FUNCTIONAL MARTINGALE COHERENCE AND AMORTIZED UNCERTAINTY FOR FROZEN FORECASTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predictive resampling quantifies uncertainty over a quantity of interest by recursively imputing future observations from a forecaster, thereby inducing a distribution over that quantity. However, a generic pretrained forecaster can systematically shift the quantity under its own imputations, conflating stochastic future variation with self-induced drift. We develop a coherence repair that preserves the forecaster's responses to augmented contexts and instead reweights the imputation law. At each context, we measure how each candidate imputation changes the quantity of interest and use these changes to reweight the imputation law. The same construction extends to controlling the local covariance of one-step changes. Local covariance, however, does not determine the uncertainty over an entire continuation. Directly estimating this quantity requires repeated future rollouts, while the current local covariance alone is insufficient to determine it. We therefore introduce *Amortized Covariance Estimation* (ACE), which predicts the finite-horizon covariance correction from current response information. At inference, ACE requires only current candidate responses and repair, avoiding future sequential rollouts while keeping the forecaster fixed. Across TabPFN and LLM forecasters, coherence repair substantially reduces self-induced drift, while ACE improves prediction of future variation beyond local covariance.

1 INTRODUCTION

Many pretrained models, including Prior-data Fitted Networks (PFNs) (Müller et al., 2022), TabPFNs (Hollmann et al., 2023), and autoregressive language models (LMs), map an observed context to a distribution over possible next outcomes. We refer to such a fixed model as a *forecaster*, and to its distribution over possible next outcomes as a *predictive distribution*. Given such a predictive distribution, a natural way to quantify uncertainty over a quantity of interest is *predictive resampling* (Fong et al., 2023; Fortini & Petrone, 2025). Starting from the observed context, predictive resampling repeatedly samples an imputation, appends it to the current context, and queries the forecaster again. Repeating this procedure generates many continuation paths, and collecting the corresponding values provides an empirical approximation to the distribution induced over the quantity of interest. Under suitable coherence and convergence conditions, the limiting distribution is called a *martingale posterior distribution* (Fong et al., 2023).

Predictive resampling is particularly attractive for pretrained forecasters because it requires only their predictions at augmented contexts, without an explicit likelihood–prior specification. For example, PFNs can condition on synthetic input-label pairs (Müller et al., 2022; Hollmann et al., 2023; 2025), while autoregressive LMs can condition on generated tokens or reasoning blocks. However, the uncertainty induced by predictive resampling depends on how the forecaster responds to its own imputations. Each imputation changes the context and can therefore change the quantity of interest itself. Ideally, these changes should reflect variation across plausible continuations, but generic pretrained forecasters can induce systematic drift in the quantity of interest instead (Ng et al., 2026; Kim et al., 2026). As a result, the uncertainty induced by predictive resampling may conflate stochastic variation across future continuations with systematic self-induced drift.

Recent work has sought to diagnose and address these issues through several complementary approaches. TabMGP directly applies predictive resampling to TabPFN to obtain uncertainty over loss-defined parameters (Ng et al., 2026), whereas Nagler & Rügamer (2026) initialize predictive

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.

resampling from a PFN predictive distribution while replacing subsequent PFN updates with a separately designed coherent predictive recursion. For Bayes-filtered transformers, Fortini et al. (2026) infer uncertainty about the limiting predictive distribution from the sequence of predictive updates, without explicitly representing a posterior. For LMs, predictive drift under self-generated continuations has been directly diagnosed (Falck et al., 2024), while it is also reported that this drift is reducible via a special prompting strategy (Kim et al., 2026).

We study how to make predictive resampling from a pretrained forecaster coherent for a chosen quantity of interest, without retraining the model or altering its responses. At each context, we quantify the effect of each candidate imputation by the change it induces in the quantity of interest. Rather than modifying these responses, we *repair* the imputation law itself, reweighting the candidate probabilities so that the expected change vanishes while remaining as close as possible to the original law. We characterize when such an exact repair is feasible and, when it is not, develop an approximate repair scheme with a bounded drift guarantee. The same framework further extends to controlling the local covariance of the one-step changes induced by the next imputation.

Controlling local covariance, however, does not determine uncertainty over an entire continuation. We therefore characterize the finite-horizon covariance accumulated across future imputations, which coincides with the endpoint covariance when exact mean repair is applied at every step and continues to represent the accumulated centered variation under approximate repair. Importantly, this future spread is not determined by the current local covariance alone, since identical current responses can evolve differently after subsequent imputations. This motivates *Amortized Covariance Estimation* (ACE), which learns to predict the finite-horizon correction from current response information. Predictive resampling supplies finite-horizon training targets, while inference requires only current candidate responses and repair, without future sequential rollouts or updates to the forecaster.

Our contributions can be summarized as follows: (i) We first develop a coherence repair for frozen forecasters that reweights the imputation law while preserving the forecaster's responses, with exact and approximate variants and an extension to local covariance control (Section 3). (ii) We characterize finite-horizon uncertainty beyond the current local covariance and introduce ACE to predict the resulting future covariance correction without sequential future rollouts (Section 4). (iii) We evaluate the framework on frozen TabPFN and LLM forecasters, demonstrating substantial reductions in self-induced drift and improved prediction of future variation (Sections 5 and 6).

2 Martingale Posterior and Functional Coherence

2.1 Predictive resampling and martingale posterior distributions

Let $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ be a standard Borel observation space. Let $(\widetilde{Z}_i)_{i\geq 1}$ be a $\mathcal{Z}$-valued stochastic process with data-generating law $\mathbb{P}_{\mathrm{data}}$. Fix $n \in \mathbb{N}$ and let $\widetilde{Z}_{1:n} = \tilde{z}_{1:n}$ denote the observed context, where $\tilde{z}_{1:n} := (\tilde{z}_1, \ldots, \tilde{z}_n)$. Separately, let $(P_k)_{k\geq 1}$ be probability kernels $P_k : \mathcal{Z}^k \times \mathcal{B}(\mathcal{Z}) \to [0,1]$, where $P_k(\cdot \mid z_{1:k})$ denotes the supplied one-step predictive distribution after a context of length k. These predictive kernels need not coincide with the conditional distributions induced by $\mathbb{P}_{\mathrm{data}}$.

Conditioning on the observed context $\tilde{z}_{1:n}$, predictive resampling (Fong et al., 2023) recursively generates a continuation according to

$$Z_1 \sim P_n(\cdot \mid \tilde{z}_{1:n}), \qquad Z_2 \sim P_{n+1}(\cdot \mid \tilde{z}_{1:n}, Z_1), \qquad \ldots. \tag{1}$$

By the Ionescu–Tulcea theorem (Ionescu-Tulcea, 1949), these kernels define a unique conditional law $\mathbb{P}^P_{\tilde{z}_{1:n}}$ over the continuation $Z_{1:\infty}$. Throughout the paper, we denote $D_t := (\tilde{z}_{1:n}, Z_{1:t})$ and $\mathcal{F}_t := \sigma(D_t)$ for $t \geq 0$ with $D_0 = \tilde{z}_{1:n}$. Thus $\mathbb{P}_{\mathrm{data}}$ describes the source of the observed context, whereas $\mathbb{P}^P_{\tilde{z}_{1:n}}$ governs its predictive continuation.

For an imputation horizon $t \geq 0$, let $F_t := \frac{1}{n+t}(\sum_{i=1}^n \delta_{\tilde{z}_i} + \sum_{s=1}^t \delta_{Z_s})$ denote the random empirical distribution. Let $\theta : \mathcal{P}(\mathcal{Z}) \to \mathbb{R}^d$ be a measurable target functional, where $\mathcal{P}(\mathcal{Z})$ denotes a space of probability measures on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$. If, under suitable coherence and regularity conditions, $F_t \to P_\infty$ and $\theta(F_t) \to \theta(P_\infty)$ almost surely, define $\theta_\infty := \theta(P_\infty)$. For any measurable set $A \subseteq \mathbb{R}^d$, the induced law

$$\Pi^\theta_\infty(A \mid \tilde{z}_{1:n}) := \mathbb{P}^P_{\tilde{z}_{1:n}}(\theta_\infty \in A) \tag{2}$$

is the *martingale posterior distribution* for θ (Fong et al., 2023). Thus $\Pi^\theta_\infty(\cdot \mid \tilde{z}_{1:n})$ is the pushforward of the predictive-resampling law on $Z_{1:\infty}$ through the limiting target θ_∞.

For the martingale posterior construction, Fong et al. (2023) impose *predictive coherence*, for which the conditionally identically distributed (c.i.d.) condition provides a sufficient condition (Berti et al., 2004). Under the c.i.d. condition, for every $A \in \mathcal{B}(\mathcal{Z})$, the predictive probabilities form a martingale with respect to $(\mathcal{F}_t)_{t\geq 0}$:

$$\mathbb{E}\left[P_{n+t+1}(Z_{t+2} \in A \,|\, D_{t+1}) \,|\, \mathcal{F}_t\right] = P_{n+t}(Z_{t+1} \in A \,|\, D_t), \qquad t \geq 0. \tag{3}$$

Intuitively, although the predictive distribution changes as future observations are imputed, these changes have no systematic conditional drift. Thus, for each $A \in \mathcal{B}(\mathcal{Z})$, the bounded martingale $P_{n+t}(A \,|\, D_t)$ converges almost surely by the martingale convergence theorem (Doob, 1953). Under suitable c.i.d. regularity conditions, these limits define a random probability measure P_∞, and the completed empirical distributions F_t converge to the same limit, yielding the martingale-posterior construction above (Berti et al., 2004; Fong et al., 2023). For Bayesian posterior predictive kernels and the corresponding target, this recovers the conventional Bayesian posterior (Fong et al., 2023).

All expectations in the remainder of this subsection are under $\mathbb{P}^P_{z_{1:n}}$. For integrable θ_∞, define the posterior-mean process $m_t := \mathbb{E}[\theta_\infty \mid \mathcal{F}_t]$. By the tower property, $(m_t)_{t\geq 0}$ is a martingale: $\mathbb{E}[m_{t+1} \mid \mathcal{F}_t] = m_t$. A finite-context target that coincides with m_t therefore inherits this martingale property. However, predictive coherence does not imply that every finite-context target is a martingale. In particular, a nonlinear plug-in target computed from the current predictive distribution or context need not coincide with the posterior-mean process m_t. We therefore impose the martingale condition directly on the chosen finite-context target by adjusting the imputation law.

2.2 Frozen forecasters and target readouts

We assume access to a fixed pretrained predictive model F, which we call a *frozen forecaster*, or simply a forecaster. Examples include PFNs and LLMs implemented with transformers (Vaswani et al., 2017). When the forecaster predicts the next observation, it induces the predictive kernels

$$P_{n+t}(A \,|\, D_t) = \mathsf{F}(D_t)(A), \qquad A \in \mathcal{B}(\mathcal{Z}). \tag{4}$$

We also allow the forecaster to predict only part of the next observation. For example, in supervised learning, let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and write an observation as $z = (x, y)$, where x is a covariate and y its response. Given a query $x_* \in \mathcal{X}$, the forecaster may return the conditional predictive distribution $\mathsf{F}(D_t, x_*) \in \mathcal{P}(\mathcal{Y})$, from which $Y_* \sim \mathsf{F}(D_t, x_*)$.

PFNs typically operate in this conditional form: they predict the response conditional on a query, but do not themselves generate the query. To define a joint continuation, write $Z_{t+1} = (X_{t+1}, Y_{t+1})$ and let $G_t(\cdot \,|\, D_t)$ be a prescribed imputation law for the next covariate. Then

$$X_{t+1} \sim G_t(\cdot \,|\, D_t), \qquad Y_{t+1} \mid X_{t+1} = x, D_t \sim \mathsf{F}(D_t, x), \tag{5}$$

$$P_{n+t}(A \times B \,|\, D_t) = \int_A \mathsf{F}(D_t, x)(B)\, G_t(\,\mathrm{d}x \,|\, D_t), \tag{6}$$

for measurable $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$. The covariate law G_t may, for example, be the empirical distribution over available covariates, a Bayesian-bootstrap distribution, or any predefined design law over the covariate space (Rubin, 1981; Fong et al., 2023; Ng et al., 2026).

