LEARNING HOW TO LOCALIZE CHANGE POINTS: IN-CONTEXT ADAPTATION ACROSS CHANGE MECHANISMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern sequential data often exhibit heterogeneous structural changes, where the appropriate localization strategy may depend on the underlying change mechanism. Change-point localization has traditionally relied on carefully designed statistics tailored to specific mechanisms, meaning that different types of changes typically require different localization procedures. We propose In-Context Change-Point Localization (ICCP), a task-level framework that enables a pretrained Transformer to learn adaptive localization strategies from previously solved change-point tasks. By using solved tasks as demonstrations, ICCP localizes the change point of a new sequence through the prompt alone, without parameter updates or retraining. We provide a theoretical analysis of this in-context adaptation process, establishing finite-sample localization error bounds that characterize the roles of pretraining data, contextual information, and sequence information. Under suitable regularity and scaling conditions, we further prove relative localization consistency. Experiments across diverse change mechanisms demonstrate that ICCP exploits contextual demonstrations and improves with increasing pretraining and contextual information.

1 INTRODUCTION

The reliability of downstream analysis often depends on appropriate data preprocessing. In sequential data, unmodeled structural changes can mix observations from distinct regimes and undermine subsequent analysis. Change-point detection and localization are therefore important preprocessing tools for identifying such changes before downstream modeling and inference. A central challenge in change-point localization is matching the statistical criterion to the underlying change mechanism (Fearnhead, 2006; Fryzlewicz, 2014; Cho & Fryzlewicz, 2015). This task dependence raises a natural question: can a localization procedure learn from previously solved change-point problems how to adapt to a new one, rather than relying on a fixed criterion specified in advance? Deep learning can reduce reliance on hand-crafted statistics by learning change-point localization rules directly from data (Li et al., 2024; Zhou et al., 2024; Li et al., 2026). However, such adaptation is largely encoded in the learned parameters, limiting test-time adaptation across heterogeneous tasks without retraining or fine-tuning.

In-context learning (ICL) offers a natural alternative by allowing a pretrained Transformer to adapt its prediction to the current prompt without parameter updates. Two distinct lines of work motivate our formulation. From the perspective of contextual decision making, Bai et al. (2023b) showed that Transformers can select among prespecified learning procedures in context; we move beyond such algorithm selection and instead learn the change-point localization rule directly from solved tasks. From the perspective of change-point adaptation, Dudley et al. (2026) studied regime changes within a single sequence under a specific change model; we instead consider a family of latent change mechanisms that vary across tasks and use complete solved change-point problems as context for localizing a new task. We call this framework *In-Context Change-Point Localization* (ICCP).

We focus on single-change-point localization and develop a Bayesian decision-theoretic analysis of ICCP, building on Wakayama & Suzuki (2026), to characterize how pretraining and contextual information jointly determine localization risk. Our main contributions are:

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.

1. We introduce *In-Context Change-Point Localization* (ICCP), a new task-level ICL framework that uses previously solved change-point problems to adapt localization to a new sequence without updating model parameters.
2. We provide finite-sample guarantees for ICCP by quantifying the effects of pretraining and contextual information on localization error. We further establish conditions under which the change mechanism can be identified from solved tasks and ICCP achieves relative localization consistency.
3. We evaluate ICCP on synthetic tasks involving different change mechanisms. Our experiments show that solved contextual tasks improve localization over a query-only baseline and examine the effects of pretraining size, context length, and sequence length.

The remainder of the paper is organized as follows. Section 3 presents the ICCP framework, Section 4 develops the theoretical guarantees, and Section 5 reports the simulation results. Section 6 concludes the paper, while additional technical details and proofs are provided in the appendices.

2 Related Work

Change-point analysis. A large statistical literature develops change-point procedures through carefully constructed statistics and segmentation criteria, including CUSUM-based and nonparametric methods, among others (Yu & Chen, 2021; Wang & Feng, 2023; Bai et al., 2023a; Fryzlewicz, 2024; Zhao et al., 2022; Song & Chen, 2022; Chen & Chu, 2023; McGonigle & Cho, 2025). To accommodate increasingly complex data structures, recent work has incorporated learned representations and data-driven detection rules into change-point analysis. This has led to a growing class of learning-based methods, including approaches based on deep neural networks (Gupta et al., 2022; Li et al., 2024; Zhou et al., 2026). By contrast, ICCP uses solved change-point tasks as in-context information to adapt localization to a new task at inference time without parameter updates.

Meta-learning. Meta-learning leverages shared structure across tasks to facilitate learning on new tasks. Finn et al. (2017) propose a model-agnostic meta-learning approach that learns an initialization enabling rapid adaptation to new tasks with a few gradient steps, whereas Conditional Neural Processes learn to predict target outputs for a new function from a small set of observed input–output pairs (Garnelo et al., 2018). Change-point methods have also been incorporated into meta-learning: Harrison et al. (2020) use Bayesian online change-point detection to infer latent task switches in nonstationary environments. In contrast, we treat change-point localization as the meta-learning problem: each context example is a solved change-point task, and these examples are used to localize the change point in a new sequence.

In-context learning. Relevant ICL theory can be broadly viewed from algorithmic and Bayesian perspectives. The former studies how Transformers implement or select learning procedures in context (Garg et al., 2022; Akyürek et al., 2023; Wang et al., 2024; Cheng et al., 2024; Guo et al., 2024; Bai et al., 2023b), while the latter views ICL as conditional prediction under task uncertainty. Panwar et al. (2024) study hierarchical meta-ICL, while Wakayama & Suzuki (2026) develop finite-sample Bayesian theory for latent task mixtures. Unlike their observation-level setting, ICCP uses complete solved change-point tasks as context. Dudley et al. (2026) study in-context detection and adaptation to regime changes within a sequence, with the change point serving as latent or side information for prediction. ICCP instead predicts the change-point location from multiple solved change-point tasks used as task-level demonstrations.

3 In-Context Change-Point Localization Framework

3.1 Task and Prompt Formulation

We first formalize the ICCP prompt underlying the model and risk analysis.

Definition 3.1 (Prompt generation for change-point tasks)**.** Suppose that there are $K < \infty$ latent change-point types, and let p denote the maximum number of context tasks. An ICCP Prompt is generated as follows.

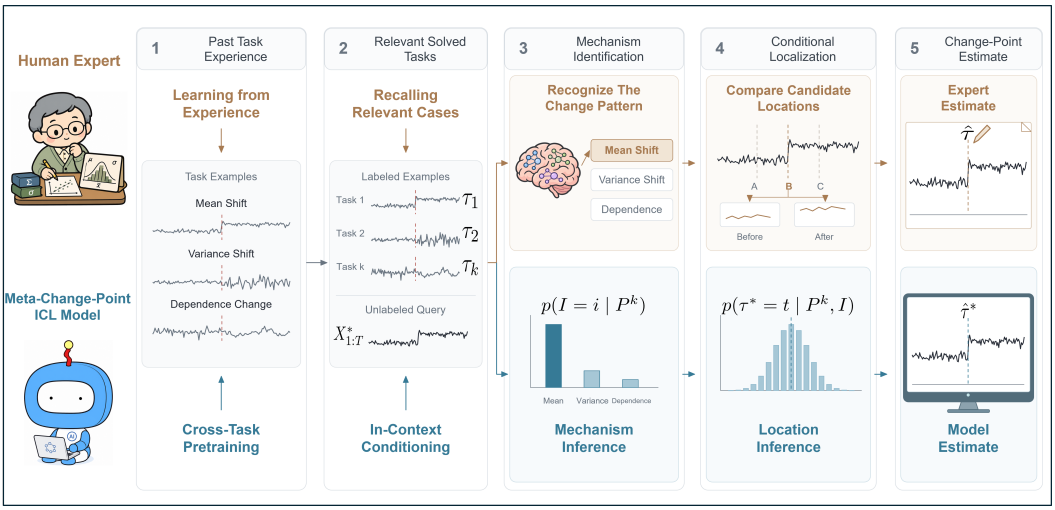


Figure 1: Schematic illustration of the proposed ICCP mechanism.

1. **Sampling the change-point type.** Sample $I \sim \Pi_I = \text{Categorical}(\alpha)$, where $\Pr(I = i) = \alpha_i > 0$ for $i = 1, \ldots, K$ and $\sum_{i=1}^{K} \alpha_i = 1$. The latent variable I specifies the underlying change mechanism.

2. **Specifying the task distribution.** Conditional on $I = i$, let $\mathcal{F}_i$ denote the class of change-point tasks associated with type i, and let $P_{\mathcal{F}_i}$ be a distribution on $\mathcal{F}_i$. Tasks drawn from $P_{\mathcal{F}_i}$ may differ in their data-generating structures, parameters, and change-point locations while sharing the same underlying change mechanism.

3. **Generating context and query tasks.** For $j = 1, \ldots, p+1$, independently sample $f^{(j)} \sim P_{\mathcal{F}_i}$. Each $f^{(j)}$ specifies a task-specific data-generating process with true change-point location $\tau_j \in \{1, \ldots, T-1\}$ and generates an observation sequence $X_{1:T}^{(j)} = (X_1^{(j)}, \ldots, X_T^{(j)})$.

4. **Constructing the prompt.** For $1 \le k \le p$, define $\mathcal{C}_k := \{(X_{1:T}^{(j)}, \tau_j)\}_{j=1}^{k}$ and $\mathcal{P}^k := (\mathcal{C}_k, X_{1:T}^{(k+1)})$. Here, $\mathcal{C}_k$ contains the k solved context tasks, while $X_{1:T}^{(k+1)}$ is the query sequence and its true change-point location τ_{k+1} is hidden from the model.

Compared with observation-level ICL, ICCP operates at the task level, with each context example representing a complete solved change-point problem. Conditional on the shared latent type I, tasks are independently drawn from the same task distribution and are exchangeable across tasks. Within each task, $X_{1:T}^{(j)}$ retains its temporal ordering.

3.2 Transformer Architecture

We consider the following model architecture.

Definition 3.2 (Uniform-attention Transformer for ICCP)**.** For any $k \in \{1, \ldots, p\}$, consider the partial prompt $\mathcal{P}^k$. Let $\theta = (\omega, \vartheta_\phi, \vartheta_\rho)$ denote the model parameters. The normalized change-point predictor is defined as

$$\widetilde{M}_\theta(\mathcal{P}^k) = \rho_{\vartheta_\rho}\left(\frac{1}{k}\sum_{j=1}^{k}\phi_{\vartheta_\phi}(\mathbf{u}_j), \mathbf{c}_{k+1}, k\right), \tag{1}$$

where the components are specified below.

Sequence encoder. Let $E_\omega : \mathcal{X}_T \to \mathcal{Z} \subset \mathbb{R}^{d_s}$ be a shared Transformer encoder with positional encoding, where $\mathcal{X}_T \subset (\mathbb{R}^{d_x})^T$, $\mathcal{Z}$ is compact, and d_s does not depend on T. For the jth context

task, define $\mathbf{z}_j = E_\omega(X_{1:T}^{(j)})$, $r_j = \tau_j/T$, and $\mathbf{u}_j = [\mathbf{z}_j^\top, r_j]^\top$. The query representation is $\mathbf{c}_{k+1} = E_\omega(X_{1:T}^{(k+1)})$. Under Assumption 3.4, $r_j \in [\kappa, 1-\kappa]$, so $\mathbf{u}_j \in \mathcal{U}_s := \mathcal{Z} \times [\kappa, 1-\kappa]$, $\mathbf{c}_{k+1} \in \mathcal{C}_s := \mathcal{Z}$, and $d_{\text{meta}} := d_s + 1$.

Task-level aggregation and decoding. Following the uniform-attention construction of Wakayama & Suzuki (2026), we use the same simplex-valued ReLU feature map and clipped ReLU decoder parameterization. Relative to that construction, the decoder additionally receives the observable context size k, allowing a single shared predictor to operate across context lengths. The exact parameterization is given in Section C.1.

The prediction on the original scale is $M_\theta(\mathcal{P}^k) := T\widetilde{M}_\theta(\mathcal{P}^k)$. Conditional on I, the context tasks are exchangeable, motivating permutation-invariant aggregation across tasks; see Section C.2. Mean aggregation does not encode the context size. For example, duplicating all context tasks leaves the aggregated representation unchanged. We therefore provide the observed context size k explicitly to the decoder. Possible information loss from the fixed-dimensional sequence representation is quantified by $\varepsilon_{\text{seq}}(p, T)$, while the statistical complexity of the model class is characterized by $V_{m,p,T}$, the pseudo-dimension of its block-level average-prefix loss class.

3.3 Risk and Training Setup

We use the normalized squared localization loss $\ell_T(u, v) := (u - v)^2/T^2$, which measures squared localization error relative to the sequence length. When $k = p$, we have $X_{1:T}^{(p+1)} = X_{1:T}^*$ and $\tau_{p+1} = \tau^*$. The localization risk at context length k is

$$R_k(M) = \mathbb{E}\big[\ell_T\big(\tau_{k+1}, M(\mathcal{P}^k)\big)\big] = \mathbb{E}\left[\frac{\{\tau_{k+1} - M(\mathcal{P}^k)\}^2}{T^2}\right]. \tag{2}$$

Suppose that pretraining uses N independent samples generated as in Definition 3.1, with the nth sample inducing $\{(\mathcal{P}_n^k, \tau_{n,k+1})\}_{k=1}^p$. A common parameter vector θ is shared across all context lengths, and the empirical risk minimizer is

$$\widehat{\theta} \in \arg\min_{\theta \in \Theta} \frac{1}{pN} \sum_{n=1}^{N} \sum_{k=1}^{p} \ell_T\big(\tau_{n,k+1}, M_\theta(\mathcal{P}_n^k)\big). \tag{3}$$

We write $\widehat{\theta}$ for any measurable empirical risk minimizer whenever one exists. The corresponding population risk is $\overline{R}(M) := p^{-1} \sum_{k=1}^p R_k(M)$, which is the population counterpart of the average-prefix ERM in equation 3, while $R_k(M)$ retains context-length-specific localization performance. In particular, $R_p(M)$ evaluates localization for the final query under the complete Prompt. For the theoretical analysis, we impose the following assumptions.

Assumption 3.3 (Conditional independence across tasks)**.** Conditional on $I = i$, the change-point tasks $\{(X_{1:T}^{(j)}, \tau_j)\}_{j \geq 1}$ are independent and identically distributed.

Assumption 3.4 (Interior change-point condition)**.** There exists $\kappa \in (0, 1/2)$ such that $\tau/T \in [\kappa, 1 - \kappa]$ almost surely under every latent change-point type.

Assumption 3.5 (Fixed-dimensional sequence representability)**.** Let $r_{k+1} := \tau_{k+1}/T$ and $\widetilde{M}_{\text{Bayes}}(\mathcal{P}^k) := \mathbb{E}[r_{k+1} \mid \mathcal{P}^k]$. Assume that the candidate sequence-encoder class in Definition 3.2 contains an order-sensitive encoder $E_T^\star$. Let $\mathbf{U}_j^\star$ and $\mathbf{C}_{k+1}^\star$ denote the corresponding context-task and query representations, and define $B_{k,T}(\mathbf{u}_{1:k}, \mathbf{c}) := \mathbb{E}[r_{k+1} \mid \mathbf{U}_{1:k}^\star = \mathbf{u}_{1:k}, \mathbf{C}_{k+1}^\star = \mathbf{c}]$. The resulting Bayes information loss is

$$\varepsilon_{\text{seq}}^2(p, T) := \max_{1 \leq k \leq p} \mathbb{E}\left[\left\{\widetilde{M}_{\text{Bayes}}(\mathcal{P}^k) - B_{k,T}(\mathbf{U}_{1:k}^\star, \mathbf{C}_{k+1}^\star)\right\}^2\right]. \tag{4}$$

Assumption 3.6 (Hölder regularity of the reduced Bayes predictor)**.** There exist constants $L > 0$ and $\alpha \in (0, 1]$, independent of T, p, and k, such that for every $k = 1, \ldots, p$, $\mathbf{u}_j, \mathbf{u}_j' \in \mathcal{U}_s$, and $\mathbf{c}, \mathbf{c}' \in \mathcal{C}_s$,

$$|B_{k,T}(\mathbf{u}_{1:k}, \mathbf{c}) - B_{k,T}(\mathbf{u}_{1:k}', \mathbf{c}')| \leq \frac{L}{k} \sum_{j=1}^{k} \big\|(\mathbf{u}_j, \mathbf{c}) - (\mathbf{u}_j', \mathbf{c}')\big\|_2^\alpha. \tag{5}$$

Assumptions 3.3–3.4 require conditional independence across tasks given I and an interior change point, while allowing arbitrary within-sequence dependence and imposing no additional boundedness or independence on X_t. Assumption 3.5 permits lossy dimension reduction, with $\varepsilon_{\mathrm{seq}}(p,T)$ measuring the resulting Bayes-relevant information loss and additional T-dependence entering through the sequence-level complexity terms. Assumption 3.6 imposes Hölder regularity on the reduced Bayes predictor to support the approximation analysis; sufficient conditions are given in Appendix D.

3.4 Bayes-optimal Contextual Predictor

Under squared localization loss, the conditional mean minimizes the prediction risk. Therefore, for each $k \in \{1,\dots,p\}$, the Bayes-optimal contextual predictor is defined as

$$M_{\mathrm{Bayes}}(\mathcal{P}^k) = \mathbb{E}\left[\tau_{k+1}\middle|\mathcal{C}_k, X_{1:T}^{(k+1)}\right]. \tag{6}$$

Thus, M_{Bayes} minimizes $R_k(M)$ for every k, and consequently also minimizes the average risk $\overline{R}(M) = p^{-1}\sum_{k=1}^{p} R_k(M)$. To reveal how the latent change-point type enters the Bayes predictor, we next express it through the posterior distribution of I.

Posterior representation. For $\mathcal{P}^k = (\mathcal{C}_k, X_{1:T}^{(k+1)})$, define $\pi_i(\mathcal{P}^k) := \Pr(I = i \mid \mathcal{P}^k)$. Under Assumption 3.3, the posterior distribution admits the mixture representation $P(\tau_{k+1} \mid \mathcal{P}^k) = \sum_{i=1}^{K} \pi_i(\mathcal{P}^k) P(\tau_{k+1} \mid X_{1:T}^{(k+1)}, I = i)$. Consequently, the Bayes predictor can be written as $M_{\mathrm{Bayes}}(\mathcal{P}^k) = \sum_{i=1}^{K} \pi_i(\mathcal{P}^k)\mathbb{E}\left[\tau_{k+1} \mid X_{1:T}^{(k+1)}, I = i\right]$.

