\rho\mu}) = k - \dfrac{z}{\mu} + O(e^{-\rho\mu})$. Corollary B.2 gives

$$B_\theta(\alpha) = q_\theta(\alpha)(1-\alpha) - \frac{1}{\mu} + O(e^{-\rho\mu}) = kQ(z) + O_\eta\left(\frac{1+|z|}{\mu}\right) + O(e^{-\rho\mu}).$$

The error is integrable after multiplication by $\mathfrak{l}(z)$, since $\int_{\mathbb{R}}(1+|z|)^m\mathfrak{l}(z)\,dz < \infty$ for every fixed $m \geq 0$.

Using $\mathrm{Cov}_\theta(X, Y \mid \alpha) = -p_\theta(\alpha)q_\theta(\alpha)d_\mu(\alpha)$, we obtain

$$B_\theta(\alpha)\mathrm{Cov}_\theta(X, Y \mid \alpha) = -k^2Q(z)\mathfrak{l}(z) + O_\eta\left(\frac{1+|z|}{\mu}\mathfrak{l}(z)\right) + O_\eta(e^{-\rho\mu}\mathfrak{l}(z)).$$

Using $d\alpha = dz/\mu$ and equation 62 to replace $I_{\mu,k}$ by $\mathbb{R}$, we find

$$\int_\rho^\ell B_\theta(\alpha)\mathrm{Cov}_\theta(X, Y \mid \alpha)\,d\alpha = -\frac{k^2}{\mu}\int_{\mathbb{R}} Q(z)\mathfrak{l}(z)\,dz + O_\eta(\mu^{-2}). \tag{63}$$

For the variance, the component-variance contribution satisfies

$$\int_\rho^\ell |B_\theta(\alpha)|\left(p_\theta(\alpha)s_{1,\mu}^2(\alpha) + q_\theta(\alpha)s_{2,\mu}^2(\alpha)\right)d\alpha = O_\eta(\mu^{-2}),$$

because the bias is uniformly bounded and $s_{1,\mu}^2(\alpha) + s_{2,\mu}^2(\alpha) = O(\mu^{-2})$ uniformly in α. The mixture contribution satisfies

$$B_\theta(\alpha)p_\theta(\alpha)q_\theta(\alpha)d_\mu(\alpha)^2 = k^3Q(z)\mathfrak{l}(z) + O_\eta\left(\frac{1+|z|}{\mu}\mathfrak{l}(z)\right) + O_\eta(e^{-\rho\mu}\mathfrak{l}(z)).$$

Consequently,

$$\int_\rho^\ell B_\theta(\alpha)\sigma_\theta^2(\alpha)\,d\alpha = \frac{k^3}{\mu}\int_{\mathbb{R}} Q(z)\mathfrak{l}(z)\,dz + O_\eta(\mu^{-2}). \tag{64}$$

With $r := \dfrac{e^z}{1+e^z}$, $dr = \mathfrak{l}(z)\,dz$, one has $Q(z) = 1 - r$, and therefore $\int_{\mathbb{R}} Q(z)\mathfrak{l}(z)\,dz = \int_0^1(1-r)\,dr = \frac{1}{2}$. Substituting into equation 63 and equation 64, and using Lemma 5.1, proves equation 60. Finally, $\dot{k} = \dfrac{\dot{\lambda} - k\dot{\mu}}{\mu}$, which gives equation 61. □

Quantitative interior drift.

Proposition C.9 (Interior slingshot). *Fix $0 < \eta < \frac{\Delta}{4}$. There exist constants $c_{\mathrm{int},\eta} > 0$ and $\mu_{\mathrm{int},\eta} > 0$ such that, whenever $\mu \geq \mu_{\mathrm{int},\eta}$, $k \in [\rho+\eta, \ell-\eta]$, one has $\dot{k} \geq \dfrac{c_{\mathrm{int},\eta}}{\mu^2} > 0$, together with $\dot{\mu} < 0$, and $\dot{\lambda} > 0$. Moreover, for a possibly larger constant $C_{\mathrm{int},\eta} > 0$,*

$$|\dot{\mu}| \leq \frac{C_{\mathrm{int},\eta}}{\mu}, \qquad |\dot{\lambda}| \leq \frac{C_{\mathrm{int},\eta}}{\mu}. \tag{65}$$

After increasing $C_{\mathrm{int},\eta}$ if necessary, one also has

$$\frac{(-\dot{\mu})_+}{\mu} \leq C_{\mathrm{int},\eta}\dot{k}. \tag{66}$$

Proof. Since $\rho + \eta \le k \le \ell - \eta$, the coefficients $k^3/2\Delta,\ k^2/2\Delta,\ (k^2+k^4)/2\Delta$ are uniformly bounded below by positive constants. The uniform remainder estimates in Lemma C.8 can therefore be absorbed into the leading terms by increasing $\mu_{\mathrm{int},\eta}$. This gives $\dot\mu < 0,\ \dot\lambda > 0,\ \dot k \ge \frac{c_{\mathrm{int},\eta}}{\mu^2}$. The size bounds equation 65 follow directly from equation 60, since k is uniformly bounded on the strip. Therefore $\frac{(-\dot\mu)_+}{\mu} \le \frac{C}{\mu^2} \le C_{\mathrm{int},\eta}\dot k$ for a possibly larger $C_{\mathrm{int},\eta}$, proving equation 66. □

Finite-time traversal of the interior strip.

Corollary C.10 (Finite-time traversal of the interior strip)**.** *Fix $0 < \eta < \dfrac{\Delta}{4}$. Suppose that, at some time t_0, $k(t_0) \in [\rho+\eta, \ell-\eta]$ and $\mu(t_0) > \mu_{\mathrm{int},\eta}$. Define*

$$\tau_{\mathrm{up}} := \inf\{t \ge t_0 : k(t) \ge \ell - \eta\}, \qquad \tau_{\mathrm{int,exit}} := \inf\{t \ge t_0 : \mu(t) \le \mu_{\mathrm{int},\eta}\}.$$

Then

$$\tau_{\mathrm{up}} \wedge \tau_{\mathrm{int,exit}} - t_0 \le \frac{\mu(t_0)^2\{\ell - \eta - k(t_0)\}}{c_{\mathrm{int},\eta}}. \tag{67}$$

In particular, if $\tau_{\mathrm{int,exit}} > \tau_{\mathrm{up}}$, then $\tau_{\mathrm{up}} < \infty$ and

$$\tau_{\mathrm{up}} - t_0 \le \frac{\mu(t_0)^2\{\ell - \eta - k(t_0)\}}{c_{\mathrm{int},\eta}}. \tag{68}$$

Proof. If $k(t_0) = \ell - \eta$, the claim is immediate. Otherwise, the lower boundary $k = \rho + \eta$ is inward-pointing by Proposition C.9; hence the trajectory remains in the interior strip until $\tau_{\mathrm{up}} \wedge \tau_{\mathrm{int,exit}}$.

On this stopped interval, with $g_{\mathrm{int}}(t) := \ell - \eta - k(t)$, $\dot g_{\mathrm{int}}(t) \le -\dfrac{c_{\mathrm{int},\eta}}{\mu(t)^2},\ \dot\mu(t) < 0$. Lemma C.1(i) gives equation 67. If $\tau_{\mathrm{int,exit}} > \tau_{\mathrm{up}}$, the stopped time equals τ_{up}, proving equation 68. □

Extension toward the upper ratio rail. The fixed-strip expansion is sufficient for finite-time traversal to $k = \ell - \eta$. The following extension remains valid when the transition point approaches the endpoint, provided the separation is at least logarithmic.

Lemma C.11 (Interior expansion with logarithmic endpoint separation)**.** *Fix $\eta_1 \in (0, \Delta)$ and $\beta > 0$. Uniformly for $\ell - \eta_1 \le k \le \ell - \dfrac{\beta \log \mu}{\mu}$, as $\mu \to \infty$,*

$$\dot\mu = -\frac{k^3}{2\Delta\mu} + O_{\eta_1,\beta}\left(\mu^{-2} + \mu^{-\beta-1}\right), \tag{69}$$

$$\dot\lambda = \frac{k^2}{2\Delta\mu} + O_{\eta_1,\beta}\left(\mu^{-2} + \mu^{-\beta-1}\right), \tag{70}$$

and

$$\dot k = \frac{k^2 + k^4}{2\Delta\mu^2} + O_{\eta_1,\beta}\left(\mu^{-3} + \mu^{-\beta-2}\right). \tag{71}$$

In particular, there exist constants $c_{\eta_1} > 0$ and $\mu_{\mathrm{sep}}(\eta_1, \beta) < \infty$ such that $\dot k \ge \dfrac{c_{\eta_1}}{\mu^2}$ and $\dot\mu < 0$, throughout this region whenever $\mu \ge \mu_{\mathrm{sep}}(\eta_1, \beta)$.

Proof. For sufficiently large μ, the interval in the statement is nonempty because $\beta \log \mu/\mu < \eta_1$. The proof follows the rescaling argument of Lemma C.8. The only difference is that the transition point $\alpha_c = 1 - k$ may now approach the lower task endpoint ρ at logarithmic distance.

Set $\alpha_c := 1 - k$ and $z := \mu(\alpha - \alpha_c)$. The rescaled task interval is $I_{\mu,k} = [a_{\mu,k}, b_{\mu,k}]$, where $a_{\mu,k} := \mu(\rho - \alpha_c) = \mu(k - \ell)$ and $b_{\mu,k} := \mu(\ell - \alpha_c) = \mu(k - \rho)$. The upper bound on k gives $a_{\mu,k} \le -\beta \log \mu$, while the lower bound on k gives $b_{\mu,k} \ge (\Delta - \eta_1)\mu$. Recall $Q(z) := \frac{1}{1+e^z}$, $\mathfrak{l}(z) := \frac{e^z}{(1+e^z)^2}$. The omitted left tail satisfies

$$\int_{-\infty}^{a_{\mu,k}} \mathfrak{l}(z)\,dz \le \int_{-\infty}^{-\beta\log\mu} \mathfrak{l}(z)\,dz = \frac{\mu^{-\beta}}{1+\mu^{-\beta}} = O(\mu^{-\beta}),$$

whereas the omitted right tail satisfies

$$\int_{b_{\mu,k}}^{\infty} \mathfrak{l}(z)\,dz \le \int_{(\Delta-\eta_1)\mu}^{\infty} \mathfrak{l}(z)\,dz = \frac{1}{1+e^{(\Delta-\eta_1)\mu}} = O_{\eta_1}(e^{-(\Delta-\eta_1)\mu}).$$

Since $0 \le Q \le 1$, the same estimates control the omitted tails of $Q(z)\mathfrak{l}(z)$. After the change of variables $d\alpha = dz/\mu$, the left endpoint truncation contributes $O(\mu^{-\beta-1})$ to the expansions of $\dot\lambda$ and $\dot\mu$, while the right endpoint contribution is exponentially small.

The local expansion errors remain $O(\mu^{-2})$ uniformly in the stated region. Indeed, the argument of Lemma C.8 gives an integrated error bounded by

$$\frac{1}{\mu}\int_{I_{\mu,k}} C_{\eta_1}\frac{1+|z|}{\mu}\mathfrak{l}(z)\,dz \le \frac{C_{\eta_1}}{\mu^2}\int_{\mathbb{R}}(1+|z|)\mathfrak{l}(z)\,dz = O_{\eta_1}(\mu^{-2}).$$

Combining the local expansion error with the endpoint truncation error gives equation 69 and equation 70. Finally, $\dot k = \frac{\dot\lambda - k\dot\mu}{\mu}$. Since k remains uniformly bounded in the stated region, substitution gives equation 71. Moreover, $k \ge \ell - \eta_1 > \rho > 0$, so the leading coefficients in $\dot\mu$ and $\dot k$ are uniformly bounded away from zero. Since $\mu^{-2} + \mu^{-\beta-1} = o(\mu^{-1}),\ \mu^{-3} + \mu^{-\beta-2} = o(\mu^{-2})$, the remainders can be absorbed into the leading terms after increasing $\mu_{\mathrm{sep}}(\eta_1,\beta)$. Hence $\dot\mu < 0,\ \dot k \ge \frac{c_{\eta_1}}{\mu^2}$. □

Upper rail-tube capture and invariance Recall $\mathcal{T}_\eta(\mu) := [\ell-\eta, U_{\mathrm{band}}(\mu)]$. Its fixed lower boundary is controlled by the interior slingshot, whereas its moving upper boundary is controlled by the outer ratio correction.

Proposition C.12 (Forward invariance of the upper rail tube)**.** *Assume $C_{\mathrm{band}} \ge 2$, and fix $0 < \eta < \Delta/4$. There exists $\mu_{\mathrm{tube},\eta} > 0$ such that, if $k(t_0) \in \mathcal{T}_\eta(\mu(t_0)),\ \mu(t_0) > \mu_{\mathrm{tube},\eta}$, and $\tau_{\mathrm{tube}} := \inf\{t \ge t_0 : \mu(t) \le \mu_{\mathrm{tube},\eta}\}$, then $k(t) \in \mathcal{T}_\eta(\mu(t))$ for every $t \in [t_0, \tau_{\mathrm{tube}})$.*

Proof. Choose $\mu_{\mathrm{tube},\eta}$ larger than $\mu_{\mathrm{int},\eta}$ and the threshold in Corollary C.4. Set $a(\mu) := \ell - \eta,\ b(\mu) := U_{\mathrm{band}}(\mu)$. At the lower boundary $k = a(\mu) = \ell - \eta$, Proposition C.9 gives $\dot k - a'(\mu)\dot\mu = \dot k > 0$. At the upper moving boundary $k = b(\mu) = U_{\mathrm{band}}(\mu)$, Corollary C.4 gives $\dot k - b'(\mu)\dot\mu = \dot k - U'_{\mathrm{band}}(\mu)\dot\mu < 0$. Lemma C.2 therefore gives $k(t) \in \mathcal{T}_\eta(\mu(t))$ for every $t \in [t_0, \tau_{\mathrm{tube}})$. □

C.1.3 Logarithmic endpoint dynamics

The upper rail tube provides dynamical trapping. A uniform $O(\mu^{-1})$ prediction bound begins after the ratio crosses a logarithmic rail slightly above ℓ: $k \ge b_c(\mu) = \ell + \frac{\log\mu + c}{\mu}$. This subsection proves the endpoint estimates underlying the logarithmic tube introduced in Section C. In particular, it proves the logarithmic-rail expansions for the functions d_k and d_{rad} defined in equation 37, verifies the choice of c_0, proves finite-time approach to the prediction rail, establishes forward invariance of the prediction-good tube, proves radial divergence after capture, and verifies preservation of the large-parameter regime during the transient stages.

Prediction and endpoint rescaling

Lemma C.13 (Uniform prediction bound above a logarithmic rail)**.** *Fix $c \in \mathbb{R}$. There exist constants $C_c, \mu_c > 0$ such that, whenever $\mu \ge \mu_c, k \ge b_c(\mu)$, one has*

$$\sup_{\alpha\in[\rho,\ell]} |F(\lambda,\mu;\alpha) - \alpha| \le \frac{C_c}{\mu}. \tag{72}$$

Proof. By Lemma B.1, $p_\theta(\alpha)/q_\theta(\alpha) = e^{\mu(k+\alpha-1)}(1-e^{-\mu\alpha})/(1-e^{-\mu(1-\alpha)})$. Since $\alpha, 1-\alpha \ge \rho$, the second factor is bounded below by $1/2$ for all sufficiently large μ, uniformly in $\alpha \in [\rho,\ell]$. If $k \ge b_c(\mu)$, then $\mu(k+\alpha-1) \ge \mu(k+\rho-1) \ge \log\mu + c$. Hence $p_\theta(\alpha)/q_\theta(\alpha) \ge e^c\mu/2$ and $q_\theta(\alpha) \le C_c/\mu$. The uniform bias expansion $B_\theta(\alpha) = q_\theta(\alpha)(1-\alpha) - \frac{1}{\mu} + O(e^{-\rho\mu})$ therefore gives equation 72. □

Lemma C.14 (Right-mass control near the lower task endpoint)**.** *Fix $c_{\min} \in \mathbb{R}$. Suppose $c \geq c_{\min},\ k = b_c(\mu)$. For $\alpha = \rho + \dfrac{u}{\mu},\ 0 \leq u \leq \Delta\mu$, one has, for every fixed $U < \infty$,*

$$\sup_{c \geq c_{\min}} \sup_{0 \leq u \leq U} \left| \mu q_\theta \left(\rho + \frac{u}{\mu} \right) - e^{-c-u} \right| \longrightarrow 0$$

as $\mu \to \infty$. Moreover, there exists $C = C(c_{\min})$ such that

$$\mu q_\theta \left(\rho + \frac{u}{\mu} \right) \leq C e^{-u} \tag{73}$$

for all sufficiently large μ, all $u \in [0, \Delta\mu]$, and all $c \geq c_{\min}$.

Proof. Write $h_\mu(\alpha) := \frac{1-e^{-\mu\alpha}}{1-e^{-\mu(1-\alpha)}}$. Uniformly for $\alpha \in [\rho, \ell]$, $h_\mu(\alpha) = 1 + O(e^{-\rho\mu})$. If $k = b_c(\mu)$ and $\alpha = \rho + u/\mu$, then $\mu(k + \alpha - 1) = \log \mu + c + u$. The odds identity gives

$$\mu q_\theta \left(\rho + \frac{u}{\mu} \right) = \frac{\mu}{1 + \mu e^{c+u} h_\mu(\rho + u/\mu)} = \frac{e^{-c-u}}{h_\mu(\rho + u/\mu) + \mu^{-1} e^{-c-u}}.$$

The first conclusion follows uniformly for $c \geq c_{\min}$ and bounded u. Since $h_\mu \geq 1/2$ for all sufficiently large μ, $\mu q_\theta(\rho + u/\mu) \leq 2e^{-c-u} \leq 2e^{-c_{\min}} e^{-u}$, which proves equation 73. □

Limiting dynamics on logarithmic rails For the endpoint analysis, define $s^2_{\mathrm{comp},\mu}(\alpha) := p_\theta(\alpha) s^2_{1,\mu}(\alpha) + q_\theta(\alpha) s^2_{2,\mu}(\alpha)$. Then

$$\sigma^2_\theta(\alpha) = s^2_{\mathrm{comp},\mu}(\alpha) + p_\theta(\alpha) q_\theta(\alpha) d_\mu(\alpha)^2.$$

Define

$$H_{\mathrm{mix}}(\alpha) := -\mu p_\theta(\alpha) q_\theta(\alpha) d_\mu(\alpha) \{1 + k d_\mu(\alpha)\}, \qquad H_{\mathrm{comp}}(\alpha) := -k\mu s^2_{\mathrm{comp},\mu}(\alpha),$$

$$S_{\mathrm{mix}} := \frac{1}{\Delta} \int_\rho^\ell B_\theta(\alpha) H_{\mathrm{mix}}(\alpha)\, d\alpha, \qquad S_{\mathrm{comp}} := \frac{1}{\Delta} \int_\rho^\ell B_\theta(\alpha) H_{\mathrm{comp}}(\alpha)\, d\alpha,$$

$$\dot\mu_{\mathrm{mix}} := -\frac{1}{\Delta} \int_\rho^\ell B_\theta(\alpha) p_\theta(\alpha) q_\theta(\alpha) d_\mu(\alpha)^2\, d\alpha, \qquad \dot\mu_{\mathrm{comp}} := -\frac{1}{\Delta} \int_\rho^\ell B_\theta(\alpha) s^2_{\mathrm{comp},\mu}(\alpha)\, d\alpha.$$

Thus, $S = S_{\mathrm{mix}} + S_{\mathrm{comp}},\ \dot\mu = \dot\mu_{\mathrm{mix}} + \dot\mu_{\mathrm{comp}}$.

Proposition C.15 (Uniform logarithmic-rail expansions)**.** *Let $I = [c_-, c_+] \subset \mathbb{R}$ be compact. Uniformly for $c \in I$ and $k = b_c(\mu)$,*

$$\sup_{c \in I} \left| \mu^2 S(\mu b_c(\mu), \mu) - d_k(c) \right| \longrightarrow 0, \tag{74}$$

$$\sup_{c \in I} \left| \mu^3 \dot\mu - d_{\mathrm{rad}}(c) \right| \longrightarrow 0, \tag{75}$$

and

$$\sup_{c \in I} \left| \mu^4 \dot k + d_k(c) \right| \longrightarrow 0. \tag{76}$$

In particular, for each fixed $c \in \mathbb{R}$,

$$S(\lambda, \mu) = \frac{d_k(c)}{\mu^2} + o_c(\mu^{-2}), \qquad \dot\mu = \frac{d_{\mathrm{rad}}(c)}{\mu^3} + o_c(\mu^{-3}),$$

and $\dot k = -\dfrac{d_k(c)}{\mu^4} + o_c(\mu^{-4})$.

Proof. Let $\alpha = \rho + u/\mu$, $k = b_c(\mu)$, and $q_{\mu,c}(u) := q_\theta(\rho + u/\mu)$. Lemma C.14, applied with $c_{\min} = c_-$, gives $\mu q_{\mu,c}(u) \to e^{-c-u}$ locally uniformly in $(c, u) \in I \times [0, \infty)$ and $\mu q_{\mu,c}(u) \leq C_I e^{-u}$; squaring the latter bound gives $\mu^2 q_{\mu,c}(u)^2 \leq C_I e^{-2u}$. Moreover, $p_\theta \to 1$, $1 - \alpha \to \ell$, $d_\mu(\alpha) \to \ell$, and $k \to \ell$ uniformly on bounded u-intervals, and the bias expansion gives $\mu B_\theta \left(\rho + \frac{u}{\mu} \right) \to$

$\ell e^{-c-u} - 1$. The rescaled mixture integrands for S_{mix} are dominated by a linear combination of e^{-u} and e^{-2u}. Uniform dominated convergence yields

$$\sup_{c\in I}\left|\mu^2 S_{\mathrm{mix}} - \left[\frac{\ell(1+\ell^2)}{\Delta}e^{-c} - \frac{\ell^2(1+\ell^2)}{2\Delta}e^{-2c}\right]\right| \longrightarrow 0.$$

Similarly,

$$\sup_{c\in I}\left|\mu^3\dot{\mu}_{\mathrm{mix}} - \left[\frac{\ell^2}{\Delta}e^{-c} - \frac{\ell^3}{2\Delta}e^{-2c}\right]\right| \longrightarrow 0.$$

Uniformly for $c \in I$, Lemma C.13 gives $|B_\theta(\alpha)| \le C_I/\mu$, while the component asymptotics give $s^2_{\mathrm{comp},\mu}(\alpha) = \mu^{-2} + o(\mu^{-2})$ uniformly in α. Moreover, the domination above gives $\int_\rho^\ell q_\theta(\alpha)\,d\alpha = O_I(\mu^{-2})$: indeed, with $\alpha = \rho + u/\mu$, one has $q_\theta(\alpha) \le C_I e^{-u}/\mu$ and $d\alpha = du/\mu$. Hence Corollary B.2 yields $\int_\rho^\ell B_\theta(\alpha)\,d\alpha = -\Delta/\mu + O_I(\mu^{-2})$. Therefore $\sup_{c\in I}|\mu^2 S_{\mathrm{comp}} - \ell| \to 0$ and $\sup_{c\in I}|\mu^3\dot{\mu}_{\mathrm{comp}} - 1| \to 0$. Adding the mixture and component contributions and using the definitions of d_k and d_{rad} proves equation 74 and equation 75. Finally, $\dot{k} = -S/\mu^2$, so equation 76 follows. The fixed-c asymptotics are the special case $I = \{c\}$. □

Selection of the prediction rail

Lemma C.16 (Choice of a prediction rail)**.** *There exists $c_0 \in \mathbb{R}$ satisfying $d_k(c_0) < 0,\ d_{\mathrm{rad}}(c_0) > 0$. Moreover, c_0 can be chosen so that*

$$d_k(c) < 0 \quad \textit{for every } c \le c_0, \qquad \text{and} \qquad \mathrm{d_{rad}(c) > 0} \quad \textit{for every } \mathrm{c \ge c_0}.$$

Proof. Set $x = e^{-c}$. Using equation 37, the functions d_k and d_{rad} are downward-opening quadratics in x:

$$d_k = \ell + \frac{\ell(1+\ell^2)}{\Delta}x - \frac{\ell^2(1+\ell^2)}{2\Delta}x^2 \quad \text{and} \quad d_{\mathrm{rad}} = 1 + \frac{\ell^2}{\Delta}x - \frac{\ell^3}{2\Delta}x^2.$$

Their positive roots are

$$x_k^\star = \frac{1}{\ell}\left(1 + \sqrt{1 + \frac{2\ell\Delta}{1+\ell^2}}\right) \quad \text{and} \quad x_\mu^\star = \frac{1}{\ell}\left(1 + \sqrt{1 + \frac{2\Delta}{\ell}}\right).$$

Since $0 < \ell < 1$, $2\ell\Delta/(1+\ell^2) < 2\Delta/\ell$, so $x_k^\star < x_\mu^\star$. Choose $x_- \in (x_k^\star, x_\mu^\star),\ c_0 := -\log x_-$. Then $d_k(c_0) < 0 < d_{\mathrm{rad}}(c_0)$. Since $c \le c_0$ is equivalent to $e^{-c} \ge x_-$, the downward-opening structure gives $d_k(c) < 0$ for every $c \le c_0$. Similarly, $c \ge c_0$ implies $e^{-c} \le x_- < x_\mu^\star$, and hence $d_{\mathrm{rad}}(c) > 0$ for every $c \ge c_0$. □

Approach to and invariance of the logarithmic tube

Proposition C.17 (Forward invariance of the prediction rail)**.** *There exist constants $a > 0$ and $\mu_{\log} > 0$ such that, whenever $\mu \ge \mu_{\log},\ k = b_{c_0}(\mu)$, one has*

$$\dot{k} - b'_{c_0}(\mu)\dot{\mu} \ge \frac{a}{\mu^4}. \tag{77}$$

Consequently, the region $k \ge b_{c_0}(\mu)$ is forward invariant from below while $\mu(t) \ge \mu_{\log}$.