For an LLM, the continuation variable Z_t may represent a single token or a larger token sequence generated autoregressively from the current context D_t. The frozen forecaster $\mathsf{F}(D_t)$ may induce a distribution over the entire sequence through its token-level conditionals, so the continuation unit may be a sentence, reasoning step, or other structured block rather than a single token. This allows the same framework to operate at different granularities without changing the underlying model.

Fong et al. (2023) primarily consider target functionals defined from the predictive distribution P_{n+t} or the empirical distribution F_t. We use a slightly broader interface that maps the current context directly to a target quantity of interest, which we call a *readout*. Formally, for a fixed evaluation rule, let $\theta_t := \Theta(D_t) \in \mathbb{R}^d$, where Θ is measurable and any auxiliary quantities defining the target are held fixed throughout the continuation. For example, for a classification PFN and a fixed query x_*, a readout may be the predictive class-probability vector $\Theta(D_t) = \mathsf{F}(D_t, x_*)$ or a scalar component corresponding to a fixed class. The completed-data target $\theta(F_t)$ is the special case $\Theta(D_t) = \theta(F_t)$. Readouts may also include loss-defined parameters such as $\Theta(D_t) = \arg\min_\beta \int \ell(z, \beta)\, \mathrm{d}F_t(z)$, assuming a uniquely specified minimizer. Thus, the readout specifies the finite-context target whose coherence we study, rather than a parameter of the pretrained forecaster.

2.3 Exact responses and functional coherence

From here onward, let q_t denote the "base" imputation law induced by the forecaster, $q_t = P_{n+t}(\cdot \mid D_t)$, and let Q_t denote an arbitrary imputation law for Z_{t+1}, not necessarily equal to q_t, with $Z_{t+1} \mid \mathcal{F}_t \sim Q_t$. For $z \in \mathcal{Z}$, write $D_t \oplus z$ for the context obtained by appending z to D_t.

For a chosen integrable readout $\theta_t = \Theta(D_t)$, we say that the process $(\theta_t)_{t\geq 0}$ is *functionally coherent* if $\mathbb{E}_{Q_t}[\theta_{t+1} \mid \mathcal{F}_t] = \theta_t$ for all $t \geq 0$, so that $(\theta_t)_{t\geq 0}$ is a martingale with respect to $(\mathcal{F}_t)_{t\geq 0}$. Unlike predictive coherence in eq. (3), which directly imposes a condition on the predictive law, functional coherence is defined for a specific readout Θ. This perspective relates to parametric martingale posteriors (Walker, 2022; Holmes & Walker, 2023), which directly construct a martingale sequence.

We now characterize a condition on the imputation law Q_t that ensures functional coherence. To this end, we first define the *in-context influence*, which quantifies the change in the readout induced by appending a candidate observation z to the current context D_t.

Definition 2.1 (In-context influence)**.** For $t \geq 0$ and a candidate imputation $z \in \mathcal{Z}$ for Z_{t+1}, define the *in-context influence* as $\psi_t(z) = (n + t + 1)\{\Theta(D_t \oplus z) - \Theta(D_t)\}$.

Thus, the in-context influence is simply a scaled one-step change in the readout. The following identity makes its connection to functional coherence explicit.

Lemma 2.1 (Response identity and coherence)**.** *For the readout process $(\theta_t)_{t\geq 0}$,*

$$\theta_{t+1} - \theta_t = \frac{\psi_t(Z_{t+1})}{n+t+1}, \qquad \mathbb{E}_{Q_t}[\theta_{t+1} - \theta_t \mid \mathcal{F}_t] = \frac{\mathbb{E}_{Q_t}[\psi_t(Z_{t+1}) \mid \mathcal{F}_t]}{n+t+1}. \tag{7}$$

In particular, $(\theta_t)_{t\geq 0}$ is functionally coherent if and only if $\mathbb{E}_{Q_t}[\psi_t(Z_{t+1}) \mid \mathcal{F}_t] = 0$ for every $t \geq 0$.

See Appendix B.1 for the proof. Hence functional coherence is exactly a zero-mean condition on the in-context influence, without requiring differentiability of Θ or a one-step response approximation.

3 Repairing Forecasters for Coherence

3.1 KL-Minimal Moment Repair

Mean repair. Lemma 2.1 shows that functional coherence at a given context can be enforced by choosing an imputation law Q_t such that $\mathbb{E}_{Q_t}[\psi_t] = 0$. In general, the base imputation law q_t need not satisfy this condition. We therefore seek a repaired imputation law that removes the conditional drift while deviating as little as possible from q_t. Specifically, we consider

$$Q_t^\star = \operatorname*{arg\,min}_{Q \ll q_t} D_{\mathrm{KL}}[Q \| q_t] \qquad \text{subject to} \qquad \mathbb{E}_Q[\psi_t] = 0. \tag{8}$$

Here, we restrict ourselves to the case where q_t has finite support, i.e., $\mathrm{supp}(q_t) = \{w_1, \ldots, w_K\}$. This setting naturally covers the finite candidate imputation laws used with the frozen forecasters of primary interest to us, including PFNs and LLMs. For imputation laws with more general support, an analogous repair can be considered under additional regularity conditions; See Appendix C.3.

Under the finite-support assumption, the repair problem above becomes a finite-dimensional information projection onto the set of distributions satisfying the coherence constraint (Csiszár, 1975). The feasibility of eq. (8) is determined by the geometry of the candidate responses $\{\psi_t(w_k)\}_{k=1}^K$.

Theorem 3.1 (KL-minimal functional coherence repair)**.** *Let $\mathcal{C}_t := \mathrm{conv}\{\psi_t(w_k)\}_{k=1}^K$ be the set of mean responses attainable by imputation laws supported on $\mathrm{supp}(q_t)$. The repair problem eq. (8) has a solution $Q_t^\star$ if and only if $0 \in \mathcal{C}_t$. When it exists, $Q_t^\star$ is unique. Furthermore, if $0 \in \mathrm{ri}(\mathcal{C}_t)$, then $Q_t^\star$ assigns positive probability to every $w_k \in \mathrm{supp}(q_t)$ and admits the exponentially tilted form*

$$Q_t^\star(w_k) = q_t(w_k) \exp\{\langle \eta_t, \psi_t(w_k)\rangle - \Lambda_t(\eta_t)\}, \quad \textit{where} \quad \Lambda_t(\eta) := \log \sum_{k=1}^K q_t(w_k) e^{\langle \eta, \psi_t(w_k)\rangle},$$

where $\eta_t \in \mathrm{span}\{\psi_t(w_k) - \psi_t(w_l)\}_{k,l=1}^K$ is the unique solution to $\nabla \Lambda_t(\eta_t) = 0$. Conversely, a finite full-support exponential tilt satisfying $\mathbb{E}_Q[\psi_t] = 0$ exists only if $0 \in \mathrm{ri}(\mathcal{C}_t)$.

See Appendix B.2 for the proof. We say that q_t is *mean repairable* if $0 \in \mathrm{ri}(\mathcal{C}_t)$, and call a sequence $(q_t)_{t=0}^{H-1}$ mean repairable if each q_t is mean repairable. When q_t is mean repairable and $Q_t^\star$ is its mean repair, $Q_t^\star$ eliminates the one-step conditional drift while retaining full support on $\mathrm{supp}(q_t)$.

Mean-covariance repair. Mean repair removes conditional drift but does not control the covariance of the resulting response, which can grow rapidly along the continuation. We therefore extend the repair problem in eq. (8) with an additional constraint on the step-wise local covariance. Specifically, let $\Gamma_t^\star \in \mathbb{S}_+^d$ be an $\mathcal{F}_t$-measurable target covariance. Then we consider

$$Q_t^\star = \arg\min_{Q \ll q_t} D_{\mathrm{KL}}[Q \| q_t] \quad \text{subject to} \quad \mathbb{E}_Q[\psi_t] = 0, \quad \mathbb{E}_Q[\psi_t \psi_t^\top] = \Gamma_t^\star. \tag{9}$$

The same information-projection argument naturally extends to the augmented moment statistic $\Psi_t(z) \coloneqq (\psi_t(z), \psi_t(z)\psi_t(z)^\top)$, yielding the following analogue of Theorem 3.1.

Theorem 3.2 (KL-minimal mean-covariance repair)**.** *Let $\mathcal{C}_t' \coloneqq \mathrm{conv}\{\Psi_t(w_k)\}_{k=1}^K$ be the augmented response set. The repair problem eq. (9) has a solution $Q_t^\star$ if and only if $(0, \Gamma_t^\star) \in \mathcal{C}_t'$. When it exists, $Q_t^\star$ is unique. If $(0, \Gamma_t^\star) \in \mathrm{ri}(\mathcal{C}_t')$, $Q_t^\star$ assigns positive probability to every $w_k \in \mathrm{supp}(q_t)$ and admits the exponentially tilted form*

$$\begin{aligned} Q_t^\star(w_k) &= q_t(w_k) \exp\left\{\langle \lambda_t, \psi_t(w_k)\rangle + \langle B_t, \psi_t(w_k)\psi_t(w_k)^\top\rangle - \Lambda_t'(\lambda_t, B_t)\right\} \\ \textit{where } \Lambda_t'(\lambda, B) &\coloneqq \log \sum_{k=1}^K q_t(w_k) \exp\left\{\langle \lambda, \psi_t(w_k)\rangle + \langle B, \psi_t(w_k)\psi_t(w_k)^\top\rangle\right\}. \end{aligned} \tag{10}$$

The natural parameter $(\lambda_t, B_t) \in \mathrm{span}\{\Psi_t(w_k) - \Psi_t(w_l)\}_{k,l=1}^K$ is the unique solution of

$$\nabla_\lambda \Lambda_t'(\lambda_t, B_t) = 0, \quad \nabla_B \Lambda_t'(\lambda_t, B_t) = \Gamma_t^\star. \tag{11}$$

Under $Z_{t+1} \mid \mathcal{F}_t \sim Q_t^\star$, $\mathrm{Cov}(\theta_{t+1} - \theta_t \mid \mathcal{F}_t) = \Gamma_t^\star/(n+t+1)^2$. Conversely, a finite full-support exponential tilt satisfying eq. (9) exists only if $(0, \Gamma_t^\star) \in \mathrm{ri}(\mathcal{C}_t')$.

See Appendix B.3 for the proof. We say that q_t is *mean-covariance repairable* if $(0, \Gamma_t^\star) \in \mathrm{ri}(\mathcal{C}_t')$, and a sequence $(q_t)_{t=0}^{H-1}$ is mean-covariance repairable if each q_t is mean-covariance repairable.

The target covariance can be chosen in several ways. We use $\Gamma_t^\star = \mathrm{Cov}_{q_t}(\psi_t)$ as the default target to preserve the response scale of the base imputation law. This target can be computed directly from the current responses under q_t before repair. If zero mean is feasible but this reference covariance is not jointly attainable, we project it onto the attainable covariance set; if the resulting target lies on the boundary of the attainable mean-covariance set and a finite full-support tilt is required, we apply the interior adjustment described in Appendix E.2. For PFN-like forecasters, the repair can preserve the prescribed covariate marginal $G_t(\cdot \mid D_t)$ while reweighting only the conditional response law. We call this *conditional repair*; See Appendix C.1 for the formulation.

3.2 Approximate Repair

Exact moment constraints can be infeasible or attainable only on the boundary. When the zero-mean constraint itself is infeasible, no imputation law $Q \ll q_t$ can eliminate the conditional drift completely. We therefore relax the moment condition in eq. (8) to control the trade-off between drift and deviation from the base imputation law. Specifically, for $\alpha > 0$, define

$$Q_{t,\alpha} = \arg\min_{Q \ll q_t} \left\{ \frac{1}{2} \|\mathbb{E}_Q[\psi_t]\|^2 + \alpha D_{\mathrm{KL}}[Q \| q_t] \right\}. \tag{12}$$

The first term penalizes the drift, while the KL term regularizes the repaired law toward the base imputation law. Thus, larger α favors staying closer to q_t, whereas smaller α favors reducing the drift. For mean-covariance repair, we additionally penalize the second-moment residual,

$$Q_{t,\alpha,\beta} = \arg\min_{Q \ll q_t} \left\{ \frac{1}{2} \|\mathbb{E}_Q[\psi_t]\|^2 + \frac{\beta}{2} \left\|\mathbb{E}_Q[\psi_t \psi_t^\top] - \Gamma_t^\star\right\|_{\mathrm{F}}^2 + \alpha D_{\mathrm{KL}}[Q \| q_t] \right\}. \tag{13}$$

Here, $\beta > 0$ weights the second-moment residual. For conditional repair, the same objectives are minimized over laws of the form $Q_t(u_a, y) = g_{t,a} R_t(y \mid u_a)$, so that the covariate marginal G_t remains unchanged. The following proposition quantifies the drift under approximate mean repair and shows that the relaxed solution approaches the best achievable repair as $\alpha \downarrow 0$.