4 Risk Analysis for In-Context Change-Point Localization

Following the Bayesian risk decomposition of Wakayama & Suzuki (2026), we specialize the *Bayes Gap* and *Posterior Variance* to normalized change-point localization. For any measurable bounded predictor M and $k \in \{1,\dots,p\}$, define $R_{\mathrm{BG},k}(M) := \mathbb{E}[\{M(\mathcal{P}^k) - M_{\mathrm{Bayes}}(\mathcal{P}^k)\}^2/T^2]$ and $R_{\mathrm{PV},k} := \mathbb{E}[\mathrm{Var}(\tau_{k+1} \mid \mathcal{P}^k)/T^2]$. Then $R_k(M) = R_{\mathrm{BG},k}(M) + R_{\mathrm{PV},k}$, with $R_{\mathrm{PV},k} = R_k(M_{\mathrm{Bayes}})$. To match the shared-prefix pretraining objective, define $\overline{R}_{\mathrm{BG}}(M) := p^{-1}\sum_{k=1}^{p} R_{\mathrm{BG},k}(M)$ and $\overline{R}_{\mathrm{PV}} := p^{-1}\sum_{k=1}^{p} R_{\mathrm{PV},k}$, so that $\overline{R}(M) = \overline{R}_{\mathrm{BG}}(M) + \overline{R}_{\mathrm{PV}}$. We next bound each term separately.

4.1 Bayes Gap: Approximation and Generalization Error

We now bound the Bayes Gap of the shared pretrained predictor. We first introduce the assumption needed for the analysis.

Assumption 4.1 (Architectural realizability and finite capacity)**.** For each (m,p,T), the model class $\Theta_{m,p,T}$ satisfies: (i) it contains a single shared parameterization whose sequence encoder realizes $E_T^\star$ and whose task feature encoder and count-aware decoder simultaneously approximate $B_{k,T}$ for all $k = 1,\dots,p$ with uniform error $O(m^{-\alpha/d_{\mathrm{meta}}})$; and (ii) $V_{m,p,T} < \infty$.

The shared approximation construction underlying Assumption 4.1 is given in Section E.3.1.

Theorem 4.2 (Per-context-length Bayes Gap for the shared ERM)**.** *Consider the prompt-generation process in Definition 3.1 and the model class in Definition 3.2. Suppose that Assumptions 3.3–3.6 and Assumption 4.1 hold, and that* $1 \le V_{m,p,T} \le N$*. Then, for any fixed* $k \in \{1,\dots,p\}$*, the shared ERM defined in equation 3 satisfies*

$$\mathbb{E}R_{\mathrm{BG},k}(M_{\widehat{\theta}}) \le Cp\left\{\varepsilon_{\mathrm{seq}}^2(p,T) + m^{-2\alpha/d_{\mathrm{meta}}} + \sqrt{\frac{V_{m,p,T}\log(eN)}{N}}\right\}, \tag{7}$$

where $d_{\mathrm{meta}} = d_s + 1$ *and* $C > 0$ *is independent of* (N,m,p,T).

The bound in equation 7 separates three sources of error. The term $\varepsilon_{\mathrm{seq}}^2(p,T)$ measures the Bayes-relevant information loss induced by the fixed-dimensional sequence representation, $m^{-2\alpha/d_{\mathrm{meta}}}$ is

the approximation error for the reduced H"older-smooth Bayes predictor, and $\sqrt{V_{m,p,T}\log(eN)/N}$ is the pretraining generalization error. Since $d_{\mathrm{meta}} = d_s + 1$ is fixed, the approximation exponent does not deteriorate with T, which enters the bound through $\varepsilon_{\mathrm{seq}}(p,T)$ and $V_{m,p,T}$. The corresponding average-prefix bound and further refinements are given in Sections E.5 and E.6, respectively.

4.2 Posterior Variance: Inference-stage Uncertainty

Unlike the Bayes Gap, the Posterior Variance is independent of the training procedure and equals the Bayes localization risk, $R_{\mathrm{PV},k} = R_k(M_{\mathrm{Bayes}})$. Under Assumption 3.3, it is nonincreasing in the context length, $R_{\mathrm{PV},k+1} \le R_{\mathrm{PV},k}$ for $k = 1, \dots, p-1$; see Lemma F.3 in Section F.4.

To benchmark the localization difficulty within each change mechanism, for $i \in \{1, \dots, K\}$ and $f \in \mathcal{F}_i$, let $\tau(f)$ denote the true change-point location and P_f the corresponding distribution of the observed sequence. Define the type-specific minimax localization risk by

$$\mathcal{R}_T(\mathcal{F}_i) := \inf_{\widehat{\tau}} \sup_{f \in \mathcal{F}_i} \mathbb{E}_f \left[\frac{\{\widehat{\tau}(X_{1:T}) - \tau(f)\}^2}{T^2} \right]. \tag{8}$$

The remaining inference-stage uncertainty comes from not knowing which latent change mechanism generated the current prompt. We next state the identifiability condition required for the following analysis.

Assumption 4.3 (Latent-type identifiability)**.** Fix the true type $i^\star$. For each i, suppose that $(X_{1:T}, \tau) \mid I = i$ has joint density $q_{i,T}(x,\tau)$ with respect to a common σ-finite measure, and that $X_{1:T} \mid I = i$ has marginal density $q^X_{i,T}(x)$ with respect to a common dominating measure. For each $j \neq i^\star$, define $D_{j,T} := \mathbb{E}[\log\{q_{i^\star,T}(X_{1:T},\tau)/q_{j,T}(X_{1:T},\tau)\} \mid I = i^\star] \in (0,\infty)$, $Z_{j,t,T} := \log\{q_{j,T}(X^{(t)}_{1:T},\tau_t)/q_{i^\star,T}(X^{(t)}_{1:T},\tau_t)\}$, and $\mathcal{G}_t := \sigma((X^{(1)}_{1:T},\tau_1),\dots,(X^{(t)}_{1:T},\tau_t))$. By conditional independence, $\mathbb{E}[Z_{j,t,T} \mid \mathcal{G}_{t-1}, I = i^\star] = -D_{j,T}$. Assume further that there exist $\nu_{j,T} > 0$ and $b_{j,T} > 0$ such that, for all $|\lambda| \le 1/b_{j,T}$,

$$\mathbb{E}\left[\exp\{\lambda(Z_{j,t,T} + D_{j,T})\} \mid \mathcal{G}_{t-1}, I = i^\star\right] \le \exp\left(\frac{\lambda^2 \nu^2_{j,T}}{2}\right). \tag{9}$$

Finally, let $D_{\min,T} := \min_{j \neq i^\star} D_{j,T} > 0$, $C_{j,T} := \sup_{0<\lambda\le 1/b_{j,T}}\{\lambda D_{j,T}/2 - \lambda^2\nu^2_{j,T}/2\} > 0$, and $C_{\mathrm{id},T} := \min_{j\neq i^\star} C_{j,T} > 0$.

The next result gives the Posterior Variance bound used in the downstream localization analysis.

Theorem 4.4 (Posterior Variance bound under latent-type identification)**.** *Suppose that Assumption 3.3 holds. If $K \ge 2$, Suppose that Assumption 4.3 holds for every true type $i \in \{1,\dots,K\}$. Let $D_{\min,i,T}$ and $C_{\mathrm{id},i,T}$ denote the corresponding identification constants. and define*

$$\varepsilon^{\mathrm{id}}_{k,i,T} := \frac{1-\alpha_i}{\alpha_i}\exp(-D_{\min,i,T}k/2) + (K-1)\exp(-C_{\mathrm{id},i,T}k), \qquad K \ge 2. \tag{10}$$

For $K = 1$, set $\varepsilon^{\mathrm{id}}_{k,1,T} := 0$ and no identifiability condition is required. Then, for any fixed $k \in \{1,\dots,p\}$,

$$R_{\mathrm{PV},k} \le \sum_{i=1}^K \alpha_i \mathcal{R}_T(\mathcal{F}_i) + 5\sum_{i=1}^K \alpha_i \varepsilon^{\mathrm{id}}_{k,i,T}. \tag{11}$$

Theorem 4.4 separates the Posterior Variance bound into a prior-weighted within-mechanism localization benchmark and the error incurred when identifying the latent mechanism from the prompt. For $K \ge 2$, the identification term decays exponentially in the number of solved context tasks whenever the latent types are sufficiently separated. A stronger type-conditional result, which explicitly establishes posterior concentration on the true latent type and convergence to the corresponding type-oracle Posterior Variance, is given in Section F.1.

4.3 Overall Risk

We now combine the Bayes Gap and Posterior Variance bounds to obtain the overall localization risk of the shared pretrained predictor.

Theorem 4.5 (Per-context-length ICCP risk bound)**.** *Suppose that the conditions of Theorems 4.2 and 4.4 hold. Then, for any $k \in \{1, \dots, p\}$,*

$$\mathbb{E}R_k(M_{\widehat{\theta}}) \le Cp\left\{\varepsilon^2_{\mathrm{seq}}(p,T) + m^{-2\alpha/d_{\mathrm{meta}}} + \sqrt{\frac{V_{m,p,T}\log(eN)}{N}}\right\} + \sum_{i=1}^{K}\alpha_i\mathcal{R}_T(\mathcal{F}_i) + 5\sum_{i=1}^{K}\alpha_i\varepsilon^{\mathrm{id}}_{k,i,T}.$$

To obtain a simpler identification term, suppose further that there exist $D_{0,T}, C_{0,T} > 0$ such that $D_{\min,i,T} \ge D_{0,T}$ and $C_{\mathrm{id},i,T} \ge C_{0,T}$ for every i, and let $c_{0,T} := \min\{D_{0,T}/2, C_{0,T}\}$. Then $\sum_{i=1}^{K}\alpha_i\varepsilon^{\mathrm{id}}_{k,i,T} \le 2(K-1)e^{-c_{0,T}k}$, and consequently

$$\mathbb{E}R_k(M_{\widehat{\theta}}) \le Cp\left\{\varepsilon^2_{\mathrm{seq}}(p,T) + m^{-2\alpha/d_{\mathrm{meta}}} + \sqrt{\frac{V_{m,p,T}\log(eN)}{N}}\right\} + \sum_{i=1}^{K}\alpha_i\mathcal{R}_T(\mathcal{F}_i) + 10(K-1)e^{-c_{0,T}k}.$$

Localized-complexity and polynomial-capacity refinements of this bound are given in Section G.3. For the complete Prompt, setting $k = p$ in Theorem 4.5 gives the localization-risk bound for the final query $(X^*_{1:T}, \tau^*)$.

4.4 Relative Change-point Localization Consistency

For an independent test Prompt $\mathcal{P}^k$, let $\widehat{\tau}_k := M_{\widehat{\theta}}(\mathcal{P}^k)$. Since the pretrained predictor is independent of the test Prompt, $\mathbb{E}R_k(M_{\widehat{\theta}}) = \mathbb{E}[(\widehat{\tau}_k - \tau_{k+1})^2/T^2]$. We now give sufficient conditions under which the risk bound in Theorem 4.5 implies relative localization consistency.

Assumption 4.6 (Sufficient conditions for relative localization consistency)**.** Consider a sequence of problems satisfying:

- (C1) **Asymptotic regime.** The dimensions d_x, d_s, and K are fixed, while $N, T \to \infty$. Let $p = p_{N,T} \ge 1$ and $k = k_{N,T} \in \{1, \dots, p\}$, so that $d_{\mathrm{meta}} = d_s + 1$ remains fixed.
- (C2) **Type identification.** For some $D_{0,T}, C_{0,T} > 0$, uniformly in i, $D_{\min,i,T} \ge D_{0,T}$ and $C_{\mathrm{id},i,T} \ge C_{0,T}$. Let $c_{0,T} := \min\{D_{0,T}/2, C_{0,T}\}$ and assume $kc_{0,T} \to \infty$.
- (C3) **Intrinsic localization.** For some $r_T \downarrow 0$, $\max_{1\le i\le K}\mathcal{R}_T(\mathcal{F}_i) \le r_T$.
- (C4) **Sequence representation.** Assume $p\,\varepsilon^2_{\mathrm{seq}}(p,T) \to 0$.
- (C5) **Learning complexity and scaling.** For some $\Gamma_T \ge 1$ and fixed constants $C_V > 0$, $\beta \ge 0$, $\gamma > 0$, and $q \ge 0$, assume $V_{m,p,T} \le C_V\Gamma_T p^\beta m^\gamma(\log m)^q$. Let $A_{N,T} := N/(\Gamma_T p^\beta)$, $a_{\mathrm{slow}} := 2\alpha/(\gamma d_{\mathrm{meta}} + 4\alpha)$, and $\mathfrak{L}_N := (\log(eN))^{q+1}$. Assume $A_{N,T} \to \infty$ and $pA_{N,T}^{-a_{\mathrm{slow}}}\mathfrak{L}_N \to 0$.

Under Assumption 4.6, the overall risk bound yields the following localization guarantee.

Theorem 4.7 (Relative change-point localization consistency under shared pretraining)**.** *Suppose that the conditions of Theorem 4.5 and Assumption 4.6 hold, and choose $m = m_{N,T} := \left\lceil A_{N,T}^{d_{\mathrm{meta}}/(\gamma d_{\mathrm{meta}}+4\alpha)}\right\rceil$. Then,*

$$\mathbb{E}\big[T^{-2}(\widehat{\tau}_k - \tau_{k+1})^2\big] \longrightarrow 0, \qquad T^{-1}|\widehat{\tau}_k - \tau_{k+1}| \xrightarrow{P} 0.$$

In particular, for the complete Prompt, taking $k = p$ and assuming $pc_{0,T} \to \infty$ gives $T^{-1}\left|M_{\widehat{\theta}}(\mathcal{P}^p) - \tau_{p+1}\right| \xrightarrow{P} 0$. A localized fast-rate variant is given in Section G.3.

5 Simulation Studies

Following the prompt-generation mechanism in Definition 3.1, we generate prompts with $p = 20$ solved tasks and one query task, all sharing the same latent change type I. We consider mean,

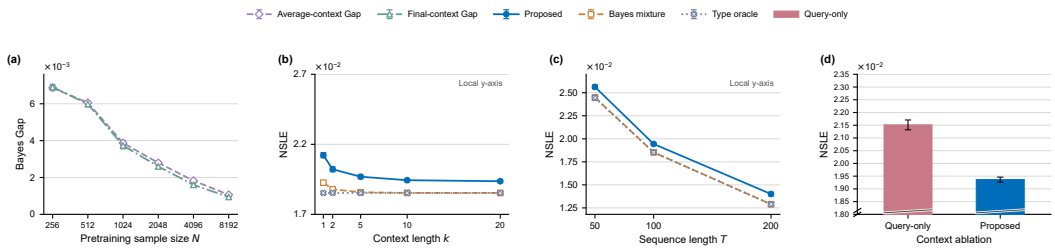


Figure 2: Simulation results with $d = 5$, $p = 20$, and $c = 0.75$. "Proposed" denotes ICCP. (a) Average shared-prefix and final-context Bayes Gaps versus pretraining size N at $T = 100$. (b) NSLE versus context length k at $N = 8192$ and $T = 100$, using the same trained models and fixed query tasks. (c) NSLE versus sequence length T at $N = 8192$ and $k = 10$. (d) Query-only and ICCP NSLE on paired test prompts at $N = 8192$ and $T = 100$, with $k = 20$ for ICCP. Points and bars show means; error bars indicate standard errors over five seeds.

marginal-variance, and cross-coordinate-correlation changes with equal probability. Conditional on I, tasks are generated independently with task-specific parameters. We set the dimension to $d = 5$ and fix the signal strength at $c = 0.75$. Performance is measured by normalized squared localization error (NSLE). The Bayes mixture and type oracle serve as theoretical references, while a Query-only baseline is used to assess the contribution of contextual information. Results are averaged over five independent seeds, with 2,000 test prompts per seed for each evaluation setting. All paired comparisons use the same test prompts within each seed. Full experimental details are provided in Appendix A.

Effects of pretraining, context, and sequence length. We examine how localization performance varies with the pretraining sample size, context length, and sequence length. As the pretraining sample size N increases, both the Bayes Gap and localization error decrease substantially (Fig. 2(a)), with the former consistent with the N-dependence established in Theorem 4.2. Holding the trained model fixed, providing more context tasks reduces ICCP localization error (Fig. 2(b)). The Bayes mixture also approaches the type oracle, suggesting improved identification of the shared latent type with additional context. Longer sequences yield lower normalized localization error (Fig. 2(c)). Together, these results support the qualitative predictions of the theory: more pretraining data reduces estimation error, additional context improves task identification, and longer sequences improve relative localization.

Context ablation. To assess the contribution of contextual information, we compare ICCP with a Query-only baseline. Both models use the same sequence-encoder architecture and optimization protocol, while Query-only is trained and evaluated using only the query sequence. For $N = 8192$ and $T = 100$, ICCP yields lower NSLE than Query-only in all five seeds (Fig. 2(d)). Complementing this ablation, the fixed-checkpoint k-sweep in Fig. 2(b) shows that the same trained ICCP model achieves lower localization error as more solved context tasks are provided while its parameters remain fixed. Together, these results indicate that contextual information improves change-point localization. Additional results are reported in Appendix B.

6 Conclusion

We introduced ICCP, a task-level framework that uses previously solved change-point tasks to localize structural changes in a new sequence without parameter updates or a prespecified localization statistic. We derived finite-sample error bounds for the proposed predictor, established relative localization consistency under suitable conditions, and provided simulations supporting the main theoretical predictions. In practice, previously diagnosed sequences and their change locations can serve directly as in-context demonstrations for new sequences. The current framework focuses on localization when a change point is assumed to be present. Natural extensions include determining whether a structural change has occurred and extending the framework to multiple change points.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and language editing, refining theoretical formulations and proof exposition, feedback on experimental design, LaTeX and code debugging, literature-search support, and figure/presentation refinement. All AI-assisted mathematical claims and proofs, experimental implementations and numerical results, and cited references were checked by the authors. The authors take full responsibility for the final content of this work, including all text, claims, code, and artifacts produced with AI assistance.