Proof. By Lemma C.16, choose $a > 0$ such that $d_k(c_0) \le -4a$. Proposition C.15 gives, on the rail $k = b_{c_0}(\mu)$,

$$\dot{k} = -\frac{d_k(c_0)}{\mu^4} + o(\mu^{-4}) \ge \frac{2a}{\mu^4} \qquad \text{and} \qquad |\dot{\mu}| \le \frac{C}{\mu^3}$$

for all sufficiently large μ. Since $b'_{c_0}(\mu) = \dfrac{1 - \log\mu - c_0}{\mu^2}$, $|b'_{c_0}(\mu)\dot{\mu}| \lesssim \frac{\log\mu}{\mu^5} = o(\mu^{-4})$. After increasing $\mu_{\log}$, equation 77 follows. □

Lemma C.18 (Endpoint shoulder estimates)**.** *Fix $\beta > 0$. There exist constants $c_{\mathrm{sh,L}} < c_0$ and $a_{\mathrm{sh}}, C_{\mathrm{sh}}, \mu_{\mathrm{sh}} > 0$, depending only on ρ, C_{band}, and β, such that, whenever $\mu \geq \mu_{\mathrm{sh}}$, $\ell - \dfrac{\beta \log \mu}{\mu} \leq k < b_{c_{\mathrm{sh,L}}}(\mu)$, one has*

$$\dot{k} - b'_{c_0}(\mu)\dot{\mu} \geq \frac{a_{\mathrm{sh}}}{\mu^4}, \qquad (\dot{\mu})_+ \leq \frac{C_{\mathrm{sh}}}{\mu^3},$$

and $\dot{k} > 0$, $\frac{(-\dot{\mu})_+}{\mu} \leq C_{\mathrm{sh}}\dot{k}$.

Proof. All implicit constants may depend on $\rho, C_{\mathrm{band}}, \beta$. Set $s := \mu(k - \ell)$, $\chi := s - \log \mu$. The assumptions imply $-(\beta + 1)\log \mu \leq \chi < c_{\mathrm{sh,L}}$. For $\alpha = \rho + u/\mu$, $0 \leq u \leq \Delta\mu$, write $q(u) := q_\theta(\rho + u/\mu)$, $p(u) := 1 - q(u)$, and $r := q(0)$. The odds identity gives

$$\frac{p(u)}{q(u)} = e^{s+u} h_\mu\left(\rho + \frac{u}{\mu}\right), \qquad h_\mu(\alpha) := \frac{1 - e^{-\mu\alpha}}{1 - e^{-\mu(1-\alpha)}}. \tag{78}$$

Define $a_\mu(u) := \frac{d}{du} \log \frac{p(u)}{q(u)}$. Differentiating equation 78 gives $a_\mu(u) = 1 + e^{-\mu\rho - u}/(1 - e^{-\mu\rho - u}) + e^{-\mu\ell + u}/(1 - e^{-\mu\ell + u})$. Since $0 \leq u \leq \Delta\mu$, both correction terms are $O(e^{-\rho\mu})$ uniformly. Hence, for all sufficiently large μ,

$$-q'(u) = a_\mu(u) p(u) q(u), \qquad \frac{1}{2} \leq a_\mu(u) \leq 2. \tag{79}$$

In particular, q is decreasing. Also, $q(\Delta\mu) \lesssim \mu^\beta e^{-\Delta\mu}$. Throughout the endpoint shoulder, k remains bounded above and below by positive constants. The bias expansion gives

$$B_\theta(\alpha) \geq \frac{\rho}{2} q_\theta(\alpha) - \frac{C}{\mu}, \qquad |B_\theta(\alpha)| \lesssim q_\theta(\alpha) + \frac{1}{\mu}. \tag{80}$$

The mixture identities, the component asymptotics, and $d_\mu(\alpha) \geq \rho/2$ give

$$J_\theta(\alpha) \gtrsim \mu p_\theta(\alpha) q_\theta(\alpha), \qquad J_\theta(\alpha) \lesssim \frac{1}{\mu} + \mu p_\theta(\alpha) q_\theta(\alpha), \tag{81}$$

and

$$\sigma^2_\theta(\alpha) \lesssim \frac{1}{\mu^2} + p_\theta(\alpha) q_\theta(\alpha). \tag{82}$$

Choose constants $A > 0$ and $C_0 > 0$ such that

$$q_\theta(\alpha) \geq \frac{A}{\mu} \Longrightarrow B_\theta(\alpha) \gtrsim q_\theta(\alpha), \qquad \{B_\theta < 0\} \subseteq \left\{q_\theta \leq \frac{C_0}{\mu}\right\}. \tag{83}$$

From equation 79,

$$\int_{\{q \leq C_0/\mu\}} p(u) q(u)\, du \lesssim \frac{1}{\mu}. \tag{84}$$

Using $d\alpha = du/\mu$, equation 80, equation 81, and equation 84,

$$\int_{\{B_\theta < 0\}} |B_\theta(\alpha)| J_\theta(\alpha)\, d\alpha \lesssim \frac{1}{\mu^2}. \tag{85}$$

Choose $R \geq 2A$ and then choose $c_{\mathrm{sh,L}} < c_0$ so negative that $4Re^{c_{\mathrm{sh,L}}} \leq 1$. At $u = 0$, $\frac{1-r}{r} = \mu e^\chi h_\mu(\rho)$. Since $h_\mu(\rho) \leq 2$ for all sufficiently large μ, the condition $\chi < c_{\mathrm{sh,L}}$ gives

$$r \geq \frac{R}{\mu} \tag{86}$$

after increasing μ_{sh}.

Because $q(\Delta\mu) < A/\mu$ for all sufficiently large μ, q is decreasing, and $r \geq R/\mu \geq 2A/\mu$, equation equation 79 yields

$$\int_{\{q \geq A/\mu\}} p(u) q(u)^2\, du \geq \frac{1}{2} \int_{\{q \geq A/\mu\}} q(u)(-q'(u))\, du = \frac{1}{4}\left(r^2 - \frac{A^2}{\mu^2}\right).$$

Hence equation 83, equation 81, and $d\alpha = du/\mu$ give

$$\int_{\{q \geq A/\mu\}} B_\theta(\alpha) J_\theta(\alpha)\, d\alpha \gtrsim r^2 - \frac{A^2}{\mu^2}.$$

Combining this with equation 85 and increasing R if necessary gives $\int_\rho^\ell B_\theta(\alpha) J_\theta(\alpha)\, d\alpha \gtrsim r^2$. Therefore, by equation 38,

$$\dot{k} \geq \frac{c_K r^2}{\mu^2} \tag{87}$$

for some constant $c_K > 0$.

From equation 79,

$$\int_0^{\Delta\mu} p(u)q(u)\, du \lesssim r, \qquad \int_0^{\Delta\mu} p(u)q(u)^2\, du \lesssim r^2. \tag{88}$$

Since $p' = -q' = a_\mu pq$ and $a_\mu \geq 1/2$, $\int_0^{\Delta\mu} q(u)\, du \lesssim \log \frac{1}{1-r}$. If $r \leq 1/2$, then $\log(1/(1-r)) \lesssim r \leq \mu^2 r^2$ by equation 86. If $r > 1/2$, then $\frac{1-r}{r} = e^s h_\mu(\rho) \gtrsim \mu^{-\beta}$, so $\log(1/(1-r)) \lesssim \log \mu \lesssim \mu^2 r^2$. Thus

$$\int_0^{\Delta\mu} q(u)\, du \lesssim \mu^2 r^2. \tag{89}$$

Using $d\alpha = du/\mu$, equation 80, equation 82, equation 88, and equation 89,

$$|\dot{\mu}| \lesssim \frac{1}{\mu} \int_0^{\Delta\mu} \left(q(u) + \frac{1}{\mu}\right)\left(\frac{1}{\mu^2} + p(u)q(u)\right) du \lesssim \frac{r^2}{\mu}. \tag{90}$$

For the positive radial part, only the set $\{B_\theta < 0\}$ can contribute positively to $\dot{\mu}$. On this set, $q_\theta \leq C_0/\mu$ and $|B_\theta| \lesssim \mu^{-1}$. Hence equation 82, equation 84, and $d\alpha = du/\mu$ give

$$(\dot{\mu})_+ \lesssim \int_{\{B_\theta < 0\}} |B_\theta(\alpha)| \sigma_\theta^2(\alpha)\, d\alpha \lesssim \frac{1}{\mu^3}. \tag{91}$$

Finally, $b'_{c_0}(\mu) = \frac{1 - \log\mu - c_0}{\mu^2}$. Equations equation 87 and equation 90 imply

$$\begin{aligned} \dot{k} - b'_{c_0}(\mu)\dot{\mu} &\geq \frac{c_K r^2}{\mu^2} - C\frac{r^2 \log\mu}{\mu^3} \\ &\geq \frac{c_K r^2}{2\mu^2} \geq \frac{c_K R^2}{2\mu^4} \end{aligned}$$

after increasing μ_{sh}. Set $a_{\mathrm{sh}} := \dfrac{c_K R^2}{2}$ and choose C_{sh} large enough to dominate the constants in equation 91 and in the comparison obtained from equation 87 and equation 90. Then

$$\dot{k} - b'_{c_0}(\mu)\dot{\mu} \geq \frac{a_{\mathrm{sh}}}{\mu^4}, \qquad (\dot{\mu})_+ \leq \frac{C_{\mathrm{sh}}}{\mu^3},$$

and $\dot{k} > 0$, $\frac{(-\dot{\mu})_+}{\mu} \leq C_{\mathrm{sh}}\dot{k}$. □

Lemma C.19 (Uniform barrier estimate in the logarithmic endpoint window)**.** *Fix $\beta > 0$. There exist constants $a_{\mathrm{win}}, C_{\mathrm{win}}, \mu_{\mathrm{win}} > 0$ such that, whenever $\mu \geq \mu_{\mathrm{win}}$, $\ell - \dfrac{\beta \log\mu}{\mu} \leq k < b_{c_0}(\mu)$, one has*

$$\dot{k} - b'_{c_0}(\mu)\dot{\mu} \geq \frac{a_{\mathrm{win}}}{\mu^4}. \tag{92}$$

Moreover, whenever $\dot{\mu} > 0$ in the same region,

$$\dot{\mu} \leq \frac{C_{\mathrm{win}}}{\mu^3}. \tag{93}$$

Finally,

$$\dot{k} > 0, \qquad \frac{(-\dot{\mu})_+}{\mu} \leq C_{\mathrm{win}}\dot{k}. \tag{94}$$

Proof. Let $c(k,\mu) := \mu(k-\ell) - \log\mu$. By Lemma C.18, there are constants $c_{\mathrm{sh,L}} < c_0$, $a_{\mathrm{sh}} > 0$, $C_{\mathrm{sh}} > 0$, and $\mu_{\mathrm{sh}} > 0$ such that the desired barrier, positive radial bound, and relative-loss estimate hold on the lower shoulder $c(k,\mu) < c_{\mathrm{sh,L}}$. It remains to treat the compact logarithmic subwindow $c(k,\mu) \in [c_{\mathrm{sh,L}}, c_0]$. On this subwindow, write $k = b_c(\mu)$ with $c \in [c_{\mathrm{sh,L}}, c_0]$. By Proposition C.15,

$$\dot{k} = -\frac{d_k(c)}{\mu^4} + o(\mu^{-4}), \qquad \dot{\mu} = \frac{d_{\mathrm{rad}}(c)}{\mu^3} + o(\mu^{-3})$$

uniformly over $c \in [c_{\mathrm{sh,L}}, c_0]$. Since $d_k(c) < 0$ for every $c \le c_0$, compactness gives $c_{\mathrm{cp}} > 0$ such that $-d_k(c) \ge 4c_{\mathrm{cp}},\ c \in [c_{\mathrm{sh,L}}, c_0]$. Therefore, for all sufficiently large μ,

$$\dot{k} \ge \frac{2c_{\mathrm{cp}}}{\mu^4}, \qquad |\dot{\mu}| \le \frac{C_{\mathrm{cp}}}{\mu^3} \tag{95}$$

for some $C_{\mathrm{cp}} > 0$. Since $|b'_{c_0}(\mu)\dot{\mu}| \lesssim \frac{\log\mu}{\mu^5} = o(\mu^{-4})$, we get

$$\dot{k} - b'_{c_0}(\mu)\dot{\mu} \ge \frac{c_{\mathrm{cp}}}{\mu^4} \tag{96}$$

after increasing the threshold.

Combining the lower-shoulder estimates with equation 95 and equation 96, define $a_{\mathrm{win}} := \min\{a_{\mathrm{sh}}, c_{\mathrm{cp}}\}$. Choose

$$C_{\mathrm{win}} \ge \max\left\{C_{\mathrm{sh}}, C_{\mathrm{cp}}, \frac{C_{\mathrm{cp}}}{2c_{\mathrm{cp}}}\right\}.$$

Taking μ_{win} larger than all thresholds used above proves equation 92, equation 93, and equation 94. □

Proposition C.20 (One-sided approach to the logarithmic rail)**.** *Fix $0 < \eta < \Delta/4$. There exist constants $a_{\mathrm{ent}}, C_{\mathrm{ent}}, \mu_{\mathrm{ent}} > 0$ such that, whenever $\mu \ge \mu_{\mathrm{ent}},\ \ell - \eta \le k < b_{c_0}(\mu)$, one has*

$$\dot{k} - b'_{c_0}(\mu)\dot{\mu} \ge \frac{a_{\mathrm{ent}}}{\mu^4}. \tag{97}$$

Moreover, whenever $\dot{\mu} > 0$ in the same region,

$$\dot{\mu} \le \frac{C_{\mathrm{ent}}}{\mu^3}. \tag{98}$$

Finally,

$$\dot{k} > 0, \qquad \frac{(-\dot{\mu})_+}{\mu} \le C_{\mathrm{ent}}\dot{k}. \tag{99}$$

Proof. Choose $0 < \eta_1 < \eta$ and $\beta_{\mathrm{sep}} > 2$. Split the region into

$$[\ell - \eta, \ell - \eta_1], \qquad \left[\ell - \eta_1, \ell - \frac{\beta_{\mathrm{sep}}\log\mu}{\mu}\right], \qquad \left[\ell - \frac{\beta_{\mathrm{sep}}\log\mu}{\mu}, b_{c_0}(\mu)\right].$$

On the first interval, Proposition C.9, applied with η_1, gives $\dot{k} \gtrsim \frac{1}{\mu^2},\ \dot{\mu} < 0,\ |\dot{\mu}| \lesssim \frac{1}{\mu}$. On the second interval, Lemma C.11 gives the same bounds. Since $|b'_{c_0}(\mu)| \lesssim \log\mu/\mu^2$, both regions satisfy

$$\dot{k} - b'_{c_0}(\mu)\dot{\mu} \gtrsim \frac{1}{\mu^2}, \qquad \frac{(-\dot{\mu})_+}{\mu} \lesssim \frac{1}{\mu^2} \lesssim \dot{k}.$$

On the endpoint window, Lemma C.19 gives $\dot{k} - b'_{c_0}(\mu)\dot{\mu} \ge \frac{a_{\mathrm{win}}}{\mu^4}$, as well as the positive radial bound and the relative-loss estimate. Combining the three regions proves equation 97, equation 98, and equation 99 after increasing μ_{ent} and enlarging C_{ent}. □

Proposition C.21 (Stopped finite-time approach to the logarithmic tube)**.** *Assume $C_{\mathrm{band}} \ge 2$, and fix $0 < \eta < \Delta/4$. There exists $\mu_{\mathrm{G,ent}} > 0$ such that the following holds. Suppose that, at some time t_0,*

$$k(t_0) \in \mathcal{T}_\eta(\mu(t_0)), \qquad k(t_0) < b_{c_0}(\mu(t_0)), \qquad \mu(t_0) > \mu_{\mathrm{G,ent}}.$$

Define $\tau_{\log} := \inf\{t \ge t_0 : k(t) \ge b_{c_0}(\mu(t))\}$ and $\tau_{\mathrm{G,exit}} := \inf\{t \ge t_0 : \mu(t) \le \mu_{\mathrm{G,ent}}\}$. Then $\tau_{\log} \wedge \tau_{\mathrm{G,exit}} < \infty$. In particular, if $\tau_{\mathrm{G,exit}} > \tau_{\log}$, then the trajectory satisfies $k(t) \in \mathcal{I}_{\log}(\mu(t))$ in finite time.

Proof. Increase $\mu_{\mathrm{G,ent}}$ so that $\mu_{\mathrm{G,ent}} \geq \max\{\mu_{\mathrm{tube},\eta}, \mu_{\mathrm{ent}}\}$. Before $\tau_{\log} \wedge \tau_{\mathrm{G,exit}}$, Proposition C.12 gives $\ell - \eta \leq k(t) < b_{c_0}(\mu(t))$. With $g_{\log}(t) := b_{c_0}(\mu(t)) - k(t)$, Proposition C.20 gives $\dot{g}_{\log}(t) \leq -\frac{a_{\mathrm{ent}}}{\mu(t)^4}$. The same proposition gives $\dot{\mu}(t) \leq C_{\mathrm{ent}}/\mu(t)^3$ whenever $\dot{\mu}(t) > 0$, and this upper bound is trivial when $\dot{\mu}(t) \leq 0$. Lemma C.1(ii) applied on the stopped interval proves that $\tau_{\log} \wedge \tau_{\mathrm{G,exit}} < \infty$. □

Corollary C.22 (Forward-invariant logarithmic prediction-good tube)**.** *Assume $C_{\mathrm{band}} \geq 2$, and fix $0 < \eta < \Delta/4$. There exist constants $C_{\mathrm{bias}}, \mu_{\log,*} > 0$ such that, for every $\mu \geq \mu_{\log,*}$, the interval $\mathcal{I}_{\log}(\mu)$ is nonempty and contained in $\mathcal{T}_\eta(\mu)$. Moreover, if $k(t_0) \in \mathcal{I}_{\log}(\mu(t_0))$, $\mu(t_0) \geq \mu_{\log,*}$, and $\tau_{\log,*} := \inf\{t \geq t_0 : \mu(t) \leq \mu_{\log,*}\}$, then $k(t) \in \mathcal{I}_{\log}(\mu(t))$ for every $t \in [t_0, \tau_{\log,*})$. Throughout the tube,*

$$\sup_{\alpha \in [\rho,\ell]} |F(\lambda, \mu; \alpha) - \alpha| \leq \frac{C_{\mathrm{bias}}}{\mu}.$$

Proof. Since $C_{\mathrm{band}} \geq 2$, $b_{c_0}(\mu) \leq U_{\mathrm{band}}(\mu)$ for all sufficiently large μ, while $b_{c_0}(\mu) > \ell - \eta$. Hence $\mathcal{I}_{\log}(\mu)$ is nonempty and contained in $\mathcal{T}_\eta(\mu)$ after increasing $\mu_{\log,*}$.

Set $a(\mu) := b_{c_0}(\mu)$, $b(\mu) := U_{\mathrm{band}}(\mu)$. At the lower boundary, Proposition C.17 gives $\dot{k} - a'(\mu)\dot{\mu} = \dot{k} - b'_{c_0}(\mu)\dot{\mu} > 0$. At the upper boundary $k = b(\mu) = U_{\mathrm{band}}(\mu)$, Corollary C.4 gives $b'(\mu)\dot{\mu} - \dot{k} = U'_{\mathrm{band}}(\mu)\dot{\mu} - \dot{k} \geq \frac{a_U}{\mu^4}$. Equivalently, $\dot{k} - b'(\mu)\dot{\mu} < 0$ on the upper boundary. Lemma C.2 therefore shows that the ratio cannot leave the moving interval $\mathcal{I}_{\log}(\mu(t))$ before $\tau_{\log,*}$. The prediction bound follows from Lemma C.13 with $c = c_0$. □

Uniform radial growth after logarithmic capture

Lemma C.23 (Uniform radial expansion in the logarithmic tube)**.** *Let $c(k, \mu) := \mu(k - \ell) - \log \mu$. Then*

$$\sup_{k \in [b_{c_0}(\mu), U_{\mathrm{band}}(\mu)]} \left|\mu^3 \dot{\mu} - d_{\mathrm{rad}}(c(k, \mu))\right| \longrightarrow 0 \tag{100}$$

as $\mu \to \infty$. Moreover,

$$m_{\mathrm{rad}} := \inf_{c \geq c_0} d_{\mathrm{rad}}(c) > 0.$$

Proof. Inside the tube, $c_0 \leq c(k, \mu) \leq (C_{\mathrm{band}} - 1) \log \mu$. Fix $c_1 > c_0$. On the compact interval $[c_0, c_1]$, Proposition C.15 gives

$$\sup_{c \in [c_0, c_1]} \left|\mu^3 \dot{\mu} - d_{\mathrm{rad}}(c)\right| \longrightarrow 0. \tag{101}$$

Now suppose $c \geq c_1$. The odds identity yields $q_\theta\left(\rho + \frac{u}{\mu}\right) \leq \frac{Ce^{-c-u}}{\mu}$. Therefore

$$\int_\rho^\ell q_\theta(\alpha)\, d\alpha \leq \frac{Ce^{-c}}{\mu^2}, \qquad \int_\rho^\ell q_\theta(\alpha)^2\, d\alpha \leq \frac{Ce^{-2c}}{\mu^3}.$$

Using the bias expansion and the uniform component asymptotics, the component-variance contribution satisfies

$$\dot{\mu}_{\mathrm{comp}} = \frac{1}{\mu^3} + o(\mu^{-3}) + O\left(\frac{e^{-c}}{\mu^4}\right)$$

uniformly for $c \geq c_1$. The mixture contribution satisfies

$$|\dot{\mu}_{\mathrm{mix}}| \leq C \int_\rho^\ell \left(\frac{1}{\mu} + q_\theta(\alpha)\right) q_\theta(\alpha)\, d\alpha \leq \frac{C}{\mu^3}\left(e^{-c} + e^{-2c}\right).$$

Hence $\left|\mu^3 \dot{\mu} - 1\right| \leq C(e^{-c} + e^{-2c}) + o(1)$ uniformly for $c \geq c_1$. Since $|d_{\mathrm{rad}}(c) - 1| \leq C(e^{-c} + e^{-2c})$, we obtain

$$\sup_{c_1 \leq c \leq (C_{\mathrm{band}} - 1)\log \mu} \left|\mu^3 \dot{\mu} - d_{\mathrm{rad}}(c)\right| \leq Ce^{-c_1} + o(1).$$

Letting first $c_1 \to \infty$ and then $\mu \to \infty$, together with equation 101, proves equation 100. By Lemma C.16, $d_{\mathrm{rad}}(c) > 0$ for every $c \geq c_0$, and $d_{\mathrm{rad}}(c) \to 1$ as $c \to \infty$. Continuity therefore gives $m_{\mathrm{rad}} > 0$. □

Proposition C.24 (Two-sided radial drift in the logarithmic tube)**.** *There exist constants $0 < c_\mu < C_\mu < \infty$ and μ_{rad} such that, whenever $\mu \geq \mu_{\mathrm{rad}}$ and $k \in \mathcal{I}_{\log}(\mu)$,*

$$\frac{c_\mu}{\mu^3} \leq \dot{\mu} \leq \frac{C_\mu}{\mu^3}.$$

Consequently, if the trajectory remains in the logarithmic tube for $t \geq t_0$, then

$$\mu(t_0)^4 + 4c_\mu(t - t_0) \leq \mu(t)^4 \leq \mu(t_0)^4 + 4C_\mu(t - t_0).$$

Proof. Lemma C.23 gives $\mu^3\dot{\mu} = d_{\mathrm{rad}}(c(k,\mu)) + o(1)$ uniformly over $k \in \mathcal{I}_{\log}(\mu)$. By Lemma C.16, $d_{\mathrm{rad}}(c) > 0$ for $c \geq c_0$, while $d_{\mathrm{rad}}(c) \to 1$ as $c \to \infty$. Hence $0 < \inf_{c \geq c_0} d_{\mathrm{rad}}(c) \leq \sup_{c \geq c_0} d_{\mathrm{rad}}(c) < \infty$, and the two-sided drift bound follows for sufficiently large μ. Multiplying by $4\mu^3$ and integrating proves the fourth-power bounds. □

Pre-capture preservation of the large-parameter regime The preceding regional transport arguments require lower bounds on μ. We now show that these lower bounds are preserved automatically when the initial radius is sufficiently large and the initial ratio is bounded away from zero.