Proposition 3.1. *Define $d_t := \min_{Q \ll q_t} \|\mathbb{E}_Q[\psi_t]\|$, where under conditional repair the minimization is restricted to laws preserving G_t. Let $Q_t^\dagger$ denote the KL-minimizing law among all laws attaining the minimum residual d_t. Then $Q_{t,\alpha}$ is unique and satisfies*

$$d_t^2 \le \|\mathbb{E}_{Q_{t,\alpha}}[\psi_t]\|^2 \le d_t^2 + 2\alpha D_{\mathrm{KL}}[Q_t^\dagger \| q_t]. \tag{14}$$

Moreover, $Q_{t,\alpha} \to Q_t^\dagger$ as $\alpha \downarrow 0$.

See Appendix B.4 for the proof and Appendix C.2 for the mean-covariance and conditional extensions. For any selected law Q_t, define its achieved moments by $b_t = \mathbb{E}_{Q_t}[\psi_t]$ and $\Gamma_t = \mathrm{Cov}_{Q_t}(\psi_t) = \mathbb{E}_{Q_t}[\psi_t \psi_t^\top] - b_t b_t^\top$. The following proposition relates the step-wise drifts to the total drift accumulated over the imputations.

Proposition 3.2. *Suppose $\theta_t \to \theta_\infty^Q$ in L^1 and $\sum_{t=0}^\infty \frac{\mathbb{E}[\|b_t\| | \mathcal{F}_0]}{n+t+1} < \infty$. Then,*

$$\mathbb{E}[\theta_\infty^Q - \theta_0 \,|\, \mathcal{F}_0] = \sum_{t=0}^\infty \frac{\mathbb{E}[b_t \,|\, \mathcal{F}_0]}{n+t+1}, \qquad \big\|\mathbb{E}[\theta_\infty^Q - \theta_0 \,|\, \mathcal{F}_0]\big\| \le \sum_{t=0}^\infty \frac{\mathbb{E}[\|b_t\| \,|\, \mathcal{F}_0]}{n+t+1}. \tag{15}$$

See Appendix B.5 for the proof. Under approximate repair, the step-wise drifts accumulate along the continuation, weighted by $(n+t+1)^{-1}$.

4 Gaussian Limits and Finite-Horizon Uncertainty

4.1 Finite-Horizon Covariance and Gaussian Limits

Let $H \ge 1$ be a finite horizon length. Define the finite-horizon centered covariance and horizon scaling factor

$$S_{n,H} := \sum_{t=0}^{H-1} \frac{\mathbb{E}[\Gamma_t \mid \mathcal{F}_0]}{(n+t+1)^2}, \qquad r_{n,H}^2 := \sum_{t=0}^{H-1} (n+t+1)^{-2}. \tag{16}$$

Thus, $S_{n,H}$ accumulates the achieved conditional covariances along the continuation.

Proposition 4.1 (Finite-horizon covariance identity)**.** *Assume that θ_0 and $(\psi_t(Z_{t+1}))_{t=0}^{H-1}$ are square integrable and exact mean repair is applied for $t = 0, \dots, H-1$. Then $\mathbb{E}[\theta_H \mid \mathcal{F}_0] = \theta_0$ and $\mathrm{Cov}(\theta_H \mid \mathcal{F}_0) = S_{n,H}$.*

See Appendix B.6 for the proof. Under approximate repair, $S_{n,H}$ still aggregates the centered conditional covariances, but need not coincide with the endpoint covariance $\mathrm{Cov}(\theta_H \mid \mathcal{F}_0)$. The endpoint covariance additionally contains contributions from the accumulated residual drift and its cross-covariance with the centered component. Proposition D.3 describes the decomposition.

If exact mean repair is applied at every step and the process $(\theta_t)_{t\ge0}$ is L^2-bounded, then $\theta_t \to \theta_\infty^Q$ almost surely and in L^2 for some random limit satisfying $\mathbb{E}[\theta_\infty^Q \mid \mathcal{F}_t] = \theta_t$. The conditional law of θ_∞^Q therefore admits a readout-level martingale posterior interpretation under the selected continuation. Moreover, $\mathrm{Cov}(\theta_\infty^Q \mid \mathcal{F}_0) = S_{n,H} + \mathbb{E}\big[\mathrm{Cov}(\theta_\infty^Q \mid \mathcal{F}_H) \mid \mathcal{F}_0\big]$. Thus, $S_{n,H}$ is the contribution of the first H updates to the covariance of θ_∞^Q. See Proposition D.2 for the full decomposition.

We also study how the readout process $(\theta_t)_{t\ge0}$ behaves as the observed context size n grows. Gaussian limits for related martingale posterior and predictive processes have been studied in prior works (Fong & Yiu, 2026; Fortini et al., 2026). The following result shows that, under suitable regularity conditions, the normalized fluctuation of the limit θ_∞^Q around θ_0 is asymptotically Gaussian.

Theorem 4.1. *Consider a sequence of observed contexts with $n \to \infty$, and define $r_n^2 := \sum_{t=0}^\infty (n+t+1)^{-2}$. Suppose that, for each n, exact mean repair is applied at every step, $\theta_t \to \theta_\infty^Q$ in L^2, and for a deterministic positive-semidefinite matrix Ω_∞,*

$$r_n^{-2} \sum_{t=0}^\infty \frac{\Gamma_t}{(n+t+1)^2} \xrightarrow{p} \Omega_\infty. \tag{17}$$

If the corresponding martingale difference array satisfies the conditional Lindeberg condition, then

$$r_n^{-1}\big(\theta_\infty^Q - \theta_0\big) \xrightarrow{d} \mathcal{N}(0, \Omega_\infty). \tag{18}$$

See Appendix D.1 for the proof and the precise Lindeberg argument.

4.2 From current to future covariance

Theorem 4.1 is encouraging: for sufficiently large context size n, the uncertainty of the limiting readout can be characterized through its Gaussian limit. Turning this asymptotic result into a practical uncertainty estimate, however, requires addressing two bottlenecks.

First, predictive resampling can only be carried out to a finite horizon H, so the practical endpoint is θ_H rather than the limiting readout θ^Q_∞. We therefore consider a scalar-valued readout and a deterministic horizon H_n that grows with n. Under suitable martingale CLT conditions,

$$S_{n,H_n}^{-1/2}(\theta_{H_n} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1), \tag{19}$$

so S_{n,H_n} provides the asymptotic variance scale for the finite-horizon readout θ_{H_n} by capturing the accumulated covariance of future martingale increments. See Proposition D.1 for precise conditions.

Second, evaluating S_{n,H_n} requires the step-wise covariances $(\Gamma_t)_{t=0}^{H_n-1}$ along the continuation, and hence requires future rollouts. At the initial context, only Γ_0 is immediately available. We therefore use $r^2_{n,H_n}\Gamma_0$ as a plug-in approximation. For $\Gamma_0 > 0$, write the discrepancy from the true finite-horizon variance in terms of the correction factor

$$\kappa_{n,H_n} := \frac{S_{n,H_n}}{r^2_{n,H_n}\Gamma_0} = \sum_{t=0}^{H_n-1} \frac{\mathbb{E}[\Gamma_t \,|\, \mathcal{F}_0]}{r^2_{n,H_n}\Gamma_0(n+t+1)^2}. \tag{20}$$

The value $\kappa_{n,H_n} = 1$ means that the weighted average future covariance agrees with the plug-in approximation, while values above or below one quantify the corresponding underestimation or overestimation. Our goal is therefore to estimate κ_{n,H_n} directly from information available at the initial context. In the following sections, we demonstrate that future covariance estimation can be amortized by learning a neural predictor of κ_{n,H_n} from rollouts generated by a pretrained forecaster. We refer to this strategy as *Amortized Covariance Estimation* (ACE).

Let $\widehat{\kappa}_{n,H_n}$ denote the ACE prediction of the correction factor. Under the finite-horizon CLT conditions, if $\widehat{\kappa}_{n,H_n}/\kappa_{n,H_n} \xrightarrow{p} 1$, Slutsky's theorem justifies the plug-in Gaussian approximation

$$\theta_{H_n} \approx \mathcal{N}\left(\theta_0,\, \widehat{\kappa}_{n,H_n} r^2_{n,H_n}\Gamma_0\right), \tag{21}$$

without explicit predictive-resampling rollouts at inference time. See Appendix D.2 for details.

Our primary ACE analysis targets $S_{n,H}$ as it is the finite-horizon variation singled out by martingale formulation and equals the endpoint covariance under exact repair. For downstream distributional prediction under approximate repair, the endpoint mean and variance can additionally depend on accumulated residual drift. In Section 5 and Appendix G, we study a practical extension that amortizes the endpoint mean and variance to construct a Gaussian approximation to the readout distribution.

5 Experiments with TabPFN Classifiers

We instantiate our framework with frozen TabPFN classifiers (Hollmann et al., 2023; 2025). The experiments address three questions: (1) whether the proposed mean and mean-covariance repairs are feasible in practice, (2) whether coherence repair improves uncertainty induced by predictive resampling, and whether the forecaster's predictive probability is preferable to a prespecified parametric projection as the target, (3) whether ACE can predict finite-horizon variation and support a Gaussian approximation to future readout uncertainty without future sequential rollouts.

5.1 Repairability and Local Moment Control

We first evaluate whether the proposed moment repairs are attainable in practical frozen TabPFN states. For each continuation state, we compute the candidate imputation law and test the corresponding moment constraints. We consider mean repair, which enforces $\mathbb{E}_Q[\psi_t] = 0$, and mean-covariance repair, which additionally matches a target local covariance. We use the base imputation covariance $\Gamma_t^{\mathrm{ref}} = \mathrm{Cov}_{q_t}(\psi_t)$ when it is jointly attainable with zero mean, and otherwise project it onto the attainable zero-mean covariance set described in Appendix E.2. The repair is performed by reweighting the candidate imputation law while keeping the forecaster frozen.

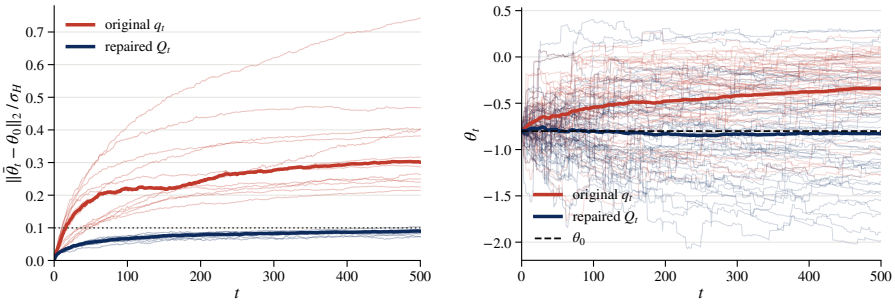

Figure 1: TabMGP-aligned coherence repair on 8 binary setups; see Figure 2 for the full figure. Left: cumulative drift for the original q_t and repaired Q_t laws, normalized by the terminal target standard deviation σ_H; the dotted line indicates the Monte Carlo floor, $1/\sqrt{100}$. Right: representative dataset (sepsis) trajectories showing reduced systematic drift after repair.

Results. Mean repair is feasible on 93.6% of evaluated states. Exact mean-covariance repair with the raw local covariance target is feasible on 71.6% of states.When the raw covariance is unattainable, projection onto the attainable zero-mean covariance set enables exact mean-covariance repair on 93.5% of states. On exact mean-covariance repair steps, both constraints are satisfied to numerical precision, with median mean and covariance residuals of 1.4×10^{-16} and 4.4×10^{-16}, respectively. The repaired law has a median $D_{\mathrm{KL}}[Q_t \| q_t]$ of 0.214.

Table 1: Repairability and local moment control across continuation states.