REPRODUCIBILITY STATEMENT

The complete prompt-generation process, model architecture, training protocol, evaluation metrics, and experimental settings are documented in Appendix A. Additional simulation results are reported in Appendix B, and the assumptions and proofs underlying the theoretical results are given in the remaining appendices. These details are intended to support independent reproduction of both the empirical and theoretical results.

REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Jushan Bai. Least squares estimation of a shift in linear processes. *Journal of Time Series Analysis*, 15(5):453–472, 1994. doi: 10.1111/j.1467-9892.1994.tb00204.x.

Peiliang Bai, Abolfazl Safikhani, and George Michailidis. Multiple change point detection in reduced rank high dimensional vector autoregressive models. *Journal of the American Statistical Association*, 118(544):2776–2792, 2023a. doi: 10.1080/01621459.2022.2079514. URL https://doi.org/10.1080/01621459.2022.2079514.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=liMSqUuVg9.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. doi: 10.1214/009053605000000282.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019. URL https://jmlr.org/papers/v20/17-612.html.

Hao Chen and Lynna Chu. Graph-based change-point analysis. *Annual Review of Statistics and Its Application*, 10(1):475–499, 2023. doi: 10.1146/annurev-statistics-122121-033817.

Xiang Cheng, Yuxin Chen, and Suvrit Sra. Transformers implement functional gradient descent to learn non-linear functions in context. 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. 8002–8037. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cheng24a.html.

Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(2):475–507, 03 2015. ISSN 1369-7412. doi: 10.1111/rssb.12079. URL https://doi.org/10.1111/rssb.12079.

Carson Dudley, Xiaofeng Liu, Samet Oymak, and Yutong Bi. In-context learning under regime change. In *2nd ICML Workshop on Foundation Models for Structured Data*, 2026. URL https://openreview.net/forum?id=o3dwy5pDAJ.

Paul Fearnhead. Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213, 2006. doi: 10.1007/s11222-006-8450-8.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/finn17a.html.

Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243 – 2281, 2014. doi: 10.1214/14-AOS1245. URL https://doi.org/10.1214/14-AOS1245.

Piotr Fryzlewicz. Narrowest significance pursuit: Inference for multiple change-points in linear models. *Journal of the American Statistical Association*, 119(546):1633–1646, 2024. doi: 10.1080/01621459.2023.2211733. URL https://doi.org/10.1080/01621459.2023.2211733.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598. Curran Associates, Inc., 2022. doi: 10.52202/068431-2217. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c529dba08a146ea8d6cf715ae8930cbe-Paper-Conference.pdf.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional neural processes. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1704–1713. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/garnelo18a.html.

Tianyu Guo, Wei Hu, Song Mei, Huan Wang, Caiming Xiong, Silvio Savarese, and Yu Bai. How do transformers learn in-context beyond simple functions? a case study on learning with representations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ikwEDva1JZ.

Muktesh Gupta, Rajesh Wadhvani, and Akhtar Rasool. Real-time change-point detection: A deep neural network-based adaptive approach for detecting changes in multivariate time series data. *Expert Systems with Applications*, 209:118260, 2022. ISSN 0957-4174. doi: 10.1016/j.eswa.2022.118260. URL https://www.sciencedirect.com/science/article/pii/S0957417422014026.

James Harrison, Apoorva Sharma, Chelsea Finn, and Marco Pavone. Continuous meta-learning without tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17571–17581. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/cc3f5463bc4d26bc38eadc8bcffbc654-Paper.pdf.

Jie Li, Paul Fearnhead, Piotr Fryzlewicz, and Tengyao Wang. Automatic change-point detection in time series via deep learning. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):273–285, 01 2024. ISSN 1369-7412. doi: 10.1093/jrsssb/qkae004. URL https://doi.org/10.1093/jrsssb/qkae004.

Yunchen Li, Guanghui Wang, Shuntuo Xu, and Zhou Yu. Integral probability metric-guided cusum-net for nonparametric changepoint detection. *Biometrika*, 113(3):asag046, 09 2026. ISSN 1464-3510. doi: 10.1093/biomet/asag046. URL https://doi.org/10.1093/biomet/asag046.

E T McGonigle and H Cho. Nonparametric data segmentation in multivariate time series via joint characteristic functions. *Biometrika*, 112(2):asaf024, 04 2025. ISSN 1464-3510. doi: 10.1093/biomet/asaf024. URL https://doi.org/10.1093/biomet/asaf024.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-context learning through the bayesian prism. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 49789–49843, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/d81cd83e7f6748af351485d73f305483-Paper-Conference.pdf.

Hoseung Song and Hao Chen. Asymptotic distribution-free changepoint detection for data with repeated observations. *Biometrika*, 109(3):783–798, 09 2022. ISSN 1464-3510. doi: 10.1093/biomet/asab048. URL https://doi.org/10.1093/biomet/asab048.

Tomoya Wakayama and Taiji Suzuki. In-context learning is provably bayesian inference: A generalization theory for meta-learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=BUFSSOuphA.

Guanghui Wang and Long Feng. Computationally efficient and data-adaptive changepoint inference in high dimension. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85 (3):936–958, 07 2023. ISSN 1369-7412. doi: 10.1093/jrsssb/qkad048. URL https://doi.org/10.1093/jrsssb/qkad048.

Zhijie Wang, Bo Jiang, and Shuai Li. In-context learning on function classes unveiled for 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. 50726–50745. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wang24ae.html.

Mengjia Yu and Xiaohui Chen. Finite sample change point inference and identification for high-dimensional mean vectors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):247–270, 04 2021. ISSN 1369-7412. doi: 10.1111/rssb.12406. URL https://doi.org/10.1111/rssb.12406.

Zifeng Zhao, Feiyu Jiang, and Xiaofeng Shao. Segmenting time series via self-normalisation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1699–1725, 11 2022. ISSN 1369-7412. doi: 10.1111/rssb.12552. URL https://doi.org/10.1111/rssb.12552.

Houlin Zhou, Hanbing Zhu, and Xuejun Wang. Change point detection via feedforward neural networks with theoretical guarantees. *Computational Statistics & Data Analysis*, 193:107913, 2024. ISSN 0167-9473. doi: 10.1016/j.csda.2023.107913. URL https://www.sciencedirect.com/science/article/pii/S0167947323002244.

Houlin Zhou, Hanbing Zhu, and Xuejun Wang. Change-point detection for multivariate nonparametric regression with deep neural networks. *Computational Statistics & Data Analysis*, 2026. doi: 10.1016/j.csda.2025.108334.

Appendix organization. Appendices A–B provide the complete simulation design and additional empirical results, so that experimental details are immediately accessible. Appendix C records architectural details. Appendix D collects sufficient conditions supporting the reduced-Bayes regularity assumption used in the approximation analysis. Appendices E–G follow the theoretical dependency chain from the Bayes Gap to Posterior Variance, overall risk, and localization consistency.

A Simulation Setup and Implementation Details

A.1 Hierarchical Data Generation

We set $d = 5$, $K = 3$, and $p = 20$. Each prompt contains $p + 1$ change-point tasks and is generated hierarchically as follows.

Step 1: Sample the shared change type. For each prompt, we first draw $I \sim \mathrm{Categorical}(1/3, 1/3, 1/3)$, where $I = 1, 2, 3$ correspond to mean, variance, and covariance changes, respectively. Conditional on I, the $p + 1$ tasks are generated independently. Thus, all tasks within a prompt share the same change type but have different task-specific parameters.

Step 2: Generate the task-specific baseline. Let $\mathbf{v} = \mathbf{1}_d/\sqrt{d}$ and $s_t = (t-1)/(T-1)$. We use three baseline profiles: $q_1(s) = s$, $q_2(s) = s^2$, and $q_3(s) = (e^{2s} - 1)/(e^2 - 1)$. Each profile is centered and normalized on the observation grid:

$$h_{a,t} = \frac{q_a(s_t) - \bar{q}_a}{\left\{T^{-1}\sum_{u=1}^{T}[q_a(s_u) - \bar{q}_a]^2\right\}^{1/2}}, \qquad \bar{q}_a = T^{-1}\sum_{u=1}^{T} q_a(s_u).$$

For task j, we independently draw $H_j \sim \mathrm{Unif}\, 1, 2, 3$ and $A_j \sim \mathrm{Unif}\, 0.1, 0.2, 0.3$, and set $\mathbf{b}_{j,t} = A_j h_{H_j,t}\mathbf{v}$. The baseline index H_j is independent of the shared change type I.

Step 3: Sample the change-point location. Each task independently draws $\tau_j \sim \mathrm{Unif}\,\lceil 0.2T\rceil, \ldots, \lfloor 0.8T\rfloor$. Hence, change points are not aligned across tasks and need not occur at the midpoint.

Step 4: Sample the change magnitude. Under the main setting $c = 0.75$, the active signal parameter is sampled independently for each task with equal probability from $\Delta_j \in 0.24, 0.3375, 0.45$ for mean changes, $\sigma_j \in 1.075, 1.1125, 1.15$ for variance changes, and $\rho_j \in 0.075, 0.1125, 0.15$ for covariance changes. For mean changes, we additionally draw an independent sign $S_j \in -1, 1$ with equal probability.

Step 5: Generate the observations. Conditional on the task parameters, observations are independent over time and satisfy

$$X_t^{(j)} \sim \begin{cases} \mathcal{N}(\mathbf{b}_{j,t}, I_d), & t \le \tau_j, \\ \mathcal{N}(\mathbf{b}_{j,t} + S_j\Delta_j\mathbf{v}, I_d), & t > \tau_j,\ I = 1, \\ \mathcal{N}(\mathbf{b}_{j,t}, \sigma_j^2 I_d), & t > \tau_j,\ I = 2, \\ \mathcal{N}\left(\mathbf{b}_{j,t}, (1-\rho_j)I_d + \rho_j\mathbf{1}_d\mathbf{1}_d^\top\right), & t > \tau_j,\ I = 3. \end{cases}$$

The baseline $\mathbf{b}_{j,t}$ persists throughout the sequence; only the change mechanism switches at τ_j. The nuisance parameters and active signal magnitude are task-specific and hidden from the predictor. The finite signal grids form part of the task distribution $P_{\mathcal{F}_i}$ rather than additional latent types. Since $|\mathbf{v}|_2 = 1$, Δ_j is the Euclidean mean-shift magnitude, σ_j is the post-change standard deviation, and ρ_j is the post-change correlation between distinct coordinates. The covariance-change construction preserves unit marginal variances.

A.2 Architecture and Training

We implemented Definition 3.2 with the same sequence-encoding, task-encoding, aggregation, and decoding structure described in the main text.

Sequence encoder. The shared sequence encoder E_ω is a bidirectional Transformer with sinusoidal positional encoding, two layers, hidden dimension 64, and four attention heads. No causal mask is used. Each layer uses pre-layer normalization and a ReLU feedforward block of width 128, followed by a final layer normalization. Temporal mean pooling and a $\tanh$ map produce the 64-dimensional sequence representation. Dropout is set to zero. We do not standardize individual sequences, since scale changes are themselves informative for the variance-change tasks.

Task feature encoder. For each solved context task, the 64-dimensional sequence representation is concatenated with its normalized change-point label τ_j/T. The resulting 65-dimensional input is

mapped by the shared task encoder ϕ_{ϑ_ϕ} through a $65 \to 64 \to 64$ network with a ReLU hidden activation. Its output is constrained to be nonnegative and normalized as

$$\phi(\mathbf{u}) = \frac{\mathrm{ReLU}(g(\mathbf{u})) + (\delta/m)\mathbf{1}_m}{\mathbf{1}_m^\top \mathrm{ReLU}(g(\mathbf{u})) + \delta}, \qquad m = 64, \quad \delta = 0.01.$$

Context aggregation and decoder. For context length k, the task features are aggregated by their uniform mean, as in Definition 3.2. The decoder ρ_{ϑ_ρ} receives the resulting 64-dimensional context representation, the 64-dimensional query representation, and the explicit context count k. It is a $129 \to 64 \to 1$ ReLU network whose output is clipped to $[0.2, 0.8]$. This gives the normalized prediction $\widetilde{M}_{\widehat{\theta}}$ and the original-scale estimate $M_{\widehat{\theta}} = T\widetilde{M}_{\widehat{\theta}}$. The output-layer weights are initialized independently from $\mathcal{N}(0, 0.01^2)$ and the output bias is initialized at 0.5.

Training objective and optimization. Training minimizes the shared-prefix squared-loss objective in equation 3, using every prefix $k = 1, \ldots, p$ from each complete training prompt. We use Adam with learning rate 5×10^{-4}, batch size 64, zero weight decay, and gradient-norm clipping at 1. Validation uses 128 independently generated prompts. Early stopping has patience 20, minimum improvement 10^{-5}, and a maximum of 150 epochs. The checkpoint selected by this validation criterion is retained, and test performance is never used for checkpoint or hyperparameter selection.

Training and evaluation protocol. Training, validation, and test prompts are generated independently from the same meta-distribution using disjoint random draws. We use five independent seeds, $1, 2, 3, 4, 5$, and evaluate on 2,000 test prompts per seed and sequence length. Within each seed, experiments with different pretraining sizes N use nested training subsets and a common validation set. For a fixed T and seed, the same test prompts are reused across models, context lengths, and the corresponding paired ablation, enabling paired comparisons without overlap between the training, validation, and test splits.

All models are frozen at test time, with no parameter updates. The query label τ^* is used only to compute evaluation metrics and is never provided as an input to the predictor.

A.3 Evaluation Metrics and Bayesian References

Localization metrics. For M test queries, we report

$$\mathrm{NSLE} = \frac{1}{M}\sum_{m=1}^{M} \frac{(\widehat{\tau}_m - \tau_m)^2}{T^2}, \qquad \mathrm{MAE} = \frac{1}{M}\sum_{m=1}^{M} |\widehat{\tau}_m - \tau_m|. \tag{12}$$

Predictions are evaluated on their continuous scale without integer rounding. NSLE is the primary metric and matches the normalized squared loss used in $R_k(M)$, whereas MAE measures the average localization error in time points.

Bayes and type-oracle references. The exact Bayes predictor is $M_{\mathrm{Bayes}} = \mathbb{E}[\tau^* \mid \mathcal{C}_k, X^*]$, obtained by marginalizing the latent type and all task-specific nuisance variables under the generative prior. The type oracle $M_{\mathrm{oracle}} = \mathbb{E}[\tau^* \mid X^*, I]$ additionally observes the true change type I, while still integrating over the unknown query-specific nuisance variables. Both are conditional-mean references rather than trained neural networks.

Bayes Gap. To mirror the shared-prefix training objective, we report the average-prefix Bayes Gap

$$\widehat{\overline{R}}_{\mathrm{BG}} = \frac{1}{Mp}\sum_{m=1}^{M}\sum_{k=1}^{p} \frac{\left\{M_{\widehat{\theta}}(\mathcal{P}_m^k) - M_{\mathrm{Bayes}}(\mathcal{P}_m^k)\right\}^2}{T^2}. \tag{13}$$

Here $\mathcal{P}_m^k$ uses task $k+1$ as the query, exactly as in the shared-prefix training objective. We also report the final-context Bayes Gap

$$\widehat{R}_{\mathrm{BG},p} = \frac{1}{M}\sum_{m=1}^{M} \frac{\left\{M_{\widehat{\theta}}(\mathcal{P}_m^p) - M_{\mathrm{Bayes}}(\mathcal{P}_m^p)\right\}^2}{T^2}. \tag{14}$$

Context-length evaluation. The context-length experiment serves a different purpose. For each test prompt, we hold the final query X^* fixed and vary only the number k of solved context tasks supplied to the frozen model. This gives paired comparisons across context lengths for the same query. At each k, we also compute the normalized squared difference between the Bayes and type-oracle predictions.

Posterior uncertainty and type identification. For each realized prompt, we compute the normalized posterior variance $\mathrm{Var}(\tau^*/T \mid \mathcal{C}_k, X^*)$ and report its empirical mean over test prompts as PV. Its type-oracle analogue is $\mathrm{Var}(\tau^*/T \mid X^*, I)$. These quantities correspond to the Bayes and type-oracle risks under normalized squared loss.

We also compute the posterior type probabilities $\pi_i = \Pr(I = i \mid \mathcal{C}_k, X^*)$. To assess identification of the shared latent type, we report the mean posterior mass assigned to the true type, π_I, and the accuracy of $\arg\max_i \pi_i$. These are Bayesian diagnostics computed from the generative model, not outputs of an auxiliary classifier.

Aggregation across seeds. All reported means and standard errors are computed from the five seed-specific estimates, with $\mathrm{SE} = \mathrm{SD}/\sqrt{5}$. Thus, the reported SE reflects both test-sampling variability and variation across fitted models; individual prompts and shared prefixes are not treated as independent replicates. Relative reductions are computed from the unrounded seed-averaged estimates, except for the paired ablation percentage, which is first computed within each seed.

A.4 Experimental Settings and Query-Only Ablation

Experimental settings. We study the effects of pretraining size, context length, and sequence length separately. For the pretraining-size experiment, we fix $T = 100$ and $p = 20$ and vary $N \in \{256, 512, 1024, 2048, 4096, 8192\}$. For the context-length experiment, we reuse the model trained with $N = 8192$, $T = 100$, and $p = 20$, and evaluate it at $k \in \{1, 2, 5, 10, 20\}$ without retraining. The final query is held fixed within each test prompt so that only the amount of available context changes. For the sequence-length experiment, we set $N = 8192$, $p = 20$, and $T \in \{50, 100, 200\}$, with evaluation at $k = 10$. The $T = 100$ checkpoint is reused from the pretraining-size experiment. Accordingly, its result differs from the $N = 8192$ final-context result because the two evaluations use $k = 10$ and $k = 20$, respectively.

Query-only ablation. To isolate the contribution of in-context information, we compare ICCP with a Query-only model that predicts the change point from the query sequence alone. Query-only uses the same sequence encoder and optimization procedure as ICCP, but replaces the context-aware decoder with a $64 \to 64 \to 1$ network that receives only the query representation. It is trained using the same 20 task sequences from each complete training prompt, treating each sequence as an independent prediction problem. No context observations, context labels, or latent-type information are provided to Query-only.