Proposition C.25 (Relative radial-loss estimate before logarithmic capture)**.** *Assume $C_{\mathrm{band}} \geq 2$, fix $0 < \eta < \Delta/4$, and fix $\kappa > 0$. There exist constants $C_{\mathrm{pres}}, \mu_{\mathrm{rel}} > 0$ such that, whenever*

$$\mu \geq \mu_{\mathrm{rel}}, \qquad \kappa \leq k \leq U_{\mathrm{band}}(\mu), \qquad k < b_{c_0}(\mu), \tag{102}$$

one has

$$\dot{k} > 0 \tag{103}$$

and

$$\frac{(-\dot{\mu})_+}{\mu} \leq C_{\mathrm{pres}}\dot{k}, \qquad (-\dot{\mu})_+ := \max\{-\dot{\mu}, 0\}. \tag{104}$$

Proof. The mixture identities and $\lambda = k\mu$ give

$$J_\theta(\alpha) = \mu p_\theta(\alpha) q_\theta(\alpha) d_\mu(\alpha) + k\mu\sigma_\theta^2(\alpha) \geq k\mu\sigma_\theta^2(\alpha). \tag{105}$$

Set $A := \int_\rho^\ell B_\theta(\alpha)\sigma_\theta^2(\alpha)\, d\alpha$, $K := \int_\rho^\ell B_\theta(\alpha) J_\theta(\alpha)\, d\alpha$. Then $-\dot{\mu} = \frac{A}{\Delta}$, $\dot{k} = \frac{K}{\Delta\mu^2}$. We cover the region in equation 102 by four subregions. **1. Lower outer region.** If $\kappa \leq k < L_{\mathrm{band}}(\mu)$, the odds identity and Corollary B.2 give $B_\theta(\alpha) \geq c_\rho > 0$, $\alpha \in [\rho, \ell]$, for all sufficiently large μ. Hence $A > 0$, and equation 105 implies

$$K = \int_\rho^\ell B_\theta(\alpha) J_\theta(\alpha)\, d\alpha \geq k\mu \int_\rho^\ell B_\theta(\alpha)\sigma_\theta^2(\alpha)\, d\alpha = k\mu A \geq \kappa\mu A.$$

Therefore $K > 0$ and

$$\frac{(-\dot{\mu})_+}{\mu} = \frac{A}{\Delta\mu} \leq \frac{K}{\Delta\kappa\mu^2} = \frac{1}{\kappa}\dot{k}.$$

2. Lower boundary layer. If $L_{\mathrm{band}}(\mu) \leq k \leq \rho + \eta$, Proposition C.6 gives $\dot{k} > 0$, $\frac{(-\dot{\mu})_+}{\mu} \leq C_{\mathrm{low},\eta}\dot{k}$. **3. Fixed interior strip.** If $\rho + \eta \leq k \leq \ell - \eta$, Proposition C.9 gives $\dot{k} > 0$, $\frac{(-\dot{\mu})_+}{\mu} \leq C_{\mathrm{int},\eta}\dot{k}$. **4. Upper approach region.** If $\ell - \eta \leq k < b_{c_0}(\mu)$, Proposition C.20 gives the desired estimates directly.

Taking C_{pres} to be the maximum of the constants in the four regions and increasing μ_{rel} so that all corresponding thresholds apply proves equation 103 and equation 104. □

Lemma C.26 (Integrated radial-loss comparison)**.** *Let $[t_0, t_1]$ be an interval on which*

$$\mu(t) > 0, \qquad \dot{k}(t) \geq 0, \qquad \frac{(-\dot{\mu}(t))_+}{\mu(t)} \leq C\dot{k}(t).$$

Then, for every $t \in [t_0, t_1]$, $\mu(t) \geq \mu(t_0)\exp\{-C(k(t) - k(t_0))\}$.

Proof. Since $\frac{d}{dt}\log\mu(t) = \dot{\mu}(t)/\mu(t) \geq -(-\dot{\mu}(t))_+/\mu(t) \geq -C\dot{k}(t)$, integration gives $\log\frac{\mu(t)}{\mu(t_0)} \geq -C\{k(t) - k(t_0)\}$. Exponentiating proves the claim. □

C.2 Completion of the proof

Proof of Theorem 5.2. Fix $\eta := \Delta/8$ and take μ_* according to the threshold convention above. Increase μ_* once more, if necessary, so that $b_{c_0}(\mu) < 2$ for every $\mu \geq \mu_*$. Let C_{pres} be the constant in Proposition C.25 and set

$$c_{\mathrm{rad}} := e^{-2C_{\mathrm{pres}}}.$$

Increase the initialization threshold so that $\mu(0) \geq 2c_{\mathrm{rad}}^{-1}\mu_*$. Write $\tau_{\mathcal{G}} := \inf\{t \geq 0 : \theta(t) \in \mathcal{G}\}$.

If $\theta(0) \in \mathcal{G}$, then $\tau_{\mathcal{G}} = 0$. If $k(0) > U_{\mathrm{band}}(\mu(0))$, Proposition C.5 gives finite-time entry through $k = U_{\mathrm{band}}(\mu)$, while Proposition C.3 shows that μ increases beforehand. Since $b_{c_0}(\mu) \leq U_{\mathrm{band}}(\mu)$ for $\mu \geq \mu_*$, this entry point already belongs to $\mathcal{G}$.

It therefore remains only to consider $k(0) \leq U_{\mathrm{band}}(\mu(0))$ with $k(0) < b_{c_0}(\mu(0))$. In this region, Proposition C.25 gives $\dot{k} > 0$ and $(-\dot{\mu})_+/\mu \leq C_{\mathrm{pres}}\dot{k}$ as long as $\mu \geq \mu_*$. Before capture, $k(t) < b_{c_0}(\mu(t)) < 2$, while $k(0) > 0$. Hence $k(t) - k(0) < 2$, and Lemma C.26 gives

$$\mu(t) \geq \mu(0)e^{-C_{\mathrm{pres}}\{k(t)-k(0)\}} \geq c_{\mathrm{rad}}\mu(0) \geq 2\mu_*.$$

Thus none of the stopped large-radius exits can occur before capture.

The regional transport results may therefore be applied successively, skipping any stage already passed: Proposition C.5 enters the dynamic band, Corollary C.7 enters the fixed interior strip, Corollary C.10 reaches the upper endpoint region, Proposition C.12 keeps the trajectory inside the upper tube, and Proposition C.21 gives finite-time crossing of the lower logarithmic rail. Hence $\tau_{\mathcal{G}} < \infty$, and the same persistence estimate gives $\inf_{0 \leq t \leq \tau_{\mathcal{G}}} \mu(t) \geq c_{\mathrm{rad}}\mu(0)$.

At capture, Corollary C.22 prevents side-rail exit and Proposition C.24 gives positive radial drift, so radial exit is impossible as well. Thus $\theta(t) \in \mathcal{G}$ for every $t \geq \tau_{\mathcal{G}}$, and Proposition C.24 gives the two-sided bound $c\mu(t)^{-3} \leq \dot{\mu}(t) \leq C\mu(t)^{-3}$ and hence equation 10 after renaming constants.

Finally, Lemma C.13 yields $\sup_{\alpha \in [\rho,\ell]} |F(\theta(t);\alpha) - \alpha| \leq C/\mu(t)$. Therefore $\mathcal{R}_\rho(\theta(t)) \leq C/\mu(t)^2$. Membership in $\mathcal{G}$ gives $k(t) = \ell + O(\{\log \mu(t)\}/\mu(t))$, while the fourth-power bounds imply $\mu(t) \asymp t^{1/4}$ after capture. The prediction and risk rates in equation 12 follow, and since $\ell > 0$, both $\mu(t)$ and $\lambda(t) = k(t)\mu(t)$ diverge. □

D Population Gradient Descent

Define $v := -\nabla\mathcal{R}_\rho$ and the population GD iterates

$$\theta_{s+1}^\gamma = \theta_s^\gamma + \gamma v(\theta_s^\gamma), \qquad \theta_0^\gamma = \theta(0). \tag{106}$$

Choose a deterministic $T_0 > \tau_{\mathcal{G}}$ for which the population flow is strictly inside $\mathcal{G}$, and choose $\tau_0 > 0$ so that positive radial and side-barrier margins persist on $[T_0, T_0 + \tau_0]$. Let $s_0(\gamma) := \lceil T_0/\gamma \rceil$.

For $R \geq \mu_*$, write

$$\mathcal{S}_R := \mathcal{G} \cap \{R \leq \mu \leq 2R\}$$

for the core logarithmic shell.

D.1 Regularity and compact Euler tracking

Lemma D.1 (Global Lipschitz regularity of the population vector field)**.** *There exists a universal constant $L_v < \infty$ such that*

$$\|v(\theta) - v(\vartheta)\| \leq L_v\|\theta - \vartheta\|, \qquad \theta, \vartheta \in \mathbb{R}^2.$$

Proof. For fixed α, write

$$g(\theta;\alpha) := \{F(\theta;\alpha) - \alpha\}\nabla F(\theta;\alpha), \qquad v(\theta) = -\mathbb{E}_\alpha g(\theta;\alpha).$$

Let $T_1 := Y$ and $T_2 := X$, with $a, b \in \{1, 2\}$ indexing (λ, μ). Exponential-family differentiation gives

$$\partial_a F = \mathrm{Cov}_\theta(X, T_a \mid \alpha), \qquad \partial_{ab}^2 F = \mathbb{E}_\theta[(X - \mathbb{E}_\theta X)(T_a - \mathbb{E}_\theta T_a)(T_b - \mathbb{E}_\theta T_b) \mid \alpha].$$

Since $X, Y \in [0,1]$, both $\|\nabla F\|$ and $\|\nabla^2 F\|_{\mathrm{op}}$ are uniformly bounded in θ and α. Also $|F-\alpha| \le 1$. Hence

$$Dg = \nabla F \nabla F^\top + (F-\alpha)\nabla^2 F$$

has uniformly bounded operator norm. Therefore $g(\cdot;\alpha)$ is globally Lipschitz with a constant independent of α. Averaging over α gives the same property for $v = -\mathbb{E}_\alpha g$. □

Lemma D.2 (Population speed in the logarithmic tube)**.** *There is a constant $C_{\mathrm{sp}} < \infty$ such that*

$$\|v(\theta)\| \le \frac{C_{\mathrm{sp}}}{\mu^3}, \qquad \theta = (\lambda,\mu) \in \mathcal{G}.$$

Proof. Inside $\mathcal{G}$, Lemma C.13 gives $|F(\theta;\alpha) - \alpha| \le \frac{C}{\mu}$ uniformly in α. Moreover, the odds formula and the lower rail imply $q_\theta(\alpha) \le \frac{C}{\mu} e^{-\mu(\alpha-\rho)}$. Using the mixture identities and the component asymptotics, $|\mathrm{Cov}_\theta(X, Y \mid \alpha)| \le C q_\theta(\alpha)$, and $\mathrm{Var}_\theta(X \mid \alpha) \le \frac{C}{\mu^2} + C q_\theta(\alpha)$. Hence $\int_\rho^\ell |\mathrm{Cov}_\theta(X, Y \mid \alpha)|\, d\alpha \le \frac{C}{\mu^2}$, and $\int_\rho^\ell \mathrm{Var}_\theta(X \mid \alpha)\, d\alpha \le \frac{C}{\mu^2}$. Substituting into the risk-gradient identity proves

$$|\partial_\lambda \mathcal{R}_\rho(\theta)| + |\partial_\mu \mathcal{R}_\rho(\theta)| \le \frac{C}{\mu^3}.$$

Since $v = -\nabla \mathcal{R}_\rho$, the claim follows. □

Lemma D.3 (Barrier regularity)**.** *There are constants $C_h, C_{hh} < \infty$ such that, for every $\theta = (\lambda,\mu) \in \mathcal{G}$, $\|\nabla h_-(\theta)\| + \|\nabla h_+(\theta)\| \le \frac{C_h}{\mu}$, and*

$$\|\nabla^2 h_-(\theta)\|_{\mathrm{op}} + \|\nabla^2 h_+(\theta)\|_{\mathrm{op}} \le \frac{C_{hh}}{\mu^2}.$$

Proof. Since $h_-(\lambda,\mu) = \frac{\lambda}{\mu} - b_{c_0}(\mu)$, and $h_+(\lambda,\mu) = U_{\mathrm{band}}(\mu) - \frac{\lambda}{\mu}$, the first derivatives of λ/μ have size $O(\mu^{-1})$ on the tube because $k = \lambda/\mu = O(1)$. Its second derivatives have size $O(\mu^{-2})$. Finally, $b'_{c_0}(\mu),\ U'_{\mathrm{band}}(\mu) = O\left(\frac{\log \mu}{\mu^2}\right)$, and $b''_{c_0}(\mu),\ U''_{\mathrm{band}}(\mu) = O\left(\frac{\log \mu}{\mu^3}\right)$. After increasing the fixed base radius μ_*, these terms are bounded by the claimed scales. □

For a vector field w on the logarithmic tube, define $\Psi_{\pm,w}(\theta) := \mu^4 \langle \nabla h_\pm(\theta), w(\theta)\rangle$.

Lemma D.4 (Normalized directional regularity)**.** *Let $R \ge \mu_*$ and suppose a C^1 vector field w on a fixed-width enlargement of $\mathcal{S}_R$ satisfies $\|w(\theta)\| \le C_w,\ \|Dw(\theta)\|_{\mathrm{op}} \le L_w$ there. Then*

$$|\Psi_{\pm,w}(\theta) - \Psi_{\pm,w}(\vartheta)| \le C(C_w + L_w) R^3 \|\theta - \vartheta\|$$

whenever θ, ϑ lie in that enlargement.

Proof. On a fixed-width enlargement of $\mathcal{S}_R$, one has $\mu \asymp R$ and $k = O(1)$. Direct differentiation of the explicit barriers therefore gives

$$\|\mu^4 \nabla h_\pm(\theta)\| \le C R^3, \qquad \left\|D\{\mu^4 \nabla h_\pm(\theta)\}\right\|_{\mathrm{op}} \le C R^2.$$

Using $\Psi_{\pm,w}(\theta) = \langle \mu^4 \nabla h_\pm(\theta), w(\theta)\rangle$, the product rule gives $\|D\Psi_{\pm,w}(\theta)\| \le C R^2 C_w + C R^3 L_w \le C(C_w + L_w) R^3$, where $R \ge \mu_* \ge 1$. The mean-value inequality proves the claim. □

D.2 Finite-horizon Euler shadowing

Lemma D.5 (Euler shadowing on a compact time interval)**.** *Let $\theta(t)$ solve $\dot\theta = v(\theta)$ on $[0,T]$, and suppose a compact set K contains a positive neighborhood of $\theta([0,T])$. Assume $\|v(\theta)\| \le V,\ \|v(\theta) - v(\vartheta)\| \le L\|\theta - \vartheta\|$ on K. Let*

$$\theta^\gamma_{s+1} = \theta^\gamma_s + \gamma v(\theta^\gamma_s), \qquad \theta^\gamma_0 = \theta(0).$$

Then, for all sufficiently small γ and all integers s satisfying $s\gamma \le T$, $\|\theta^\gamma_s - \theta(s\gamma)\| \le C_{T,L,V}\gamma$.

Proof. For $t_s = s\gamma$, $\theta(t_{s+1}) = \theta(t_s) + \gamma v(\theta(t_s)) + \tau_s$, where

$$\begin{aligned}\|\tau_s\| &\le \int_{t_s}^{t_{s+1}} \|v(\theta(u)) - v(\theta(t_s))\|\, du \\ &\le L \int_{t_s}^{t_{s+1}} \|\theta(u) - \theta(t_s)\|\, du \\ &\le \frac{1}{2} LV\gamma^2.\end{aligned}$$

Set $e_s = \|\theta_s^\gamma - \theta(t_s)\|$. Before either trajectory leaves K, $e_{s+1} \le (1+\gamma L)e_s + \frac{1}{2}LV\gamma^2$. Since $e_0 = 0$, $e_s \le (V\gamma/2)\{(1+\gamma L)^s - 1\} \le (V\gamma/2)(e^{LT} - 1)$. For sufficiently small γ this is smaller than the clearance of the reference path from ∂K, so a stopped argument rules out exit from K. This proves the bound. □

Applying Lemma D.5 on $[0, T_0 + \tau_0]$ gives, for all sufficiently small γ,

$$\max_{0 \le s \le s_0(\gamma)} \|\theta_s^\gamma - \theta(s\gamma)\| \le C_{\mathrm{gd}}\gamma. \tag{107}$$

Since $s_0(\gamma)\gamma \in [T_0, T_0 + \gamma]$ and the reference flow has strict tube margin there, reducing the learning-rate threshold if necessary gives

$$\theta^\gamma_{s_0(\gamma)} \in \mathrm{int}(\mathcal{G}). \tag{108}$$

D.3 One-step barrier stability

Lemma D.6 (No-overshoot barrier lemma)**.** *Fix $R \ge \mu_*$. Choose a fixed $K_{\mathrm{tube}} > 0$ larger than $\sup\{|\lambda/\mu| : (\lambda, \mu) \in \mathcal{G}\} + 1$, and define the two-sided side-rail enlargement*

$$\mathcal{U}_R := \{(\lambda, \mu) : R \le \mu \le 4R, \quad |\lambda/\mu| \le K_{\mathrm{tube}}\}.$$

Let w be a C^1 vector field on an open neighborhood of $\mathcal{U}_R$. Suppose

$$\frac{c_r}{\mu^3} \le w_\mu(\theta) \le \frac{C_r}{\mu^3}, \qquad \|w(\theta)\| \le \frac{C_w}{\mu^3}, \tag{109}$$

for $\theta \in \mathcal{S}_R$, and suppose further that

$$\langle \nabla h_-(\theta), w(\theta)\rangle \ge \frac{a'_-}{\mu^4} \quad \text{for } \theta \in \mathcal{U}_R \text{ with } h_-(\theta) = 0, \tag{110}$$

and

$$\langle \nabla h_+(\theta), w(\theta)\rangle \ge \frac{a'_+}{\mu^4} \quad \text{for } \theta \in \mathcal{U}_R \text{ with } h_+(\theta) = 0.$$

Assume that the normalized directional functions $\Psi_{\pm,w}$ are CR^3-Lipschitz on $\mathcal{U}_R$. Then there exists a constant $\gamma_{\mathrm{bar}} > 0$, depending only on the constants in these bounds and on μ_*, such that for every $0 < \gamma \le \gamma_{\mathrm{bar}}$ the update $\theta^+ = \theta + \gamma w(\theta)$ cannot cross either side boundary of $\mathcal{G}$ from a starting point $\theta \in \mathcal{S}_R$.

Proof. We treat the lower rail; the upper rail is identical. Suppose for contradiction that $h_-(\theta) \ge 0$, $h_-(\theta^+) < 0$. By continuity, the segment $[\theta, \theta^+]$ contains a point ζ satisfying $h_-(\zeta) = 0$. By the speed bound, $\|\theta - \zeta\| \le \|\theta^+ - \theta\| \le \frac{C\gamma}{R^3}$. Because $w_\mu(\theta) > 0$ and the step length is $O(\gamma R^{-3})$, for sufficiently small fixed γ the segment stays in $\mathcal{U}_R$. At ζ, $\Psi_{-,w}(\zeta) \ge c\,a'_-$ after absorbing the comparability of $\mu(\zeta)$ and R. The Lipschitz bound gives $\Psi_{-,w}(\theta) \ge c\,a'_- - C\gamma$. Choose γ so that the right-hand side is at least $c\,a'_-/2$. Then

$$\langle \nabla h_-(\theta), w(\theta)\rangle \ge \frac{c\,a'_-}{2R^4}. \tag{111}$$

Taylor's theorem gives

$$h_-(\theta^+) = h_-(\theta) + \gamma\langle \nabla h_-(\theta), w(\theta)\rangle + \frac{\gamma^2}{2} w(\theta)^\top \nabla^2 h_-(\xi) w(\theta)$$

for a point ξ on the segment. Since the whole segment lies in $\mathcal{U}_R$, where $\mu \asymp R$ and $k = O(1)$, direct differentiation of h_- gives $\|\nabla^2 h_-(\xi)\|_{\mathrm{op}} \le \frac{C}{R^2}$. Together with the speed bound,

$$\left|\frac{\gamma^2}{2} w(\theta)^\top \nabla^2 h_-(\xi) w(\theta)\right| \le C\frac{\gamma^2}{R^8}.$$

The first-order term in equation 111 is at least $c\gamma/R^4$. Since $R \ge \mu_*$, a further fixed reduction of γ_{bar} makes the remainder at most half of this term. Therefore $h_-(\theta^+) > h_-(\theta) \ge 0$, a contradiction. □

Proposition D.7 (Discrete logarithmic-tube invariance)**.** *There exists a constant $\gamma_{\mathrm{tube}} > 0$, depending only on the fixed population and tube constants, such that for every $0 < \gamma \le \gamma_{\mathrm{tube}}$, if $\theta_s^\gamma \in \mathcal{G}$, then $\theta_{s+1}^\gamma \in \mathcal{G}$ and*

$$\mu_{s+1}^\gamma - \mu_s^\gamma \ge \frac{c_\mu \gamma}{(\mu_s^\gamma)^3}. \tag{112}$$

Moreover, after increasing a fixed constant $C_{\mu,\mathrm{gd}}$,

$$(\mu_{s+1}^\gamma)^4 - (\mu_s^\gamma)^4 \le C_{\mu,\mathrm{gd}}\gamma. \tag{113}$$

Proof. Apply Lemma D.6 with $w = v$. The radial bounds follow from Proposition C.24; the lower- and upper-rail margins follow from Proposition C.17 and Corollary C.4; the speed bound is Lemma D.2; and normalized directional regularity follows from Lemmas D.4 and D.1.