Metric	Value
Mean feasibility	93.6%
Mean-cov. feasibility, raw Γ	71.6%
Mean-cov. feasibility, projected	93.5%
Mean residual, median	1.4×10^{-16}
Covariance residual, median	4.4×10^{-16}
$D_{\mathrm{KL}}[Q_t \| q_t]$, median	0.214

5.2 Comparison with TabMGP and Predictive-Probability Readouts

We evaluate coherence repair under the same loss-defined functional considered in TabMGP (Ng et al., 2026). For the completed empirical distribution F_t, we use $\theta_t = \arg\min_\vartheta \mathbb{E}_{F_t}[\ell(Z, \vartheta)]$, with logistic loss for binary classification. The base imputation law q_t generates the continuation, whereas the logistic fit defines only the readout θ_t. We use the same target functional and empirical covariate law G_t as TabMGP and compare the continuation under q_t with its repaired Q_t. Rate denotes 95% joint coverage, Size the variance trace of the distribution, and Drift the endpoint displacement.

Results. Under the base continuation law q_t, the TabMGP-aligned loss-defined readout exhibits substantial accumulated displacement. The repaired law Q_t reduces the mean normalized endpoint drift from 0.38 to 0.08, with the reduction observed across all eight datasets. At the same time, 95% joint coverage and distributional size remain broadly comparable between q_t and Q_t. Thus, repair reduces self-induced movement of the chosen readout while largely preserving the overall scale of the induced uncertainty.

Table 2: Results on the eight real-data.

	TabMGP q_t			Repaired Q_t		
Dataset	Rate	Size	Drift	Rate	Size	Drift
rice	0.83	4.91	0.40	**0.97**	**4.82**	**0.07**
sepsis	**0.98**	**1.21**	0.74	1.00	1.91	**0.09**
banknote	**1.00**	**1.19**	0.32	1.00	1.22	**0.08**
mozilla	**0.98**	**1.33**	0.26	0.96	1.38	**0.07**
skin	1.00	1.45	0.40	**1.00**	**1.23**	**0.09**
blood	0.93	1.08	0.47	**0.98**	**1.29**	**0.09**
phoneme	0.96	1.53	0.23	**0.95**	**1.37**	**0.09**
telescope	1.00	7.86	0.22	**0.97**	**6.38**	**0.09**
Mean	0.96	2.57	0.38	**0.98**	**2.45**	**0.08**

5.3 ACE for Finite-Horizon Gaussian Uncertainty

We compare ACE with Margin, the best output-only heuristic; Hidden-only, a learned predictor from TabPFN's final-layer representation; and Local covariance, the one-step baseline $r^2_{n,H}\Gamma_0$. Hidden + Local combines the latter two learned inputs, while $S_{n,H}$ is a high-cost future-resampling reference. We evaluate fixed-query class probabilities across 30 classification datasets, tracking $\theta_t = p_F(Y = c^\star \mid x^\star, D_t)$ over repaired continuations. ACE predicts the finite-horizon variation $S_{n,H}$ from current prediction, candidate-response, repair, and horizon information.

Table 3: Prediction of finite-horizon variation at $H = 32$.

Method	Spearman	AUROC
Margin	.558	.757
Hidden-only	.699	.828
Local covariance	.789	.892
Hidden + Local	.792	.890
ACE	**.844**	**.931**
$S_{n,32}$ reference	.890	.955

Results. ACE outperforms heuristic, current local covariance, and learned hidden-state baselines, improving macro Spearman from 0.789 to 0.844 and AUROC from 0.892 to 0.931 at $H = 32$. This indicates ACE captures predictive signal about future uncertainty beyond both local covariance and hidden representations. See Appendix F.3 for details and per-dataset results. ACE above targets the centered finite-horizon variation $S_{n,H}$, which equals the endpoint variance under exact repair. Under approximate repair, accumulated residual drift can shift the endpoint mean and variance away from $(\theta_0, S_{n,H})$. We therefore consider an extension that predicts the finite-horizon endpoint mean and variance from current-state information and forms $\widehat{P}_{n,H} = \mathcal{N}(\widehat{\mu}_{n,H}, \widehat{V}_{n,H})$. The Gaussian family is CLT-motivated, but we evaluate it as a finite-horizon approximation rather than a test of the terminal martingale CLT. Centered ACE uses $\mathcal{N}(\theta_0, \widehat{S}_{n,H})$, while Amortized endpoint uses $\mathcal{N}(\widehat{\mu}_{n,H}, \widehat{V}_{n,H})$.

The 32-path empirical reference uses the rollout endpoints directly, while all methods are scored against 128 independent evaluation endpoints per state. We report the continuous ranked probability score (CRPS), 90% interval score (IS90), empirical 90% coverage (Cover), and per-instance inference time.

Table 4: Finite-horizon distribution approximation. Bold indicates the best rollout-free method.

Method	CRPS ↓	IS90 ↓	Cover	Time
Centered ACE	0.0603	0.558	0.931	∼0.7 s
Amortized endpoint	**0.0441**	**0.369**	0.841	0.74 s
32-path empirical	0.0388	0.297	0.853	1,693 s

Results. The endpoint extension of ACE amortizes the finite-horizon mean and variance from current-state information. It reduces CRPS from 0.0603 for centered ACE to 0.0441, approaching the 32-path empirical reference at 0.0388 without future sequential rollouts. This is consistent with the martingale-CLT intuition that the accumulated repaired increments admit a useful finite-horizon Gaussian approximation based on their first two moments. See Appendix F.4 for controls, moment ablations, per-dataset results, and secondary horizons, and Appendix F.5 for computational details.

6 Experiments with Reasoning LLMs

We now instantiate our framework for reasoning LLMs on multiple-choice question answering (MCQA) tasks. We treat each generated reasoning block as an imputation Z_t. Starting from an initial reasoning prefix D_0, we sample a reasoning block, append it to the prefix, and repeat this procedure for H steps. At each step, a separate answer branch produces the distribution π_t over answer options without modifying the reasoning prefix. We fix the initially most probable answer $a^\star := \arg\max_a \pi_0(a)$ and use its probability $\theta_t := \pi_t(a^\star)$ as the scalar readout. See Section 6 for the full experimental protocol. Our first goal is to test whether the finite-horizon variance S_H captures the variation accumulated over multiple future reasoning steps. At each prefix, we sample 16 candidate reasoning blocks with mean repair applied. Let e_H denote the centered endpoint fluctuation of θ_H after removing the accumulated residual drift. By construction, our finite-horizon variance predicts $\mathbb{E}[e_H^2 \mid D_0] = S_H$. We therefore compare the realized second moment $\mathbb{E}[e_H^2]$ against $\mathbb{E}[S_H]$ over independent continuations. The result is shown in Table 5. The realized centered variation closely matches S_H for both models and horizons, confirming that S_H captures variation accumulated over future reasoning. We next compare ACE against a baseline that extrapolates S_H from the local covariance of the same 16 candidate responses. With the same probing budget and no future rollouts, ACE reduces the squared prediction error against e_H^2 by 3.0-4.4%, indicating that current responses contain information about future variation beyond their local covariance. See Section 6 for additional results and details.

Table 5: Finite-horizon uncertainty over LLM reasoning.

	Llama	OLMo
$E[e_2^2]/E[S_2]$	0.97	1.00
$E[e_4^2]/E[S_4]$	1.02	0.94
ACE reduction, $H = 2$	**3.5%**	**3.0%**
ACE reduction, $H = 4$	**4.0%**	**4.4%**

7 Conclusion and Limitation

We introduced a framework for coherent predictive resampling of a chosen quantity of interest while keeping the forecaster fixed. Mean repair reweights the imputation law using exact in-context responses to control self-induced drift, with a natural extension to local covariance. We further characterized finite-horizon covariance and introduced *Amortized Covariance Estimation* (ACE) to predict future variation from current responses without future rollouts. Across frozen TabPFN and LLM forecasters, repair reduces drift and ACE improves uncertainty estimation. However, functional coherence is readout-specific and support-dependent, and does not imply distribution-level coherence or external calibration. ACE remains requires current candidate probing and specific to forecaster. Broader supports, larger models, and more transferable amortization remain are left for future work.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, draft portions of the manuscript, and assist with derivations and proofs of theorems and propositions. All AI-assisted mathematical arguments were checked by the authors against the stated assumptions, and all AI-assisted text was reviewed and revised by the authors. References, technical claims, and mathematical statements were verified by the authors against the original sources or independent derivations rather than accepted from AI outputs without verification. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies uncertainty estimation and coherence repair for frozen predictive models using existing benchmark datasets and pretrained models. It does not involve human-subject data collection or deployment in high-stakes settings. The uncertainty quantities studied here characterize behavior under specified continuation laws and should not be interpreted as guarantees of real-world calibration, reliability, or safety. We explicitly discuss limitations arising from the choice of readout, candidate support, forecaster, and evaluation setting, and do not claim that the proposed methods remove broader biases or risks of the underlying models.

REPRODUCIBILITY STATEMENT

We provide source code and experiment configurations for the reported results as supplementary material. Section 2-4 specify the framework, algorithms, assumptions, and uncertainty quantities, while Appendices B-E provide complete proofs and technical extensions. The experimental sections and appendices document the datasets, frozen forecasters, continuation settings, baselines, metrics, and additional per-dataset and ablation results for the TabPFN and LLM experiments. Together with the accompanying code, these materials are intended to support reproduction of the main theoretical and empirical results.

REFERENCES

Laura Battaglia, Stefano Cortinovis, Chris Holmes, David T Frazier, and Jack Jewson. Variational predictive resampling. *arXiv preprint arXiv:2605.11168*, 2026.

Patrizia Berti, Luca Pratelli, and Pietro Rigo. Limit theorems for a class of identically distributed random variables. *The Annals of Probability*, 32(3), 2004. ISSN 0091-1798. doi: 10.1214/009117904000000676. URL http://dx.doi.org/10.1214/009117904000000676.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pp. 146–158, 1975.

J. L. Doob. *Stochastic processes*, volume 101. New York Wiley, 1953.

F. Falck, Z. Wang, and C. Holmes. Is in-context learning in large language models bayesian? a martingale perspective. In *Proceedings of The 41st International Conference on Machine Learning (ICML 2024)*, 2024.

Edwin Fong and Andrew Yiu. Asymptotics for a class of parametric martingale posteriors. *Biometrika*, 113(2):asag007, 2026.

Edwin Fong, Chris Holmes, and Stephen G Walker. Martingale posterior distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1357–1391, 11 2023. ISSN 1369-7412. doi: 10.1093/jrsssb/qkad005. URL https://doi.org/10.1093/jrsssb/qkad005.

S. Fortini and S. Petrone. Exchangeability, prediction and predictive modeling in Bayesian statistics. *Statistical Science*, 40(1):40–67, 2025.

Sandra Fortini, Kenyon Ng, Sonia Petrone, Judith Rousseau, and Susan Wei. Uncertainty decomposition for bayes-filtered transformers via bayesian predictive inference, 2026. *URL https://arxiv.org/abs/2602.04596 v2*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *International conference on machine learning*, pp. 2404–2414. PMLR, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Peter Hall, Christopher C Heyde, ZW Birnbaum, and E Lukacs. *Martingale limit theory and its application*, volume 12. Elsevier, 1980.

N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter. TabPFN: a transformer that solves small tabular classification problems in a second. In *Proceedings of The 11th International Conference on Learning Representations (ICLR 2023)*, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

C. C. Holmes and S. G. Walker. Statistical inference with exchangeability and martingales. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2023.

C. T. Ionescu-Tulcea. Mesures dans les espaces produits. *Atti Accad. Naz. Lincei Rend.*, 7:208–211, 1949.

Tom Joy, Francesco Pinto, Ser-Nam Lim, Philip HS Torr, and Puneet K Dokania. Sample-dependent adaptive temperature scaling for improved calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14919–14926, 2023.

S. Kim, S. Lee, E. Fong, H. Lee, and J. Lee. From drift to coherence: stabilizing beliefs in LLMs. In *Proceedings of The 43rd International Conference on Machine Learning (ICML 2026)*, 2026.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Hyungi Lee, Eunggu Yun, Giung Nam, Edwin Fong, and Juho Lee. Martingale posterior neural processes. *arXiv preprint arXiv:2304.09431*, 2023.

S. Müller, N. Hollmann, S. P. Arango, J. Grabocka, and F. Hutter. Transformers can do bayesian inference. In *Proceedings of The 10th International Conference on Learning Representations (ICLR 2022)*, 2022.

Thomas Nagler and David Rügamer. Uncertainty quantification for prior-data fitted networks using martingale posteriors, 2026. URL https://arxiv.org/abs/2505.11325.