For the paired comparison, both models are evaluated on the same final queries at $N = 8192$ and $T = 100$, with ICCP using the full context $k = 20$. The two models are not parameter matched; the purpose of the ablation is to remove context information while retaining the same sequence encoder and training budget.

Bayesian references and paired comparison. Because the two predictors receive different information, their corresponding Bayes references also differ. For Query-only, the reference is $\mathbb{E}[\tau^* \mid X^*]$, whereas for ICCP it is $\mathbb{E}[\tau^* \mid \mathcal{C}_{20}, X^*]$. Their Bayes Gaps therefore measure approximation relative to different information sets and should not be directly interpreted as the total benefit of context.

We instead quantify the predictive benefit of context through a paired comparison of NSLE. For seed s, the relative reduction is

$$100 \frac{\widehat{R}_s^{\mathrm{query}} - \widehat{R}_s^{\mathrm{ICCP}}}{\widehat{R}_s^{\mathrm{query}}}.$$

The reported mean and SE are computed from the five within-seed relative reductions rather than from a ratio of pooled errors.

B Additional Simulation Results

Unless otherwise stated, table entries are means ± standard errors over five seeds, with 2,000 test prompts per seed. MAE is reported in time steps, and BG denotes the normalized squared prediction difference from the corresponding Bayes predictor.

B.1 Pretraining Size

Table 1 reports the effect of the number of meta-training prompts N. Both the average and final-context Bayes Gaps decrease with N, together with the final-context localization error. The Bayes and type-oracle references do not depend on N; at $T = 100$ and $k = 20$, both give NSLE 0.018530 and MAE 10.767 on the shared test sets. A positive Bayes Gap remains at the largest training size.

Table 1: Effect of meta-training size N at $T = 100$, $p = 20$, and $k = 20$.

N	Average BG $\times 10^3$	Final BG $\times 10^3$	NSLE	MAE
256	6.8638 ± 0.1189	6.9103 ± 0.1397	0.025666 ± 0.000192	13.352 ± 0.055
512	6.0576 ± 0.0625	5.9863 ± 0.1505	0.024612 ± 0.000137	12.992 ± 0.064
1024	3.8745 ± 0.1368	3.7328 ± 0.1764	0.022412 ± 0.000167	12.274 ± 0.052
2048	2.7955 ± 0.0846	2.6017 ± 0.1338	0.021165 ± 0.000163	11.886 ± 0.058
4096	1.8273 ± 0.0724	1.6022 ± 0.0615	0.020090 ± 0.000181	11.472 ± 0.090
8192	1.0627 ± 0.0169	0.9415 ± 0.0369	0.019364 ± 0.000098	11.174 ± 0.068

B.2 Context Length and Type Identification

Table 2 reports the fixed-query context-length experiment. As additional solved tasks are supplied to the frozen ICCP model, its localization error decreases and the Bayes mixture approaches the type oracle. The oracle risk does not depend on k because the query is conditionally independent of the context given the true type.

Table 2: Effect of context length k at $N = 8192$ and $T = 100$. The final query is held fixed across context lengths.

k	ICCP NSLE	ICCP MAE	Bayes NSLE
1	0.021224 ± 0.000179	11.738 ± 0.075	0.019264 ± 0.000089
2	0.020217 ± 0.000170	11.452 ± 0.077	0.018791 ± 0.000099
5	0.019687 ± 0.000136	11.270 ± 0.074	0.018576 ± 0.000072
10	0.019439 ± 0.000112	11.197 ± 0.066	0.018531 ± 0.000062
20	0.019364 ± 0.000098	11.174 ± 0.068	0.018530 ± 0.000062

k	Oracle NSLE	Bayes–oracle gap
1	0.018530 ± 0.000062	$(6.66 \pm 0.41) \times 10^{-4}$
2	0.018530 ± 0.000062	$(2.52 \pm 0.23) \times 10^{-4}$
5	0.018530 ± 0.000062	$(2.14 \pm 0.85) \times 10^{-5}$
10	0.018530 ± 0.000062	$(3.50 \pm 2.17) \times 10^{-8}$
20	0.018530 ± 0.000062	$(2.37 \pm 2.36) \times 10^{-25}$

Table 3 provides complementary Bayesian diagnostics. As k increases, both the posterior mass assigned to the true latent type and the corresponding type-identification accuracy approach one. This is consistent with the convergence of the Bayes mixture toward the type oracle.

The averaged trends do not imply strict monotonicity in every seed. In particular, the ICCP–Bayes Gap increases slightly from $k = 10$ to 20 in two seeds, while empirical Bayes NSLE exhibits small increases in three seeds. The nearly vanishing Bayes–oracle gap at large k indicates numerical agreement on these samples and is not used to infer a convergence rate. Empirical Bayes risk and average posterior variance need not coincide on a finite test set.

Table 3: Posterior uncertainty and latent-type identification across context lengths.

k	BG $\times 10^3$	PV	π_I	Type accuracy
1	2.1416 ± 0.0336	0.019687 ± 0.000067	0.9047 ± 0.0012	0.9371 ± 0.0012
2	1.5227 ± 0.0357	0.019249 ± 0.000065	0.9635 ± 0.0007	0.9760 ± 0.0014
5	1.1301 ± 0.0277	0.019025 ± 0.000068	0.9974 ± 0.0002	0.9978 ± 0.0003
10	0.9603 ± 0.0255	0.018999 ± 0.000066	0.9999 ± 0.0001	0.9999 ± 0.0001
20	0.9415 ± 0.0369	0.018998 ± 0.000066	1.0000 ± 0.0000	1.0000 ± 0.0000

B.3 Sequence Length

Table 4 reports the sequence-length experiment. Increasing T decreases normalized localization error, while the absolute ICCP MAE increases with the number of time points. The ICCP–Bayes Gap decreases from $T = 50$ to $T = 100$ and increases slightly at $T = 200$, showing that improved relative localization with longer sequences does not necessarily imply a smaller approximation gap at fixed training size and architecture.

The $T = 100$ checkpoint is reused from the pretraining-size experiment. Its result differs from the $N = 8192$ final-context result because evaluation here uses $k = 10$ rather than $k = 20$.

Table 4: Effect of sequence length T at $N = 8192$, $p = 20$, and $k = 10$.

T	ICCP NSLE	ICCP MAE
50	0.025633 ± 0.000212	6.628 ± 0.033
100	0.019439 ± 0.000112	11.197 ± 0.066
200	0.014006 ± 0.000200	18.184 ± 0.129

T	Bayes NSLE	Oracle NSLE	BG $\times 10^3$
50	0.024480 ± 0.000278	0.024452 ± 0.000273	1.1316 ± 0.0780
100	0.018531 ± 0.000062	0.018530 ± 0.000062	0.9603 ± 0.0255
200	0.012894 ± 0.000203	0.012894 ± 0.000203	1.0587 ± 0.0124

B.4 Paired Context Ablation

Table 5 compares ICCP with the Query-only ablation on identical final queries. ICCP achieves lower NSLE and MAE in all five paired seeds. The mean within-seed relative NSLE reduction is 9.977%, with an SE of 0.579 percentage points.

The two methods condition on different information sets: the Query-only Bayes reference is $\mathbb{E}[\tau^* \mid X^*]$, whereas the ICCP reference is $\mathbb{E}[\tau^* \mid \mathcal{C}_{20}, X^*]$. Their Bayes Gaps are therefore reported relative to their respective Bayes targets and should not be compared directly as a measure of the total benefit of context.

Table 5: Paired comparison of ICCP and Query-only at $N = 8192$ and $T = 100$. BG is computed relative to each method's information-matched Bayes predictor.

Method	NSLE	MAE	BG $\times 10^3$
Query-only	0.021514 ± 0.000193	11.937 ± 0.081	1.0849 ± 0.0125
ICCP	$\mathbf{0.019364 \pm 0.000098}$	$\mathbf{11.174 \pm 0.068}$	0.9415 ± 0.0369

These comparisons support the use of solved context tasks within the specified meta-distribution, rather than establishing that context is universally beneficial.

B.5 Scope of the Simulation Evidence

All runs terminated under the validation-based early-stopping rule, with selected checkpoints occurring between epochs 57 and 136. No test outcomes were used for checkpoint or hyperparameter selection.

The experiments cover a finite grid within a three-type Gaussian meta-distribution and a fixed signal regime. They do not establish out-of-distribution robustness or the asymptotic rates of the theoretical bounds. Increasing N changes the learned predictor, whereas increasing k changes only the amount of context supplied to a frozen predictor; the corresponding finite-sample trends therefore probe different aspects of the theory.

C Architectural Details and Invariance

This appendix records implementation details that support Definition 3.2, including the count-aware decoder, uniform task-level aggregation, and the resulting permutation invariance across solved context tasks.

C.1 Feature Map and Count-Aware Decoder

Following Wakayama & Suzuki (2026), let $g_{\vartheta_\phi} : \mathbb{R}^{d_{\text{meta}}} \to \mathbb{R}^m$ be a feedforward ReLU network and define

$$\phi_{\vartheta_\phi}(\mathbf{u}) = \mathrm{Renorm}_\delta\big(g_{\vartheta_\phi}(\mathbf{u})\big), \qquad \mathrm{Renorm}_\delta(\mathbf{s}) = \frac{\sigma(\mathbf{s}) + (\delta/m)\mathbf{1}}{\mathbf{1}^\top \sigma(\mathbf{s}) + \delta},$$

where $\delta \in (0, 1]$ and $\sigma(x) = \max\{x, 0\}$ is applied coordinatewise. Hence $\phi_{\vartheta_\phi}(\mathbf{u}) \in \Delta^{m-1}$.

Let $h_{\vartheta_\rho} : \mathbb{R}^{m+d_s+1} \to \mathbb{R}$ be a feedforward ReLU network and define the count-aware decoder by

$$\rho_{\vartheta_\rho}(\mathbf{s}, \mathbf{c}, k) = \mathrm{clip}_{[\kappa, 1-\kappa]}\left\{h_{\vartheta_\rho}([\mathbf{s}^\top, \mathbf{c}^\top, k]^\top)\right\},$$

where $\mathrm{clip}_{[a,b]}(x) := a + \sigma(x - a) - \sigma(x - b)$. Relative to the construction of Wakayama & Suzuki (2026), the decoder additionally receives the observable context size k.

C.2 Task-Level Uniform Aggregation

The aggregation in Definition 3.2 is an instance of uniform attention over the solved context tasks. Use the value vectors $\phi_{\vartheta_\phi}(\mathbf{u}_1), \ldots, \phi_{\vartheta_\phi}(\mathbf{u}_k)$ and set the attention query and key projections to zero. All k attention logits are then equal, so their softmax weights are exactly $1/k$. With the identity value readout, the attention output is

$$\mathbf{s}_k = \frac{1}{k} \sum_{j=1}^{k} \phi_{\vartheta_\phi}(\mathbf{u}_j).$$

The attention domain consists only of these k context-task values; padding and the query task are excluded. The query representation $\mathbf{c}_{k+1}$ is supplied separately to the decoder, as in equation 1. This construction implements the aggregation already specified in the model definition.

C.3 Permutation Invariance and Observable Context Size

For every permutation σ of $\{1, \ldots, k\}$,

$$\frac{1}{k} \sum_{j=1}^{k} \phi_{\vartheta_\phi}(\mathbf{u}_{\sigma(j)}) = \frac{1}{k} \sum_{j=1}^{k} \phi_{\vartheta_\phi}(\mathbf{u}_j).$$

Thus the aggregation is invariant to the order of the context tasks. This symmetry agrees with their conditional exchangeability under Assumption 3.3; it does not impose invariance to permutations of observations within a task. Each sequence is first processed by the order-sensitive encoder E_ω with positional encoding.

Mean aggregation does not generally determine cardinality: repeating every task leaves the mean unchanged. The observable integer k therefore remains a separate input to $\rho_{\vartheta_\rho}(\mathbf{s}_k, \mathbf{c}_{k+1}, k)$. Section E.3.1 uses this input to combine all context-length-specific approximation slices into one shared decoder.

D Hölder Regularity of the Reduced Bayes Predictor

This section verifies Assumption 3.6 through primitive conditions on the reduced latent-type mixture. The key decomposition writes the reduced Bayes predictor as a mixture of type-specific query means, weighted by posterior type probabilities. We first show that stable posterior gating together with regular type-specific means yields Hölder regularity, and then give a high-probability version in which failure of regularity contributes only an additional bad-set probability term. Simple examples at the end illustrate conditions under which the type-specific means are regular.

For $x = (\mathbf{u}_{1:k}, \mathbf{c})$ and $x' = (\mathbf{u}'_{1:k}, \mathbf{c}')$, write

$$d_k(x, x') := \frac{1}{k} \sum_{j=1}^{k} \|(\mathbf{u}_j, \mathbf{c}) - (\mathbf{u}'_j, \mathbf{c}')\|_2^\alpha. \tag{15}$$

For $0 < \alpha \leq 1$, d_k is a metric on $\mathcal{U}_s^k \times \mathcal{C}_s$ and is invariant under simultaneous permutations of the context coordinates. Unless stated otherwise, all regularity constants and compact representation domains are uniform in (k, p, T).

D.1 Reduced Posterior Mixture Representation

For a generic task, let

$$R = \tau/T, \qquad \mathbf{C}^\star = E_T^\star(X_{1:T}), \qquad \mathbf{U}^\star = [(\mathbf{C}^\star)^\top, R]^\top.$$

Let $q_{i,T}^\star$ and $r_{i,T}^\star$ denote densities of $\mathbf{U}^\star \mid I = i$ and $\mathbf{C}^\star \mid I = i$, respectively, with respect to common dominating measures, and define

$$\mu_{i,T}^\star(\mathbf{c}) := \mathbb{E}[R \mid \mathbf{C}^\star = \mathbf{c}, I = i] \in [\kappa, 1 - \kappa].$$

At points with positive denominator, the reduced posterior weight of type i is

$$\pi_{i,k,T}^\star(\mathbf{u}_{1:k}, \mathbf{c}) = \frac{\alpha_i r_{i,T}^\star(\mathbf{c}) \prod_{j=1}^k q_{i,T}^\star(\mathbf{u}_j)}{\sum_{\ell=1}^K \alpha_\ell r_{\ell,T}^\star(\mathbf{c}) \prod_{j=1}^k q_{\ell,T}^\star(\mathbf{u}_j)}.$$

Because the deterministic representation preserves conditional task independence, conditioning on the latent type gives

$$B_{k,T}(\mathbf{u}_{1:k}, \mathbf{c}) = \sum_{i=1}^{K} \pi_{i,k,T}^\star(\mathbf{u}_{1:k}, \mathbf{c}) \mu_{i,T}^\star(\mathbf{c}). \tag{16}$$

We use bounded, context-permutation-invariant versions of $B_{k,T}$ on null sets. Thus, regularity of $B_{k,T}$ reduces to controlling the posterior weights and the type-specific query means.

D.2 Hölder Regularity on Stable Regions

Lemma D.1 (Posterior gating and mixture regularity on stable regions)**.** *Use the reduced mixture representation above and write $\pi_{k,T}^\star(x) = (\pi_{1,k,T}^\star(x), \ldots, \pi_{K,k,T}^\star(x))$. For a reference type $i^\star$, let $\mathcal{H}_{i^\star,k,T} \subseteq \mathcal{U}_s^k \times \mathcal{C}_s$ be context-permutation-invariant. On its corresponding coordinate projections, suppose that all $q_{i,T}^\star$ and $r_{i,T}^\star$ are finite and strictly positive and satisfy*

$$|\log q_{i,T}^\star(\mathbf{u}) - \log q_{i,T}^\star(\mathbf{u}')| \leq L_q \|\mathbf{u} - \mathbf{u}'\|_2^\alpha, \qquad |\log r_{i,T}^\star(\mathbf{c}) - \log r_{i,T}^\star(\mathbf{c}')| \leq L_r \|\mathbf{c} - \mathbf{c}'\|_2^\alpha.$$

Define

$$a_i(x) = \log \alpha_i + \log r_{i,T}^\star(\mathbf{c}) + \sum_{j=1}^{k} \log q_{i,T}^\star(\mathbf{u}_j),$$

and suppose that, for some $C_0, c_0 > 0$, all $x, x' \in \mathcal{H}_{i^\star,k,T}$ and $t \in [0,1]$ satisfy

$$1 - \mathrm{softmax}((1-t)a(x) + ta(x'))_{i^\star} \le C_0 e^{-c_0 k}.$$

Then there is a constant $L_\pi < \infty$, uniform in (k, p, T), such that

$$\|\pi^\star_{k,T}(x) - \pi^\star_{k,T}(x')\|_1 \le L_\pi \left\{ \frac{1}{k} \sum_{j=1}^{k} \|\mathbf{u}_j - \mathbf{u}'_j\|_2^\alpha + \|\mathbf{c} - \mathbf{c}'\|_2^\alpha \right\}. \tag{17}$$

If, in addition,

$$|\mu^\star_{i,T}(\mathbf{c}) - \mu^\star_{i,T}(\mathbf{c}')| \le L_\mu \|\mathbf{c} - \mathbf{c}'\|_2^\alpha$$

uniformly over the relevant query projections, then

$$|B_{k,T}(x) - B_{k,T}(x')| \le (2L_\pi + L_\mu) d_k(x, x'), \qquad x, x' \in \mathcal{H}_{i^\star,k,T}.$$

Proof. For $\pi = \mathrm{softmax}(a)$,

$$\sum_{r,s} \left| \frac{\partial \pi_r}{\partial a_s} \right| = 2 \left(1 - \sum_r \pi_r^2 \right) \le 4(1 - \pi_{i^\star}).$$

Integrating along the logit segment and using the assumed posterior concentration gives

$$\|\,\mathrm{softmax}(a(x)) - \mathrm{softmax}(a(x'))\|_1 \le 4C_0 e^{-c_0 k} \|a(x) - a(x')\|_\infty.$$

The log-density bounds imply

$$\|a(x) - a(x')\|_\infty \le L_q \sum_{j=1}^{k} \|\mathbf{u}_j - \mathbf{u}'_j\|_2^\alpha + L_r \|\mathbf{c} - \mathbf{c}'\|_2^\alpha.$$

Since $\sup_{k \ge 1} k e^{-c_0 k} < \infty$, this proves equation 17 with a uniform constant.