The radial update is exact: $\mu_{s+1}^\gamma - \mu_s^\gamma = \gamma v_\mu(\theta_s^\gamma)$, which proves equation 112. For the upper fourth-power increment, write

$$\delta_s := \gamma v_\mu(\theta_s^\gamma), \qquad 0 \le \delta_s \le \frac{C_\mu \gamma}{(\mu_s^\gamma)^3}.$$

Then

$$\begin{aligned}(\mu_s^\gamma + \delta_s)^4 - (\mu_s^\gamma)^4 &= 4(\mu_s^\gamma)^3\delta_s + 6(\mu_s^\gamma)^2\delta_s^2 + 4\mu_s^\gamma\delta_s^3 + \delta_s^4 \\ &\le 4C_\mu\gamma + C\gamma^2,\end{aligned}$$

where the last constant is uniform because $\mu_s^\gamma \ge \mu_*$. For $\gamma \le \gamma_0$, absorb $C\gamma^2$ into $C_{\mu,\mathrm{gd}}\gamma$. □

D.4 Population-GD reference trajectory

Proposition D.8 (Population-GD reference trajectory)**.** *Under Assumption 2.1, there exists a deterministic $\gamma_0 > 0$, depending only on the fixed population and initialization constants, such that for every $0 < \gamma \le \gamma_0$ the population GD iterates equation 106 enter $\mathcal{G}$ after finitely many iterations. More precisely, $\theta_s^\gamma \in \mathcal{G}$ for every $s \ge s_0(\gamma)$, and*

$$(\mu_{s_0}^\gamma)^4 + 4c_\mu\gamma(s - s_0) \le (\mu_s^\gamma)^4 \le (\mu_{s_0}^\gamma)^4 + C_{\mu,\mathrm{gd}}\gamma(s - s_0). \tag{114}$$

Consequently,

$$\sup_{\alpha\in[\rho,\ell]} |F(\theta_s^\gamma;\alpha) - \alpha| \le \frac{C_{\mathrm{bias}}}{\{(\mu_{s_0}^\gamma)^4 + 4c_\mu\gamma(s - s_0)\}^{1/4}}, \tag{115}$$

Thus the threshold-estimation error is $O((\gamma s)^{-1/4})$ and the risk is $O((\gamma s)^{-1/2})$.

Proof. Finite-horizon shadowing gives $\theta_{s_0}^\gamma \in \operatorname{int}\mathcal{G}$ for every sufficiently small γ. Proposition D.7 then gives tube invariance for all subsequent iterations. By convexity of x^4 and equation 112,

$$(\mu_{s+1}^\gamma)^4 \ge (\mu_s^\gamma)^4 + 4(\mu_s^\gamma)^3(\mu_{s+1}^\gamma - \mu_s^\gamma) \ge (\mu_s^\gamma)^4 + 4c_\mu\gamma.$$

The upper fourth-power increment in Proposition D.7 iterates in the same way, so together the two estimates prove equation 114. The prediction bound follows from Lemma C.13, and the risk bound follows by squaring the uniform prediction bound and averaging over α. □

E EMPIRICAL-PROCESS ESTIMATES

This appendix proves the two empirical-process inputs used in Section F: compact approximation before capture and scale-adapted approximation inside $\mathcal{G}$. All constants $C, c > 0$ may change from line to line. Compact constants are universal; logarithmic-tube constants may additionally depend on ρ, c_0, C_{band}, and the fixed population thresholds.

For an i.i.d. sample $Z_1, \dots, Z_N$ with law P, write $P_N f := N^{-1} \sum_{i=1}^N f(Z_i)$ and $Pf := \mathbb{E} f(Z)$. For $R \geq 1$, write $K_R := [0, R]^2$.

For one task block $Z = (\alpha, \mathcal{D})$, define

$$g(\theta; \alpha) := \{F(\theta; \alpha) - \alpha\} \nabla F(\theta; \alpha), \qquad g_n(\theta; Z) := \{F_n(\theta; \mathcal{D}) - \alpha\} \nabla F_n(\theta; \mathcal{D}). \tag{116}$$

Retain $v = -\nabla \mathcal{R}_\rho$ and write

$$v_n(\theta; Z) := -g_n(\theta; Z), \qquad \bar{v}_n(\theta) := \mathbb{E}_Z v_n(\theta; Z) = -\nabla \mathcal{R}_{\rho,n}(\theta), \tag{117}$$

$$\widehat{v}_{m,n}(\theta) := \frac{1}{m} \sum_{j=1}^m v_n(\theta; Z_j) = -\nabla \widehat{\mathcal{R}}_{m,n}(\theta).$$

For $R \geq \mu_*$, recall the core shell and define its statistical enlargement by

$$\mathcal{S}_R := \mathcal{G} \cap \{R \leq \mu \leq 2R\}, \tag{118}$$

$$\widetilde{\mathcal{S}}_R := \mathcal{G} \cap \left\{ \frac{R}{2} \leq \mu \leq 4R \right\}. \tag{119}$$

Also write $\partial_\pm \widetilde{\mathcal{S}}_R := \{\theta \in \widetilde{\mathcal{S}}_R : h_\pm(\theta) = 0\}$.

Define the normalized empirical shell errors

$$\begin{aligned}
\mathcal{E}_\mu(R) &:= \sup_{\theta \in \widetilde{\mathcal{S}}_R} \mu^3 |[\widehat{v}_{m,n}(\theta)]_\mu - v_\mu(\theta)|, \\
\mathcal{E}_-(R) &:= \sup_{\theta \in \partial_- \widetilde{\mathcal{S}}_R} \mu^4 |\langle \nabla h_-(\theta), \widehat{v}_{m,n}(\theta) - v(\theta) \rangle|, \\
\mathcal{E}_+(R) &:= \sup_{\theta \in \partial_+ \widetilde{\mathcal{S}}_R} \mu^4 |\langle \nabla h_+(\theta), \widehat{v}_{m,n}(\theta) - v(\theta) \rangle|, \\
\mathcal{E}_{\mathrm{sp}}(R) &:= \sup_{\theta \in \widetilde{\mathcal{S}}_R} \mu^3 \|\widehat{v}_{m,n}(\theta) - v(\theta)\|, \\
\mathcal{E}(R) &:= \max\{\mathcal{E}_\mu(R), \mathcal{E}_-(R), \mathcal{E}_+(R), \mathcal{E}_{\mathrm{sp}}(R)\}.
\end{aligned} \tag{120}$$

E.1 BASIC EMPIRICAL-PROCESS TOOLS

Lemma E.1 (One-task gradient regularity). *There exist universal constants $B_G, L_G < \infty$ such that, for every n, every task block $Z = (\alpha, \mathcal{D})$, and every $\theta, \vartheta \in \mathbb{R}^2$,*

$$\|g_n(\theta; Z)\| \leq B_G, \qquad \|g_n(\theta; Z) - g_n(\vartheta; Z)\| \leq L_G \|\theta - \vartheta\|.$$

Proof. This is the empirical analogue of Lemma D.1. Under the empirical softmax law, boundedness of $X, Y \in [0, 1]$ gives uniform bounds on ∇F_n and $\nabla^2 F_n$. Since also $F_n, \alpha \in [0, 1]$, both $\|g_n\|$ and $\|Dg_n\|_{\mathrm{op}}$ are uniformly bounded. The mean-value inequality gives the claim. □

Lemma E.2 (Dudley bound for bounded Lipschitz parametric classes). *Let $\Theta \subset \mathbb{R}^d$ have Euclidean diameter at most D_Θ, and suppose $|f_\theta(z)| \leq B$, $|f_\theta(z) - f_\vartheta(z)| \leq L\|\theta - \vartheta\|$ for every z and every $\theta, \vartheta \in \Theta$. Then*

$$\mathbb{E} \sup_{\theta \in \Theta} |(P_N - P) f_\theta| \leq \frac{CB}{\sqrt{N}} \sqrt{d \log\left(C \left(1 + \frac{L D_\Theta}{B} \right) \right)}. \tag{121}$$

Moreover, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\sup_{\theta \in \Theta} |(P_N - P) f_\theta| \leq CB \sqrt{\frac{d \log\left(C \left(1 + \frac{L D_\Theta}{B} \right) \right) + \log(C/\delta)}{N}}. \tag{122}$$

Proof. By symmetrization,

$$\mathbb{E}\sup_{\theta\in\Theta}|(P_N-P)f_\theta| \le 2\mathbb{E}\mathbb{E}_\varepsilon \sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i f_\theta(Z_i)\right|.$$

Conditional on the sample, the Rademacher process is sub-Gaussian with respect to

$$d_N(\theta,\vartheta) := \left\{\frac{1}{N}\sum_{i=1}^N \big(f_\theta(Z_i)-f_\vartheta(Z_i)\big)^2\right\}^{1/2}.$$

By the Lipschitz assumption, $d_N(\theta,\vartheta)\le L\|\theta-\vartheta\|$, while $\operatorname{diam}(\Theta,d_N)\le 2B$. Hence

$$N(\epsilon,\Theta,d_N)\le\left(1+\frac{CLD_\Theta}{\epsilon}\right)^d.$$

Dudley's entropy integral therefore gives

$$\begin{aligned}\mathbb{E}_\varepsilon\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^N\varepsilon_i f_\theta(Z_i)\right| &\le \frac{C}{\sqrt{N}}\left[B+\int_0^{2B}\sqrt{d\log\left(1+\frac{CLD_\Theta}{\epsilon}\right)}\,d\epsilon\right]\\ &\le \frac{CB}{\sqrt{N}}\sqrt{d\log\left(C\left(1+\frac{LD_\Theta}{B}\right)\right)}.\end{aligned}$$

Together with symmetrization, this proves equation 121.

For the high-probability bound, define $\Phi(Z_1,\dots,Z_N):=\sup_{\theta\in\Theta}|(P_N-P)f_\theta|$. If only the ith observation is replaced by Z_i', then

$$\big|\Phi(Z_1,\dots,Z_i,\dots,Z_N)-\Phi(Z_1,\dots,Z_i',\dots,Z_N)\big|\le\sup_{\theta\in\Theta}\frac{1}{N}|f_\theta(Z_i)-f_\theta(Z_i')|\le\frac{2B}{N}.$$

Thus McDiarmid's inequality, with $c_i=2B/N$, gives for every $t>0$,

$$\mathbb{P}(\Phi-\mathbb{E}\Phi\ge t)\le\exp\left(-\frac{2t^2}{\sum_{i=1}^N c_i^2}\right)=\exp\left(-\frac{Nt^2}{2B^2}\right).$$

Taking $t=B\sqrt{2\log(1/\delta)/N}$ shows that, with probability at least $1-\delta$,

$$\Phi\le\mathbb{E}\Phi+B\sqrt{\frac{2\log(1/\delta)}{N}}.$$

Combining this with equation 121 and enlarging the universal constant C yields equation 122. □

Lemma E.3 (Moments of bounded empirical averages)**.** *Let $Z_1,\dots,Z_N$ be i.i.d. with $\mathbb{E}Z_1=0$, $|Z_1|\le b$, and $\mathbb{E}Z_1^2\le v$. For $\bar{Z}_N:=N^{-1}\sum_{i=1}^N Z_i$, one has $\mathbb{E}\bar{Z}_N^2\le v/N$, $\mathbb{E}\bar{Z}_N^4\le C\{v^2/N^2+b^2v/N^3\}$, and $\mathbb{E}\bar{Z}_N^8\le C\{v^4/N^4+b^6v/N^7\}$.*

Proof. The second-moment bound is immediate. For $r\in\{4,8\}$, Rosenthal's inequality gives $\mathbb{E}|\sum_{i=1}^N Z_i|^r\le C_r\{(Nv)^{r/2}+N\mathbb{E}|Z_1|^r\}$. Since $\mathbb{E}|Z_1|^r\le b^{r-2}v$, division by N^r gives the stated fourth- and eighth-moment bounds. □

Lemma E.4 (Self-normalized moment and bias bounds)**.** *Let $A>0$ almost surely and let ϕ satisfy $PA=1$, $P\phi=0$, $\|A\|_\infty\le B$, $P(A-1)^2\le\sigma_A^2$, and $\|\phi\|_\infty\le b$, $P\phi^2\le s_2$, $|\phi|\le dA$ a.s. Set $e_N:=\frac{P_N\phi}{P_NA}$. There are universal constants $C,c>0$ such that, whenever $N\ge C(B+\sigma_A^2)$,*

$$\mathbb{E}e_N^2\le C\frac{s_2}{N}+Cd^2e^{-cN/(B+\sigma_A^2)}, \tag{123}$$

$$\mathbb{E}e_N^4\le C\left(\frac{s_2^2}{N^2}+\frac{b^2s_2}{N^3}\right)+Cd^4e^{-cN/(B+\sigma_A^2)},$$

$$\mathbb{E}e_N^8\le C\left(\frac{s_2^4}{N^4}+\frac{b^6s_2}{N^7}\right)+Cd^8e^{-cN/(B+\sigma_A^2)}, \tag{124}$$

$$|\mathbb{E}e_N|\le C\frac{\sqrt{s_2\sigma_A^2}}{N}+C(b+d)e^{-cN/(B+\sigma_A^2)}. \tag{125}$$

Proof. Write $U := P_N\phi$, $T := P_N A - 1$, and $\mathcal{H} := \{|T| \le 1/2\}$. Bernstein's inequality gives $\mathbb{P}(\mathcal{H}^c) \le 2e^{-cN/(B+\sigma_A^2)}$. On $\mathcal{H}$, $|e_N| \le 2|U|$, while globally $|e_N| \le d$. Lemma E.3, applied to $Z_i = \phi_i$ with $v = s_2$, gives the required second, fourth, and eighth moments of U. Splitting over $\mathcal{H}$ and $\mathcal{H}^c$ proves equation 123–equation 124.

For the bias, on $\mathcal{H}$, $\frac{U}{1+T} = U - \frac{UT}{1+T}$. Since $\mathbb{E}U = 0$, $|\mathbb{E}[U\mathbf{1}_{\mathcal{H}}]| \le b\mathbb{P}(\mathcal{H}^c)$, while

$$\mathbb{E}\left[\frac{|UT|}{1+T}\mathbf{1}_{\mathcal{H}}\right] \le 2\sqrt{\mathbb{E}U^2\mathbb{E}T^2} \le 2\frac{\sqrt{s_2\sigma_A^2}}{N}.$$

The contribution of $\mathcal{H}^c$ is at most $d\mathbb{P}(\mathcal{H}^c)$, proving equation 125. □

Lemma E.5 (Self-normalized high-probability deviation)**.** *Let $A > 0$ almost surely and let ϕ satisfy*

$$PA = 1, \qquad P\phi = 0, \qquad \|A\|_\infty \le B, \qquad P(A-1)^2 \le \sigma_A^2,$$

and

$$\|\phi\|_\infty \le b, \qquad P\phi^2 \le s_2.$$

For an i.i.d. sample of size N, define

$$e_N := \frac{P_N\phi}{P_N A}.$$

There exist universal constants $C, c > 0$ such that, for every

$$1 \le u \le c\,\frac{N}{B + \sigma_A^2 + 1},$$

with probability at least $1 - 4e^{-u}$,

$$|e_N| \le C\left\{\sqrt{\frac{s_2 u}{N}} + \frac{bu}{N}\right\}. \tag{126}$$

Proof. Write

$$U := P_N\phi, \qquad T := P_N A - 1.$$

Since $P(A-1) = 0$, $P(A-1)^2 \le \sigma_A^2$, and $|A-1| \le B+1$, Bernstein's inequality gives

$$\mathbb{P}\left(|T| > \frac{1}{2}\right) \le 2\exp\left\{-c\frac{N}{B+\sigma_A^2+1}\right\}.$$

Hence, for $u \le cN/(B+\sigma_A^2+1)$, after adjusting the universal constant c if necessary,

$$\mathbb{P}\left(|T| > \frac{1}{2}\right) \le 2e^{-u}. \tag{127}$$

Since $P\phi = 0$, $P\phi^2 \le s_2$, and $|\phi| \le b$, Bernstein's inequality also gives, for every $u \ge 1$,

$$\mathbb{P}\left(|U| > C\left\{\sqrt{\frac{s_2 u}{N}} + \frac{bu}{N}\right\}\right) \le 2e^{-u}. \tag{128}$$

On the event $|T| \le 1/2$, $P_N A = 1 + T \ge 1/2$, and therefore $|e_N| \le 2|U|$. Combining equation 127 and equation 128 by a union bound proves equation 126. □

E.2 Compact vector-field approximation

The compact-entry argument separates the finite-context bias from the finite-task fluctuation.

Proposition E.6 (Uniform finite-context bias on compact rectangles)**.** *There is a universal constant $C > 0$ such that, for every $R \ge 1$ and $n \ge 2$,*

$$\sup_{\theta \in K_R} \|\nabla\mathcal{R}_{\rho,n}(\theta) - \nabla\mathcal{R}_\rho(\theta)\| \le Ce^{2R}\sqrt{\frac{\log\big(C(1+R)\big)}{n}}. \tag{129}$$

Proof. Let $\mathcal{A}_{\mathrm{mom}} := \{(0,0),(1,0),(0,1),(2,0),(1,1)\}$ and $\psi_{a,b}(x,y) := x^a y^b$. Define $Q^{(n)}_{a,b}(\theta) := n^{-1}\sum_{i=1}^n \psi_{a,b}(X_i,Y_i)e^{\lambda Y_i+\mu X_i}$ and $Q_{a,b}(\theta;\alpha) := \mathbb{E}[\psi_{a,b}(X,Y)e^{\lambda Y+\mu X} \mid \alpha]$. For fixed α and $(a,b) \in \mathcal{A}_{\mathrm{mom}}$, the class $f_\theta(x,y) := \psi_{a,b}(x,y)e^{\lambda y+\mu x},\ \theta \in K_R$, has envelope e^{2R}, parameter-Lipschitz constant at most $\sqrt{2}e^{2R}$, and parameter diameter $\sqrt{2}R$. Lemma E.2 therefore gives

$$\mathbb{E}\left[\sup_{\theta\in K_R}\max_{(a,b)\in\mathcal{A}_{\mathrm{mom}}}|Q^{(n)}_{a,b}(\theta)-Q_{a,b}(\theta;\alpha)|\,\middle|\,\alpha\right] \le Ce^{2R}\sqrt{\frac{\log\big(C(1+R)\big)}{n}}. \tag{130}$$

For $q=(q_{0,0},q_{1,0},q_{0,1},q_{2,0},q_{1,1}),\ q_{0,0}>0$, define

$$\Gamma_\alpha(q) := \left(\frac{q_{1,0}}{q_{0,0}}-\alpha\right)\begin{pmatrix}\dfrac{q_{1,1}}{q_{0,0}}-\dfrac{q_{1,0}q_{0,1}}{q_{0,0}^2}\\[2ex] \dfrac{q_{2,0}}{q_{0,0}}-\dfrac{q_{1,0}^2}{q_{0,0}^2}\end{pmatrix}.$$

Then $g_n(\theta;Z)=\Gamma_\alpha((Q^{(n)}_{a,b}(\theta))_{(a,b)\in\mathcal{A}_{\mathrm{mom}}})$ and $g(\theta;\alpha)=\Gamma_\alpha((Q_{a,b}(\theta;\alpha))_{(a,b)\in\mathcal{A}_{\mathrm{mom}}})$.

Because the raw weight is at least one on K_R, the relevant moment vectors satisfy $q_{0,0}\ge 1, 0\le q_{a,b}\le q_{0,0}$, and Γ_α is uniformly Lipschitz on this set. Hence

$$\mathbb{E}_{\mathcal{D}|\alpha}\left[\sup_{\theta\in K_R}\|g_n(\theta;Z)-g(\theta;\alpha)\|\right] \le Ce^{2R}\sqrt{\frac{\log\big(C(1+R)\big)}{n}}. \tag{131}$$

Since $\nabla\mathcal{R}_{\rho,n}(\theta)=\mathbb{E}_{\alpha,\mathcal{D}}g_n(\theta;Z)$ and $\nabla\mathcal{R}_\rho(\theta)=\mathbb{E}_\alpha g(\theta;\alpha)$, we obtain

$$\sup_{\theta\in K_R}\|\nabla\mathcal{R}_{\rho,n}(\theta)-\nabla\mathcal{R}_\rho(\theta)\| \le \mathbb{E}_\alpha\mathbb{E}_{\mathcal{D}|\alpha}\left[\sup_{\theta\in K_R}\|g_n(\theta;Z)-g(\theta;\alpha)\|\right],$$

and equation 129 follows from equation 131. □

Proposition E.7 (Uniform finite-task concentration)**.** *There is a universal constant $C>0$ such that, for every $R\ge 1$, $m\ge 2$, and $\delta\in(0,1)$, with probability at least $1-\delta$,*

$$\sup_{\theta\in K_R}\|\nabla\widehat{\mathcal{R}}_{m,n}(\theta)-\nabla\mathcal{R}_{\rho,n}(\theta)\| \le C\sqrt{\frac{\log\big(C(1+R)/\delta\big)}{m}}. \tag{132}$$

Proof. Write $g_j(\theta) := g_n(\theta;Z_j)$ and $\bar g_n(\theta) := \mathbb{E}g_j(\theta) = \nabla\mathcal{R}_{\rho,n}(\theta)$. By Lemma E.1, each coordinate class $\{Z\mapsto[g_n(\theta;Z)]_r : \theta\in K_R\},\ r\in\{1,2\}$, is uniformly bounded and Lipschitz in θ. Lemma E.2, followed by a union bound over the two coordinates, proves equation 132. □

Proposition E.8 (Dudley compact vector-field approximation)**.** *Let $K\subset[0,R]^2$ be compact with $R\ge 1$. For $m,n\ge 2$ and $\delta\in(0,1)$, define*

$$\varepsilon_{m,n}(R,\delta) := C\sqrt{\frac{\log(C(1+R)/\delta)}{m}} + Ce^{2R}\sqrt{\frac{\log(C(1+R))}{n}}. \tag{133}$$

Then, with probability at least $1-\delta$, $\sup_{\theta\in K}\|\widehat v_{m,n}(\theta)-v(\theta)\| \le \varepsilon_{m,n}(R,\delta)$. For fixed R, the error is $O_R\{\sqrt{\log(C/\delta)/m}+n^{-1/2}\}$.

Proof of Proposition E.8. Use the decomposition $\widehat v_{m,n}-v = -(\nabla\widehat{\mathcal{R}}_{m,n}-\nabla\mathcal{R}_{\rho,n}) - (\nabla\mathcal{R}_{\rho,n}-\nabla\mathcal{R}_\rho)$. Since $K\subset K_R$, Propositions E.7 and E.6 give the result. □

E.3 One-task estimates in the logarithmic tube

For fixed $\theta=(\lambda,\mu)$ and $\alpha\in[\rho,\ell]$, define $W_\theta(X,Y) := e^{\lambda Y+\mu X}$, $\mathcal{Z}_{\theta,\alpha} := \mathbb{E}[W_\theta(X,Y)\mid\alpha]$, and $W_{\theta,\alpha} := W_\theta/\mathcal{Z}_{\theta,\alpha}$. Let P_α denote conditional expectation given α. In particular, $P_\alpha W_{\theta,\alpha}=1$.

Lemma E.9 (Endpoint weight geometry)**.** *After increasing μ_* if necessary, the following bounds hold uniformly over $\theta = (\lambda, \mu) \in \mathcal{G}$ and $\alpha \in [\rho, \ell]$. Let $s := \mu(\alpha - \rho)$, $p := \mathbb{E}_\theta[Y \mid \alpha]$, $q := 1 - p$, $F := F(\theta; \alpha)$, $\sigma_X^2 := \mathrm{Var}_\theta(X \mid \alpha)$, and $\mathsf{C} := \mathrm{Cov}_\theta(X, Y \mid \alpha)$. Then*

$$q \lesssim \frac{e^{-s}}{\mu}, \qquad |F - \alpha| \lesssim \frac{1}{\mu}, \qquad \sigma_X^2 \lesssim \frac{1}{\mu^2} + \frac{e^{-s}}{\mu}, \qquad |\mathsf{C}| \lesssim \frac{e^{-s}}{\mu}.$$

Moreover, the normalized population weight satisfies

$$W_{\theta,\alpha}(x, 1) \lesssim \mu e^{-\mu(\alpha - x)}, \quad 0 \le x \le \alpha, \qquad W_{\theta,\alpha}(x, 0) \lesssim e^{-s} e^{-\mu(1-x)}, \quad \alpha < x \le 1.$$

In particular, $P_\alpha W_{\theta,\alpha}^2 \lesssim \mu$.