K. Ng, E. Fong, D. T. Frazier, J. Knoblauch, and S. Wei. TabMGP: martingale posterior with TabPFN. In *Proceedings of The 43rd International Conference on Machine Learning (ICML 2026)*, 2026.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd international conference on artificial intelligence and statistics*, pp. 837–848. PMLR, 2019.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

D. B. Rubin. The Bayesian bootstrap. *Annals of Statistics*, 9(1):130–134, 1981.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

S. G. Walker. A new look at Bayesian uncertainty. In *Handbook of Statistics, Chapter 4*, volume 47, pp. 83–101. Elsevier, 2022.

Xi Wang, James McInerney, Lequn Wang, and Nathan Kallus. Entropy after ¡/think¿ for reasoning model early exiting, 2026. URL https://arxiv.org/abs/2509.26522.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

Johnathan Xie, Annie S Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18128–18138, 2024.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. Dynamic early exit in reasoning models, 2025. URL https://arxiv.org/abs/2504.15895.

A Related Work

A.1 Predictive resampling and martingale posterior uncertainty

Predictive resampling provides a general framework for uncertainty quantification by repeatedly imputing future observations from a predictive distribution and evaluating the resulting variation of a quantity of interest. Martingale posterior distributions provide a theoretical framework for predictive-resampling uncertainty by defining a distribution over limiting functionals under coherent predictive sequences (Fong et al., 2023). Martingale posterior neural processes (Lee et al., 2023) extend this perspective to neural processes by learning predictive mechanisms for functional uncertainty. Parametric martingale posterior theory further studies centered parameter updates and their asymptotic properties (Fong & Yiu, 2026). Variational predictive resampling combines predictive resampling with variational predictive distributions to obtain scalable posterior approximations of uncertainty induced by predictive resampling (Battaglia et al., 2026).

These works provide principled uncertainty constructions from predictive continuations, typically under coherent predictive mechanisms or conditions such as conditionally identically distributed (c.i.d.) condition that enable martingale interpretations. In contrast, we consider a generic frozen forecaster whose own imputations may systematically move a chosen readout, without assuming that the original continuation law is coherent. Our work instead focuses on repairing the continuation law of a frozen forecaster to enforce functional coherence while preserving its response.

A.2 Coherence and uncertainty in frozen forecasters

Recent works have studied uncertainty and coherence issues arising from predictive resampling with pretrained forecasters. For tabular foundation models, TabMGP (Ng et al., 2026) applies martingale posterior ideas to TabPFN (Hollmann et al., 2023) by using predictive resampling to infer uncertainty over loss-defined parameters. Related work has also explored uncertainty quantification for prior-data fitted networks through martingale posterior constructions (Nagler & Rügamer, 2026). These approaches demonstrate that uncertainty induced by predictive resampling can be meaningfully studied in pretrained models, while focusing on specific parameterizations or predictive targets.

Similar questions about predictive stability and uncertainty have recently been investigated in language models. Falck et al. (2024) analyze in-context learning from a martingale perspective, while Kim et al. (2026) study predictive drift in LLM reasoning and proposes stabilization strategies for sequential belief updates. Fortini et al. (2026) further investigate uncertainty decomposition in transformer models through Bayesian predictive inference. In contrast to approaches that stabilize or interpret predictive behavior through prompting, sampling strategies, or model-specific constructions, we keep the frozen forecaster unchanged and repair the continuation law associated with a chosen readout to enforce functional coherence.

A.3 Amortized uncertainty estimation and calibration

Amortized inference reduces the cost of repeated uncertainty estimation by learning mappings from observations to approximate inference quantities. Neural posterior estimation and simulation-based inference learn approximate posterior quantities from simulated observations, avoiding repeated inference procedures at deployment (Papamakarios et al., 2019; Greenberg et al., 2019). Amortized variational inference similarly replaces repeated optimization with learned inference networks that approximate posterior distributions (Kingma & Welling, 2013). In contrast, ACE does not amortize a posterior approximation, but learns the finite-horizon covariance correction induced by repaired predictive continuations from current response information.

Calibration methods address a complementary problem of aligning predictive confidence with empirical reliability. Temperature scaling calibrates neural network confidence through global logit rescaling (Guo et al., 2017), while adaptive temperature methods extend this idea by predicting input-dependent scaling factors for individual examples or language models (Joy et al., 2023; Xie et al., 2024). These approaches improve confidence calibration but preserve prediction rankings because they only rescale logits. Our directional LLM experiments instead use signed local responses to capture option-specific belief changes, enabling corrections beyond confidence rescaling.

B PROOFS

We use the notation of Sections 2 and 4. All statements about conditional expectations hold almost surely. We use the Euclidean norm for vectors and the Frobenius inner product $\langle A, B\rangle_{\mathrm{F}} = \mathrm{tr}(A^\top B)$ for matrices. The selected laws are assumed to be measurable probability kernels.

B.1 PROOF OF LEMMA 2.1

Proof. Since $D_{t+1} = D_t \oplus Z_{t+1}$, Definition 2.1 gives

$$\psi_t(Z_{t+1}) = (n+t+1)\{\Theta(D_t \oplus Z_{t+1}) - \Theta(D_t)\} = (n+t+1)(\theta_{t+1} - \theta_t). \tag{22}$$

This proves the pathwise identity. By integrability of the readout process,

$$\mathbb{E}[\theta_{t+1} - \theta_t \mid \mathcal{F}_t] = \frac{\mathbb{E}[\psi_t(Z_{t+1}) \mid \mathcal{F}_t]}{n+t+1} = \frac{\mathbb{E}_{Q_t}[\psi_t]}{n+t+1}. \tag{23}$$

Because θ_t is $\mathcal{F}_t$-measurable, this drift vanishes exactly when $\mathbb{E}[\theta_{t+1} \mid \mathcal{F}_t] = \theta_t$. Imposing this condition at every step therefore makes $(\theta_t)_{t\geq 0}$ an integrable martingale. □

B.2 PROOF OF THEOREM 3.1: MEAN REPAIR

Proof. (1) Feasibility and uniqueness.

A law $Q \ll q_t$ is identified with its probability vector $(Q(w_k))_{k=1}^K$. Its mean response is $\mathbb{E}_Q[\psi_t] = \sum_{k=1}^K Q(w_k)\psi_t(w_k)$, so the attainable mean set is exactly $\mathcal{C}_t = \mathrm{conv}\{\psi_t(w_k)\}_{k=1}^K$. Hence the zero-mean constraint is feasible if and only if $0 \in \mathcal{C}_t$.

When feasible, the constraint set is a nonempty compact convex subset of the finite probability simplex. Since $q_t(w_k) > 0$ on its support, $D_{\mathrm{KL}}[Q\|q_t]$ is continuous and strictly convex on this simplex, with the usual convention $0 \log 0 = 0$. Therefore, the constrained KL problem has a unique minimizer $Q_t^\star$.

(2) Relative interior and full support.

We first note that $u \in \mathcal{C}_t$ admits a representation $u = \sum_{k=1}^K p_k \psi_t(w_k)$ with $p_k > 0$ and $\sum_{k=1}^K p_k = 1$ if and only if $u \in \mathrm{ri}(\mathcal{C}_t)$. For the forward direction, a strictly positive convex combination cannot lie on a proper supporting face of $\mathcal{C}_t$.

For the reverse direction, let $\bar{\psi} = K^{-1} \sum_{k=1}^K \psi_t(w_k)$. Since $u \in \mathrm{ri}(\mathcal{C}_t)$ and $\bar{\psi} \in \mathcal{C}_t$, for sufficiently small $\varepsilon > 0$,

$$u' = \frac{u - \varepsilon\bar{\psi}}{1-\varepsilon} \in \mathcal{C}_t. \tag{24}$$

If p' is any probability vector attaining u', then

$$p = (1-\varepsilon)p' + \frac{\varepsilon}{K}\mathbf{1} \tag{25}$$

has strictly positive entries and attains u.

Now suppose $0 \in \mathrm{ri}(\mathcal{C}_t)$ and let R be a strictly positive feasible law. The KL minimizer $Q_t^\star$ must also have full support. Otherwise, for $Q_\varepsilon = (1-\varepsilon)Q_t^\star + \varepsilon R$, feasibility is preserved and

$$D_{\mathrm{KL}}[Q_\varepsilon\|q_t] - D_{\mathrm{KL}}[Q_t^\star\|q_t] = \varepsilon \log \varepsilon \sum_{k:Q_t^\star(w_k)=0} R(w_k) + O(\varepsilon) < 0 \tag{26}$$

for sufficiently small $\varepsilon > 0$, contradicting optimality.

(3) Exponential form and parameter uniqueness.

Since $Q_t^\star$ has full support, first-order optimality under the normalization and mean constraints gives finite $c \in \mathbb{R}$ and $\eta_t \in \mathbb{R}^d$ such that

$$\log \frac{Q_t^\star(w_k)}{q_t(w_k)} + 1 = c + \langle \eta_t, \psi_t(w_k)\rangle \qquad \text{for all } k. \tag{27}$$

After normalization,

$$Q_t^\star(w_k) = q_t(w_k)\exp\{\langle \eta_t, \psi_t(w_k)\rangle - \Lambda_t(\eta_t)\}. \tag{28}$$

For any η, let $Q_{t,\eta}(w_k) = q_t(w_k)\exp\{\langle \eta, \psi_t(w_k)\rangle - \Lambda_t(\eta)\}$. Differentiating Λ_t gives

$$\nabla\Lambda_t(\eta) = \mathbb{E}_{Q_{t,\eta}}[\psi_t], \qquad \nabla^2\Lambda_t(\eta) = \mathrm{Cov}_{Q_{t,\eta}}(\psi_t). \tag{29}$$

Thus the zero-mean constraint for $Q_t^\star$ is equivalent to $\nabla\Lambda_t(\eta_t) = 0$.

Let $L_t = \mathrm{span}\{\psi_t(w_k) - \psi_t(w_\ell) : k, \ell \in [K]\}$. Adding any vector in $L_t^\perp$ to η_t changes all exponents by the same constant and therefore leaves the normalized law unchanged. Hence η_t may be restricted to L_t. For every nonzero $v \in L_t$,

$$v^\top\nabla^2\Lambda_t(\eta)v = \mathrm{Var}_{Q_{t,\eta}}(\langle v, \psi_t\rangle) > 0. \tag{30}$$

Therefore Λ_t is strictly convex on the identifiable subspace L_t, so η_t is unique there. If $L_t = \{0\}$, feasibility implies $\psi_t(w_k) = 0$ for every k, and the unique restricted parameter is $\eta_t = 0$.

(4) Converse.

Conversely, any finite exponential tilt has full support. If it also satisfies $\mathbb{E}_Q[\psi_t] = 0$, then zero is a strictly positive convex combination of the candidate responses. By part (2), $0 \in \mathrm{ri}(\mathcal{C}_t)$. □

B.3 Proof of Theorem 3.2: Mean-Covariance Repair

Proof. (1) Feasibility and uniqueness.

Define the augmented statistic $\Psi_t(z) = (\psi_t(z), \psi_t(z)\psi_t(z)^\top)$ in $\mathbb{R}^d \times \mathbb{S}^d$, equipped with the product of the Euclidean and Frobenius inner products. The constraints in eq. (9) are equivalent to $\mathbb{E}_Q[\Psi_t] = (0, \Gamma_t^\star)$. Hence their attainable set is $\mathcal{C}_t' = \mathrm{conv}\{\Psi_t(w_k)\}_{k=1}^K$, and feasibility is equivalent to $(0, \Gamma_t^\star) \in \mathcal{C}_t'$. As in the proof of Theorem 3.1, compactness of the feasible set and strict convexity of KL give a unique minimizer $Q_t^\star$ whenever the constraint is feasible.

(2) Full-support quadratic tilt.