For the full mixture, add and subtract $\sum_i \pi^\star_{i,k,T}(x') \mu^\star_{i,T}(\mathbf{c})$ in equation 16. Boundedness and the Hölder condition on the means give

$$|B_{k,T}(x) - B_{k,T}(x')| \le \|\pi^\star_{k,T}(x) - \pi^\star_{k,T}(x')\|_1 + L_\mu \|\mathbf{c} - \mathbf{c}'\|_2^\alpha.$$

Both terms on the right are controlled by $d_k(x, x')$, which proves the claim. □

Posterior gating controls the mixture weights only; regularity of the type-specific means must be verified separately. Section D.4 gives simple sufficient conditions for this second component.

D.3 High-Probability Hölder Regularity

Corollary D.2 (Common good sets from type-specific stable regions)**.** *Suppose that Assumptions 3.3, 3.4, and 3.5 hold. For every i and $k \le p$, let $\mathcal{H}_{i,k,T}$ be measurable and context-permutation-invariant, with*

$$\Pr\{(\mathbf{U}^\star_{1:k}, \mathbf{C}^\star_{k+1}) \notin \mathcal{H}_{i,k,T} \mid I = i\} \le \delta_{i,k,T}.$$

Suppose that $B_{k,T}$ is L-Lipschitz for d_k on each region, with uniform L, and that for some $\eta_0 > 0$,

$$\inf_{x \in \mathcal{H}_{i,k,T},\, y \in \mathcal{H}_{j,k,T}} d_k(x, y) \ge \eta_0, \qquad i \ne j,$$

whenever both regions are nonempty. Then $\mathcal{H}_{k,T} := \bigcup_i \mathcal{H}_{i,k,T}$ satisfies

$$\Pr\{(\mathbf{U}^\star_{1:k}, \mathbf{C}^\star_{k+1}) \notin \mathcal{H}_{k,T}\} \le \sum_i \alpha_i \delta_{i,k,T},$$

and $B_{k,T}$ is d_k-Lipschitz on $\mathcal{H}_{k,T}$ with constant $\max\{L, \eta_0^{-1}\}$.

Proof. Conditional on $I = i$, the complement of the union is contained in $\mathcal{H}_{i,k,T}^c$, so averaging over the type prior gives the probability bound. Within one region use the assumed Hölder bound; across two distinct regions, boundedness gives

$$|B_{k,T}(x) - B_{k,T}(y)| \le 1 \le \eta_0^{-1} d_k(x, y).$$

□

For a nonempty common good set $\mathcal{H}_{k,T}$ on which $B_{k,T}$ is L-Lipschitz, define the bounded symmetric extension

$$F_{k,T}(x) := \mathrm{clip}_{[\kappa,1-\kappa]} \left\{ \frac{1}{k!} \sum_{\sigma \in S_k} \inf_{y \in \mathcal{H}_{k,T}} \left[B_{k,T}(y) + L d_k(\sigma x, y) \right] \right\}; \tag{18}$$

if $\mathcal{H}_{k,T}$ is empty, set $F_{k,T} \equiv \kappa$. This extension agrees with $B_{k,T}$ on the good set and preserves the common Hölder modulus and permutation symmetry.

Proposition D.3 (High-probability Hölder regularity)**.** *Suppose that Assumptions 3.3, 3.4, and 3.5 hold. For each $k \le p$, let $\mathcal{H}_{k,T}$ be a measurable, context-permutation-invariant set such that*

$$\Pr\{(\mathbf{U}_{1:k}^\star, \mathbf{C}_{k+1}^\star) \notin \mathcal{H}_{k,T}\} \le \delta_{H,k,T},$$

and suppose that the same bounded predictor $B_{k,T}$ is L-Lipschitz for d_k on $\mathcal{H}_{k,T}$, uniformly in (k, p, T). Assume further that the extension family $(F_{k,T})_{k=1}^p$ in equation 18, composed with $E_T^\star$, satisfies the shared realizability and finite-capacity requirement of Assumption 4.1, with any implementation error of order $m^{-\alpha/d_{\mathrm{meta}}}$ uniformly over k.

With $\overline{\delta}_{H,T} := p^{-1} \sum_{k=1}^p \delta_{H,k,T}$,

$$\inf_{\theta \in \Theta_{m,p,T}} \overline{R}_{\mathrm{BG}}(M_\theta) \le C \left\{ \varepsilon_{\mathrm{seq}}^2(p, T) + m^{-2\alpha/d_{\mathrm{meta}}} + \overline{\delta}_{H,T} \right\}. \tag{19}$$

If the shared ERM also satisfies the conditions of Lemma E.4, then

$$\mathbb{E}\overline{R}_{\mathrm{BG}}(M_{\widehat{\theta}}) \le C \left\{ \varepsilon_{\mathrm{seq}}^2(p, T) + m^{-2\alpha/d_{\mathrm{meta}}} + \overline{\delta}_{H,T} + \sqrt{\frac{V_{m,p,T} \log(eN)}{N}} \right\}.$$

Here C is independent of (N, m, p, T).

The proof is given in Section E.6.1. Thus, the global Hölder condition in Assumption 3.6 may be replaced by regularity on high-probability common regions, at the cost of the additional term $\overline{\delta}_{H,T}$.

D.4 Examples of Regular Type-Specific Means

The following examples verify only the regularity of $\mu_{i,T}^\star$ for the reduced query representation $\mathbf{Z} = \mathbf{C}^\star$; posterior gating is handled separately above.

Discrete linear location model. Let

$$\mathcal{G}_T = \{r/T : \lceil \kappa T \rceil \le r \le \lfloor (1-\kappa)T \rfloor\}, \qquad R \in \mathcal{G}_T.$$

Conditional on $I = i$ and $A = (a, b)$, suppose that

$$\Pr(R = r \mid \mathbf{Z} = \mathbf{z}, A = (a, b), I = i) \propto \exp\left[-\frac{\{r - (a^\top \mathbf{z} + b)\}^2}{2\sigma_i^2} \right], \qquad r \in \mathcal{G}_T,$$

where $\|a\|_2 \le A_0$ and $0 < \underline{\sigma} \le \sigma_i \le \overline{\sigma} < \infty$. If $\mu_{i,A}$ denotes the corresponding conditional mean, differentiation with respect to the location parameter gives $\mathrm{Var}(R \mid \mathbf{z}, A, I = i)/\sigma_i^2$, which is at most $(1 - 2\kappa)^2/(4\underline{\sigma}^2)$. Hence

$$|\mu_{i,A}(\mathbf{z}) - \mu_{i,A}(\mathbf{z}')| \le \frac{(1-2\kappa)^2 A_0}{4\underline{\sigma}^2} \|\mathbf{z} - \mathbf{z}'\|_2.$$

If $A \perp \mathbf{Z} \mid I = i$, integration over $A \mid I = i$ preserves the same uniform Lipschitz bound for $\mu_{i,T}^\star$.

Finite Hölder expansions. Suppose that $g_a(\mathbf{z}) = \sum_{r=1}^{R_0} a_r g_r(\mathbf{z})$, where $\|a\|_1 \le A_0$ and each g_r is uniformly α-Hölder with constant L_g. If

$$\mathbb{E}[R \mid \mathbf{Z} = \mathbf{z}, A = a, I = i] = h_i(g_a(\mathbf{z})),$$

where h_i is uniformly L_h-Lipschitz, then

$$|h_i(g_a(\mathbf{z})) - h_i(g_a(\mathbf{z}'))| \le L_h A_0 L_g \|\mathbf{z} - \mathbf{z}'\|_2^{\alpha}.$$

Under $A \perp \mathbf{Z} \mid I = i$, the same bound holds after mixing over A. A finite convex dictionary is the special case $a \in \Delta^{R_0-1}$. If the mixing law of A depends on $\mathbf{Z}$, its variation must be controlled separately.

E Bayes-Gap Analysis

This section proves the Bayes-Gap bound by separating three sources of error. A fixed-dimensional sequence representation first incurs the information-loss term $\varepsilon^2_{\text{seq}}(p, T)$; Hölder approximation of the reduced Bayes targets contributes $m^{-2\alpha/d_{\text{meta}}}$; and block-level ERM contributes the pretraining estimation term $\sqrt{V_{m,p,T} \log(eN)/N}$. We establish these ingredients in this order and then combine them in the main bound. High-probability regularity, localized fast rates, and polynomial-capacity rates are collected at the end as optional refinements.

E.1 Bayes Optimality and Risk Decomposition

E.1.1 Bayes Optimality and Posterior Representation

Fix k and let $\mathcal{S}_k := \sigma(\mathcal{C}_k, X_{1:T}^{(k+1)})$. For any bounded $\mathcal{S}_k$-measurable prediction a, conditional squared-error orthogonality gives

$$\mathbb{E}\left[\frac{(\tau_{k+1} - a)^2}{T^2} \,\middle|\, \mathcal{S}_k\right] = \frac{\mathrm{Var}(\tau_{k+1} \mid \mathcal{S}_k)}{T^2} + \frac{(a - \mathbb{E}[\tau_{k+1} \mid \mathcal{S}_k])^2}{T^2}. \tag{20}$$

Hence the Bayes predictor is the conditional mean, and in normalized coordinates $R = \tau_{k+1}/T$,

$$\widetilde{M}_{\text{Bayes}}(\mathcal{P}^k) = \mathbb{E}[R \mid \mathcal{S}_k] = T^{-1} M_{\text{Bayes}}(\mathcal{P}^k).$$

Conditioning additionally on the finite latent type gives, for any measurable set D,

$$\Pr(\tau_{k+1} \in D \mid \mathcal{S}_k) = \sum_{i=1}^{K} \pi_i(\mathcal{P}^k) \Pr(\tau_{k+1} \in D \mid \mathcal{S}_k, I = i).$$

By Assumption 3.3, the last conditional probability equals $\Pr(\tau_{k+1} \in D \mid X_{1:T}^{(k+1)}, I = i)$ almost surely. This yields the posterior-mixture representation in Section 3.4; no unconditional independence of query and context tasks is used.

E.1.2 Risk Decomposition

Proof of the risk decomposition in Section 4. Applying equation 20 with $a = M(\mathcal{P}^k)$ and taking expectations gives

$$\begin{aligned} R_k(M) &= \mathbb{E}_{\mathcal{P}^k} \frac{\mathrm{Var}(\tau_{k+1} \mid \mathcal{P}^k)}{T^2} + \mathbb{E}_{\mathcal{P}^k} \frac{\{M(\mathcal{P}^k) - M_{\text{Bayes}}(\mathcal{P}^k)\}^2}{T^2} \\ &= R_{\text{PV},k} + R_{\text{BG},k}(M). \end{aligned}$$

Setting $M = M_{\text{Bayes}}$ identifies the Bayes risk with $R_{\text{PV},k}$, and averaging over k gives the corresponding decomposition for $\overline{R}$.

For the trained predictor, condition first on the independent training blocks and then average over them:

$$\begin{aligned} \mathbb{E}R_k(M_{\widehat{\theta}}) &= R_{\text{PV},k} + \mathbb{E}R_{\text{BG},k}(M_{\widehat{\theta}}), \\ \mathbb{E}\overline{R}(M_{\widehat{\theta}}) &= \overline{R}_{\text{PV}} + \mathbb{E}\overline{R}_{\text{BG}}(M_{\widehat{\theta}}). \end{aligned}$$

□

E.2 Sequence Representation and Information Loss

This subsection gives a concrete sufficient condition for the sequence-information term $\varepsilon_{\mathrm{seq}}(p,T)$ to vanish. A pilot statistic recoverable from the fixed-dimensional representation transfers its localization accuracy directly to the reduced Bayes predictor.

Proposition E.1 (Pilot control of sequence information loss)**.** *Under Assumptions 3.3, 3.4, and 3.5, suppose that a measurable map $g_T : \mathcal{Z} \to [\kappa, 1-\kappa]$ recovers a pilot statistic $S_T(X_{1:T}) = g_T(E_T^\star(X_{1:T}))$. Then, for every $k \le p$,*

$$\mathbb{E}\left[\left\{\widetilde{M}_{\mathrm{Bayes}}(\mathcal{P}^k) - B_{k,T}(\mathbf{U}_{1:k}^\star, \mathbf{C}_{k+1}^\star)\right\}^2\right] \le \mathbb{E}\left[\{r_{k+1} - S_T(X_{1:T}^{(k+1)})\}^2\right]. \tag{21}$$

Consequently, if the generic-query pilot mean squared error is e_T, then $\varepsilon_{\mathrm{seq}}^2(p,T) \le e_T$.

Proof. Let $\mathcal{S}_k = \sigma(\mathcal{P}^k)$ and $\mathcal{S}_k^{\mathrm{red}} = \sigma(\mathbf{U}_{1:k}^\star, \mathbf{C}_{k+1}^\star)$. Since $\mathcal{S}_k^{\mathrm{red}} \subseteq \mathcal{S}_k$,

$$B_{k,T} = \mathbb{E}[r_{k+1} \mid \mathcal{S}_k^{\mathrm{red}}] = \mathbb{E}[\widetilde{M}_{\mathrm{Bayes}}(\mathcal{P}^k) \mid \mathcal{S}_k^{\mathrm{red}}].$$

Conditional-expectation orthogonality yields

$$\mathbb{E}(r_{k+1} - B_{k,T})^2 = \mathbb{E}(r_{k+1} - \widetilde{M}_{\mathrm{Bayes}})^2 + \mathbb{E}(\widetilde{M}_{\mathrm{Bayes}} - B_{k,T})^2.$$

Because the pilot is $\mathcal{S}_k^{\mathrm{red}}$-measurable, the left-hand side is no larger than $\mathbb{E}\{r_{k+1} - S_T(X_{1:T}^{(k+1)})\}^2$. Dropping the first nonnegative term proves equation 21. □

Mean-shift example. Consider $X_t = \mu_1 + \Delta \mathbf{1}\{t > \tau\} + \varepsilon_t$ with an interior change point. If the clipped least-squares pilot

$$S_T(X) = \mathrm{clip}_{[\kappa,1-\kappa]}(\widehat{\tau}_{\mathrm{LS}}(X)/T)$$

is recoverable from an admissible fixed-dimensional representation and satisfies $\mathbb{E}(S_T - R)^2 \to 0$, then Proposition E.1 gives $\varepsilon_{\mathrm{seq}}^2(p,T) \to 0$; classical consistency results for mean-shift localization include Bai (1994). If p grows, condition (C4) requires the stronger rate $p\mathbb{E}(S_T - R)^2 \to 0$. Recoverability from $E_T^\star$ is an explicit representation requirement and does not follow from statistical consistency of the pilot alone.

E.3 Shared Predictor Approximation

E.3.1 Simultaneous Approximation Across Context Lengths

Lemma E.2 (Simultaneous Bayes approximation with the prescribed architecture)**.** *Suppose that Assumptions 3.3, 3.4, 3.5, and 3.6 hold. Use bounded, context-permutation-invariant versions of $B_{k,T}$ on the fixed compact domains $\mathcal{U}_s^k \times \mathcal{C}_s$, and let $d_{\mathrm{meta}} = d_s + 1$.*

Fix any $\delta \in (0,1]$. For every sufficiently large integer m, there are finite feedforward ReLU networks $g_m : \mathbb{R}^{d_{\mathrm{meta}}} \to \mathbb{R}^m$ and $h_{m,p} : \mathbb{R}^{m+d_s+1} \to \mathbb{R}$ such that, with

$$\phi_m = \mathrm{Renorm}_\delta \circ g_m, \qquad \rho_{m,p}(\mathbf{s}, \mathbf{c}, t) = \mathrm{clip}_{[\kappa,1-\kappa]}\{h_{m,p}([\mathbf{s}^\top, \mathbf{c}^\top, t]^\top)\},$$

the encoder takes values in Δ^{m-1} and

$$\max_{1 \le k \le p} \sup_{\mathbf{u}_{1:k} \in \mathcal{U}_s^k,\, \mathbf{c} \in \mathcal{C}_s} \left| B_{k,T}(\mathbf{u}_{1:k}, \mathbf{c}) - \rho_{m,p}\left(\frac{1}{k}\sum_{j=1}^k \phi_m(\mathbf{u}_j), \mathbf{c}, k\right)\right| \le C m^{-\alpha/d_{\mathrm{meta}}}. \tag{22}$$

The same encoder and decoder serve every $k = 1, \dots, p$ and do not observe I.

If Assumption 4.1 holds for this construction, including its composition with $E_T^\star$, there is one shared $\theta_{m,p,T} \in \Theta_{m,p,T}$ satisfying

$$\overline{R}_{\mathrm{BG}}(M_{\theta_{m,p,T}}) \le 2\varepsilon_{\mathrm{seq}}^2(p,T) + C m^{-2\alpha/d_{\mathrm{meta}}}. \tag{23}$$

In particular,

$$\inf_{\theta \in \Theta_{m,p,T}} \overline{R}_{\mathrm{BG}}(M_\theta) \le C\{\varepsilon_{\mathrm{seq}}^2(p,T) + m^{-2\alpha/d_{\mathrm{meta}}}\}. \tag{24}$$

Here and below, C is independent of (N, m, p, T). Enlarging C extends the risk bounds to the finitely many admissible widths below the explicit grid threshold, and all components of the construction are counted in $V_{m,p,T}$.

Proof. We first construct a shared approximation to the reduced targets and then compose it with the sequence representation.

Step 1: A shared grid and exact ReLU tents. Write $d = d_{\mathrm{meta}}$ within this proof. Compactness gives an $h/2$-net in $\mathcal{U}_s$ for the ℓ_∞ metric with at most $C_{\mathcal{U}} h^{-d}$ points. For all sufficiently large m, choose $h \asymp m^{-1/d} \in (0, 1)$ with at most m net points, and add distinct points of $\mathcal{U}_s$ if needed to obtain $\mathbf{r}_1, \dots, \mathbf{r}_m$. All representative points lie in the domain of $B_{k,T}$.