Proof. Write $I_L := \int_0^\alpha e^{\mu x}\, dx$ and $I_R := \int_\alpha^1 e^{\mu x}\, dx$. Since $\alpha, 1 - \alpha \ge \rho$, $I_L \asymp e^{\mu\alpha}/\mu$ and $I_R \asymp e^{\mu}/\mu$. Hence, with left and right tilted masses $L = e^\lambda I_L$ and $R = I_R$,

$$\frac{L}{R} \asymp e^{\mu(k+\alpha-1)} \ge \mu e^{c_0 + s},$$

because $k \ge b_{c_0}(\mu)$. Thus $q = R/(L + R) \lesssim e^{-s}/\mu$.

Let X_L and X_R denote draws from the left and right tilted component laws, respectively. By the explicit truncated-exponential formulas in Lemma B.1, $\mathbb{E}(\alpha - X_L)^r + \mathbb{E}(1 - X_R)^r \lesssim \mu^{-r}$ for every fixed integer $r \ge 1$. Combining these component-moment bounds with the two-component formulas gives

$$|F - \alpha| \lesssim \mu^{-1}, \qquad \sigma_X^2 \lesssim \mu^{-2} + \frac{e^{-s}}{\mu}, \qquad |\mathsf{C}| \lesssim \frac{e^{-s}}{\mu}.$$

Finally, the normalizing constant satisfies $\mathcal{Z}_{\theta,\alpha} \gtrsim e^{\lambda + \mu\alpha}/\mu$. The two stated weight envelopes follow directly. Squaring them and integrating over the two intervals gives $P_\alpha W_{\theta,\alpha}^2 \lesssim \mu$. □

Corollary E.10 (Influence-function envelopes)**.** *Under the conditions of Lemma E.9, define*

$$\phi_F := W_{\theta,\alpha}(X - F), \qquad \phi_p := W_{\theta,\alpha}(Y - p),$$
$$\phi_{\sigma^2} := W_{\theta,\alpha}\big((X - F)^2 - \sigma_X^2\big), \quad \phi_{\mathsf{C}} := W_{\theta,\alpha}\big((X - F)(Y - p) - \mathsf{C}\big).$$

Each influence function is P_α-centered, and uniformly over $\theta \in \mathcal{G}$ and $\alpha \in [\rho, \ell]$,

	$\|\phi\|_\infty$	$P_\alpha \phi^2$
ϕ_F	$\lesssim 1$	$\lesssim \mu^{-1}$
ϕ_p	$\lesssim e^{-s}$	$\lesssim e^{-2s}\mu^{-1}$
ϕ_{σ^2}	$\lesssim \mu^{-1} + e^{-s}$	$\lesssim \mu^{-3} + e^{-2s}\mu^{-1}$
ϕ_{C}	$\lesssim e^{-s}$	$\lesssim e^{-2s}\mu^{-1}$.

Moreover, $|\phi_\diamond| \le 2W_{\theta,\alpha}$ for $\diamond \in \{F, p, \sigma^2, \mathsf{C}\}$.

Proof. Centering follows directly from the definitions. On the left interval set $u = \alpha - X$, and on the right set $v = 1 - X$. Lemma E.9 reduces all bounds to

$$\sup_{u \ge 0} \mu e^{-\mu u} \left(u + \frac{1}{\mu}\right)^r \lesssim \mu^{-(r-1)}$$

and

$$\mu^2 \int_0^\infty e^{-2\mu u} \left(u + \frac{1}{\mu}\right)^{2r} du \lesssim \mu^{-(2r-1)}.$$

Applying these with $r = 1$ for ϕ_F and ϕ_{C} and $r = 2$ for ϕ_{σ^2}, together with $q \lesssim e^{-s}/\mu$ and the population bounds of Lemma E.9, gives the table. The domination $|\phi_\diamond| \le 2W_{\theta,\alpha}$ follows from $X, Y, F, p \in [0, 1]$. □

For a context $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, define

$$F_n := F_n(\theta; \mathcal{D}), \qquad p_n := \mathbb{E}_{\theta,n} Y, \qquad \sigma^2_{X,n} := \mathrm{Var}_{\theta,n}(X), \qquad \mathsf{C}_n := \mathrm{Cov}_{\theta,n}(X, Y),$$

and

$$e_F := \frac{P_n \phi_F}{P_n W_{\theta,\alpha}}, \qquad e_p := \frac{P_n \phi_p}{P_n W_{\theta,\alpha}}, \qquad e_{\sigma^2} := \frac{P_n \phi_{\sigma^2}}{P_n W_{\theta,\alpha}}, \qquad e_{\mathsf{C}} := \frac{P_n \phi_{\mathsf{C}}}{P_n W_{\theta,\alpha}}.$$

Then

$$F_n - F = e_F, \qquad p_n - p = e_p, \tag{134}$$

and

$$\sigma^2_{X,n} - \sigma^2_X = e_{\sigma^2} - e_F^2, \qquad \mathsf{C}_n - \mathsf{C} = e_{\mathsf{C}} - e_F e_p. \tag{135}$$

Lemma E.11 (One-task gradient moment and bias bounds)**.** *There exist constants $C, c > 0$ such that, uniformly over $\theta = (\lambda, \mu) \in \mathcal{G}$ whenever $n \geq C\mu$,*

$$\mathbb{E}_\alpha \|g(\theta; \alpha)\|^2 \leq \frac{C}{\mu^5}, \tag{136}$$

$$\mathbb{E}_{\alpha,\mathcal{D}} \|g_n(\theta; Z) - g(\theta; \alpha)\|^2 \leq \frac{C}{\mu^4 n} + Ce^{-cn/\mu}, \tag{137}$$

$$\|\mathbb{E}_{\alpha,\mathcal{D}} g_n(\theta; Z) - \mathbb{E}_\alpha g(\theta; \alpha)\| \leq \frac{C}{\mu^2 n} + Ce^{-cn/\mu}. \tag{138}$$

If $\mathcal{H} \subseteq \{P_n W_{\theta,\alpha} \geq 1/2\}$, then

$$\mathbb{E}_{\alpha,\mathcal{D}} \left[\|g_n(\theta; Z) - g(\theta; \alpha)\|^2 \mathbf{1}_{\mathcal{H}} \right] \leq \frac{C}{\mu^4 n}. \tag{139}$$

Proof. Apply Lemma E.4 with $A = W_{\theta,\alpha}$ and $\phi \in \{\phi_F, \phi_p, \phi_{\sigma^2}, \phi_{\mathsf{C}}\}$. By Lemma E.9, $\|A\|_\infty + P_\alpha(A-1)^2 \lesssim \mu$, while Corollary E.10 gives the required envelope, second-moment, and domination bounds, with $d = 2$. Hence

$$\mathbb{E}[e_F^2 \mid \alpha] \leq \frac{C}{\mu n} + Ce^{-cn/\mu}, \qquad \mathbb{E}[e_p^2 \mid \alpha] \leq \frac{Ce^{-2s}}{\mu n} + Ce^{-cn/\mu}, \tag{140}$$

and

$$\mathbb{E}[e_{\sigma^2}^2 \mid \alpha] \leq \frac{C}{n} \left(\frac{1}{\mu^3} + \frac{e^{-2s}}{\mu} \right) + Ce^{-cn/\mu},$$

$\mathbb{E}[e_{\mathsf{C}}^2 \mid \alpha] \leq \frac{Ce^{-2s}}{\mu n} + Ce^{-cn/\mu}$.

The fourth- and eighth-moment bounds in Lemma E.4, together with Corollary E.10 and $n \geq C\mu$, give by Cauchy–Schwarz

$$\mathbb{E}[(e_F e_p)^2 + (e_F e_{\mathsf{C}})^2 \mid \alpha] \leq \frac{Ce^{-2s}}{\mu^2 n^2} + Ce^{-cn/\mu},$$

$$\mathbb{E}[e_F^2 e_{\sigma^2}^2 \mid \alpha] \leq C \left\{ \frac{1}{\mu^4 n^2} + \frac{e^{-2s}}{\mu^2 n^2} \right\} + Ce^{-cn/\mu},$$

$$\mathbb{E}[e_F^6 + (e_F^2 e_p)^2 \mid \alpha] \leq C \left\{ \frac{1}{\mu^3 n^3} + \frac{e^{-2s}}{\mu^3 n^3} \right\} + Ce^{-cn/\mu}.$$

These are the only higher-order products needed below.

Write $B := F - \alpha$. Then $g(\theta; \alpha) = B \begin{pmatrix} \mathsf{C} \\ \sigma_X^2 \end{pmatrix}$, while

$$(F_n - \alpha)\mathsf{C}_n - B\mathsf{C} = \mathsf{C}\, e_F + B(\mathsf{C}_n - \mathsf{C}) + e_F(\mathsf{C}_n - \mathsf{C}), \tag{141}$$

$$(F_n - \alpha)\sigma^2_{X,n} - B\sigma^2_X = \sigma^2_X e_F + B(\sigma^2_{X,n} - \sigma^2_X) + e_F(\sigma^2_{X,n} - \sigma^2_X). \tag{142}$$

Using Lemma E.9, the ratio identities, and Cauchy–Schwarz gives

$$\mathbb{E}_{\mathcal{D}|\alpha} \|g_n - g\|^2 \leq C \left\{ \frac{1}{\mu^5 n} + \frac{e^{-s}}{\mu^4 n} + \frac{e^{-2s}}{\mu^3 n} \right\} + Ce^{-cn/\mu}. \tag{143}$$

Also, $\|g(\theta;\alpha)\|^2 \le C\{\mu^{-6} + e^{-s}\mu^{-5} + e^{-2s}\mu^{-4}\}$. Since

$$s = \mu(\alpha - \rho), \qquad \int_\rho^\ell e^{-qs}\, d\alpha \le \frac{C_q}{\mu}, \qquad q > 0, \tag{144}$$

integrating over α proves equation 136 and equation 137.

For the bias, equation 125, the identities equation 134–equation 135, and the expansion equation 141 give

$$\left\|\mathbb{E}_{\mathcal{D}|\alpha}[g_n - g]\right\| \le C\left\{\frac{1}{\mu^2 n} + \frac{e^{-s}}{\mu n}\right\} + Ce^{-cn/\mu}.$$

Since equation 144 gives $\int_\rho^\ell e^{-s}\, d\alpha \le \frac{C}{\mu}$, we obtain $\|\mathbb{E}_{\alpha,\mathcal{D}} g_n(\theta;Z) - \mathbb{E}_\alpha g(\theta;\alpha)\| \le C/(\mu^2 n) + Ce^{-cn/\mu}$, which proves equation 138.

Finally, on any $\mathcal{H} \subseteq \{P_n W_{\theta,\alpha} \ge 1/2\}$, $|e_\diamond| \le 2|P_n \phi_\diamond|$. Repeating the same moment calculation on $\mathcal{H}$ removes the denominator-failure term and gives equation 139. □

Lemma E.12 (Fixed-point control at shell resolution)**.** *There are constants $C, c > 0$ such that, for every $R \ge \mu_*$, $\theta \in \widetilde{\mathcal{S}}_R$, $\alpha \in [\rho, \ell]$, and $n \ge CR$, conditional on α, with probability at least $1 - Ce^{-cn/R}$, one has*

$$P_n W_{\theta,\alpha} \ge \frac{3}{4} \tag{145}$$

and

$$|F_n - F| + |p_n - p| + |\sigma^2_{X,n} - \sigma^2_X| + |\mathsf{C}_n - \mathsf{C}| \le \frac{C}{R}.$$

Proof. Since $\mu \asymp R$ on $\widetilde{\mathcal{S}}_R$, Lemma E.9 gives $\|W_{\theta,\alpha}\|_\infty + P_\alpha(W_{\theta,\alpha} - 1)^2 \le CR$. Bernstein's inequality therefore yields $P_n W_{\theta,\alpha} \ge 3/4$ with conditional failure probability at most $Ce^{-cn/R}$.

Apply Lemma E.5 to each influence function in Corollary E.10, with $u = c_1 n/R$. Since $B + \sigma_A^2 + 1 \lesssim R$, choosing $c_1 > 0$ sufficiently small places u in the admissible range of the lemma, while $n \ge CR$ ensures $u \ge 1$. Thus, with conditional failure probability at most $Ce^{-cn/R}$,

$$|e_F| + |e_p| + |e_{\sigma^2}| + |e_{\mathsf{C}}| \le \frac{C}{R}.$$

Intersecting this event with the denominator event above and using a union bound preserves the same failure order. The identities equation 134–equation 135 then prove the claim. □

E.4 Uniform shellwise concentration

Lemma E.13 (Uniform good-denominator event on one shell)**.** *There are constants $C, c > 0$ such that, for every $R \ge \mu_*$ and $n \ge CR$, there is an event $E_R(Z)$ satisfying*

$$\mathbb{P}\{E_R(Z)^c\} \le C\exp\left(-\frac{cn}{R} + C\log(CR)\right). \tag{146}$$

On $E_R(Z)$,

$$\inf_{\theta \in \widetilde{\mathcal{S}}_R} P_n W_{\theta,\alpha} \ge \frac{1}{2}, \tag{147}$$

and

$$\sup_{\theta \in \widetilde{\mathcal{S}}_R} \mu^2 \|g_n(\theta;Z)\| \le C. \tag{148}$$

Proof. Fix a sufficiently small universal $c_{\mathrm{net}} > 0$ and choose an η_R-net $\mathcal{N}_R$ of $\widetilde{\mathcal{S}}_R$ with $\eta_R := \frac{c_{\mathrm{net}}}{R}$. Since $\widetilde{\mathcal{S}}_R$ has diameter $O(R)$ in two dimensions, $|\mathcal{N}_R| \le \left(1 + \frac{CR}{\eta_R}\right)^2 \le CR^4$. Lemma E.12 and a union bound over $\mathcal{N}_R$ give equation 146 and, at every net point, $P_n W_{\theta,\alpha} \ge \frac{3}{4}$, together with $O(R^{-1})$ control of

$$F_n - F, \qquad p_n - p, \qquad \sigma^2_{X,n} - \sigma^2_X, \qquad \mathsf{C}_n - \mathsf{C}.$$

The maps $F_n, F, p_n, p, \sigma^2_{X,n}, \sigma^2_X, \mathsf{C}_n, \mathsf{C}$ have universal parameter-Lipschitz constants. Hence, if $\vartheta \in \mathcal{N}_R$ is nearest to $\theta \in \widetilde{\mathcal{S}}_R$, $|H(\theta) - H(\vartheta)| \le C\eta_R = \frac{Cc_{\mathrm{net}}}{R}$ for each such map H. Thus the mesh $\eta_R = c_{\mathrm{net}}/R$ produces exactly the $O(R^{-1})$ extension error required here.

Moreover, $\nabla_\theta \log P_n W_{\theta,\alpha} = \mathbb{E}_{\theta,n}(Y, X) - \mathbb{E}_\theta(Y, X)$, so $\theta \mapsto \log P_n W_{\theta,\alpha}$ is universally Lipschitz. Choosing c_{net} sufficiently small extends the denominator lower bound from $3/4$ on the net to $1/2$ on the full shell, proving equation 147.

Finally, the uniform $O(R^{-1})$ deviations established above, together with Lemma E.9 and $\mu \asymp R$, give $|F_n - \alpha| \le C/\mu$ and $\sigma^2_{X,n} + |\mathsf{C}_n| \le C/\mu$ uniformly over $\widetilde{\mathcal{S}}_R$. Hence $\|g_n(\theta; Z)\| \le C/\mu^2$, proving equation 148. □

For $\diamond \in \{\mu, -, +\}$, define

$$
\begin{aligned}
f_{\mu,\theta}(Z) &:= \mu^3 v_{n,\mu}(\theta; Z), && \theta \in \widetilde{\mathcal{S}}_R,\\
f_{-,\theta}(Z) &:= \mu^4 \langle \nabla h_-(\theta), v_n(\theta; Z)\rangle, && \theta \in \partial_- \widetilde{\mathcal{S}}_R,\\
f_{+,\theta}(Z) &:= \mu^4 \langle \nabla h_+(\theta), v_n(\theta; Z)\rangle, && \theta \in \partial_+ \widetilde{\mathcal{S}}_R.
\end{aligned}
$$

Lemma E.14 (Directional envelopes, variance, and regularity)**.** *There is a constant $C > 0$ such that, for every $R \ge \mu_*$ and $n \ge CR$, if $\widetilde{f}_{\diamond,\theta}(Z) := f_{\diamond,\theta}(Z)\mathbf{1}_{E_R(Z)}$, then, on the relevant parameter domain for each $\diamond \in \{\mu, -, +\}$,*

$$\sup_{\theta,Z} |f_{\diamond,\theta}(Z)| \le CR^3, \tag{149}$$

$$\sup_{\theta,Z} |\widetilde{f}_{\diamond,\theta}(Z)| \le CR, \tag{150}$$

$$\sup_{\theta} \mathbb{E}\widetilde{f}_{\diamond,\theta}(Z)^2 \le C\left(R + \frac{R^2}{n}\right), \tag{151}$$

$$|f_{\diamond,\theta}(Z) - f_{\diamond,\vartheta}(Z)| \le CR^3\|\theta - \vartheta\|. \tag{152}$$

Proof. Lemma E.1 gives $\|g_n\| \le C$ globally. Since $\mu \asymp R$ on $\widetilde{\mathcal{S}}_R$ and $\mu^4\|\nabla h_\pm(\theta)\| \le C\mu^3$, the global envelope is CR^3. On E_R, Lemma E.13 improves $\|g_n\|$ to $C\mu^{-2}$, giving the truncated envelope CR.

For fixed θ, $g_n = g + (g_n - g)$. By equation 136 and equation 139, $\mathbb{E}_\alpha\|g\|^2 \le C\mu^{-5}$ and $\mathbb{E}_{\alpha,\mathcal{D}}[\|g_n - g\|^2\mathbf{1}_{E_R}] \le C/(\mu^4 n)$. Each directional scaling has magnitude at most $C\mu^3$, so $\mathbb{E}\widetilde{f}^2_{\diamond,\theta} \le C\mu^6\{\mu^{-5} + (\mu^4 n)^{-1}\} \le C(R + R^2/n)$.

Finally, $\|\nabla(\mu^3)\| \le CR^2$, and Lemma D.3 gives $\|D(\mu^4\nabla h_\pm(\theta))\|_{\mathrm{op}} \le CR^2$ and $\mu^4\|\nabla h_\pm(\theta)\| \le CR^3$. Together with the global bounds on g_n and Dg_n, these imply equation 152. Since $E_R(Z)$ does not depend on the particular θ inside the enlarged shell, the same bound holds for the truncated classes. □

Lemma E.15 (Finite-dimensional Lipschitz Bernstein bound)**.** *Let $K \subset \mathbb{R}^d$ have diameter at most D_K, and suppose*

$$|f_\theta(Z)| \le b, \qquad \mathbb{E}f_\theta(Z)^2 \le s_2, \qquad |f_\theta(Z) - f_\vartheta(Z)| \le L_K\|\theta - \vartheta\|.$$

Then, for every $\eta > 0$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\sup_{\theta \in K} |(P_m - P)f_\theta| \le C\left\{\sqrt{\frac{s_2 L_\eta(\delta)}{m}} + \frac{bL_\eta(\delta)}{m} + L_K\eta\right\},$$

where $L_\eta(\delta) := d\log\left(1 + \frac{CD_K}{\eta}\right) + \log\frac{C}{\delta}$.

Proof. Take an η-net of K with cardinality at most $\left(1 + \frac{CD_K}{\eta}\right)^d$. Bernstein's inequality and a union bound give the first two terms on the net. If ϑ is nearest to θ, then $|(P_m - P)(f_\theta - f_\vartheta)| \le P_m|f_\theta - f_\vartheta| + P|f_\theta - f_\vartheta| \le 2L_K\eta$. Absorb the factor two into C. □

Theorem E.16 (Shellwise directional approximation)**.** *There is a fixed constant $C_{\mathrm{res}} \geq 1$ such that the following holds. For $R \geq \mu_*$ and $\delta \in (0,1)$, set $L_R(\delta) := \log\left(\frac{C_{\mathrm{res}}(1+R)mn}{\delta}\right)$. If $n \geq C_{\mathrm{res}} R L_R(\delta)$, then, with probability at least $1-\delta$,*

$$\max\{\mathcal{E}_\mu(R), \mathcal{E}_-(R), \mathcal{E}_+(R)\} \leq C_{\mathrm{res}} \left\{ \sqrt{\frac{RL_R(\delta)}{m}} + \frac{RL_R(\delta)}{m} + \frac{R}{n} \right\}. \tag{153}$$

Proof of Theorem E.16. Fix $\diamond \in \{\mu, -, +\}$. By Lemma E.14, the truncated class satisfies

$$d = 2, \qquad D_K \leq CR, \qquad b = CR, \qquad s_2 = C\left(R + \frac{R^2}{n}\right), \qquad L_K = CR^3.$$

The $O(R^{-1})$ net in Lemma E.13 is sufficient there because the quantities being extended have $O(1)$ Lipschitz constants. Here the scaled directional class has $L_K = O(R^3)$, so we choose the finer sample-dependent mesh

$$\eta := \frac{1}{CmR^3}. \tag{154}$$

Then $L_K \eta \leq CR^3 \frac{1}{CmR^3} \leq \frac{C}{m}$. Moreover,

$$\begin{aligned} L_\eta(\delta) &= 2\log\left(1 + \frac{CD_K}{\eta}\right) + \log\frac{C}{\delta} \\ &\leq C\log(CmR^4) + \log\frac{C}{\delta} \\ &\leq CL_R(\delta). \end{aligned}$$

Thus the net-extension contribution is $O(m^{-1})$, while the finer covering changes only the logarithmic factor.

Under the context condition in Theorem E.16, $n \geq CRL_R(\delta)$, so $s_2 \leq CR$. Lemma E.15 therefore yields

$$\begin{aligned} \sup_\theta |(P_m - P)\widetilde{f}_{\diamond,\theta}| &\leq C\left\{ \sqrt{\frac{RL_R(\delta)}{m}} + \frac{RL_R(\delta)}{m} + \frac{1}{m} \right\} \\ &\leq C\left\{ \sqrt{\frac{RL_R(\delta)}{m}} + \frac{RL_R(\delta)}{m} \right\}. \end{aligned}$$

Applying this bound to the three classes with failure probability $\delta/9$ each and taking a union bound gives, with probability at least $1 - \delta/3$,

$$\max_{\diamond\in\{\mu,-,+\}} \sup_\theta |(P_m - P)\widetilde{f}_{\diamond,\theta}| \leq C\left\{ \sqrt{\frac{RL_R(\delta)}{m}} + \frac{RL_R(\delta)}{m} \right\}. \tag{155}$$

Set $p_R^{\mathrm{bad}}(n) := C\exp\left(-\frac{cn}{R} + C\log(CR)\right)$. By Lemma E.13, $\mathbb{P}\left(\bigcup_{j=1}^m E_R(Z_j)^c\right) \leq mp_R^{\mathrm{bad}}(n)$. The context condition ensures, after increasing its numerical constant if necessary, $mp_R^{\mathrm{bad}}(n) \leq \delta/3$, $R^3 p_R^{\mathrm{bad}}(n) \leq CR/n$. On the event that all observed task blocks belong to E_R, $P_m f_{\diamond,\theta} = P_m \widetilde{f}_{\diamond,\theta}$, while equation 149 gives $|P(f_{\diamond,\theta} - \widetilde{f}_{\diamond,\theta})| \leq CR^3 p_R^{\mathrm{bad}}(n) \leq CR/n$. Together with equation 155, this controls the finite-task fluctuation $\widehat{v}_{m,n} - \bar{v}_n$ in all three directions.