Suppose $(0, \Gamma_t^\star) \in \mathrm{ri}(\mathcal{C}_t')$. Applying the relative-interior and stationarity arguments of Theorem 3.1 to the augmented statistic Ψ_t shows that $Q_t^\star$ has full support and admits finite parameters $\lambda_t \in \mathbb{R}^d$ and $B_t \in \mathbb{S}^d$ such that

$$Q_t^\star(w_k) = q_t(w_k)\exp\big\{\langle \lambda_t, \psi_t(w_k)\rangle + \langle B_t, \psi_t(w_k)\psi_t(w_k)^\top\rangle_{\mathrm{F}} - \Lambda_t'(\lambda_t, B_t)\big\}. \tag{31}$$

For generic (λ, B), let $Q_{t,\lambda,B}$ denote this normalized quadratic tilt. Differentiating Λ_t' gives

$$\nabla_\lambda\Lambda_t'(\lambda, B) = \mathbb{E}_{Q_{t,\lambda,B}}[\psi_t], \qquad \nabla_B\Lambda_t'(\lambda, B) = \mathbb{E}_{Q_{t,\lambda,B}}[\psi_t\psi_t^\top], \tag{32}$$

where the matrix derivative is taken with respect to the Frobenius pairing. Thus the two moment constraints are equivalent to

$$\nabla_\lambda\Lambda_t'(\lambda_t, B_t) = 0, \qquad \nabla_B\Lambda_t'(\lambda_t, B_t) = \Gamma_t^\star. \tag{33}$$

(3) Parameter uniqueness.

Let $L_t' = \mathrm{span}\{\Psi_t(w_k) - \Psi_t(w_\ell) : k, \ell \in [K]\}$. Parameters orthogonal to L_t' change every exponent by the same constant and therefore leave the normalized law unchanged. Hence (λ_t, B_t) may be restricted to L_t'. For every nonzero $(u, A) \in L_t'$,

$$\big\langle (u, A), \nabla^2\Lambda_t'(\lambda, B)(u, A)\big\rangle = \mathrm{Var}_{Q_{t,\lambda,B}}\big(\langle u, \psi_t\rangle + \langle A, \psi_t\psi_t^\top\rangle_{\mathrm{F}}\big) > 0. \tag{34}$$

Therefore Λ_t' is strictly convex on the identifiable subspace L_t', so the natural parameter is unique there. If $L_t' = \{0\}$, the parameter space is the singleton $\{(0, 0)\}$, so uniqueness is immediate.

(4) Converse.

Conversely, any finite quadratic tilt has full support. If it satisfies both moment constraints, then $(0, \Gamma_t^\star)$ is a strictly positive convex combination of the augmented support points. By the relative-interior characterization used above, $(0, \Gamma_t^\star) \in \mathrm{ri}(\mathcal{C}_t')$.

(5) Increment covariance.

Under exact mean-covariance repair,

$$\mathbb{E}_{Q_t^\star}[\psi_t] = 0, \qquad \mathbb{E}_{Q_t^\star}[\psi_t \psi_t^\top] = \Gamma_t^\star. \tag{35}$$

Lemma 2.1 therefore gives

$$\mathrm{Cov}(\theta_{t+1} - \theta_t \mid \mathcal{F}_t) = \frac{\mathrm{Cov}_{Q_t^\star}(\psi_t)}{(n+t+1)^2} = \frac{\Gamma_t^\star}{(n+t+1)^2}. \tag{36}$$

□

B.4 Proof of Proposition 3.1: Approximate Repair

Proof. Let $\mathcal{Q}_t$ be the finite simplex of laws $Q \ll q_t$. Its attainable mean set is $\mathcal{C}_t = \mathrm{conv}\{\psi_t(w_k)\}_{k=1}^K$. Since $\mathcal{C}_t$ is compact and convex, the strictly convex function $m \mapsto \|m\|^2$ has a unique minimizer $m_t^\dagger \in \mathcal{C}_t$, with $d_t = \|m_t^\dagger\|$. The set $\{Q \in \mathcal{Q}_t : \mathbb{E}_Q[\psi_t] = m_t^\dagger\}$ is nonempty, compact, and convex. Since KL is strictly convex on $\mathcal{Q}_t$, it has a unique minimizer $Q_t^\dagger$ on this set.

For $\alpha > 0$, the objective in eq. (12) is continuous and strictly convex on $\mathcal{Q}_t$, so it has a unique minimizer $Q_{t,\alpha}$. Optimality relative to $Q_t^\dagger$ gives

$$\frac{1}{2}\|\mathbb{E}_{Q_{t,\alpha}}[\psi_t]\|^2 + \alpha D_{\mathrm{KL}}[Q_{t,\alpha}\|q_t] \le \frac{1}{2}d_t^2 + \alpha D_{\mathrm{KL}}[Q_t^\dagger\|q_t]. \tag{37}$$

Since $\|\mathbb{E}_{Q_{t,\alpha}}[\psi_t]\| \ge d_t$, this yields the two bounds in eq. (14). It also gives

$$D_{\mathrm{KL}}[Q_{t,\alpha}\|q_t] \le D_{\mathrm{KL}}[Q_t^\dagger\|q_t]. \tag{38}$$

For convergence, let $\alpha_j \downarrow 0$ and let $Q_{t,\alpha_{j_\ell}} \to \bar{Q}$ be any convergent subsequence, which exists by compactness. The residual bound and continuity imply $\|\mathbb{E}_{\bar{Q}}[\psi_t]\| = d_t$, while continuity of KL and the KL bound give

$$D_{\mathrm{KL}}[\bar{Q}\|q_t] \le D_{\mathrm{KL}}[Q_t^\dagger\|q_t]. \tag{39}$$

By uniqueness of $Q_t^\dagger$ on the minimum-residual set, $\bar{Q} = Q_t^\dagger$. Thus every subsequential limit is $Q_t^\dagger$, and hence $Q_{t,\alpha} \longrightarrow Q_t^\dagger$ as $\alpha \downarrow 0$.

Finally, $d_t = 0$ if and only if $0 \in \mathcal{C}_t$, including when the origin lies on the boundary of $\mathcal{C}_t$. Thus, an irreducible positive residual occurs exactly when $0 \notin \mathcal{C}_t$. □

B.5 Proof of Proposition 3.2: Residual Drift Accumulation

Proof. For every finite H, Lemma 2.1 and the tower property give

$$\mathbb{E}[\theta_H - \theta_0 \mid \mathcal{F}_0] = \sum_{t=0}^{H-1} \frac{\mathbb{E}[b_t \mid \mathcal{F}_0]}{n+t+1}. \tag{40}$$

By conditional Jensen's inequality and the assumed summability,

$$\sum_{t=0}^{\infty} \frac{\|\mathbb{E}[b_t \mid \mathcal{F}_0]\|}{n+t+1} \le \sum_{t=0}^{\infty} \frac{\mathbb{E}[\|b_t\| \mid \mathcal{F}_0]}{n+t+1} < \infty \qquad \text{a.s.} \tag{41}$$

Hence the series on the right converges almost surely.

Since $\theta_H \to \theta_\infty^Q$ in L^1, contractivity of conditional expectation gives

$$\mathbb{E}[\theta_H - \theta_0 \mid \mathcal{F}_0] \longrightarrow \mathbb{E}[\theta_\infty^Q - \theta_0 \mid \mathcal{F}_0] \qquad \text{in } L^1. \tag{42}$$

The almost-sure convergence of the series and uniqueness of limits in probability therefore yield

$$\mathbb{E}[\theta_\infty^Q - \theta_0 \mid \mathcal{F}_0] = \sum_{t=0}^{\infty} \frac{\mathbb{E}[b_t \mid \mathcal{F}_0]}{n+t+1}. \tag{43}$$

Finally, the triangle inequality and conditional Jensen's inequality give

$$\big\|\mathbb{E}[\theta_\infty^Q - \theta_0 \mid \mathcal{F}_0]\big\| \le \sum_{t=0}^{\infty} \frac{\mathbb{E}[\|b_t\| \mid \mathcal{F}_0]}{n+t+1}. \tag{44}$$

□

B.6 Proof of Proposition 4.1: Finite-Horizon Covariance Identity

Proof. Under exact mean repair, $b_t = 0$ for $t = 0, \dots, H-1$. Define $\Delta_{t+1} = \theta_{t+1} - \theta_t = \psi_t(Z_{t+1})/(n+t+1)$. The assumptions make these increments square integrable, and

$$\mathbb{E}[\Delta_{t+1} \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\Delta_{t+1}\Delta_{t+1}^\top \mid \mathcal{F}_t] = \frac{\Gamma_t}{(n+t+1)^2}. \tag{45}$$

For $0 \le s < t < H$, adaptedness and the tower property imply

$$\mathbb{E}[\Delta_{s+1}\Delta_{t+1}^\top \mid \mathcal{F}_0] = \mathbb{E}\big[\Delta_{s+1}\mathbb{E}[\Delta_{t+1}^\top \mid \mathcal{F}_t] \mid \mathcal{F}_0\big] = 0. \tag{46}$$

Thus the increments are conditionally orthogonal given $\mathcal{F}_0$. Since $\theta_H - \theta_0 = \sum_{t=0}^{H-1} \Delta_{t+1}$, $\mathbb{E}[\theta_H \mid \mathcal{F}_0] = \theta_0$. Since θ_0 is $\mathcal{F}_0$-measurable, conditional orthogonality gives

$$\begin{aligned} \mathrm{Cov}(\theta_H \mid \mathcal{F}_0) &= \sum_{t=0}^{H-1} \mathbb{E}[\Delta_{t+1}\Delta_{t+1}^\top \mid \mathcal{F}_0] \\ &= \sum_{t=0}^{H-1} \frac{\mathbb{E}[\Gamma_t \mid \mathcal{F}_0]}{(n+t+1)^2} \\ &= S_{n,H}. \end{aligned} \tag{47}$$

□

C Repair Extensions

C.1 Conditional Repair with Fixed Covariate Marginals

Fix a context D_t. Let $G_t(\cdot \mid D_t)$ have support $\{u_a\}_{a=1}^A$ with masses $g_{t,a} > 0$ summing to one, and let $\mathcal{Y}_{t,a}$ be the finite support of $\mathsf{F}(D_t, u_a)$. The base joint law is

$$q_t(u_a, y) = g_{t,a}\mathsf{F}(D_t, u_a)(y). \tag{48}$$

Let $\mathcal{Q}_{t|G}$ consist of laws $Q(u_a, y) = g_{t,a}R(y \mid u_a)$, where each $R(\cdot \mid u_a)$ is supported on $\mathcal{Y}_{t,a}$. Their KL divergence decomposes as

$$D_{\mathrm{KL}}[Q\|q_t] = \sum_{a=1}^A g_{t,a} D_{\mathrm{KL}}[R(\cdot \mid u_a)\|\mathsf{F}(D_t, u_a)], \tag{49}$$

and the attainable mean set is

$$\mathcal{C}_{t|G} = \sum_{a=1}^A g_{t,a}\,\mathrm{conv}\{\psi_t(u_a, y) : y \in \mathcal{Y}_{t,a}\}. \tag{50}$$

Proposition C.1 (Conditional mean repair)**.** *The KL-minimal zero-mean repair over $\mathcal{Q}_{t|G}$ exists if and only if $0 \in \mathcal{C}_{t|G}$ and is unique whenever it exists. It has full support if and only if $0 \in \mathrm{ri}(\mathcal{C}_{t|G})$. In that case, $Q_t^\star(u_a, y) = g_{t,a}R_t^\star(y \mid u_a)$, where*

$$R_t^\star(y \mid u_a) = \frac{\mathsf{F}(D_t, u_a)(y)\exp\{\langle \eta_t, \psi_t(u_a, y)\rangle\}}{\sum_{y' \in \mathcal{Y}_{t,a}} \mathsf{F}(D_t, u_a)(y')\exp\{\langle \eta_t, \psi_t(u_a, y')\rangle\}}. \tag{51}$$

The parameter satisfies $\nabla\Lambda_{t|G}(\eta_t) = 0$ for

$$\Lambda_{t|G}(\eta) = \sum_{a=1}^A g_{t,a}\log \sum_{y \in \mathcal{Y}_{t,a}} \mathsf{F}(D_t, u_a)(y)\exp\{\langle \eta, \psi_t(u_a, y)\rangle\}, \tag{52}$$

and is unique on the subspace

$$L_{t|G} = \mathrm{span}\{\psi_t(u_a, y) - \psi_t(u_a, y') : a = 1, \dots, A,\ y, y' \in \mathcal{Y}_{t,a}\}. \tag{53}$$

Proof. (1) Feasibility and uniqueness.