Define

$$\chi_a(\mathbf{u}) = \sigma\left(1 - \frac{\|\mathbf{u} - \mathbf{r}_a\|_\infty}{h}\right), \quad S(\mathbf{u}) = \sum_{a=1}^{m} \chi_a(\mathbf{u}), \quad \psi_a(\mathbf{u}) = \frac{\chi_a(\mathbf{u})}{S(\mathbf{u})}.$$

The net property gives $S(\mathbf{u}) \geq 1/2$. Hence $\psi(\mathbf{u}) \in \Delta^{m-1}$, and $\psi_a(\mathbf{u}) > 0$ implies $\|\mathbf{u} - \mathbf{r}_a\|_2 \leq \sqrt{d}\, h$. The normalized histogram ψ is used only analytically; each unnormalized tent χ_a is exactly ReLU-realizable using absolute values and finite maxima.

Step 2: The actual positive-smoothing encoder. Take $g_{m,a} = h^{-\alpha}\chi_a$. This is a nonnegative ReLU output, so the encoder prescribed in Definition 3.2 satisfies

$$\phi_m(\mathbf{u}) = \mathrm{Renorm}_\delta(g_m(\mathbf{u})) = (1 - \epsilon(\mathbf{u}))\psi(\mathbf{u}) + \epsilon(\mathbf{u})\frac{\mathbf{1}_m}{m}, \qquad \epsilon(\mathbf{u}) = \frac{\delta}{h^{-\alpha} S(\mathbf{u}) + \delta} \leq 2h^\alpha. \tag{25}$$

Thus $\phi_m(\mathbf{u}) \in \Delta^{m-1}$, while the smoothing contributes only $O(h^\alpha)$ mass uniformly over m, k, and $\delta \in (0, 1]$; the same encoder is used for every context length.

Step 3: Decoder slices on histograms. For $\mathbf{s}, \mathbf{t} \in \Delta^{m-1}$, let

$$W_\alpha(\mathbf{s}, \mathbf{t}) := \min_{\Gamma \geq 0:\, \Gamma \mathbf{1} = \mathbf{s},\, \Gamma^\top \mathbf{1} = \mathbf{t}} \sum_{a,b} \Gamma_{ab} \|\mathbf{r}_a - \mathbf{r}_b\|_2^\alpha.$$

The ground cost is a metric because $0 < \alpha \leq 1$; W_α denotes the displayed transport cost, without an additional $1/\alpha$ power. Set

$$\Delta_k = \{\mathbf{n}/k : \mathbf{n} \in \{0, \dots, k\}^m,\ \sum_a n_a = k\}.$$

For $\mathbf{v} = \mathbf{n}/k \in \Delta_k$, define $\rho_{k,\mathbf{c}}(\mathbf{v})$ to be $B_{k,T}$ evaluated on the multiset containing n_a copies of $\mathbf{r}_a$ and query representation $\mathbf{c}$. Permutation invariance makes this definition unambiguous.

For two such multisets, the transport problem can be expanded into a matching problem on k equally weighted points. A minimizing coupling can be chosen as a matching, by the extreme-point characterization of doubly stochastic matrices. Applying equation 5 along that matching gives

$$|\rho_{k,\mathbf{c}}(\mathbf{v}) - \rho_{k,\mathbf{c}}(\mathbf{w})| \leq L W_\alpha(\mathbf{v}, \mathbf{w}).$$

Extend this slice to the full simplex by

$$\bar{\rho}_k(\mathbf{s}, \mathbf{c}) := \inf_{\mathbf{v} \in \Delta_k} \{\rho_{k,\mathbf{c}}(\mathbf{v}) + L W_\alpha(\mathbf{s}, \mathbf{v})\}.$$

It agrees with $\rho_{k,\mathbf{c}}$ on Δ_k and is L-Lipschitz in its first argument for W_α. Moreover,

$$|\bar{\rho}_k(\mathbf{s}, \mathbf{c}) - \bar{\rho}_k(\mathbf{s}, \mathbf{c}')| \leq L\|\mathbf{c} - \mathbf{c}'\|_2^\alpha,$$

because every term in the infimum has this query-regularity bound. This establishes joint continuity.

Step 4: Transport control for the prescribed encoder. For each $\mathbf{u}_j$, choose a net point $\mathbf{r}_{a_j}$ within $h/2$ in ℓ_∞, and write

$$\mathbf{v}_k = \frac{1}{k}\sum_{j=1}^{k} \mathbf{e}_{a_j}, \qquad \mathbf{s}_k = \frac{1}{k}\sum_{j=1}^{k} \phi_m(\mathbf{u}_j).$$

If $\psi_a(\mathbf{u}_j) > 0$, then $\|\mathbf{r}_a - \mathbf{r}_{a_j}\|_2 \le (3\sqrt{d}/2)h$. Couple the local mass in equation 25 to $\mathbf{r}_{a_j}$, and send its remaining uniform mass to the same point. With $D = \operatorname{diam}(\mathcal{U}_s)$, averaging these couplings gives

$$W_\alpha(\mathbf{s}_k, \mathbf{v}_k) \le \frac{1}{k}\sum_{j=1}^{k}\{(3\sqrt{d}/2)^\alpha h^\alpha + D^\alpha \epsilon(\mathbf{u}_j)\} \le Ch^\alpha.$$

Also, equation 5 gives

$$|B_{k,T}(\mathbf{u}_{1:k}, \mathbf{c}) - \rho_{k,\mathbf{c}}(\mathbf{v}_k)| \le CLh^\alpha.$$

Combining these bounds yields

$$|B_{k,T}(\mathbf{u}_{1:k}, \mathbf{c}) - \bar{\rho}_k(\mathbf{s}_k, \mathbf{c})| \le CLh^\alpha, \tag{26}$$

uniformly over $k \le p$ and the reduced inputs.

Step 5: ReLU slices and clipping. Let $F_k = \operatorname{clip}_{[\kappa,1-\kappa]} \bar{\rho}_k$ on $\mathcal{K}_m = \Delta^{m-1} \times \mathcal{C}_s$. Clipping preserves equation 26. Since $W_\alpha(\mathbf{s}, \mathbf{t}) \le (D^\alpha/2)\|\mathbf{s} - \mathbf{t}\|_1$, the functions F_k are uniformly α-Hölder for the ambient ℓ_1 distance on $\mathcal{K}_m$, with some fixed constant $H > 0$. The ℓ_1 diameter of this product set is bounded independently of m.

For completeness, a finite ReLU approximation follows directly from this regularity. Take an η-net $\{z_b\}$ of $\mathcal{K}_m$ in ℓ_1 and define

$$f_{m,k}(z) = \min_b\{F_k(z_b) + H\eta^{\alpha-1}\|z - z_b\|_1\}.$$

This is a finite ReLU network because absolute values and finite minima are ReLU-representable. For any term whose center has distance at least η, its linear distance penalty dominates $H\|z - z_b\|_1^\alpha$; a nearer center can undershoot by at most $H\eta^\alpha$. A center within η gives an overshoot of at most $2H\eta^\alpha$. Thus

$$F_k(z) - H\eta^\alpha \le f_{m,k}(z) \le F_k(z) + 2H\eta^\alpha \quad (z \in \mathcal{K}_m).$$

Choose η to make this error at most h^α, and set $b_{m,k} = \operatorname{clip}_{[\kappa,1-\kappa]} f_{m,k}$. Then $b_{m,k}$ is a ReLU network with the valid output range, and

$$\sup_{\mathbf{u}_{1:k},\mathbf{c}} |B_{k,T}(\mathbf{u}_{1:k}, \mathbf{c}) - b_{m,k}(\mathbf{s}_k, \mathbf{c})| \le Ch^\alpha. \tag{27}$$

The finite net may require many decoder parameters. No polynomial bound on that number is used.

Step 6: One count-aware decoder. First normalize each clipped slice:

$$a_{m,q}(\mathbf{s}, \mathbf{c}) := \frac{b_{m,q}(\mathbf{s}, \mathbf{c}) - \kappa}{1 - 2\kappa} \in [0, 1].$$

Define ReLU gates

$$G_q(t) = \sigma(1 - \sigma(t - q) - \sigma(q - t)), \qquad S_0(a, g) = a - \sigma(a - g).$$

For integer $k \in \{1, \dots, p\}$, $G_q(k) = \mathbf{1}\{q = k\}$. For $a \in [0, 1]$, $S_0(a, 0) = 0$ and $S_0(a, 1) = a$. Consequently the single decoder

$$\rho_{m,p}(\mathbf{s}, \mathbf{c}, t) = \operatorname{clip}_{[\kappa,1-\kappa]}\left\{\kappa + (1 - 2\kappa)\sum_{q=1}^{p} S_0(a_{m,q}(\mathbf{s}, \mathbf{c}), G_q(t))\right\} \tag{28}$$

agrees exactly with $b_{m,k}$ when $t = k$. Its interior is a finite ReLU network; its outer clipping is precisely the operation in Definition 3.2. The outer clip also enforces the output range at noninteger inputs. No continuity of the target as a function of k is assumed. All slices are stored in the same parameter vector, and the only selection input is the observed count. Equations equation 27 and equation 28 prove equation 22.

Step 7: Composition and Bayes-Gap risk. Compose the shared modules with $E_T^\star$. Assumption 4.1 places this construction, or a realization with the same uniform approximation order, in $\Theta_{m,p,T}$. Neither the grid nor the runtime selector observes I. For the resulting normalized output $\widehat{B}_{m,k}$,

$$\begin{aligned} R_{\mathrm{BG},k}(M_{\theta_{m,p,T}}) &= \mathbb{E}(\widetilde{M}_{\mathrm{Bayes}} - \widehat{B}_{m,k})^2 \\ &\le 2\mathbb{E}(\widetilde{M}_{\mathrm{Bayes}} - B_{k,T})^2 + 2\mathbb{E}(B_{k,T} - \widehat{B}_{m,k})^2 \\ &\le 2\varepsilon_{\mathrm{seq}}^2(p, T) + Cm^{-2\alpha/d_{\mathrm{meta}}}. \end{aligned}$$

The equality uses $M_\theta = T\widetilde{M}_\theta$ and $M_{\mathrm{Bayes}} = T\widetilde{M}_{\mathrm{Bayes}}$. Averaging over k gives equation 23, and taking the infimum gives equation 24. For the finitely many admissible widths below the grid threshold, boundedness gives the same risk order after enlarging C. Hence the conclusion holds for every admissible m, with the full construction included in the class defining $V_{m,p,T}$. □

E.4 Global Generalization for Shared Pretraining

An observation is a complete labeled block $Z_{1:p+1}$.

$$\Lambda_\theta(Z_{1:p+1}) := \frac{1}{p}\sum_{k=1}^{p} \ell_T\big(\tau_{k+1}, M_\theta(\mathcal{P}^k)\big)\,,$$

and define

$$\mathcal{L}_{m,p,T} := \{\Lambda_\theta : \theta \in \Theta_{m,p,T}\}, \qquad V_{m,p,T} := \mathrm{Pdim}(\mathcal{L}_{m,p,T}).$$

The N blocks are independent, whereas the p prefix losses within a block may be dependent. All empirical-process classes below are assumed measurable and separable, and empirical minimizers are measurable.

Lemma E.3 (Global ERM bound for a bounded class)**.** *Let $\mathcal{G}$ be a measurable, separable class with $|g| \le B$ and $1 \le \mathrm{Pdim}(\mathcal{G}) \le V \le N$. If $\widehat{g}$ is an empirical minimizer based on N independent observations, then*

$$\mathbb{E}P\widehat{g} \le \inf_{g\in\mathcal{G}} Pg + C_B\sqrt{\frac{V\log(eN)}{N}}, \tag{29}$$

where C_B depends only on B.

Proof. For any deterministic $g_0 \in \mathcal{G}$, the ERM inequality gives

$$P\widehat{g} - Pg_0 \le 2\sup_{g\in\mathcal{G}} |(P_N - P)g|.$$

Symmetrization and the standard pseudo-dimension bound for the Rademacher complexity of a bounded class yield

$$\mathbb{E}\sup_{g\in\mathcal{G}} |(P_N - P)g| \le C_B\sqrt{\frac{V\log(eN)}{N}}.$$

Taking Pg_0 down to the infimum proves equation 29. For block losses the sample size is N, not Np. □

Define the fixed Bayes block loss and the block excess loss by

$$\Lambda_{\mathrm{Bayes}}(Z_{1:p+1}) := \frac{1}{p}\sum_{k=1}^{p} \ell_T(\tau_{k+1}, M_{\mathrm{Bayes}}(\mathcal{P}^k)), \qquad L_\theta := \Lambda_\theta - \Lambda_{\mathrm{Bayes}}.$$

Here P denotes expectation over a new complete labeled block, and P_N is the empirical measure of the N training blocks.

Lemma E.4 (Global oracle inequality for shared-prefix ERM)**.** *Suppose the shared ERM is measurable, is based on N independent complete blocks, and $1 \le V_{m,p,T} \le N$. Then*

$$\mathbb{E}\overline{R}_{\mathrm{BG}}(M_{\widehat{\theta}}) \le \inf_{\theta\in\Theta_{m,p,T}} \overline{R}_{\mathrm{BG}}(M_\theta) + C\sqrt{\frac{V_{m,p,T}\log(eN)}{N}}. \tag{30}$$

Proof. Subtracting the fixed block loss Λ_{Bayes} leaves the empirical minimizer unchanged. The excess-loss class has envelope one, pseudo-dimension $V_{m,p,T}$, and population mean $PL_\theta = \overline{R}_{\mathrm{BG}}(M_\theta)$. Lemma E.3 therefore gives equation 30 directly; no independence across prefixes is required. □

E.5 BAYES-GAP BOUND

The main bound follows by combining the shared approximation error with block-level generalization. Its three terms correspond to sequence information loss, approximation of the reduced Bayes targets, and estimation from N independent pretraining blocks.

Theorem E.5 (Average Bayes Gap bound for the shared ERM)**.** *Consider the prompt-generation process in Definition 3.1 and the model class in Definition 3.2. Suppose that Assumptions 3.3–3.6 and Assumption 4.1 hold, and that* $1 \le V_{m,p,T} \le N$*. Then the shared ERM defined in equation 3 satisfies*

$$\mathbb{E}\overline{R}_{\mathrm{BG}}(M_{\widehat{\theta}}) \le C\left\{\varepsilon^2_{\mathrm{seq}}(p,T) + m^{-2\alpha/d_{\mathrm{meta}}} + \sqrt{\frac{V_{m,p,T}\log(eN)}{N}}\right\}, \tag{31}$$

where $d_{\mathrm{meta}} = d_s + 1$ *and* $C > 0$ *is independent of* (N, m, p, T)*.*

Proof of Theorem E.5. Lemma E.2, including its extension to every admissible width, gives

$$\inf_{\theta\in\Theta_{m,p,T}} \overline{R}_{\mathrm{BG}}(M_\theta) \le C_1\{\varepsilon^2_{\mathrm{seq}}(p,T) + m^{-2\alpha/d_{\mathrm{meta}}}\}.$$

The global inequality equation 30 in Lemma E.4, applied to the N independent complete training blocks, gives

$$\mathbb{E}\overline{R}_{\mathrm{BG}}(M_{\widehat{\theta}}) \le \inf_{\theta\in\Theta_{m,p,T}} \overline{R}_{\mathrm{BG}}(M_\theta) + C_2\sqrt{\frac{V_{m,p,T}\log(eN)}{N}}.$$

Combining the two displays proves equation 31. The three terms are, respectively, sequence information loss, approximation of the reduced targets, and estimation from independent training blocks. Both constants are independent of (N, m, p, T). Only the global oracle inequality is used. □

Specified context length.

Proof of Theorem 4.2. For every realization of the training data, each population Bayes Gap is nonnegative. Thus, for a specified k,

$$R_{\mathrm{BG},k}(M_{\widehat{\theta}}) \le \sum_{s=1}^{p} R_{\mathrm{BG},s}(M_{\widehat{\theta}}) = p\overline{R}_{\mathrm{BG}}(M_{\widehat{\theta}}).$$

Taking expectations over the training data and applying Theorem E.5 proves equation 7. The factor p converts the average-prefix bound into a guarantee at the specified context length; no independence across prefixes is required. □

E.6 REFINEMENTS AND EXPLICIT RATES

The results below strengthen or relax the main Bayes-Gap analysis but are not needed for Theorem E.5: the first replaces global Hölder regularity by a high-probability version, the second gives a localized fast rate, and the third translates abstract capacity into explicit width choices.

E.6.1 HIGH-PROBABILITY HÖLDER REGULARITY

Proof of Proposition D.3. **Step 1: One bounded symmetric extension at each length.** The function d_k in equation 15 is a metric: the triangle inequality follows from the Euclidean triangle inequality and $(a+b)^\alpha \le a^\alpha + b^\alpha$. It is invariant under simultaneous permutations of the context coordinates.

For a nonempty $\mathcal{H}_{k,T}$, the infimum in equation 18 is the McShane extension of the restriction of $B_{k,T}$. It is finite on the bounded domain, equals $B_{k,T}$ on the good set, and is L-Lipschitz for d_k. These facts do not require the good set to be closed. Averaging its permuted copies preserves its Lipschitz constant. Since both the set and the chosen version of $B_{k,T}$ are permutation-invariant, the average still agrees with $B_{k,T}$ there. Clipping to $[\kappa, 1-\kappa]$ preserves these properties. The extension

is continuous, hence measurable. If the good set is empty, the prescribed constant extension has the same required bounds; its complement probability is one.

Step 2: A shared realization of the extensions. The function-level construction in Lemma E.2 uses only boundedness, permutation invariance, and the common Hölder modulus of its reduced targets. Apply that construction to the entire family $(F_{k,T})_{k=1}^{p}$. Its grid and ReLU-plus-Renorm_{δ} encoder are shared, while equation 28 combines all clipped decoder slices. The proposition's explicit extension-realizability requirement puts the resulting composition in the same model class, with the same approximation order. Write its outputs as $\widehat{B}_{m,k}$.