It remains to add the finite-context bias. By equation 138, $\|\bar{v}_n(\theta) - v(\theta)\| \leq C\left(\frac{1}{\mu^2 n} + e^{-cn/\mu}\right)$. Multiplying by μ^3 in the radial direction and by $\mu^4\|\nabla h_\pm(\theta)\| \leq C\mu^3$ in the barrier directions gives $C\left(\frac{R}{n} + R^3 e^{-cn/R}\right) \leq C\frac{R}{n}$ under the same context condition. The exceptional events have total probability at most $2\delta/3 < \delta$, proving equation 153. □

Proposition E.17 (Shellwise speed approximation)**.** *After increasing C_{res} if necessary, under the same assumptions as Theorem E.16, with probability at least $1-\delta$,*

$$\mathcal{E}_{\mathrm{sp}}(R) \leq C_{\mathrm{res}} \left\{ \sqrt{\frac{RL_R(\delta)}{m}} + \frac{RL_R(\delta)}{m} + \frac{R}{n} \right\}. \tag{156}$$

Proof. For $r \in \{1,2\}$, define $f_{\mathrm{sp},r,\theta}(Z) := \mu^3[v_n(\theta;Z)]_r$ and $\widetilde{f}_{\mathrm{sp},r,\theta} := f_{\mathrm{sp},r,\theta}\mathbf{1}_{E_R(Z)}$. The argument of Lemma E.14 applies verbatim to these coordinate classes: on $\widetilde{\mathcal{S}}_R$,

$$|\widetilde{f}_{\mathrm{sp},r,\theta}| \le CR, \qquad \mathbb{E}\widetilde{f}^2_{\mathrm{sp},r,\theta} \le C\left(R + \frac{R^2}{n}\right),$$

while the untruncated class is CR^3-Lipschitz in θ. Applying Lemma E.15 with the same mesh $\eta = (CmR^3)^{-1}$ as in Theorem E.16 yields

$$\sup_{\theta\in\widetilde{\mathcal{S}}_R} |(P_m - P)\widetilde{f}_{\mathrm{sp},r,\theta}| \le C\left\{\sqrt{\frac{RL_R(\delta)}{m}} + \frac{RL_R(\delta)}{m}\right\}.$$

The bad-task truncation and finite-context bias are exactly those in the proof of Theorem E.16 and contribute CR/n. A union bound over $r = 1,2$ and conversion to the Euclidean norm prove equation 156. □

F Proof of the Finite-Pretraining Theorem

The empirical-process section already defined the one-task gradients, the vector fields v, $\bar{v}_n$, and $\widehat{v}_{m,n}$, and the core and enlarged shells. We now analyze empirical GD,

$$\widehat{\theta}^\gamma_{s+1} = \widehat{\theta}^\gamma_s + \gamma\widehat{v}_{m,n}(\widehat{\theta}^\gamma_s), \qquad \widehat{\theta}^\gamma_0 = \theta^\gamma_0. \tag{157}$$

F.1 Two-stage perturbation principle

The proof has two stages, with population GD as the reference dynamics. Before capture, empirical GD tracks population GD on a fixed compact set for a fixed physical time using compact Dudley approximation and discrete Grönwall control. After capture, trajectory tracking is replaced by scale-adapted shell estimates that preserve the radial drift, the two barrier margins, and the speed bound required for discrete no-overshoot.

F.2 Regime I: compact tracking of population GD

Let $s_0 = s_0(\gamma)$ be the deterministic population-GD entry index from Section D. Reduce the population learning-rate threshold if necessary and fix $\gamma_0 > 0$ so that equation 107 and equation 108 hold for every $0 < \gamma \le \gamma_0$. Choose one deterministic compact set K_0 containing, with positive clearance, every θ^γ_s with $0 \le s \le s_0(\gamma)$ and $0 < \gamma \le \gamma_0$. Such a common compact set exists by uniform finite-horizon shadowing on $[0, T_0 + \tau_0]$. Let L_0 be a Lipschitz constant of v on K_0 and $L_{h,0}$ a common Lipschitz constant of h_- and h_+ there.

Define the uniform compact and entry clearances

$$\begin{aligned}
d^{\mathrm{gd}}_{K,0} &:= \inf_{\substack{0<\gamma\le\gamma_0\\ 0\le s\le s_0(\gamma)}} \mathrm{dist}(\theta^\gamma_s, \partial K_0),\\
d^{\mathrm{gd}}_{\mu,0} &:= \inf_{0<\gamma\le\gamma_0} \{\mu^\gamma_{s_0(\gamma)} - \mu_*\},\\
d^{\mathrm{gd}}_{-,0} &:= \inf_{0<\gamma\le\gamma_0} h_-(\theta^\gamma_{s_0(\gamma)}),\\
d^{\mathrm{gd}}_{+,0} &:= \inf_{0<\gamma\le\gamma_0} h_+(\theta^\gamma_{s_0(\gamma)}).
\end{aligned}$$

By the strict post-capture population-flow margin and the uniform Euler shadowing, all four quantities are strictly positive after reducing γ_0 if necessary. Set

$$d^{\mathrm{gd}}_{\mathrm{ent}} := \min\left\{d^{\mathrm{gd}}_{K,0}, d^{\mathrm{gd}}_{\mu,0}, \frac{d^{\mathrm{gd}}_{-,0}}{L_{h,0}}, \frac{d^{\mathrm{gd}}_{+,0}}{L_{h,0}}\right\}. \tag{158}$$

Lemma F.1 (Discrete compact tracking)**.** *Let $K \subset \mathbb{R}^2$ be compact and suppose the population and empirical GD sequences start from the same point. Assume v is L-Lipschitz on K and*

$$\sup_{\theta\in K} \|\widehat{v}_{m,n}(\theta) - v(\theta)\| \le \varepsilon.$$

Then, before either sequence leaves K,

$$\|\widehat{\theta}_s^\gamma - \theta_s^\gamma\| \le \varepsilon \begin{cases} \dfrac{(1+\gamma L)^s - 1}{L}, & L > 0, \\ s\gamma, & L = 0. \end{cases}$$

If this upper bound is strictly smaller than the clearance of the reference sequence from ∂K, then neither sequence exits K over the stated index range.

Proof. Set $e_s := \|\widehat{\theta}_s^\gamma - \theta_s^\gamma\|$. Then, before exit,

$$\begin{aligned} e_{s+1} &\le e_s + \gamma \|\widehat{v}_{m,n}(\widehat{\theta}_s^\gamma) - v(\theta_s^\gamma)\| \\ &\le (1+\gamma L) e_s + \gamma\varepsilon. \end{aligned}$$

Iterating from $e_0 = 0$ proves the displayed bound. The final statement follows by the standard stopped argument: at a first exit time the empirical iterate would have to be at least the reference clearance away from the corresponding population iterate, contradicting the tracking bound. □

Applying Lemma F.1 on K_0, define

$$\Delta_{0,\gamma} := \begin{cases} \dfrac{(1+\gamma L_0)^{s_0} - 1}{L_0}, & L_0 > 0, \\ s_0\gamma, & L_0 = 0. \end{cases}$$

Since $s_0\gamma \le T_0 + \gamma$, the deterministic constant $\Delta_0^{\mathrm{gd}} := \sup_{0<\gamma\le\gamma_0} \Delta_{0,\gamma}$ is finite. Set

$$\varepsilon_{\mathrm{ent}}^{\mathrm{gd}} := \frac{d_{\mathrm{ent}}^{\mathrm{gd}}}{2\Delta_0^{\mathrm{gd}}}.$$

If $\sup_{\vartheta\in K_0} \|\widehat{v}_{m,n}(\vartheta) - v(\vartheta)\| \le \varepsilon_{\mathrm{ent}}^{\mathrm{gd}}$, the stopped argument prevents exit from K_0 and yields $\widehat{\theta}_{s_0}^\gamma \in \mathrm{int}(\mathcal{G})$.

For later uniform radial bounds, define

$$\underline{\mu}_0^{\mathrm{gd}} := \inf_{0<\gamma\le\gamma_0} \mu_{s_0(\gamma)}^\gamma, \qquad \overline{\mu}_0^{\mathrm{gd}} := \sup_{0<\gamma\le\gamma_0} \mu_{s_0(\gamma)}^\gamma. \tag{159}$$

Then $\mu_* < \underline{\mu}_0^{\mathrm{gd}} \le \overline{\mu}_0^{\mathrm{gd}} < \infty$.

F.3 Regime II: discrete continuation with overlapping statistical shells

Recall the core shell $\mathcal{S}_R$, its enlarged statistical shell $\widetilde{\mathcal{S}}_R$, and the normalized errors $\mathcal{E}_\mu(R)$, $\mathcal{E}_-(R)$, $\mathcal{E}_+(R)$, and $\mathcal{E}_{\mathrm{sp}}(R)$ from Appendix E. The core shell governs the dyadic continuation, whereas concentration is proved on the enlargement so that the complete landing step and any potential side-rail crossing point are covered by the same statistical event.

Applying Theorem E.16 and Proposition E.17 with failure probability $\delta/2$ each, taking a union bound, and absorbing the resulting constant inside the logarithm into C_{res} gives, with probability at least $1-\delta$,

$$\max\{\mathcal{E}_\mu(R), \mathcal{E}_-(R), \mathcal{E}_+(R), \mathcal{E}_{\mathrm{sp}}(R)\} \le C_{\mathrm{res}} \left\{ \sqrt{\frac{RL_R(\delta)}{m}} + \frac{RL_R(\delta)}{m} + \frac{R}{n} \right\}, \tag{160}$$

where $L_R(\delta) := \log\left(\frac{C_{\mathrm{res}}(1+R)mn}{\delta}\right)$. Thus the enlarged shell needed for discrete handoff does not change the statistical order of the shell resolution.

Let $a_- > 0$ and $a_+ > 0$ denote the population inward-margin constants from Proposition C.17 and Corollary C.4, respectively. Thus, after increasing μ_* if necessary,

$$\langle \nabla h_\pm(\theta), v(\theta)\rangle \ge \frac{a_\pm}{\mu^4} \qquad \text{on } h_\pm(\theta) = 0.$$

Set

$$c_{\mathrm{dir}} := \frac{1}{4}\min\{c_\mu, a_-, a_+\}, \qquad \overline{c}_\mu := C_\mu + c_{\mathrm{dir}},$$

and fix $C_{\mathrm{sp},0} > 0$. On a shell event satisfying

$$\mathcal{E}_\mu(R), \mathcal{E}_-(R), \mathcal{E}_+(R) \le c_{\mathrm{dir}}, \qquad \mathcal{E}_{\mathrm{sp}}(R) \le C_{\mathrm{sp},0}, \tag{161}$$

the empirical field inherits, throughout the enlarged shell where relevant,

$$\frac{c_\mu}{2\mu^3} \le [\widehat{v}_{m,n}(\theta)]_\mu \le \frac{\overline{c}_\mu}{\mu^3}, \tag{162}$$

the side-rail margins

$$\langle \nabla h_\pm(\theta), \widehat{v}_{m,n}(\theta)\rangle \ge \frac{a_\pm}{2\mu^4} \quad \text{on } \partial_\pm \widetilde{\mathcal{S}}_R, \tag{163}$$

and, with $C_{\mathrm{step}} := C_{\mathrm{sp}} + C_{\mathrm{sp},0}$, the speed bound

$$\|\widehat{v}_{m,n}(\theta)\| \le \frac{C_{\mathrm{step}}}{\mu^3}. \tag{164}$$

Proposition F.2 (Discrete empirical continuation through one shell)**.** *There exists a fixed $\gamma_{\mathrm{shell}} > 0$ such that the following holds. Suppose the enlarged-shell event equation 161 holds at scale $R \ge \mu_*$. If $0 < \gamma \le \gamma_{\mathrm{shell}}$, $\widehat{\theta}^\gamma_{s_*} \in \mathcal{S}_R$, and $\tau_R := \inf\{s \ge s_* : \widehat{\mu}^\gamma_s \ge 2R\}$, then every iterate with $s_* \le s \le \tau_R$ belongs to $\mathcal{G}$. Moreover,*

$$(\widehat{\mu}^\gamma_s)^4 \ge (\widehat{\mu}^\gamma_{s_*})^4 + 2c_\mu\gamma(s - s_*), \qquad s_* \le s \le \tau_R, \tag{165}$$

and for a fixed constant $c^{\mathrm{gd}}_+ > 0$,

$$(\widehat{\mu}^\gamma_s)^4 \le (\widehat{\mu}^\gamma_{s_*})^4 + c^{\mathrm{gd}}_+\gamma(s - s_*), \qquad s_* \le s \le \tau_R. \tag{166}$$

Moreover, $\tau_R < \infty$; more precisely,

$$\tau_R \le s_* + \left\lceil \frac{(2R)^4 - (\widehat{\mu}^\gamma_{s_*})^4}{2c_\mu\gamma} \right\rceil + 1.$$

The landing iterate satisfies

$$\widehat{\theta}^\gamma_{\tau_R} \in \mathcal{S}_{2R}; \tag{167}$$

in particular, $2R \le \widehat{\mu}^\gamma_{\tau_R} < 4R$.

Proof. On equation 161, equations equation 162–equation 164 hold throughout the enlarged shell. For fixed pretraining data, Lemma E.1 makes $\widehat{v}_{m,n}$ globally bounded and globally Lipschitz with universal constants. Hence Lemma D.4 applies on the two-sided enlargement $\mathcal{U}_R$ from Lemma D.6. Every side-rail point in $\mathcal{U}_R$ has $R \le \mu \le 4R$ and lies in $\partial_\pm \widetilde{\mathcal{S}}_R$, so the empirical inward margins are available on the entire rail portion required by the no-overshoot argument.

Consider an iterate before τ_R. It lies radially in the core range $R \le \widehat{\mu}^\gamma_s < 2R$. The speed bound gives a step length $O(\gamma R^{-3})$. After reducing γ_{shell}, the entire segment joining $\widehat{\theta}^\gamma_s$ to $\widehat{\theta}^\gamma_{s+1}$ has radial coordinate in $[R/2, 4R]$. If the segment attempted to cross a side rail, the crossing point would therefore lie in $\partial_\pm \widetilde{\mathcal{S}}_R$, where the empirical inward margin equation 163 is available. The no-overshoot Lemma D.6 then rules out such a crossing. Thus tube membership is preserved, including on the landing step.

The radial update satisfies

$$\widehat{\mu}^\gamma_{s+1} - \widehat{\mu}^\gamma_s = \gamma[\widehat{v}_{m,n}(\widehat{\theta}^\gamma_s)]_\mu \ge \frac{c_\mu\gamma}{2(\widehat{\mu}^\gamma_s)^3}.$$

Convexity of x^4 gives $(\widehat{\mu}^\gamma_{s+1})^4 - (\widehat{\mu}^\gamma_s)^4 \ge 2c_\mu\gamma$, which proves equation 165. Similarly, from

$$0 \le \widehat{\mu}^\gamma_{s+1} - \widehat{\mu}^\gamma_s \le \frac{\overline{c}_\mu\gamma}{(\widehat{\mu}^\gamma_s)^3},$$

expanding the fourth power and using $\widehat{\mu}^\gamma_s \ge \mu_*$ yields equation 166 for a fixed c^{gd}_+.

If τ_R were infinite, equation 165 would hold for every $s \geq s_*$ and would eventually force $(\widehat{\mu}_s^\gamma)^4 > (2R)^4$, contradicting the definition of τ_R. Solving equation 165 for the first index at which $\widehat{\mu}_s^\gamma \geq 2R$ gives the stated upper bound on τ_R.

If $\tau_R = s_*$, then $\widehat{\mu}_{s_*}^\gamma = 2R$ because the starting point lies in $\mathcal{S}_R$, and the handoff conclusion is immediate. Assume therefore that $\tau_R > s_*$. At the landing time, $\widehat{\mu}_{\tau_R-1}^\gamma < 2R$, $\widehat{\mu}_{\tau_R}^\gamma \geq 2R$. The upper speed bound gives $\widehat{\mu}_{\tau_R}^\gamma < 2R + \frac{C\gamma}{R^3} < 4R$ for sufficiently small fixed γ. The preceding no-overshoot argument also gives $\widehat{\theta}_{\tau_R}^\gamma \in \mathcal{G}$. Therefore

$$\widehat{\theta}_{\tau_R}^\gamma \in \mathcal{G} \cap \{2R \leq \mu < 4R\} \subseteq \mathcal{S}_{2R},$$

which proves equation 167. □

F.4 Statistical radius and a sufficient pretraining regime

Choose $R_{\mathrm{cmp}} \geq 1$ with $K_0 \subset [0, R_{\mathrm{cmp}}]^2$. Fix $C_{\mathrm{cmp}} \geq 1$ sufficiently large and $c_{\mathrm{cmp}} > 0$ sufficiently small, and define

$$p_{\mathrm{ent}}(m,n) := C_{\mathrm{cmp}} \exp\{-c_{\mathrm{cmp}} m (\varepsilon_{\mathrm{ent}}^{\mathrm{gd}})^2\}. \tag{168}$$

Set

$$\varepsilon_{\mathrm{cmp}}(m,n) := \varepsilon_{m,n}(R_{\mathrm{cmp}}, p_{\mathrm{ent}}(m,n)), \qquad e_{\mathrm{cmp}}^{\mathrm{gd}}(m,n) := \Delta_0^{\mathrm{gd}} \varepsilon_{\mathrm{cmp}}(m,n),$$

and

$$r_-^{\mathrm{gd}} := \underline{\mu}_0^{\mathrm{gd}} - e_{\mathrm{cmp}}^{\mathrm{gd}}(m,n), \qquad r_+^{\mathrm{gd}} := \overline{\mu}_0^{\mathrm{gd}} + e_{\mathrm{cmp}}^{\mathrm{gd}}(m,n). \tag{169}$$

Let $n_{\mathrm{eff}} := m \wedge n$ and fix $C_{\mathrm{stat}} > 1$. Choose the shell constants $C_{\mathrm{shell}} \geq 1$ and $c_{\mathrm{shell}} > 0$ from the concentration bounds. Then choose $c_{\mathrm{stat},0} > 0$ sufficiently small, depending only on the fixed population and concentration constants and on the enlargement factors in equation 119, and fix

$$0 < c_{\mathrm{stat}} \leq c_{\mathrm{stat},0}. \tag{170}$$

Define $\overline{M}_{\mathrm{stat}}(m,n) := c_{\mathrm{stat}} n_{\mathrm{eff}} / \log(C_{\mathrm{stat}} mn)$. Whenever $\overline{M}_{\mathrm{stat}} \geq 2\mu_*$, let J_{stat} be the largest integer satisfying $2R_{J_{\mathrm{stat}}} \leq \overline{M}_{\mathrm{stat}}$, $R_j := 2^j \mu_*$, and set $M_{\mathrm{stat}} := 2R_{J_{\mathrm{stat}}}$. Then

$$\frac{1}{2}\overline{M}_{\mathrm{stat}} < M_{\mathrm{stat}} \leq \overline{M}_{\mathrm{stat}}. \tag{171}$$

For $0 \leq j \leq J_{\mathrm{stat}}$ define

$$p_j(m,n) := C_{\mathrm{shell}}(1+R_j)mn \exp\left\{-c_{\mathrm{shell}} \frac{n_{\mathrm{eff}}}{R_j}\right\}, \tag{172}$$

and

$$p_{\mathrm{train}}^{\mathrm{gd}}(m,n) := p_{\mathrm{ent}}(m,n) + \sum_{j=0}^{J_{\mathrm{stat}}} p_j(m,n). \tag{173}$$

For the fixed sufficiently small choice of c_{stat}, there exist fixed constants $C_{\mathrm{tr}}, c_{\mathrm{tr}}, c_{\mathrm{prob}}, q_{\mathrm{prob}} > 0$ and deterministic thresholds $m_{\mathrm{prob}}, n_{\mathrm{prob}}$ such that

$$p_{\mathrm{train}}^{\mathrm{gd}}(m,n) \leq C_{\mathrm{tr}} e^{-c_{\mathrm{tr}} m (\varepsilon_{\mathrm{ent}}^{\mathrm{gd}})^2} + C_{\mathrm{tr}}(mn)^{-c_{\mathrm{prob}}} \tag{174}$$

whenever

$$m \geq m_{\mathrm{prob}}, \qquad n \geq n_{\mathrm{prob}}, \qquad \frac{m \wedge n}{\log(C_{\mathrm{stat}} mn)} \geq q_{\mathrm{prob}}.$$

Proposition F.3 (Simple sufficient empirical-GD pretraining regime)**.** *Fix the population initialization and the constants above. There exist deterministic constants $m_0, n_0, q_0 > 0$ such that, whenever*

$$m \geq m_0, \qquad n \geq n_0, \qquad A_{m,n} := \frac{m \wedge n}{\log(C_{\mathrm{stat}} mn)} \geq q_0, \tag{175}$$

one has

$$p_{\mathrm{ent}}(m,n) \in (0,1), \qquad e_{\mathrm{cmp}}^{\mathrm{gd}}(m,n) \leq \frac{d_{\mathrm{ent}}^{\mathrm{gd}}}{2}, \qquad r_-^{\mathrm{gd}} > \mu_*,$$

and $M_{\mathrm{stat}} > 2r_+^{\mathrm{gd}}$, $p_{\mathrm{train}}^{\mathrm{gd}}(m,n) \leq \frac{1}{2}$. Moreover, for every certified shell $0 \leq j \leq J_{\mathrm{stat}}$, $p_j(m,n) \in (0,1)$, $n \geq C_{\mathrm{res}} R_j L_{R_j}(p_j/2)$, and

$$C_{\mathrm{res}} \left\{ \sqrt{\frac{R_j L_{R_j}(p_j/2)}{m}} + \frac{R_j L_{R_j}(p_j/2)}{m} + \frac{R_j}{n} \right\} \leq \min\{c_{\mathrm{dir}}, C_{\mathrm{sp},0}\}.$$

Proof of Proposition F.3. All population-GD entry quantities are fixed once the initialization and the learning-rate ceiling γ_0 are fixed. Because $\varepsilon_{\mathrm{ent}}^{\mathrm{gd}} > 0$ is fixed, the choice $p_{\mathrm{ent}} = C_{\mathrm{cmp}} e^{-c_{\mathrm{cmp}} m (\varepsilon_{\mathrm{ent}}^{\mathrm{gd}})^2}$ gives

$$\frac{1}{m} \log \left(\frac{C(1+R_{\mathrm{cmp}})}{p_{\mathrm{ent}}} \right) \le c_{\mathrm{cmp}} (\varepsilon_{\mathrm{ent}}^{\mathrm{gd}})^2 + \frac{C}{m}.$$

Choose $c_{\mathrm{cmp}} > 0$ sufficiently small and then choose m_0, n_0 sufficiently large that $p_{\mathrm{ent}} \in (0,1)$, $e_{\mathrm{cmp}}^{\mathrm{gd}} \le \frac{d_{\mathrm{ent}}^{\mathrm{gd}}}{2}$. In particular $r_-^{\mathrm{gd}} > \mu_*$.