Each conditional probability vector ranges over a simplex, so its mean ranges over the corresponding convex hull. Taking the weighted sum gives eq. (50), proving the feasibility condition. The family $\mathcal{Q}_{t|G}$ is compact and convex. The cancellation of $g_{t,a}$ in the joint density ratio gives eq. (49), which is continuous and strictly convex because all base probabilities and weights $g_{t,a}$ are positive. Hence the constrained problem has a unique minimizer whenever feasible.

(2) Relative interior and full support.

Let $\bar{Q}(u_a, y) = g_{t,a}/|\mathcal{Y}_{t,a}|$ and $\bar{m} = \mathbb{E}_{\bar{Q}}[\psi_t]$. For $u \in \mathrm{ri}(\mathcal{C}_{t|G})$ and sufficiently small $\varepsilon \in (0, 1)$,

$$u' = \frac{u - \varepsilon \bar{m}}{1 - \varepsilon} \in \mathcal{C}_{t|G}. \tag{54}$$

If $Q' \in \mathcal{Q}_{t|G}$ attains u', then $(1-\varepsilon)Q' + \varepsilon\bar{Q}$ is strictly positive and attains u. Conversely, a positive conditional law can attain a supporting maximum only if every response within each covariate support attains its corresponding maximum. The supporting functional would then be constant on $\mathcal{C}_{t|G}$, so the supporting face cannot be proper. Thus a strictly positive feasible law exists exactly when $0 \in \mathrm{ri}(\mathcal{C}_{t|G})$. In that case, the boundary perturbation argument in Appendix B.2 shows that the KL minimizer is also strictly positive.

(3) Exponential form and parameter uniqueness.

First-order optimality uses a separate normalization multiplier for each covariate and a shared multiplier for the global mean constraint. This gives eq. (51). For generic η, let $R_{t,\eta}$ denote the corresponding conditional tilt. Differentiation gives

$$\begin{aligned} \nabla\Lambda_{t|G}(\eta) &= \sum_{a=1}^{A} g_{t,a}\mathbb{E}_{R_{t,\eta}(\cdot \mid u_a)}[\psi_t(u_a, Y)], \\ \nabla^2\Lambda_{t|G}(\eta) &= \sum_{a=1}^{A} g_{t,a}\mathrm{Cov}_{R_{t,\eta}(\cdot \mid u_a)}(\psi_t(u_a, Y)). \end{aligned} \tag{55}$$

The mean constraint is therefore $\nabla\Lambda_{t|G}(\eta_t) = 0$. If the Hessian quadratic form vanishes for $v \in L_{t|G}$, then $\langle v, \psi_t(u_a, y)\rangle$ is constant over y at every u_a. Thus $v \in L_{t|G}^{\perp}$, forcing $v = 0$. The Hessian is therefore positive definite on $L_{t|G}$. Components in $L_{t|G}^{\perp}$ cancel in the separate normalizers, so η_t may be restricted to $L_{t|G}$ and is unique there. If $L_{t|G} = \{0\}$, feasibility makes the mean constraint redundant and the unique restricted parameter is $\eta_t = 0$. □

Conditional mean-covariance repair. Replace ψ_t by Ψ_t and the target zero by $(0, \Gamma_t^\star)$. The attainable augmented moment set is

$$\mathcal{C}'_{t|G} = \sum_{a=1}^{A} g_{t,a}\,\mathrm{conv}\{\Psi_t(u_a, y) : y \in \mathcal{Y}_{t,a}\}. \tag{56}$$

Feasibility and finite full-support representability correspond to membership in $\mathcal{C}'_{t|G}$ and its relative interior, respectively. In the latter case, the conditional law is

$$R_t^\star(y \mid u_a) \propto \mathsf{F}(D_t, u_a)(y)\exp\left\{\langle \lambda_t, \psi_t(u_a, y)\rangle + \langle B_t, \psi_t(u_a, y)\psi_t(u_a, y)^\top\rangle_{\mathrm{F}}\right\}, \tag{57}$$

with normalization over y separately at each u_a. The same argument applies to $\Psi_t - (0, \Gamma_t^\star)$ in $\mathbb{R}^d \times \mathbb{S}^d$, as in Appendix B.3. The moment constraints are global under G_t, rather than imposed at every covariate separately. A single normalizer over all pairs (u_a, y) would generally change G_t.

Randomized candidate pools. Repair is relative to the realized candidate law, which may be an empirical law obtained from a native proposal rather than the proposal itself. For randomized pools, let $\mathcal{F}_t^+$ augment $\mathcal{F}_t$ with all candidate-generation randomness revealed before drawing Z_{t+1}. Conditional on $\mathcal{F}_t^+$, the realized candidate law, repaired law Q_t, and achieved moments are fixed, so the preceding conditional identities hold with $\mathcal{F}_t^+$ in place of $\mathcal{F}_t$. Under exact mean repair,

$$\mathbb{E}[\psi_t(Z_{t+1}) \mid \mathcal{F}_t^+] = 0 \quad \Longrightarrow \quad \mathbb{E}[\psi_t(Z_{t+1}) \mid \mathcal{F}_t] = 0 \tag{58}$$

by the tower property. Covariance identities based on realized-pool moments should likewise be interpreted conditionally on $\mathcal{F}_t^+$, including $\mathcal{F}_0^+$ at the initial state. Averaging over candidate-pool randomness is then a separate application of the law of total covariance.

C.2 Approximate Mean-Covariance and Conditional Repair

We extend Appendix B.4 to the joint objective in eq. (13). Let $\mathcal{Q}_t$ be either the unrestricted finite simplex or $\mathcal{Q}_{t|G}$, and fix $\alpha, \beta > 0$. With the Euclidean and Frobenius product norm, define $\chi_t(z) := \big(\psi_t(z), \sqrt{\beta}\{\psi_t(z)\psi_t(z)^\top - \Gamma_t^\star\}\big)$. Then

$$\|\mathbb{E}_Q[\chi_t]\|^2 = \|\mathbb{E}_Q[\psi_t]\|^2 + \beta\|\mathbb{E}_Q[\psi_t\psi_t^\top] - \Gamma_t^\star\|_{\mathrm{F}}^2. \tag{59}$$

Let $d_{t,\beta} = \min_{Q\in\mathcal{Q}_t} \|\mathbb{E}_Q[\chi_t]\|$, and let $Q_{t,\beta}^\dagger$ minimize KL among laws attaining this residual.

Proposition C.2 (Regularized joint moment repair)**.** *The minimizer $Q_{t,\alpha,\beta}$ exists, is unique, and is positive on the retained support. It satisfies*

$$d_{t,\beta}^2 \le \|\mathbb{E}_{Q_{t,\alpha,\beta}}[\chi_t]\|^2 \le d_{t,\beta}^2 + 2\alpha D_{\mathrm{KL}}[Q_{t,\beta}^\dagger \| q_t], \tag{60}$$

and $Q_{t,\alpha,\beta} \to Q_{t,\beta}^\dagger$ as $\alpha \downarrow 0$ with β fixed. Let $m = \mathbb{E}_{Q_{t,\alpha,\beta}}[\psi_t]$ and $E_2 = \mathbb{E}_{Q_{t,\alpha,\beta}}[\psi_t\psi_t^\top] - \Gamma_t^\star$. The unrestricted optimizer has the form

$$Q_{t,\alpha,\beta}(w_k) \propto q_t(w_k)\exp\left\{-\frac{\langle m, \psi_t(w_k)\rangle + \beta\langle E_2, \psi_t(w_k)\psi_t(w_k)^\top\rangle_{\mathrm{F}}}{\alpha}\right\}. \tag{61}$$

For conditional repair, the same exponent is normalized separately over responses at each covariate, preserving G_t.

Proof. (1) Existence, residual bound, and convergence.

The compactness and strict-convexity arguments in Appendix B.4 apply to χ_t on either family $\mathcal{Q}_t$. They give unique minimizers $Q_{t,\alpha,\beta}$ and $Q_{t,\beta}^\dagger$. Optimality gives

$$\frac{1}{2}\|\mathbb{E}_{Q_{t,\alpha,\beta}}[\chi_t]\|^2 + \alpha D_{\mathrm{KL}}[Q_{t,\alpha,\beta}\|q_t] \le \frac{1}{2}d_{t,\beta}^2 + \alpha D_{\mathrm{KL}}[Q_{t,\beta}^\dagger\|q_t]. \tag{62}$$

Since $\|\mathbb{E}_{Q_{t,\alpha,\beta}}[\chi_t]\| \ge d_{t,\beta}$, this proves eq. (60) and gives

$$D_{\mathrm{KL}}[Q_{t,\alpha,\beta}\|q_t] \le D_{\mathrm{KL}}[Q_{t,\beta}^\dagger\|q_t]. \tag{63}$$

The same compactness and continuity argument proves convergence to $Q_{t,\beta}^\dagger$ as in Appendix B.4.

(2) Full support.

If the minimizer has a zero coordinate, consider $Q_\varepsilon = (1-\varepsilon)Q_{t,\alpha,\beta} + \varepsilon q_t \in \mathcal{Q}_t$. The moment penalty changes by $O(\varepsilon)$. The total change in the objective is therefore

$$\alpha\varepsilon\log\varepsilon\left(\sum_{w:Q_{t,\alpha,\beta}(w)=0} q_t(w)\right) + O(\varepsilon) < 0 \tag{64}$$

for sufficiently small $\varepsilon > 0$, contradicting optimality.

(3) Exponential form.

For unrestricted repair, first-order optimality gives a scalar c such that

$$\langle m, \psi_t(w_k)\rangle + \beta\langle E_2, \psi_t(w_k)\psi_t(w_k)^\top\rangle_{\mathrm{F}} + \alpha\left(1 + \log\frac{Q_{t,\alpha,\beta}(w_k)}{q_t(w_k)}\right) = c. \tag{65}$$

Rearranging and normalizing gives eq. (61). For conditional repair, differentiation with respect to $R(y \mid u_a)$ multiplies both the moment and KL derivatives by $g_{t,a}$. Dividing by $g_{t,a} > 0$ gives the same exponent with a separate normalizer at each covariate. The moments m and E_2 remain global moments of the joint law. □

For mean repair alone, the same argument gives

$$Q_{t,\alpha}(w_k) \propto q_t(w_k)\exp\left\{-\frac{\langle \mathbb{E}_{Q_{t,\alpha}}[\psi_t], \psi_t(w_k)\rangle}{\alpha}\right\}. \tag{66}$$

The residual bound and convergence in Appendix B.4 also hold over $\mathcal{Q}_{t|G}$, using its minimum residual and KL-minimizing law. For joint repair, the achieved covariance satisfies

$$\begin{aligned} \mathrm{Cov}_{Q_{t,\alpha,\beta}}(\psi_t) - \Gamma_t^\star &= E_2 - mm^\top, \\ \|\mathrm{Cov}_{Q_{t,\alpha,\beta}}(\psi_t) - \Gamma_t^\star\|_{\mathrm{F}} &\le \|E_2\|_{\mathrm{F}} + \|m\|^2. \end{aligned} \tag{67}$$

The bound follows from $\|mm^\top\|_{\mathrm{F}} = \|m\|^2$. This centered covariance, not the uncentered second moment, enters $S_{n,H}$. In contrast, covariance projection in Appendix E.2 retains the zero-mean constraint and first changes the covariance target.

C.3 A General-Support Extension

The sequential response identity does not require finite support. For the KL representation, we give a sufficient extension for bounded responses.

Proposition C.3 (General-support mean repair)**.** *Suppose ψ_t is measurable and essentially bounded under q_t. Let $\mathcal{C}_t$ be the closed convex hull of its q_t-essential range, and let L_t be the linear space parallel to its affine hull. If $0 \in \mathrm{ri}(\mathcal{C}_t)$, there exists a unique $\eta_t \in L_t$ such that*

$$\frac{\mathrm{d}Q_t^\star}{\mathrm{d}q_t}(z) = \exp\{\langle \eta_t, \psi_t(z)\rangle - \Lambda_t(\eta_t)\}, \quad \textit{where} \quad \Lambda_t(\eta) = \log \int e^{\langle \eta, \psi_t(z)\rangle}\, q_t(\,\mathrm{d}z), \tag{68}$$

with $\mathbb{E}_{Q_t^\star}[\psi_t] = 0$, and $Q_t^\star$ uniquely minimizes $D_{\mathrm{KL}}[Q\|q_t]$ over all $Q \ll q_t$ satisfying $\mathbb{E}_Q[\psi_t] = 0$.

Proof. (1) Existence of a finite parameter.