On $\mathcal{H}_{k,T}$, $F_{k,T} = B_{k,T}$ and the error is at most $Cm^{-\alpha/d_{\mathrm{meta}}}$. On the complement, both $B_{k,T}$ and $\widehat{B}_{m,k}$ lie in $[\kappa, 1-\kappa]$. Therefore

$$\mathbb{E}(B_{k,T} - \widehat{B}_{m,k})^2 \le Cm^{-2\alpha/d_{\mathrm{meta}}} + (1-2\kappa)^2\delta_{H,k,T}.$$

Combining this with equation 4 gives

$$R_{\mathrm{BG},k}(M_\theta) \le 2\varepsilon_{\mathrm{seq}}^2(p,T) + Cm^{-2\alpha/d_{\mathrm{meta}}} + 2\delta_{H,k,T}.$$

The same θ applies to every k. Averaging the display proves equation 19; no union bound over prefix bad events is needed.

Step 3: Shared ERM. Apply Lemma E.4 to this common comparison class. Substituting equation 19 gives the stated ERM bound. This use of the global oracle inequality requires no global regularity of the unextended $B_{k,T}$ and no localized-complexity assumption. □

E.6.2 Localized Fast-Rate Bound

Bernstein condition for block excess losses.

Lemma E.6 (A Bernstein condition for block excess losses)**.** *Use the prompt distribution and bounded normalized predictors of Definition 3.2, under Assumption 3.4. Then, for every $\theta \in \Theta_{m,p,T}$,*

$$|L_\theta| \le 1, \qquad PL_\theta = \overline{R}_{\mathrm{BG}}(M_\theta) \ge 0, \qquad PL_\theta^2 \le 10PL_\theta. \tag{32}$$

No independence between prefixes within a block is required.

Proof. Both Λ_θ and Λ_{Bayes} lie in $[0,1]$, proving the envelope bound. Set

$$\Delta_{\theta,k} = \frac{M_\theta(\mathcal{P}^k) - M_{\mathrm{Bayes}}(\mathcal{P}^k)}{T}, \qquad \xi_k = \frac{\tau_{k+1} - M_{\mathrm{Bayes}}(\mathcal{P}^k)}{T}.$$

Then

$$L_\theta = \frac{1}{p}\sum_{k=1}^{p}(\Delta_{\theta,k}^2 - 2\xi_k\Delta_{\theta,k}).$$

For every k, $\Delta_{\theta,k}$ is $\mathcal{P}^k$-measurable and $\mathbb{E}[\xi_k \mid \mathcal{P}^k] = 0$. Taking expectations therefore gives $PL_\theta = p^{-1}\sum_k \mathbb{E}\Delta_{\theta,k}^2 = \overline{R}_{\mathrm{BG}}(M_\theta)$. Since $|\Delta_{\theta,k}|, |\xi_k| \le 1$, Jensen's inequality gives

$$L_\theta^2 \le \frac{1}{p}\sum_{k=1}^{p}(\Delta_{\theta,k}^2 - 2\xi_k\Delta_{\theta,k})^2 \le \frac{9}{p}\sum_{k=1}^{p}\Delta_{\theta,k}^2.$$

Taking expectations proves the stated bound with constant 10. Jensen's inequality controls the entire within-block average directly; no prefix independence was used. □

Localized empirical-process condition. Write

$$\mathcal{G}_{m,p,T}^{\mathrm{exc}} := \{L_\theta : \theta \in \Theta_{m,p,T}\}.$$

The offset Λ_{Bayes} is a fixed function of the block, independent of θ and the training sample. Subtracting it preserves pseudo-dimension: on any proposed shattered sample, replace each threshold t_n by $t_n + \Lambda_{\mathrm{Bayes}}(z_n)$. The reverse translation gives the converse. Thus

$$\mathrm{Pdim}(\mathcal{G}_{m,p,T}^{\mathrm{exc}}) = \mathrm{Pdim}(\mathcal{L}) = V_{m,p,T}.$$

This identity concerns block losses, not the pseudo-dimension of the predictor class before forming losses.

Assumption E.7 (Additional localized empirical-process control)**.** Let $a_N := V_{m,p,T}\log(eN)/N$, where P_N is based on N independent complete blocks with common law P. Assume that the localized excess-loss classes are separable and their empirical-process suprema are measurable. For $0 < r \le 1$, define

$$D_r := \sup_{g\in\{0\}\cup\{g\in\mathcal{G}^{\mathrm{exc}}_{m,p,T}:Pg\le r\}} |(P_N - P)g|.$$

In particular, $D_r = 0$ if the localized excess-loss class is empty. There is a constant $C_{\mathrm{loc}} < \infty$, independent of (N, m, p, T), such that, whenever $1 \le V_{m,p,T} \le N$,

$$\mathbb{E}D_r \le C_{\mathrm{loc}}\{\sqrt{ra_N} + a_N\}, \qquad 0 < r \le 1. \tag{33}$$

This is an additional condition; it is not inferred solely from finite pseudo-dimension.

Lemma E.8 (Localized oracle inequality for shared-prefix ERM)**.** *Under the conditions of Lemma E.4, if Assumption E.7 also holds, then*

$$\mathbb{E}\overline{R}_{\mathrm{BG}}(M_{\widehat{\theta}}) \le C\left\{\inf_{\theta\in\Theta_{m,p,T}} \overline{R}_{\mathrm{BG}}(M_\theta) + \frac{V_{m,p,T}\log(eN)}{N}\right\}. \tag{34}$$

The constant depends only on the fixed envelope, Bernstein, and localized-complexity constants.

Proof. Write $a = a_N$. If $a \ge 1$, equation 34 follows from $0 \le \overline{R}_{\mathrm{BG}} \le 1$. We therefore take $0 < a < 1$.

Step 1: Concentration at a deterministic radius. For $r \in (0, 1]$, apply Bousquet's bounded empirical-process inequality in the form of Theorem A.1 of Bartlett et al. (2005) to the centered signed class

$$\{\pm(g - Pg) : g \in \mathcal{G}^{\mathrm{exc}}_{m,p,T},\ Pg \le r\} \cup \{0\}.$$

Its members have mean zero, absolute value at most 2, and variance at most $10r$ by Lemma E.6. Its sample supremum is ND_r; the class is measurable and separable by assumption. The variance proxy in that inequality is bounded by $N(10r + 4\mathbb{E}D_r)$. Consequently, with probability at least $1 - e^{-x}$,

$$D_r \le \mathbb{E}D_r + \sqrt{\frac{2x(10r + 4\mathbb{E}D_r)}{N}} + \frac{2x}{3N}, \qquad x > 0.$$

Using $\sqrt{8(\mathbb{E}D_r)x/N} \le \mathbb{E}D_r + 2x/N$ and enlarging constants gives

$$D_r \le C\left\{\sqrt{ra} + a + \sqrt{\frac{rx}{N}} + \frac{x}{N}\right\} \tag{35}$$

with the same probability. This step uses an empirical-process concentration inequality; no star-shaped class or fixed-point theorem is assumed.

Step 2: Excluding deterministic risk shells. Fix a deterministic comparator $g_0 = L_{\theta_0}$ and let $b = Pg_0$ and $u = PL_{\widehat{\theta}}$. Both belong to $[0, 1]$. Choose a sufficiently large fixed constant C_* and let $t \ge C_*(b + a)$. For $s_j = 2^j t < 1$, consider the event $s_j < u \le 2s_j$, with localization radius $r_j = \min(2s_j, 1)$. Both g_0 and $L_{\widehat{\theta}}$ lie in the radius-r_j class on this event. Since $P_N L_{\widehat{\theta}} \le P_N g_0$,

$$u \le b + 2D_{r_j}.$$

Take $C_* \ge 4$, so $b \le s_j/4$. If $D_{r_j} \le s_j/4$, the preceding display gives $u \le 3s_j/4$, contradicting the shell event.

In equation 35, choose $x = c_* N s_j$ with a sufficiently small fixed $c_* > 0$. Since $r_j \le 2s_j$ and $s_j \ge C_* a$, first choosing C_* large and then c_* small makes its right-hand side at most $s_j/4$. Therefore

$$\Pr(s_j < u \le 2s_j) \le e^{-c_* N s_j}.$$

All radii are deterministic; the random ERM enters only through its shell membership.

Step 3: Tail integration and the comparator infimum. The shells cover $\{u > t\}$. Also $Na = V_{m,p,T}\log(eN) \ge 1$, so $Nt \ge C_*$. Summing their probabilities gives

$$\Pr(u > t) \le \sum_{j\ge 0} e^{-c_* N 2^j t} \le C' e^{-c_* N t}, \qquad t \ge C_*(b + a),$$

after enlarging C_* if needed. If $C_*(b+a) \geq 1$, boundedness already proves the desired expectation bound. Otherwise, integrating from this threshold to one yields

$$\mathbb{E}u \leq C_*(b+a) + \frac{C'}{c_* N} \leq C''(b+a),$$

because $a \geq 1/N$. The argument holds for every deterministic comparator; taking its population risk down to the infimum and using $PL_\theta = \overline{R}_{\mathrm{BG}}(M_\theta)$ proves equation 34. Empty localized classes were covered by the zero-function convention. Neither a realizable Bayes target nor within-block prefix independence is required.

□

Corollary E.9 (Bayes Gap under additional localized complexity)**.** *Under the conditions of Theorem E.5, if, in addition, Assumption E.7 holds, then*

$$\mathbb{E}\overline{R}_{\mathrm{BG}}(M_{\widehat{\theta}}) \leq C\left\{\varepsilon^2_{\mathrm{seq}}(p,T) + m^{-2\alpha/d_{\mathrm{meta}}} + \frac{V_{m,p,T}\log(eN)}{N}\right\}. \tag{36}$$

The expectation is over the independent pretraining blocks; the population Bayes Gap already averages over a fresh test Prompt. The constant C is independent of (N, m, p, T) and may depend on the fixed approximation, envelope, Bernstein, and localized-complexity constants.

Proof of Corollary E.9. Under Assumption E.7, the fast inequality equation 34 in Lemma E.8 gives

$$\mathbb{E}\overline{R}_{\mathrm{BG}}(M_{\widehat{\theta}}) \leq C\left\{\inf_{\theta\in\Theta_{m,p,T}} \overline{R}_{\mathrm{BG}}(M_\theta) + \frac{V_{m,p,T}\log(eN)}{N}\right\}.$$

Substituting equation 24 from Lemma E.2 proves equation 36. No further realizability of the Bayes predictor is required. □

E.6.3 Polynomial Capacity and Explicit Rates

Assumption E.10 (Polynomial capacity growth)**.** There are $\Gamma_T \geq 1$ and fixed constants $C_V > 0$, $\beta \geq 0$, $\gamma > 0$, and $q \geq 0$, independent of (N, m, p, T), such that, for every admissible integer $m \geq 2$,

$$V_{m,p,T} \leq C_V \Gamma_T p^\beta m^\gamma (\log m)^q. \tag{37}$$

Here $V_{m,p,T}$ is the pseudo-dimension of the complete block-loss class in Section 4.1. It includes the sequence encoder, task encoder, all decoder slices, count-aware selection, and clipping. The factor Γ_T accommodates capacity growth with sequence length.

This is an additional condition on the chosen model classes. The finite-network construction in Lemma E.2 does not establish equation 37; approximation realizability and the capacity bound must both hold for the same class.

Capacity interpretation. Assumption E.10 is an abstract capacity envelope for the full block-loss class rather than an architecture-specific pseudo-dimension calculation. Standard neural-network capacity bounds motivate polynomial dependence on model size up to depth and logarithmic factors (Bartlett et al., 2019); sharp exponents for the complete Transformer architecture are not needed here.

Width choices and explicit rates. Define

$$\begin{aligned} A_{N,T} &:= \frac{N}{\Gamma_T p^\beta}, \qquad a_{\mathrm{slow}} := \frac{2\alpha}{\gamma d_{\mathrm{meta}} + 4\alpha}, \\ a_{\mathrm{fast}} &:= \frac{2\alpha}{\gamma d_{\mathrm{meta}} + 2\alpha}. \end{aligned} \tag{38}$$

The following widths balance the polynomial orders of the approximation and statistical terms; logarithmic factors are retained in the resulting bounds.

Corollary E.11 (Explicit rates under polynomial capacity growth)**.** *Suppose that Assumption E.10 holds and $A_{N,T} > 1$. For each branch below, require the conditions of Theorem E.5 at the stated width, including architectural realizability and $1 \le V_{m,p,T} \le N$.*

For the global branch, choose

$$m^\star_{\mathrm{slow}} = \left\lceil A_{N,T}^{d_{\mathrm{meta}}/(\gamma d_{\mathrm{meta}}+4\alpha)} \right\rceil. \tag{39}$$

The shared ERM at this width satisfies

$$\mathbb{E}\overline{R}_{\mathrm{BG}}(M_{\widehat{\theta}}) \le C\left\{\varepsilon^2_{\mathrm{seq}}(p,T) + A_{N,T}^{-a_{\mathrm{slow}}}(\log(eA_{N,T}))^{q/2}(\log(eN))^{1/2}\right\}. \tag{40}$$

If, in addition, Assumption E.7 holds at the width

$$m^\star_{\mathrm{fast}} = \left\lceil A_{N,T}^{d_{\mathrm{meta}}/(\gamma d_{\mathrm{meta}}+2\alpha)} \right\rceil, \tag{41}$$

then the shared ERM at this width satisfies

$$\mathbb{E}\overline{R}_{\mathrm{BG}}(M_{\widehat{\theta}}) \le C\left\{\varepsilon^2_{\mathrm{seq}}(p,T) + A_{N,T}^{-a_{\mathrm{fast}}}(\log(eA_{N,T}))^{q}\log(eN)\right\}. \tag{42}$$

Constants are independent of (N, m, p, T), with additional dependence on C_V, β, γ, q allowed. Since $A_{N,T} \le N$, the displayed logarithmic factors are bounded respectively by $(\log(eN))^{(q+1)/2}$ and $(\log(eN))^{q+1}$. Both are therefore bounded by the common envelope $\mathfrak{L}_N = (\log(eN))^{q+1}$ used in the main text.

The restriction $A_{N,T} > 1$ ensures that the rounded widths are at least two. Realizability and $1 \le V_{m,p,T} \le N$ remain separate admissibility requirements.

Proof of Corollary E.11. Write $d = d_{\mathrm{meta}}$ and $A = A_{N,T} > 1$ within this proof. For any fixed $b > 0$, rounding gives $A^b \le \lceil A^b \rceil \le 2A^b$ and $\log\lceil A^b \rceil \le \log 2 + b \log A \le C_b \log(eA)$. The constants depend only on the fixed exponents, not on (N, m, p, T).

Step 1: Global statistical term and slow width. Assumption E.10 gives

$$\sqrt{\frac{V_{m,p,T}\log(eN)}{N}} \le \sqrt{C_V}\, A^{-1/2} m^{\gamma/2} (\log m)^{q/2} \sqrt{\log(eN)}.$$

Balancing only powers, $m^{-2\alpha/d} = A^{-1/2}m^{\gamma/2}$ gives $m = A^{d/(\gamma d+4\alpha)}$. At its rounded value equation 39,

$$m^{-2\alpha/d} \le A^{-a_{\mathrm{slow}}}, \qquad A^{-1/2}m^{\gamma/2} \le CA^{-1/2+\gamma d/[2(\gamma d+4\alpha)]} = CA^{-a_{\mathrm{slow}}}.$$

The logarithmic bound above controls $(\log m)^{q/2}$ by $C(\log(eA))^{q/2}$. Substitution into Theorem E.5 proves equation 40, absorbing the approximation term because the displayed logarithmic factor is at least one.

Step 2: Local statistical term and fast width. Under the additional Assumption E.7, use Corollary E.9 and

$$\frac{V_{m,p,T}\log(eN)}{N} \le C_V A^{-1} m^{\gamma} (\log m)^q \log(eN).$$

The power balance $m^{-2\alpha/d} = A^{-1}m^{\gamma}$ gives $m = A^{d/(\gamma d+2\alpha)}$. At equation 41,

$$m^{-2\alpha/d} \le A^{-a_{\mathrm{fast}}}, \qquad A^{-1}m^{\gamma} \le CA^{-1+\gamma d/(\gamma d+2\alpha)} = CA^{-a_{\mathrm{fast}}}.$$

Combining these inequalities with $(\log m)^q \le C(\log(eA))^q$ proves equation 42. Finally, $\Gamma_T \ge 1$, $p \ge 1$, and $\beta \ge 0$ imply $A \le N$, giving the stated common logarithmic envelope. The calculations apply only at widths satisfying the corollary's realizability and capacity requirements. □

E.6.4 Consequences at a Specified Context Length

For any specified $k \in \{1, \ldots, p\}$, nonnegativity gives

$$R_{\mathrm{BG},k}(M) \le p\, \overline{R}_{\mathrm{BG}}(M). \tag{43}$$

Therefore each average-prefix bound in this subsection immediately yields the corresponding fixed-k bound after multiplication by p, including the case $k = p$. No simultaneous probability guarantee over all context lengths is asserted.

F Posterior-Variance Analysis

This section analyzes the inference-stage uncertainty that remains after pretraining. The key mechanism is latent-type identification: solved context tasks concentrate the posterior mass on the true type, which drives the mixture Posterior Variance toward the corresponding type-oracle variance. The oracle variance is in turn bounded by the intrinsic localization difficulty $\mathcal{R}_T(\mathcal{F}_i)$. We first establish this type-conditional result, then average over the latent-type prior to obtain the main bound, and finally show that the expected Posterior Variance is nonincreasing with the number of solved context tasks.

F.1 Type-Conditional Posterior Variance

Fix $i^\star \in \{1, \dots, K\}$ and define

$$R_{\mathrm{PV},k}^{(i^\star)} := \mathbb{E}\left[T^{-2}\,\mathrm{Var}(\tau_{k+1} \mid \mathcal{P}^k) \,\middle|\, I = i^\star\right].$$

This is the mixture posterior variance averaged under the true-type sampling law. If the type is revealed by an oracle, define

$$R_{\mathrm{PV}}^{\mathrm{oracle},(i^\star)} := \mathbb{E}\left[T^{-2}\,\mathrm{Var}(\tau \mid X_{1:T}, I = i^\star) \,\middle|\, I = i^\star\right],$$

for an independent query task $(X_{1:T}, \tau)$ from type $i^\star$; by Assumption 3.3, this quantity does not depend on k.