For the shell events,

$$L_{R_j}(p_j/2) = \log \left(\frac{2C_{\mathrm{res}}(1+R_j)mn}{p_j} \right) \le C + c_{\mathrm{shell}} \frac{n_{\mathrm{eff}}}{R_j}.$$

Hence $R_j L_{R_j}(p_j/2) \le C R_j + c_{\mathrm{shell}} n_{\mathrm{eff}}$. Choose c_{shell} sufficiently small, then $c_{\mathrm{stat}} \le c_{\mathrm{stat},0}$ sufficiently small, and finally increase the fixed thresholds m_0, n_0, q_0. Uniformly over all certified shells, this gives $n \ge C_{\mathrm{res}} R_j L_{R_j}(p_j/2)$ and

$$C_{\mathrm{res}} \left\{ \sqrt{\frac{R_j L_{R_j}(p_j/2)}{m}} + \frac{R_j L_{R_j}(p_j/2)}{m} + \frac{R_j}{n} \right\} \le \min\{c_{\mathrm{dir}}, C_{\mathrm{sp},0}\}$$

after increasing the deterministic sample thresholds. The uniform upper entry radius $\overline{\mu}_0^{\mathrm{gd}}$ is fixed. Increasing q_0 if necessary makes

$$\overline{M}_{\mathrm{stat}} \ge 4 \left(\overline{\mu}_0^{\mathrm{gd}} + \frac{d_{\mathrm{ent}}^{\mathrm{gd}}}{2} \right).$$

Since $M_{\mathrm{stat}} > \overline{M}_{\mathrm{stat}}/2$ and $r_+^{\mathrm{gd}} \le \overline{\mu}_0^{\mathrm{gd}} + d_{\mathrm{ent}}^{\mathrm{gd}}/2$, this implies $M_{\mathrm{stat}} > 2 r_+^{\mathrm{gd}}$. Finally, the shell failure sum satisfies

$$\sum_{j=0}^{J_{\mathrm{stat}}} p_j \le Cmn(1+M_{\mathrm{stat}}) \log \left(\frac{2M_{\mathrm{stat}}}{\mu_*} \right) \exp \left\{ -c \frac{n_{\mathrm{eff}}}{M_{\mathrm{stat}}} \right\} \le C(mn)^{-c_{\mathrm{prob}}}$$

for sufficiently small c_{stat}. Increasing m_0, n_0, q_0 once more gives $p_{\mathrm{train}}^{\mathrm{gd}} \le 1/2$ and $p_j \in (0,1)$ for every certified shell. This proves all claims. □

F.5 Deterministic training-time window

Let $c_-^{\mathrm{gd}} := 2c_\mu$, and let $c_+^{\mathrm{gd}} > 0$ be the fixed upper fourth-power growth constant in Proposition F.2. For the actual GD iterates, retain the integer entry index $s_0 = s_0(\gamma)$. Introduce the post-capture training clock $T_s := T_0 + \gamma(s - s_0), \quad s \ge s_0$, so that $T_{s_0} = T_0$. Define $\mathbb{T}_\gamma := \{T_0 + \gamma j : j = 0, 1, 2, \ldots\}$, and write

$$\widehat{\theta}^\gamma(T_s) := \widehat{\theta}_s^\gamma, \qquad \widehat{\mu}^\gamma(T_s) := \widehat{\mu}_s^\gamma. \tag{176}$$

Define the deterministic radial envelopes directly on the post-capture clock:

$$M_-^{\mathrm{gd}}(T) := \left\{ (r_-^{\mathrm{gd}})^4 + c_-^{\mathrm{gd}}(T - T_0) \right\}^{1/4}, \tag{177}$$

$$M_+^{\mathrm{gd}}(T) := \left\{ (r_+^{\mathrm{gd}})^4 + c_+^{\mathrm{gd}}(T - T_0) \right\}^{1/4}. \tag{178}$$

Thus

$$M_\pm^{\mathrm{gd}}(T_s) = \left\{ (r_\pm^{\mathrm{gd}})^4 + c_\pm^{\mathrm{gd}} \gamma (s - s_0) \right\}^{1/4}.$$

Set

$$S_{\mathrm{stat}}(m,n,\gamma) := s_0 + \left\lfloor \frac{M_{\mathrm{stat}}^4 - (r_+^{\mathrm{gd}})^4}{c_+^{\mathrm{gd}} \gamma} \right\rfloor, \tag{179}$$

and define the corresponding deterministic statistical time

$$T_{\mathrm{stat}}^{\mathrm{gd}}(m,n,\gamma) := T_0 + \gamma \left\{ S_{\mathrm{stat}}(m,n,\gamma) - s_0 \right\}. \tag{180}$$

F.6 Completion of the proof

Proof of Theorem A.1. Increase A_0 if necessary so that $A_{m,n} \geq A_0$ implies $m \geq m_0$, $n \geq n_0$, and $A_{m,n} \geq q_0$. Hence Proposition F.3 applies.

Let $\mathcal{E}_{\mathrm{ent}}$ denote the compact approximation event from Proposition E.8 on K_0. On $\mathcal{E}_{\mathrm{ent}}$, the compact error is below $\varepsilon_{\mathrm{ent}}^{\mathrm{gd}}$, so Lemma F.1 gives

$$\widehat{\theta}_{s_0}^{\gamma} \in \mathrm{int}(\mathcal{G}), \qquad r_{-}^{\mathrm{gd}} \leq \widehat{\mu}_{s_0}^{\gamma} \leq r_{+}^{\mathrm{gd}}.$$

Since $r_{-}^{\mathrm{gd}} > \mu_*$ and $M_{\mathrm{stat}} > 2r_{+}^{\mathrm{gd}}$, there is a unique $j_{\mathrm{ent}} \in \{0, \ldots, J_{\mathrm{stat}}\}$ such that

$$R_{j_{\mathrm{ent}}} \leq \widehat{\mu}_{s_0}^{\gamma} < 2R_{j_{\mathrm{ent}}},$$

hence $\widehat{\theta}_{s_0}^{\gamma} \in \mathcal{S}_{R_{j_{\mathrm{ent}}}}$.

For every deterministic shell $0 \leq j \leq J_{\mathrm{stat}}$, let $\mathcal{E}_j^{\mathrm{dir}}$ and $\mathcal{E}_j^{\mathrm{sp}}$ denote the events from Theorem E.16 and Proposition E.17, respectively, each with failure probability $p_j/2$. Proposition F.3 verifies the required deterministic smallness conditions. Define

$$\mathcal{E}_{\mathrm{shell}} := \bigcap_{j=0}^{J_{\mathrm{stat}}} \left(\mathcal{E}_j^{\mathrm{dir}} \cap \mathcal{E}_j^{\mathrm{sp}}\right), \qquad \mathcal{E}_{\mathrm{train}} := \mathcal{E}_{\mathrm{ent}} \cap \mathcal{E}_{\mathrm{shell}}.$$

Then the union bound gives

$$\mathbb{P}(\mathcal{E}_{\mathrm{train}}) \geq 1 - p_{\mathrm{train}}^{\mathrm{gd}}(m,n) \geq 1 - Ce^{-cm} - C(mn)^{-c}.$$

On $\mathcal{E}_{\mathrm{train}}$, all certified shells are available before the data-dependent entry index j_{ent} is identified. Proposition F.2 therefore applies at $R_{j_{\mathrm{ent}}}$. Its handoff property places the landing iterate in $\mathcal{S}_{R_{j_{\mathrm{ent}}+1}}$ whenever $j_{\mathrm{ent}} < J_{\mathrm{stat}}$. Iterating this argument propagates the empirical path through every subsequent certified shell while preserving $\mathcal{G}$, up to the terminal radius $M_{\mathrm{stat}} = 2R_{J_{\mathrm{stat}}}$.

Iterating the one-shell fourth-power bounds from the entry radii gives, at every certified grid time T, $M_{-}^{\mathrm{gd}}(T) \leq \widehat{\mu}^{\gamma}(T) \leq M_{+}^{\mathrm{gd}}(T)$. By the definition of $T_{\mathrm{stat}}^{\mathrm{gd}}$, this holds simultaneously for every $T \in [T_0, T_{\mathrm{stat}}^{\mathrm{gd}}] \cap \mathbb{T}_{\gamma}$, proving equation 19.

It remains to identify the late statistical-scale window. Evaluating the lower envelope at the deterministic endpoint $T_{\mathrm{stat}}^{\mathrm{gd}}$ and using $\lfloor x \rfloor \geq x - 1$ gives

$$\{M_{-}^{\mathrm{gd}}(T_{\mathrm{stat}}^{\mathrm{gd}})\}^4 \geq (r_{-}^{\mathrm{gd}})^4 + \frac{c_{-}^{\mathrm{gd}}}{c_{+}^{\mathrm{gd}}}\{M_{\mathrm{stat}}^4 - (r_{+}^{\mathrm{gd}})^4\} - c_{-}^{\mathrm{gd}}\gamma.$$

The entry radii and γ are bounded independently of (m,n), so after increasing the fixed threshold A_0, the right-hand side is at least $\frac{c_{-}^{\mathrm{gd}}}{2c_{+}^{\mathrm{gd}}} M_{\mathrm{stat}}^4$. Thus $M_{-}^{\mathrm{gd}}(T_{\mathrm{stat}}^{\mathrm{gd}}) \geq c_{\mathrm{term}} M_{\mathrm{stat}}$ with $c_{\mathrm{term}} := (c_{-}^{\mathrm{gd}}/(2c_{+}^{\mathrm{gd}}))^{1/4}$. Choosing the main-text constant $c_{\mathrm{rate}} < c_{\mathrm{term}}$ makes $T_{\mathrm{rate}}^{\mathrm{gd}}$ well defined. For every $T \in \mathcal{T}_{\mathrm{stat}}(m,n,\gamma)$, monotonicity of the envelopes gives $c_{\mathrm{rate}} M_{\mathrm{stat}} \leq \widehat{\mu}^{\gamma}(T) \leq M_{\mathrm{stat}}$. Since $M_{\mathrm{stat}} \asymp A_{m,n}$, this is equation 20; Lemma C.13 gives the accompanying $C/A_{m,n}$ population prediction bound. Finally, the fourth-power envelopes imply $T_{\mathrm{rate}}^{\mathrm{gd}} - T_0 \asymp T_{\mathrm{stat}}^{\mathrm{gd}} - T_0 \asymp A_{m,n}^4$. □

G Fresh-Context Transfer from Frozen Checkpoints

The empirical-GD theorem above controls every iterate in the complete deterministic post-capture training-time window. We keep the post-capture time T_0 exactly as it was fixed in Section F; in particular, none of the quantities s_0, K_0, $r_{\pm}^{\mathrm{gd}}$, the compact-entry constants, or the shellwise pretraining thresholds are redefined in this section.

We formulate fresh transfer for an arbitrary checkpoint selected from a later deterministic subwindow. The checkpoint is chosen using pretraining information only and is frozen before the fresh threshold, the fresh-context size, the confidence level, or the fresh observations are revealed.

Fix

$$\eta \in (0, \Delta). \tag{181}$$

Consider the safe one-head family

$$\alpha^{\mathrm{fr}} \in [\rho + \eta, \ell], \qquad \ell := 1 - \rho. \tag{182}$$

Conditional on α^{fr}, draw an independent clean fresh context

$$X_i^{\mathrm{fr}} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,1), \qquad Y_i^{\mathrm{fr}} := \mathbf{1}\{X_i^{\mathrm{fr}} \le \alpha^{\mathrm{fr}}\}, \qquad i = 1, \dots, N. \tag{183}$$

For any frozen $\theta = (\lambda, \mu)$, define

$$F_N^{\mathrm{fr}}(\theta) := \frac{\sum_{i=1}^N X_i^{\mathrm{fr}} e^{\lambda Y_i^{\mathrm{fr}} + \mu X_i^{\mathrm{fr}}}}{\sum_{i=1}^N e^{\lambda Y_i^{\mathrm{fr}} + \mu X_i^{\mathrm{fr}}}}. \tag{184}$$

G.1 Fixed-parameter fresh-context bounds

The fresh-context proof uses two complementary arguments.

Below the self-normalized resolution scale, the fixed positive margin $\alpha^{\mathrm{fr}} - \rho \ge \eta$ implies that a positive observation sufficiently close to the threshold has a fixed normalized score advantage over every negative observation. Order statistics and deterministic softmax geometry then give an all-N bound with no effective-sample-size assumption.

For sufficiently large contexts, the normalized weighted empirical moments resolve the frozen concentration scale. A self-normalized concentration inequality then gives the sharper approximation

$$F_N^{\mathrm{fr}}(\theta) = F(\theta; \alpha^{\mathrm{fr}}) + O_{\mathbb{P}}\left(\sqrt{\frac{1}{\mu N}} + \frac{1}{N}\right).$$

The second argument removes the artificial growth of the deterministic $\log N/\mu$ softmax approximation when N is very large.

G.1.1 Deterministic softmax geometry

Lemma G.1 (Softmax approximation to a maximum)**.** *Let $z_1, \dots, z_M \in [0,1]$, $M \ge 1$, and define*

$$\overline{z}_\mu := \frac{\sum_{i=1}^M z_i e^{\mu z_i}}{\sum_{i=1}^M e^{\mu z_i}}.$$

Then

$$0 \le \max_{1 \le i \le M} z_i - \overline{z}_\mu \le \frac{\log M}{\mu}. \tag{185}$$

Proof. Let $w_i := e^{\mu z_i} / \sum_{j=1}^M e^{\mu z_j}$. The Shannon entropy $H(w) := -\sum_i w_i \log w_i$ satisfies $0 \le H(w) \le \log M$, while

$$H(w) = \log\left(\sum_{j=1}^M e^{\mu z_j}\right) - \mu \sum_{i=1}^M w_i z_i.$$

Since $\log\left(\sum_{j=1}^M e^{\mu z_j}\right) \ge \mu \max_i z_i$, the upper bound follows. The lower bound holds because $\overline{z}_\mu$ is a convex combination of the z_i. □

Lemma G.2 (One-head class-margin bound)**.** *Fix a clean labeled context $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with at least one positive-label observation, and define $M_1 := \max_{y_i = 1} x_i$. For $\theta = (\lambda, \mu)$ with $\mu > 0$, write $k = \lambda/\mu$. If the negative-label class is nonempty, define $M_0 := \max_{y_i = 0} x_i$, $\Gamma_+ := k + M_1 - M_0$. If the negative-label class is empty, set $\Gamma_+ := +\infty$. Then, for every threshold α consistent with the clean labels,*

$$|F_N(\theta; \mathcal{D}) - \alpha| \le (\alpha - M_1) + \frac{\log(N+1)}{\mu} + \min\left\{1, N e^{-\mu \Gamma_+}\right\}. \tag{186}$$

Proof. Let W_1 and W_0 denote the total unnormalized attention weights in the positive and negative classes, respectively. Within the positive class, the common factor e^{λ} cancels after normalization. Hence Lemma G.1 gives

$$0 \le M_1 - \frac{\sum_{y_i=1} x_i e^{\mu x_i}}{\sum_{y_i=1} e^{\mu x_i}} \le \frac{\log N}{\mu}.$$

The positive-class softmax average therefore lies in $\left[M_1 - \frac{\log N}{\mu},\, M_1\right] \subseteq (-\infty, \alpha]$.

If the negative class is nonempty, $W_1 \ge e^{\lambda+\mu M_1},\ W_0 \le N e^{\mu M_0}$, so $\frac{W_0}{W_1} \le N e^{-\mu\Gamma_+}$. Thus the total negative-class softmax mass is at most $\min\left\{1, N e^{-\mu\Gamma_+}\right\}$. Since every feature value lies in $[0,1]$, adding the negative class can move the output by at most this mass. Combining this with $\alpha - M_1$ and the within-positive-class softmax error proves equation 186. If the negative class is empty, the last term vanishes and the same argument applies. □

Lemma G.3 (Absorption of the wrong-class term)**.** *For every fixed $c > 0$, there exist constants $C_c, \mu_c > 0$ such that, for every $N \ge 1$ and every $\mu \ge \mu_c$,*

$$\min\left\{1, N e^{-c\mu}\right\} \le C_c \frac{1 + \log(N+1)}{\mu}. \tag{187}$$

Proof. If $\log(N+1) \ge \frac{c\mu}{2}$, then $\frac{1+\log(N+1)}{\mu} \ge \frac{c}{2}$, so the right-hand side dominates 1 after choosing $C_c \ge 2/c$. If instead $\log(N+1) < \frac{c\mu}{2}$, then $N e^{-c\mu} \le e^{-c\mu/2} \le \frac{C_c}{\mu}$ after increasing μ_c and C_c. This proves the claim. □

Proposition G.4 (Order statistic one-head bound)**.** *Fix $\eta \in (0, \Delta)$. There exist constants $C_\eta, \mu_\eta > 0$, depending only on the fixed structural constants and η, such that the following holds. For every*

$$\theta = (\lambda, \mu) \in \mathcal{G}, \qquad \mu \ge \mu_\eta, \qquad \alpha \in [\rho + \eta, \ell],$$

every integer $N \ge 1$, and every $u \ge 1$,

$$\mathbb{P}_{\mathrm{fr}}\left(\left|F_N^{\mathrm{fr}}(\theta) - \alpha\right| > C_\eta \left\{\frac{u}{N} + \frac{1 + \log(N+1)}{\mu}\right\}\right) \le e^{-u}. \tag{188}$$

The constants are uniform over the displayed parameter and threshold ranges, and no lower-bound condition on N is imposed.

Proof of Proposition G.4. Fix $\alpha \in [\rho + \eta, \ell]$. Let $\mu_{\eta/2}^{\mathrm{wc}}$ denote the threshold μ_c from Lemma G.3 with $c = \eta/2$. Increase μ_η once and for all so that

$$\mu \ge \mu_\eta, \qquad \theta \in \mathcal{G} \quad \Longrightarrow \quad k \ge \ell, \tag{189}$$

and

$$\mu_\eta \ge \mu_{\eta/2}^{\mathrm{wc}}. \tag{190}$$

For example, it is enough to impose $\mu_\eta \ge \max\left\{\mu_*, e^{-c_0}, \mu_{\eta/2}^{\mathrm{wc}}\right\}$, because $k \ge b_{c_0}(\mu) = \ell + \frac{\log\mu + c_0}{\mu}$.

First suppose

$$\frac{u}{N} \le \frac{\eta}{2}. \tag{191}$$

Define the threshold-neighborhood event

$$\mathcal{E}_u := \left\{\exists i \le N : X_i^{\mathrm{fr}} \in \left(\alpha - \frac{u}{N}, \alpha\right]\right\}. \tag{192}$$

The interval in equation 192 has length u/N and is contained in $(0, \alpha]$ under equation 191. Therefore

$$\mathbb{P}_{\mathrm{fr}}(\mathcal{E}_u^c) = \left(1 - \frac{u}{N}\right)^N \le e^{-u}. \tag{193}$$

On $\mathcal{E}_u$, the positive-label class is nonempty. Only now define $M_1 := \max\{X_i^{\mathrm{fr}} : Y_i^{\mathrm{fr}} = 1\}$. Then

$$0 \le \alpha - M_1 \le \frac{u}{N}. \tag{194}$$

If the negative class is nonempty, let $M_0 := \max\{X_i^{\mathrm{fr}} : Y_i^{\mathrm{fr}} = 0\}$. Since $M_0 \le 1$, equation 189 and equation 194 imply

$$\begin{aligned}\Gamma_+ &= k + M_1 - M_0 \\ &\ge \ell + \alpha - \frac{u}{N} - 1 \\ &= \alpha - \rho - \frac{u}{N} \\ &\ge \frac{\eta}{2}.\end{aligned} \tag{195}$$

If the negative class is empty, set $\Gamma_+ = +\infty$ as in Lemma G.2.

Applying Lemmas G.2 and G.3 with $c = \eta/2$ gives, on $\mathcal{E}_u$,

$$\left|F_N^{\mathrm{fr}}(\theta) - \alpha\right| \le C_\eta \left\{ \frac{u}{N} + \frac{1 + \log(N+1)}{\mu} \right\}.$$

Together with equation 193, this proves the desired probability bound in the regime equation 191.

It remains to consider $u/N > \eta/2$. Since $F_N^{\mathrm{fr}}(\theta), \alpha \in (0,1)$, one always has $\left|F_N^{\mathrm{fr}}(\theta) - \alpha\right| \le 1$. After taking $C_\eta \ge 2/\eta$, $C_\eta \frac{u}{N} > 1$, so the bad event in equation 188 is empty. This completes the proof for every $N \ge 1$. □

Proposition G.5 (Large-context self-normalized one-head bound)**.** *There exist constants $C_{\mathrm{fr}}, \mu_{\mathrm{fr}} > 0$, depending only on the fixed structural constants, such that the following holds. For every*

$$\theta = (\lambda, \mu) \in \mathcal{G}, \qquad \mu \ge \mu_{\mathrm{fr}}, \qquad \alpha \in [\rho, \ell],$$

every $u \ge 1$, and every

$$N \ge C_{\mathrm{fr}} \mu u, \tag{196}$$

one has

$$\mathbb{P}_{\mathrm{fr}}\left(\left|F_N^{\mathrm{fr}}(\theta) - \alpha\right| > C_{\mathrm{fr}} \left\{ \frac{1}{\mu} + \sqrt{\frac{u}{\mu N}} + \frac{u}{N} \right\} \right) \le 4e^{-u}. \tag{197}$$

In particular, under equation 196,

$$\left|F_N^{\mathrm{fr}}(\theta) - \alpha\right| \le C_{\mathrm{fr}} \left\{ \frac{1}{\mu} + \frac{u}{N} \right\} \tag{198}$$

on the same event after increasing C_{fr}. Unlike Proposition G.4, this proposition is valid on the full closed threshold interval $[\rho, \ell]$, but it requires the fresh context to resolve the scale μ.

Proof of Proposition G.5. Fix $\theta = (\lambda, \mu) \in \mathcal{G},\ \alpha \in [\rho, \ell]$. Let

$$W_{\theta,\alpha}(X, Y) := \frac{e^{\lambda Y + \mu X}}{\mathbb{E}[e^{\lambda Y + \mu X} \mid \alpha]},$$

and define $\phi_F := W_{\theta,\alpha}\big(X - F(\theta;\alpha)\big)$. Then $P_\alpha W_{\theta,\alpha} = 1,\ P_\alpha \phi_F = 0$, and the fresh-context error has the exact self-normalized representation

$$F_N^{\mathrm{fr}}(\theta) - F(\theta;\alpha) = \frac{P_N \phi_F}{P_N W_{\theta,\alpha}}. \tag{199}$$

By Lemma E.9, uniformly over $\theta \in \mathcal{G}$ and $\alpha \in [\rho, \ell]$, $\|W_{\theta,\alpha}\|_\infty \le C\mu$, $P_\alpha\big(W_{\theta,\alpha} - 1\big)^2 \le C\mu$, $\|\phi_F\|_\infty \le C,\ P_\alpha \phi_F^2 \le \frac{C}{\mu}$. Thus, in Lemma E.5, we may take $B \le C\mu$, $\sigma_A^2 \le C\mu$, $b \le C$, and $s_2 \le C/\mu$. After increasing C_{fr}, the condition $N \ge C_{\mathrm{fr}} \mu u$ implies

$$1 \le u \le c \frac{N}{B + \sigma_A^2 + 1},$$

so Lemma E.5 applies to equation 199. Hence, with probability at least $1 - 4e^{-u}$,

$$\left|F_N^{\mathrm{fr}}(\theta) - F(\theta;\alpha)\right| \le C_{\mathrm{fr}} \left\{ \sqrt{\frac{u}{\mu N}} + \frac{u}{N} \right\}. \tag{200}$$

The population prediction bound in $\mathcal{G}$ gives $|F(\theta;\alpha) - \alpha| \le \frac{C_{\mathrm{bias}}}{\mu}$. The triangle inequality proves equation 197.

Finally, $N \ge C_{\mathrm{fr}}\mu u$ implies $\sqrt{\frac{u}{\mu N}} \le \frac{C}{\mu}$, which gives equation 198 after enlarging C_{fr}. □

Define $L_\mu(u) := 1 + \log(1 + C_{\mathrm{fr}}\mu u)$ and $\Phi(\mu, u) := L_\mu(u)/\mu$. For every fixed $u \ge 1$, the map $\mu \mapsto \Phi(\mu, u)$ is strictly decreasing on $(0, \infty)$.

G.1.2 Proof of the main-text fixed-estimator proposition

Proof of Proposition 4.1. Fix $\alpha \in [\rho + \eta, \ell],\ u \ge 1$. There are two cases.