Theorem F.1 (Type-conditional posterior concentration and oracle Posterior Variance bound)**.** *Consider the prompt-generation process in Definition 3.1. Fix $i^\star \in \{1, \dots, K\}$ and suppose that Assumption 3.3 holds. If $K \geq 2$, also suppose that Assumption 4.3 holds and define, for any fixed $k \in \{1, \dots, p\}$,*

$$\varepsilon_{k,T}^{\mathrm{id}} := \frac{1 - \alpha_{i^\star}}{\alpha_{i^\star}} \exp(-D_{\min,T} k/2) + (K-1)\exp(-C_{\mathrm{id},T} k). \tag{44}$$

For $K = 1$, set $\varepsilon_{k,T}^{\mathrm{id}} := 0$ and no identifiability condition is required. Then

$$\mathbb{E}\big[1 - \pi_{i^\star}(\mathcal{P}^k) \mid I = i^\star\big] \leq \varepsilon_{k,T}^{\mathrm{id}}, \tag{45}$$

$$\left|R_{\mathrm{PV},k}^{(i^\star)} - R_{\mathrm{PV}}^{\mathrm{oracle},(i^\star)}\right| \leq 5\varepsilon_{k,T}^{\mathrm{id}}, \tag{46}$$

$$R_{\mathrm{PV},k}^{(i^\star)} \leq \mathcal{R}_T(\mathcal{F}_{i^\star}) + 5\varepsilon_{k,T}^{\mathrm{id}}. \tag{47}$$

Thus the inference-stage uncertainty separates into the type-oracle localization difficulty and a latent-type identification error. The unconditional bound in Theorem 4.4 follows by averaging this result over the type prior.

F.2 Ingredients for the Type-Conditional Bound

F.2.1 Oracle Benchmark

Lemma F.2 (Oracle posterior variance and minimax localization risk)**.** *Fix a true type $i^\star$ and let $\mathcal{A}_T$ be the measurable predictors from the observed sequence space to $[1, T-1]$. Under $P_{\mathcal{F}_{i^\star}}$, the conditional mean*

$$\mu_{i^\star}(X) = \mathbb{E}[\tau \mid X, I = i^\star], \qquad X = X_{1:T},$$

minimizes the type-oracle squared localization risk, and

$$\begin{aligned} R_{\mathrm{PV}}^{\mathrm{oracle},(i^\star)} &= \inf_{\widehat{\tau} \in \mathcal{A}_T} \mathbb{E}\left[\frac{(\widehat{\tau}(X) - \tau)^2}{T^2} \,\middle|\, I = i^\star\right] \\ &\leq \mathcal{R}_T(\mathcal{F}_{i^\star}). \end{aligned} \tag{48}$$

Allowing arbitrary real-valued measurable rules gives the same infimum after projection onto $[1, T-1]$.

Proof. For any $\widehat{\tau} \in \mathcal{A}_T$,

$$\mathbb{E}[(\widehat{\tau}(X) - \tau)^2 \mid X, I = i^\star] = \mathrm{Var}(\tau \mid X, I = i^\star) + (\widehat{\tau}(X) - \mu_{i^\star}(X))^2,$$

so the conditional mean attains the first equality in equation 48. Moreover, for every fixed $\widehat{\tau}$,

$$\begin{aligned}\mathbb{E}\left[\frac{(\widehat{\tau}(X) - \tau)^2}{T^2} \,\middle|\, I = i^\star\right] &= \int_{\mathcal{F}_{i^\star}} \mathbb{E}_f\left[\frac{(\widehat{\tau}(X) - \tau(f))^2}{T^2}\right] dP_{\mathcal{F}_{i^\star}}(f) \\ &\leq \sup_{f \in \mathcal{F}_{i^\star}} \mathbb{E}_f\left[\frac{(\widehat{\tau}(X) - \tau(f))^2}{T^2}\right].\end{aligned}$$

Taking the infimum over the same rule class proves the inequality; equality with the worst-case minimax risk is not required. □

F.2.2 Latent-Type Posterior Concentration

Proof of the concentration assertion in Theorem F.1. Fix $K \geq 2$, $I = i^\star$, and $k \in \{1, \ldots, p\}$. All probabilities and expectations below are conditional on $I = i^\star$ unless stated otherwise.

Step 1: Context likelihood concentration. For $j \neq i^\star$, let

$$Z_{j,t,T} = \log \frac{q_{j,T}(X^{(t)}_{1:T}, \tau_t)}{q_{i^\star,T}(X^{(t)}_{1:T}, \tau_t)}, \qquad S_{j,k,T} = \sum_{t=1}^{k} Z_{j,t,T}.$$

Iterating the conditional MGF bound equation 9 gives, for $0 < \lambda \leq 1/b_{j,T}$,

$$\mathbb{E}\exp\{\lambda(S_{j,k,T} + kD_{j,T})\} \leq \exp(k\lambda^2\nu^2_{j,T}/2).$$

Hence exponential Markov inequality and the definition of $C_{j,T}$ yield

$$\Pr(S_{j,k,T} > -kD_{j,T}/2) \leq e^{-kC_{j,T}}.$$

With

$$\mathcal{E}_{k,T} := \bigcap_{j \neq i^\star} \{S_{j,k,T} \leq -kD_{j,T}/2\},$$

a union bound gives

$$\Pr(\mathcal{E}^c_{k,T} \mid I = i^\star) \leq (K-1)e^{-kC_{\mathrm{id},T}}. \tag{49}$$

Step 2: Query likelihood and posterior mass. Write $X = X^{(k+1)}_{1:T}$. Conditional task independence and Bayes' rule give

$$\frac{\pi_j(\mathcal{P}^k)}{\pi_{i^\star}(\mathcal{P}^k)} = \frac{\alpha_j}{\alpha_{i^\star}} \frac{q^X_{j,T}(X)}{q^X_{i^\star,T}(X)} \exp(S_{j,k,T}). \tag{50}$$

Although the query likelihood ratio need not be bounded pointwise, under the true type

$$\mathbb{E}\left[\frac{q^X_{j,T}(X)}{q^X_{i^\star,T}(X)} \,\middle|\, I = i^\star\right] = \int_{\{q^X_{i^\star,T} > 0\}} q^X_{j,T}\, d\mu_{X,T} \leq 1. \tag{51}$$

On $\mathcal{E}_{k,T}$,

$$1 - \pi_{i^\star}(\mathcal{P}^k) \leq e^{-kD_{\min,T}/2} \sum_{j \neq i^\star} \frac{\alpha_j}{\alpha_{i^\star}} \frac{q^X_{j,T}(X)}{q^X_{i^\star,T}(X)}.$$

Since $\mathcal{E}_{k,T}$ depends only on the solved context tasks and is conditionally independent of the query, equation 51 and equation 49 imply

$$\mathbb{E}[1 - \pi_{i^\star}(\mathcal{P}^k) \mid I = i^\star] \leq \frac{1 - \alpha_{i^\star}}{\alpha_{i^\star}} e^{-kD_{\min,T}/2} + (K-1)e^{-kC_{\mathrm{id},T}} = \varepsilon^{\mathrm{id}}_{k,T},$$

which proves equation 45. □

F.2.3 Mixture-to-Oracle Variance Comparison

Completion of the proof of Theorem F.1. Take $K \geq 2$, set $X = X_{1:T}^{(k+1)}$, and define

$$\mu_i(X) := \mathbb{E}[\tau_{k+1} \mid X, I = i], \qquad v_i(X) := \mathrm{Var}(\tau_{k+1} \mid X, I = i).$$

By conditional task independence, these are also the type-specific conditional moments given the full prompt. The conditional law of total variance gives

$$\mathrm{Var}(\tau_{k+1} \mid \mathcal{P}^k) = \sum_{i=1}^{K} \pi_i(\mathcal{P}^k) v_i(X) + \mathrm{Var}_{I|\mathcal{P}^k}(\mu_I(X)). \tag{52}$$

Since $0 \leq v_i \leq T^2$ and $|\mu_i - \mu_{i^\star}| \leq 2T$,

$$\left| \sum_i \pi_i v_i - v_{i^\star} \right| \leq T^2(1 - \pi_{i^\star}), \qquad \mathrm{Var}_{I|\mathcal{P}^k}(\mu_I) \leq 4T^2(1 - \pi_{i^\star}).$$

Therefore

$$\left| \mathrm{Var}(\tau_{k+1} \mid \mathcal{P}^k) - v_{i^\star}(X) \right| \leq 5T^2(1 - \pi_{i^\star}). \tag{53}$$

Dividing by T^2, taking expectations conditional on $I = i^\star$, and using equation 45 proves equation 46. Lemma F.2 then gives equation 47. For $K = 1$, $\pi_1 \equiv 1$ and $\varepsilon_{k,T}^{\mathrm{id}} = 0$, so the oracle comparison is exact. □

Remark. The quantity $R_{\mathrm{PV},k}^{(i^\star)}$ averages the mixture posterior variance under the true-type sampling law. It is not the type-conditional prediction risk of the mixture Bayes predictor; the latter contains an additional squared deviation from the type-specific conditional mean.

F.3 Unconditional Posterior-Variance Bound

Proof of Theorem 4.4. For fixed k, average Theorem F.1 over $I \sim \mathrm{Categorical}(\alpha)$:

$$R_{\mathrm{PV},k} = \sum_{i=1}^{K} \alpha_i R_{\mathrm{PV},k}^{(i)} \leq \sum_{i=1}^{K} \alpha_i \mathcal{R}_T(\mathcal{F}_i) + 5 \sum_{i=1}^{K} \alpha_i \varepsilon_{k,i,T}^{\mathrm{id}}.$$

This is equation 11; for $K = 1$, the identification term is zero. □

F.4 Monotonicity with Context Length

Lemma F.3 (Monotonicity of Posterior Variance in the context length)**.** *Under Assumption 3.3,* $R_{\mathrm{PV},k+1} \leq R_{\mathrm{PV},k}$ *for* $k = 1, \ldots, p-1$.

Proof of Lemma F.3. On one probability space, draw the latent type, conditionally i.i.d. solved tasks $Z_j = (X_{1:T}^{(j)}, \tau_j)$, and an additional query pair (X°, τ°) conditionally independent of the solved tasks. Let

$$\mathcal{H}_k^\circ := \sigma(Z_1, \ldots, Z_k, X^\circ), \qquad \mathcal{H}_k^\circ \subseteq \mathcal{H}_{k+1}^\circ.$$

By conditional task exchangeability, $R_{\mathrm{PV},k} = T^{-2}\mathbb{E}\,\mathrm{Var}(\tau^\circ \mid \mathcal{H}_k^\circ)$. The conditional law of total variance yields

$$\mathrm{Var}(\tau^\circ \mid \mathcal{H}_k^\circ) = \mathbb{E}[\mathrm{Var}(\tau^\circ \mid \mathcal{H}_{k+1}^\circ) \mid \mathcal{H}_k^\circ] + \mathrm{Var}(\mathbb{E}[\tau^\circ \mid \mathcal{H}_{k+1}^\circ] \mid \mathcal{H}_k^\circ).$$

Taking expectations and discarding the nonnegative second term proves $R_{\mathrm{PV},k+1} \leq R_{\mathrm{PV},k}$. The claim concerns expected posterior variance, not a pointwise conditional-variance ordering. □

G Overall Risk and Localization Consistency

This section closes the theoretical argument by combining the pretraining-stage Bayes Gap with the inference-stage Posterior Variance. The resulting risk bound separates four sources of localization error: sequence information loss, statistical learning error, intrinsic within-type localization difficulty, and latent-type identification error. Under the scaling conditions in Assumption 4.6, each contribution vanishes, yielding relative change-point localization consistency. We then record refined Bayes-Gap variants, fixed-context consequences, and the effect of integer rounding.

G.1 Overall ICCP Risk Bound

Proof of Theorem 4.5. Condition on the pretraining data. By the risk decomposition in Section 4,

$$R_k(M_{\widehat{\theta}}) = R_{\mathrm{PV},k} + R_{\mathrm{BG},k}(M_{\widehat{\theta}}).$$

Averaging over the pretraining data and applying Theorems 4.2 and 4.4 gives the first bound in Theorem 4.5; only the Bayes-Gap term carries the factor p.

For the simplified identification term, suppose $K \geq 2$ and use the uniform lower bounds in the theorem. Since $\sum_i \alpha_i = 1$,

$$\sum_{i=1}^{K} \alpha_i \frac{1-\alpha_i}{\alpha_i} = K - 1.$$

Thus, with $c_{0,T} = \min\{D_{0,T}/2, C_{0,T}\}$,

$$\sum_i \alpha_i \varepsilon_{k,i,T}^{\mathrm{id}} \leq (K-1)e^{-kD_{0,T}/2} + (K-1)e^{-kC_{0,T}} \leq 2(K-1)e^{-kc_{0,T}}.$$

Multiplying by the coefficient 5 in the Posterior-Variance bound yields the term $10(K-1)e^{-kc_{0,T}}$. For $K = 1$, the identification contribution is identically zero. □

G.2 Relative Localization Consistency

Proof of Theorem 4.7. Let

$$A = A_{N,T} = \frac{N}{\Gamma_T p^{\beta}}, \qquad \mathfrak{L}_N = (\log(eN))^{q+1}, \qquad a = a_{\mathrm{slow}} = \frac{2\alpha}{\gamma d_{\mathrm{meta}} + 4\alpha}.$$

Under (C5), $A \to \infty$, so the prescribed width $m = m^{\star}_{\mathrm{slow}}$ in equation 39 is eventually admissible. The polynomial-capacity bound and the rounding estimates from Section E.6.3 imply

$$\frac{V_{m,p,T}}{N} \leq CA^{-1}m^{\gamma}(\log m)^q \leq CA^{-2a}(\log(eA))^q \longrightarrow 0,$$

so $V_{m,p,T} \leq N$ eventually, together with the architectural realizability conditions already assumed in Theorem 4.5.

Combining the slow polynomial-capacity Bayes-Gap bound with Theorem 4.4 and the uniform identification simplification gives, for $K \geq 2$,

$$\mathbb{E}R_k(M_{\widehat{\theta}}) \leq Cp\varepsilon_{\mathrm{seq}}^2(p,T) + CpA^{-a}\mathfrak{L}_N + r_T + 10(K-1)e^{-kc_{0,T}}. \tag{54}$$

The four terms vanish by (C4), (C5), (C3), and (C2), respectively. For $K = 1$, the identification term is absent and the remaining three terms are unchanged. Hence

$$\mathbb{E}R_k(M_{\widehat{\theta}}) = \mathbb{E}\left[\frac{(\widehat{\tau}_k - \tau_{k+1})^2}{T^2}\right] \longrightarrow 0.$$

For any fixed $\eta > 0$, Markov's inequality gives

$$\Pr\left(\frac{|\widehat{\tau}_k - \tau_{k+1}|}{T} > \eta\right) \leq \eta^{-2}\mathbb{E}R_k(M_{\widehat{\theta}}) \longrightarrow 0,$$

which proves the claimed relative localization consistency. □

G.3 Refined Bounds and Consequences

Localized and polynomial-capacity refinements. Each refined overall-risk bound has the form

$$\mathbb{E}R_k(M_{\widehat{\theta}}) \leq Cp\,\mathcal{B}_{N,m,p,T} + \sum_i \alpha_i \mathcal{R}_T(\mathcal{F}_i) + 5\sum_i \alpha_i \varepsilon_{k,i,T}^{\mathrm{id}}, \tag{55}$$

where $\mathcal{B}_{N,m,p,T}$ is the corresponding Bayes-Gap bound. Specifically,

$$\mathcal{B}_{N,m,p,T} = \varepsilon_{\mathrm{seq}}^2(p,T) + m^{-2\alpha/d_{\mathrm{meta}}} + \frac{V_{m,p,T}\log(eN)}{N}$$

under Corollary E.9; under the slow and fast branches of Corollary E.11, respectively,

$$\mathcal{B}_{N,m,p,T} = \varepsilon^2_{\text{seq}}(p,T) + A^{-a_{\text{slow}}}_{N,T}(\log(eA_{N,T}))^{q/2}(\log(eN))^{1/2},$$

and

$$\mathcal{B}_{N,m,p,T} = \varepsilon^2_{\text{seq}}(p,T) + A^{-a_{\text{fast}}}_{N,T}(\log(eA_{N,T}))^{q}\log(eN).$$

Equation equation 55 follows by combining the corresponding fixed-k Bayes-Gap bound from Section E.6.4 with the same Posterior-Variance bound. Under the uniform identification lower bounds, its last term is again bounded by $10(K-1)e^{-kc_{0,T}}$.

Localized fast-rate consistency. If Assumption E.7 also holds, use $m = m^\star_{\text{fast}}$ from equation 41 and replace the slow-rate requirement in (C5) by

$$pA^{-a_{\text{fast}}}_{N,T}\mathfrak{L}_N \longrightarrow 0, \qquad a_{\text{fast}} = \frac{2\alpha}{\gamma d_{\text{meta}} + 2\alpha}.$$

Then the fast branch of equation 55 yields the same four-term decomposition as equation 54, with a_{slow} replaced by a_{fast}, and the preceding Markov argument again gives relative localization consistency.

Fixed-Context and Full-Prompt Consequences. If $k = k_0$ is fixed and $K \geq 2$, the displayed sufficient bound requires $k_0 c_{0,T} \to \infty$ for the identification term to vanish; increasing N alone does not enforce this condition. This condition is sufficient for the bound above and is not claimed to be necessary.

For the complete Prompt, take $k = p$. Under $pc_{0,T} \to \infty$ and the remaining global consistency conditions, equation 54 gives $\mathbb{E}R_p(M_{\widehat{\theta}}) \to 0$, and hence

$$T^{-1}\left|M_{\widehat{\theta}}(\mathcal{P}^p) - \tau_{p+1}\right| \xrightarrow{P} 0.$$

For $K = 1$, no context-based identification growth condition is required.

Integer-valued localization. For any real-valued estimate $\widehat{\tau}$, let $\widehat{\tau}^{\,c}$ be its projection onto $[1, T-1]$ and let $\widehat{\tau}^{\,\text{int}}$ be a nearest integer in $\{1,\ldots,T-1\}$. Since τ already lies in this set,

$$\frac{|\widehat{\tau}^{\,\text{int}} - \tau|}{T} \leq \frac{|\widehat{\tau} - \tau|}{T} + \frac{1}{2T}.$$

Thus integer rounding preserves relative localization consistency as $T \to \infty$.