If $N \ge C_{\mathrm{fr}}\mu u$, Proposition G.5 gives, with failure probability at most $4e^{-u}$,

$$\left|F_N^{\mathrm{fr}}(\theta) - \alpha\right| \le C \left\{ \frac{1}{\mu} + \frac{u}{N} \right\} \le C_\eta \left\{ \frac{L_\mu(u)}{\mu} + \frac{u}{N} \right\}.$$

If instead $N < C_{\mathrm{fr}}\mu u$, Proposition G.4 gives, with failure probability at most e^{-u},

$$\left|F_N^{\mathrm{fr}}(\theta) - \alpha\right| \le C_\eta \left\{ \frac{u}{N} + \frac{1 + \log(N+1)}{\mu} \right\}.$$

Since $N + 1 \le 1 + C_{\mathrm{fr}}\mu u$, after a harmless enlargement of the constant, $1 + \log(N+1) \le L_\mu(u)$. Thus the bound in equation 8 holds in both regimes. In either case the failure probability is at most $4e^{-u}$. □

Lemma G.6 (Freeze-then-transfer conditioning)**.** *Let $\mathcal{F}_{\mathrm{pre}}$ be the pretraining sigma-field and let $\widehat{\theta}$ be $\mathcal{F}_{\mathrm{pre}}$-measurable. Suppose that on a pretraining event $\mathcal{E}_{\mathrm{train}}$ one has $\widehat{\theta} \in \Theta_{\mathrm{good}}$, and suppose that for every deterministic $\theta \in \Theta_{\mathrm{good}}$ a fresh-context event $\mathcal{A}(\theta)$ satisfies $\mathbb{P}_{\mathrm{fr}}(\mathcal{A}(\theta)) \ge 1 - \delta$, with constants uniform over Θ_{good}. Then, on $\mathcal{E}_{\mathrm{train}}$, $\mathbb{P}_{\mathrm{fr}}(\mathcal{A}(\widehat{\theta}) \mid \mathcal{F}_{\mathrm{pre}}) \ge 1 - \delta$.*

Proof. Condition on $\mathcal{F}_{\mathrm{pre}}$. The value of $\widehat{\theta}$ is then fixed and belongs to Θ_{good} on $\mathcal{E}_{\mathrm{train}}$, so the deterministic fixed-parameter bound applies directly. □

Lemma G.6 is the only probabilistic step needed to pass from a fixed-parameter fresh-context theorem to a deterministic or pretraining-measurable checkpoint. This makes the quantifier order explicit and avoids claiming a single fresh-context event that is simultaneous over all checkpoints.

G.2 Freeze-then-transfer theorem

Proof of Theorem A.2. Work on the training event $\mathcal{E}_{\mathrm{train}}$ from Theorem A.1. For every $T \in \mathcal{T}_{\mathrm{stat}}(m, n, \gamma)$, that theorem gives $\widehat{\theta}^\gamma(T) \in \mathcal{G}$ and $c_A A_{m,n} \le \widehat{\mu}^\gamma(T) \le C_A A_{m,n}$. Increasing the fixed threshold q_η^{fr} ensures that $c_A A_{m,n}$ satisfies the sufficiently-large-μ condition in Proposition 4.1.

Let $\widehat{T}$ be deterministic or $\mathcal{F}_{\mathrm{pre}}$-measurable with values in $\mathcal{T}_{\mathrm{stat}}$ on $\mathcal{E}_{\mathrm{train}}$, and set $\widehat{\theta} := \widehat{\theta}^\gamma(\widehat{T})$. Conditional on $\mathcal{F}_{\mathrm{pre}}$, this checkpoint is fixed. Lemma G.6 and Proposition 4.1 therefore give, for every fixed $\alpha^{\mathrm{fr}} \in [\rho + \eta, \ell]$, every $N \ge 1$, and every $u \ge 1$, conditional fresh probability at least $1 - 4e^{-u}$ and

$$|F_N^{\mathrm{fr}}(\widehat{\theta}) - \alpha^{\mathrm{fr}}| \le C_\eta \left[\frac{1 + \log(1 + C_{\mathrm{fr}}\widehat{\mu}^\gamma(\widehat{T})u)}{\widehat{\mu}^\gamma(\widehat{T})} + \frac{u}{N} \right].$$

This is equation 21. □

G.3 FULL CLOSED INTERVAL ONCE THE FRESH CONTEXT RESOLVES THE SCALE

Corollary G.7 (Full-boundary transfer along the certified time window)**.** *Under the assumptions of Theorem A.2, fix $T \in \mathcal{T}_{\mathrm{stat}}(m,n,\gamma)$ and $u \geq 1$. If*

$$N \geq C_{\mathrm{fr}} M_{+}^{\mathrm{gd}}(T) u, \tag{201}$$

then, for every fixed $\alpha^{\mathrm{fr}} \in [\rho, \ell]$, with joint probability at least $1 - p_{\mathrm{train}}^{\mathrm{gd}}(m,n) - 4e^{-u}$,

$$\left| F_N^{\mathrm{fr}}\left(\widehat{\theta}^{\gamma}(T)\right) - \alpha^{\mathrm{fr}} \right| \leq C \left[\frac{1}{M_{-}^{\mathrm{gd}}(T)} + \sqrt{\frac{u}{M_{-}^{\mathrm{gd}}(T) N}} + \frac{u}{N} \right]. \tag{202}$$

If additionally T lies in the statistical-scale training-time window and

$$N \geq C_{\mathrm{fr}} C_A A_{m,n} u, \tag{203}$$

then

$$\left| F_N^{\mathrm{fr}}\left(\widehat{\theta}^{\gamma}(T)\right) - \alpha^{\mathrm{fr}} \right| \leq C \left[\frac{1}{A_{m,n}} + \sqrt{\frac{u}{A_{m,n} N}} + \frac{u}{N} \right]. \tag{204}$$

Thus the frozen estimator is consistent at the lower endpoint when the certified scale and the fresh-context size diverge under the stated relation. The remaining obstruction concerns a uniform all-N vanishing theorem for one frozen raw head.

Proof. Fix $T \in \mathcal{T}_{\mathrm{stat}}(m,n,\gamma)$ and work on $\mathcal{E}_{\mathrm{train}}$. Then $\widehat{\theta}^{\gamma}(T) \in \mathcal{G}$ and $M_{-}^{\mathrm{gd}}(T) \leq \widehat{\mu}^{\gamma}(T) \leq M_{+}^{\mathrm{gd}}(T)$. If $N \geq C_{\mathrm{fr}} M_{+}^{\mathrm{gd}}(T) u$, Proposition G.5 applies uniformly for every fixed $\alpha^{\mathrm{fr}} \in [\rho, \ell]$. Replacing $\widehat{\mu}^{\gamma}(T)$ by the lower envelope gives equation 202. On the statistical-scale window, $M_{-}^{\mathrm{gd}}(T) \geq c_A A_{m,n}$ and $M_{+}^{\mathrm{gd}}(T) \leq C_A A_{m,n}$, which gives equation 204 under equation 203. □

Remark G.8 (Heuristic lower-endpoint obstruction for the all-context-length theorem)**.** *Fix a certified time T and write*

$$\widehat{\theta}^{\gamma}(T) = (\widehat{\lambda}^{\gamma}(T), \widehat{\mu}^{\gamma}(T)), \qquad \widehat{k}^{\gamma}(T) := \frac{\widehat{\lambda}^{\gamma}(T)}{\widehat{\mu}^{\gamma}(T)}.$$

Suppose both fresh label classes are present and define

$$M_1 := \max\{X_i^{\mathrm{fr}} : Y_i^{\mathrm{fr}} = 1\}, \qquad M_0 := \max\{X_i^{\mathrm{fr}} : Y_i^{\mathrm{fr}} = 0\},$$

together with $G_L := \alpha^{\mathrm{fr}} - M_1$, $G_R := 1 - M_0$. The normalized positive-minus-negative maximum-score gap is

$$\Delta_N(T) := \widehat{k}^{\gamma}(T) + M_1 - M_0 = (\alpha^{\mathrm{fr}} - \rho) + (\widehat{k}^{\gamma}(T) - \ell) + G_R - G_L. \tag{205}$$

On the safe interior, $\alpha^{\mathrm{fr}} - \rho \geq \eta$ provides a fixed deterministic margin.

At the lower endpoint $\alpha^{\mathrm{fr}} = \rho$, this fixed term vanishes. Inside the logarithmic tube,

$$\widehat{k}^{\gamma}(T) - \ell = O\left(\frac{\log \widehat{\mu}^{\gamma}(T)}{\widehat{\mu}^{\gamma}(T)} \right),$$

whereas fresh endpoint spacings fluctuate on the N^{-1} scale. In the intermediate regime $1 \ll N \ll \widehat{\mu}^{\gamma}(T)/\log \widehat{\mu}^{\gamma}(T)$, the stochastic term $\widehat{\mu}^{\gamma}(T)(G_R - G_L)$ has scale $\widehat{\mu}^{\gamma}(T)/N$ and can dominate the learned logarithmic advantage. This comparison indicates that the negative class can win the maximum-score competition with nonvanishing probability and can therefore place order-one mass on the wrong class and potentially produce order-one prediction error. We use this observation only as a heuristic obstruction here; a formal impossibility statement would require a separate lower-bound proposition.

If instead $\alpha^{\mathrm{fr}} \sim \mathrm{Unif}([\rho, \ell])$ and probability is also taken over the fresh task, then $\mathbb{P}\left(\alpha^{\mathrm{fr}} - \rho \leq \eta\right) = \eta/\Delta$. Thus one may exclude a boundary layer in a joint-over-task theorem, but that boundary-layer probability must explicitly enter the final failure probability or risk. This obstruction also motivates studying reflected multi-expert models in future work. Reflection can repair physical-boundary coverage, but a routed construction must additionally control finite-context fluctuations.

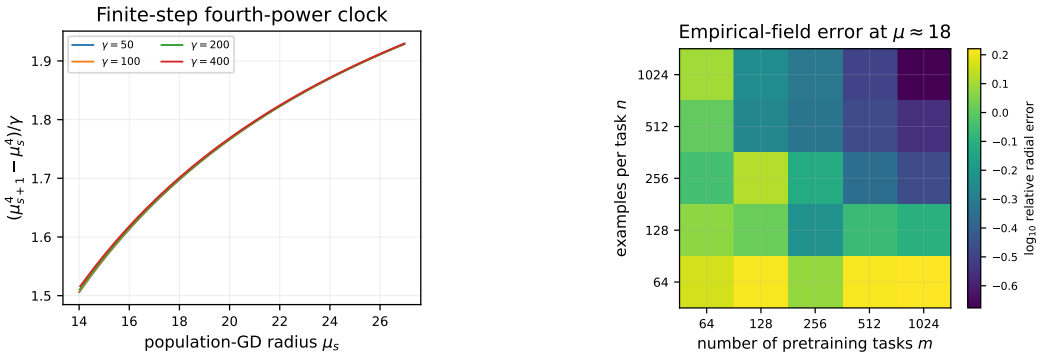

Figure 3: Additional finite-step and finite-pretraining diagnostics. Population GD preserves the normalized fourth-power increment across step sizes (left), while a full (m, n) grid separates task and within-task sampling effects at a common shell (right). The omitted continuation panel is reported in Figure 1.

H Experimental Details and Additional Diagnostics

H.1 Reproducibility Protocol

Population calculations use analytic feature-coordinate integrals and 160-point Gauss-Legendre quadrature over the task threshold. Empirical experiments use exact full-batch gradients on independently sampled pretraining corpora. Every finite-pretraining iteration index is first computed from the deterministic population-GD reference path; it is therefore fixed before the corresponding empirical corpus is sampled. For transfer, one checkpoint is frozen per pretraining corpus and reused for every fresh threshold, context realization, and context size N.

Unless stated otherwise, a "worst-grid pointwise 95th-percentile error" is computed by taking the 95th percentile over independent fresh contexts at each tested threshold and then maximizing over the finite threshold grid. The metric provides pointwise coverage on that grid; simultaneous coverage over the continuum lies outside its definition. Code is available at `https://anonymous.4open.science/r/ICl_Decision_Stump-B5F3`.

H.2 Detailed finite-step and transfer diagnostics

The sufficient-stopping test uses $(128, 128)$, $(256, 256)$, $(512, 512)$, $(128, 512)$, $(512, 128)$, $(256, 512)$, and $(512, 256)$, with 30 independent corpora per configuration. The aggregate radius and prediction-error outcomes are reported in Section 6. A stricter diagnostic based on one concrete finite-radius instantiation of the asymptotic logarithmic tube holds throughout the window in 104 of 210 runs. The theorem leaves its finite-radius threshold and comparison constants existential, so this operational tube serves as a descriptive diagnostic. An auxiliary axis sweep with six corpora per configuration remains reliable under the operational tracking criterion up to the largest tested reference radius, $\mu_{\mathrm{ref}} = 60$, for all seven tested pairs, including values well beyond $A_{m,n}^{\mathrm{proxy}}$.

The separate end-to-end sweep used to generate Figure 1 uses 10, 8, 6, and 4 corpora for $m = n \in \{64, 128, 256, 512\}$, respectively, and 6, 6, 5, and 5 corpora for $(m, n) \in \{(128, 512), (512, 128), (256, 512), (512, 256)\}$. These are the corpus counts for the displayed trajectory and frozen-transfer experiment; the sufficient-stopping diagnostic above uses 210 corpora.

Figure 3 supplements the main continuation plot with the discrete-time clock and a common-shell heatmap comparing task-level and within-task sampling error. The continuation panel in Figure 1 also records behavior beyond the certified proxy radius; failure-threshold estimation lies outside the scope of this sweep.

Figure 4 expands the transfer summary. Thin curves retain corpus-level pretraining variability, while each dotted line marks the population-resolution floor of the same frozen checkpoint. The large-context panel isolates the inverse learned-resolution dependence. At $N = 4096$, the axis swaps

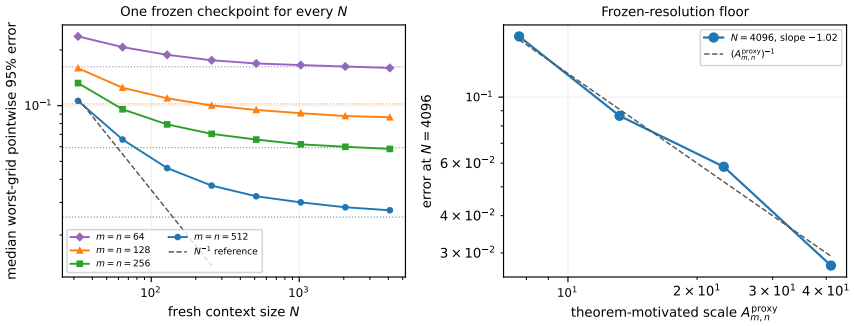

Figure 4: Detailed freeze-then-transfer results. One checkpoint is selected without fresh-task information and reused for every N (left). At large N, the remaining error decreases approximately inversely with the theorem-motivated pretraining scale (right).

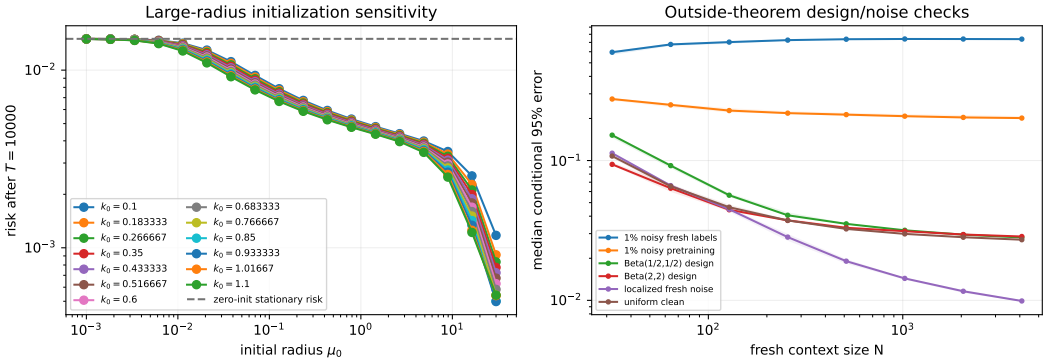

Figure 5: Scope checks. The common-time initialization sweep exhibits the zero-gradient delay at small radius (left). Two nonuniform designs remain close to the uniform baseline, while one-percent global label corruption changes the behavior substantially (right). Noise and design experiments are exploratory deviations from the theorem.

$(128, 512)$ and $(512, 128)$ give errors 0.128 and 0.124, while $(256, 512)$ and $(512, 256)$ give 0.0592 and 0.0586. After fitting an additive floor, the pre-floor excess-error slopes against N range from -0.87 to -0.94 across the four balanced corpus sizes.

The initialization sweep uses 13 evenly spaced ratios $k_0 \in [0.1, 1.1]$ and 19 initial radii: zero and 18 positive values geometrically spaced from 10^{-3} to 30. The design-and-noise scope check uses six independently sampled pretraining corpora per condition, eight threshold values, and 300 fresh contexts per threshold. The two nonuniform designs are $\mathrm{Beta}(2, 2)$ and $\mathrm{Beta}(1/2, 1/2)$ and are used for both pretraining and fresh-context evaluation. Global noise flips each affected label independently with probability 0.01, while threshold-localized fresh noise flips with probability $1/2$ on $|X - \alpha| \le 1/\widehat{\mu}$. The right panel of Figure 5 displays these matched-design and label-corruption conditions. Bands retain variation across the six independently sampled pretraining corpora.

Additional population and initialization checks. The main population experiment follows one complete transient. Figure 6 repeats the post-capture diagnostic across several task interiors and initial states and separately checks numerical quadrature and ODE convergence. The stabilization of $\mu \sup_\alpha |F - \alpha|$ is insensitive to these tested choices, while the numerical discrepancies are negligible at the quadrature order used in the main experiment.

The large-radius assumption has a visible dynamical role. At the origin the attention weights are uniform and symmetry makes the gradient vanish. Figure 7 shows linear vanishing of the gradient near this stationary point and delayed escape from nearby initializations. Within the fixed training budget, performance is highly sensitive to the initial radius; the asymptotic behavior of every nonzero small initialization remains unresolved.

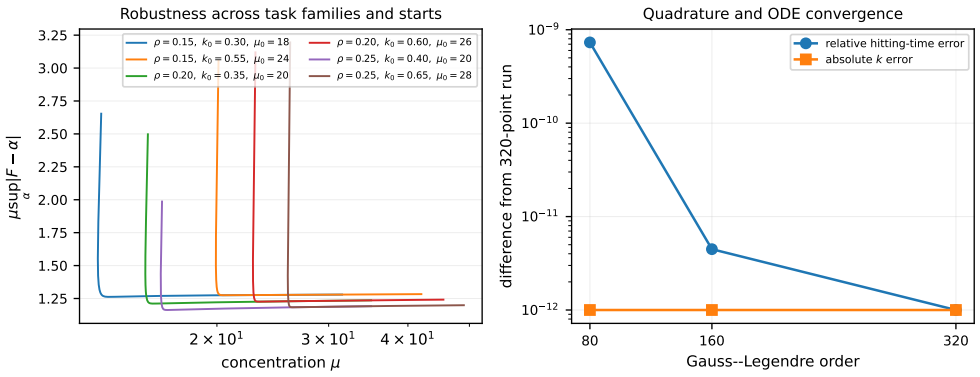

Figure 6: Population robustness and numerical convergence. The scaled prediction error stabilizes across the tested task interiors and initial states (left). Increasing the Gauss–Legendre order confirms convergence of the hitting time and ratio trajectory (right).

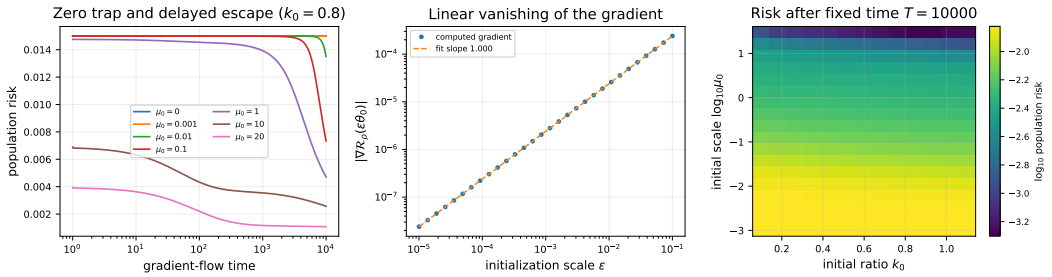

Figure 7: Initialization sensitivity. Exactly zero initialization is stationary and nearby starts escape slowly (left); the gradient norm vanishes linearly with initialization scale (center); and the risk after a common time budget depends strongly on the initial radius (right).

H.3 Exploratory label-noise

To test the robustness of a pretrained model, we use a frozen-checkpoint protocol: the checkpoint is selected before the noisy fresh context is observed and the same parameters are reused for every context size N. This is our primary robustness experiment. As an auxiliary mechanism-level diagnostic, we also evaluate a resolution-matched sequence with $\mu = N/8$ to ask whether corruption localized at the learned resolution preserves the predicted scaling. In this sequence, localized noise flips labels with probability $1/2$ on $|X - \alpha| \le 1/\mu$, the tapered condition uses flip probability $\frac{1}{2}\{1 - (\mu|X - \alpha|)^2\}_+$, and global Massart noise flips labels independently with probability 0.01. This sequence changes the parameters with N and therefore tracks a family of models; the frozen-checkpoint protocol evaluates a single pretrained model.

For the frozen one-head test, we use $\widehat{\mu} = 40$ and $\widehat{k} = 0.86472$ for every context size. Clean and threshold-localized conditions approach checkpoint-dependent population floors, whereas global Massart flips produce order-one error (Figure 8).

Along the resolution-matched sequence, clean, localized, and Tsybakov errors decrease approximately as N^{-1} on the safe interior, while the fixed one-head orientation retains its physical-boundary obstruction. We use this coupled scaling as a mechanism-level diagnostic of resolution-compatible corruption across the model sequence.

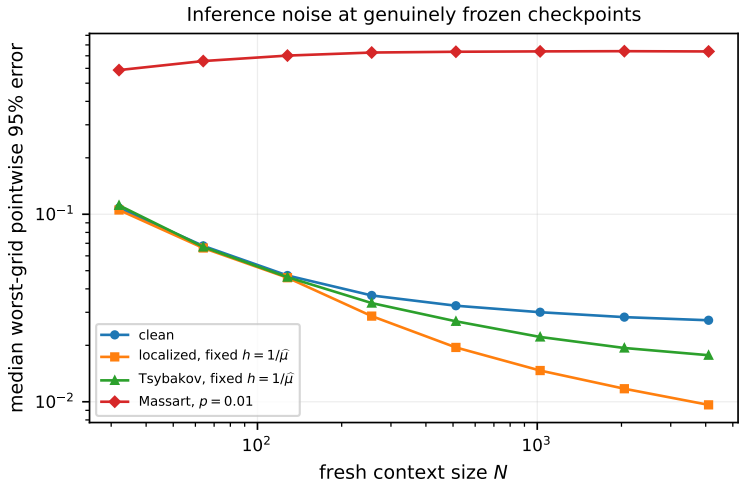

Figure 8: Inference-time noise at a genuinely frozen one-head checkpoint. Clean and threshold-localized conditions improve with fresh-context size and then approach finite-resolution floors. Homogeneous one-percent Massart flips cause persistent order-one error.

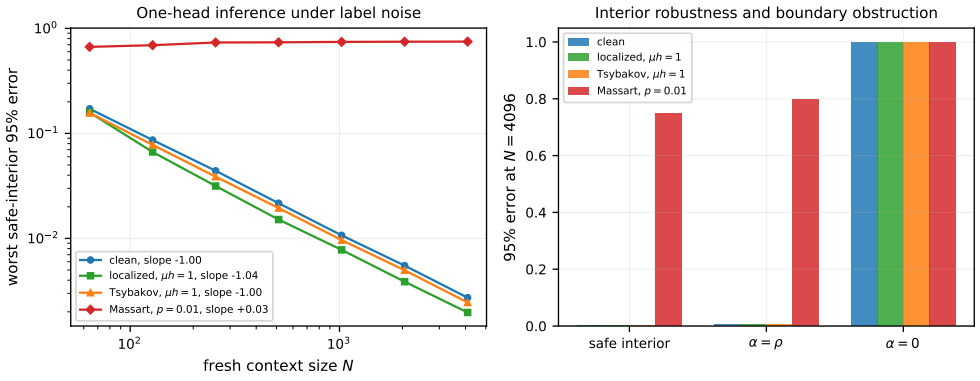

Figure 9: Resolution-matched one-head noise stress test. Localized and Tsybakov corruption preserve the clean interior trend when $\mu = N/8$ (left), while the exact one-head boundary obstruction persists across context sizes and localized corruption (right).