Since zero belongs to the affine hull, $\psi_t(z) \in L_t$ almost surely under q_t. If $L_t = \{0\}$, then $\psi_t = 0$ almost surely, $\eta_t = 0$, and $Q_t^\star = q_t$. Otherwise, choose $M > 0$ with $\|\psi_t\| \le M$ almost surely and $\delta > 0$ such that the radius-δ ball in L_t is contained in $\mathcal{C}_t$. For each unit $v \in L_t$, $\delta v \in \mathcal{C}_t$. Hence the essential-range definition gives $q_t\{\langle v, \psi_t\rangle > \delta/2\} > 0$. Choose a finite $\delta/(4M)$-net $\{v_j\}_{j=1}^J$ of the unit sphere and set

$$c = \min_{1 \le j \le J} q_t\{\langle v_j, \psi_t\rangle > \delta/2\} > 0. \tag{69}$$

For any unit v, a net point v_j satisfies $|\langle v - v_j, \psi_t\rangle| \le \delta/4$ almost surely. Hence

$$q_t\{\langle v, \psi_t\rangle > \delta/4\} \ge c, \qquad \Lambda_t(rv) \ge r\delta/4 + \log c \quad (r > 0). \tag{70}$$

This lower bound is uniform over unit v. Thus Λ_t is coercive on L_t and, being continuous by boundedness, attains a finite minimum η_t there.

(2) Stationarity and parameter uniqueness.

Let $Q_{t,\eta}$ denote the normalized tilt at η. Boundedness permits differentiation under the integral, giving

$$\nabla\Lambda_t(\eta) = \mathbb{E}_{Q_{t,\eta}}[\psi_t], \qquad \nabla^2\Lambda_t(\eta) = \mathrm{Cov}_{Q_{t,\eta}}(\psi_t). \tag{71}$$

The gradient belongs to L_t, so restricted stationarity gives $\nabla\Lambda_t(\eta_t) = 0$. For nonzero $v \in L_t$, the projection $\langle v, \psi_t\rangle$ cannot be constant on the essential range, since otherwise $v \in L_t^\perp$. Because $Q_{t,\eta}$ and q_t are equivalent, the Hessian quadratic form is therefore strictly positive along every nonzero $v \in L_t$. Therefore Λ_t is strictly convex on L_t, proving uniqueness of η_t.

(3) KL minimality.

For any feasible $Q \ll q_t$ of finite KL, the bounded log density ratio gives

$$\begin{aligned} D_{\mathrm{KL}}[Q\|q_t] &= D_{\mathrm{KL}}[Q\|Q_t^\star] + \mathbb{E}_Q\left[\log \frac{\mathrm{d}Q_t^\star}{\mathrm{d}q_t}\right] \\ &= D_{\mathrm{KL}}[Q\|Q_t^\star] - \Lambda_t(\eta_t) \\ &= D_{\mathrm{KL}}[Q\|Q_t^\star] + D_{\mathrm{KL}}[Q_t^\star\|q_t]. \end{aligned} \tag{72}$$

Thus $Q_t^\star$ is the unique KL minimizer. Laws of infinite KL cannot improve its finite objective. □

The mean-covariance analogue follows by applying the argument to $\Psi_t - (0, \Gamma_t^\star)$ in $\mathbb{R}^d \times \mathbb{S}^d$ under the corresponding relative-interior condition. The finite-support boundary criterion does not extend in general. For example, if q_t is uniform on $(0, 1)$ and $\psi_t(z) = z$, zero lies in the closed response hull but no law $Q \ll q_t$ has zero mean. For unbounded responses, existence of a finite tilt requires additional exponential-integrability and dual-attainment assumptions.

D GAUSSIAN LIMITS AND COVARIANCE DECOMPOSITION

We use the selected continuation law and the notation of Section 4. For asymptotic statements, n indexes the initial contexts, and the dependence of θ_t, ψ_t, Γ_t, and $\mathcal{F}_t$ on n is suppressed. A CLT conditional on a fixed observed context or realized initial pool requires its assumptions under that conditional law, not merely after averaging over this information.

D.1 PROOF OF THEOREM 4.1

Proof. Let $X_{n,t+1} = \psi_t(Z_{t+1})/\{r_n(n+t+1)\}$. Exact mean repair gives

$$\mathbb{E}[X_{n,t+1} \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[X_{n,t+1}X_{n,t+1}^\top \mid \mathcal{F}_t] = \frac{\Gamma_t}{r_n^2(n+t+1)^2}. \tag{73}$$

The conditional Lindeberg assumption means that, for every $\varepsilon > 0$,

$$\sum_{t=0}^{\infty} \mathbb{E}\big[\|X_{n,t+1}\|^2 \mathbf{1}\{\|X_{n,t+1}\| > \varepsilon\} \mid \mathcal{F}_t\big] \xrightarrow{p} 0. \tag{74}$$

By Lemma 2.1 and the assumed L^2 convergence,

$$\sum_{t=0}^{\infty} X_{n,t+1} = \frac{\theta_\infty^Q - \theta_0}{r_n} \quad \text{in } L^2, \qquad \sum_{t=0}^{\infty} \mathbb{E}\|X_{n,t+1}\|^2 = \frac{\mathbb{E}\|\theta_\infty^Q - \theta_0\|^2}{r_n^2} < \infty. \tag{75}$$

The second identity follows from martingale-increment orthogonality.

For each n, choose a deterministic integer $m_n \geq n$ such that $\sum_{t=m_n}^{\infty} \mathbb{E}\|X_{n,t+1}\|^2 \leq n^{-2}$. The preceding summability makes this possible, and orthogonality gives

$$\mathbb{E}\left\| \sum_{t=m_n}^{\infty} X_{n,t+1} \right\|^2 \leq n^{-2}. \tag{76}$$

The omitted predictable covariance is positive semidefinite. Its expected trace is at most n^{-2}, so its operator norm also converges to zero in probability. Consequently, eq. (17) implies

$$\sum_{t=0}^{m_n-1} \mathbb{E}[X_{n,t+1}X_{n,t+1}^\top \mid \mathcal{F}_t] \xrightarrow{p} \Omega_\infty. \tag{77}$$

The truncated array also satisfies eq. (74), since its Lindeberg sum is bounded by the full sum.

For fixed $v \neq 0$, the projected array has predictable variance converging to $v^\top \Omega_\infty v$. Its Lindeberg condition follows from

$$\begin{aligned} &\sum_{t=0}^{m_n-1} \mathbb{E}\big[(v^\top X_{n,t+1})^2 \mathbf{1}\{|v^\top X_{n,t+1}| > \varepsilon\} \mid \mathcal{F}_t\big] \\ &\qquad \leq \|v\|^2 \sum_{t=0}^{m_n-1} \mathbb{E}\big[\|X_{n,t+1}\|^2 \mathbf{1}\{\|X_{n,t+1}\| > \varepsilon/\|v\|\} \mid \mathcal{F}_t\big] \xrightarrow{p} 0. \end{aligned} \tag{78}$$

If $v^\top \Omega_\infty v > 0$, the scalar martingale-array CLT gives the corresponding normal limit (Hall et al., 1980). If $v^\top \Omega_\infty v = 0$, write $V_n(v) = \sum_{t=0}^{m_n-1} \mathbb{E}[(v^\top X_{n,t+1})^2 \mid \mathcal{F}_t]$. Retaining only increments whose accumulated predictable variance, including that increment, is at most $a > 0$ gives a martingale sum with second moment at most a. It agrees with the original sum when $V_n(v) \leq a$, so

$$\Pr\left\{ \left| \sum_{t=0}^{m_n-1} v^\top X_{n,t+1} \right| > \varepsilon \right\} \leq \Pr\{V_n(v) > a\} + \frac{a}{\varepsilon^2}. \tag{79}$$

Taking $n \to \infty$ and then $a \downarrow 0$ proves convergence to zero in this case. The case $v = 0$ is immediate. Thus Cramér–Wold gives

$$\sum_{t=0}^{m_n-1} X_{n,t+1} \xrightarrow{d} \mathcal{N}(0, \Omega_\infty). \tag{80}$$

The omitted tail converges to zero in L^2, so Slutsky's theorem proves eq. (18). □

Integral comparison gives $(n+1)^{-1} \le r_n^2 \le n^{-1}$, hence $nr_n^2 \to 1$. This identifies the normalization scale but does not establish the quadratic-variation or Lindeberg assumptions.

D.2 Finite-Horizon Gaussian Approximation and ACE

Consider a scalar readout with deterministic finite horizons $H_n \to \infty$. For each n, assume exact mean repair through H_n, square integrability of θ_0 and increments, $0 < S_{n,H_n} < \infty$ almost surely.

Proposition D.1 (Finite-horizon martingale CLT)**.** *Suppose*

$$\frac{1}{S_{n,H_n}} \sum_{t=0}^{H_n-1} \frac{\Gamma_t}{(n+t+1)^2} \xrightarrow{p} 1, \tag{81}$$

and, for every $\varepsilon > 0$,

$$\frac{1}{S_{n,H_n}} \sum_{t=0}^{H_n-1} \frac{1}{(n+t+1)^2} \mathbb{E}\Big[\psi_t(Z_{t+1})^2 \mathbf{1}\Big\{|\psi_t(Z_{t+1})| > \varepsilon(n+t+1)\sqrt{S_{n,H_n}}\Big\} \mid \mathcal{F}_t\Big] \xrightarrow{p} 0. \tag{82}$$

Then $S_{n,H_n}^{-1/2}(\theta_{H_n} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1)$.

Proof. Let $X_{n,t+1} = \psi_t(Z_{t+1})/\{(n+t+1)\sqrt{S_{n,H_n}}\}$ for $0 \le t < H_n$. Since S_{n,H_n} is $\mathcal{F}_0$-measurable, its definition and exact repair give

$$\mathbb{E}\left[\sum_{t=0}^{H_n-1} X_{n,t+1}^2 \mid \mathcal{F}_0\right] = \frac{1}{S_{n,H_n}} \sum_{t=0}^{H_n-1} \frac{\mathbb{E}[\Gamma_t \mid \mathcal{F}_0]}{(n+t+1)^2} = 1. \tag{83}$$

The normalized increments are square integrable, and exact repair gives $\mathbb{E}[X_{n,t+1} \mid \mathcal{F}_t] = 0$. Their predictable variance and Lindeberg sums are precisely eqs. (81) and (82). The scalar martingale-array CLT (Hall et al., 1980) applies, and telescoping identifies their sum with $S_{n,H_n}^{-1/2}(\theta_{H_n} - \theta_0)$. □

The predictable variance sum in eq. (81) is random, whereas S_{n,H_n} is its conditional expectation given $\mathcal{F}_0$. This expectation identity alone does not imply the concentration or Lindeberg conditions.

Estimated covariance. Under Proposition D.1, suppose $\Gamma_0 > 0$ and let $\widehat{\kappa}_{n,H_n} > 0$ be the ACE prediction from the initial information. Set $\kappa_{n,H_n} = S_{n,H_n}/(r_{n,H_n}^2 \Gamma_0)$ and $\widehat{S}_{n,H_n} = r_{n,H_n}^2 \Gamma_0 \widehat{\kappa}_{n,H_n}$. If $\widehat{\kappa}_{n,H_n}/\kappa_{n,H_n} \xrightarrow{p} 1$, then Slutsky's theorem gives

$$\frac{\theta_{H_n} - \theta_0}{\sqrt{\widehat{S}_{n,H_n}}} = \left(\frac{S_{n,H_n}}{\widehat{S}_{n,H_n}}\right)^{1/2} \frac{\theta_{H_n} - \theta_0}{\sqrt{S_{n,H_n}}} \xrightarrow{d} \mathcal{N}(0,1). \tag{84}$$

The condition $\log \widehat{\kappa}_{n,H_n} - \log \kappa_{n,H_n} \xrightarrow{p} 0$ also suffices by continuity of the exponential function. This relative accuracy is an additional condition on ACE, not a consequence of coherence repair or rollout supervision alone. The Gaussian approximation characterizes uncertainty under the selected continuation law rather than providing coverage guarantees under $\mathbb{P}_{\mathrm{data}}$.

Limits of current-response information. For fixed n, let $\theta_0 = 0$, $Q_t = q_t = \frac{1}{2}\delta_{\theta_t - 1} + \frac{1}{2}\delta_{\theta_t + 1}$, and define the readout update by $\theta_{t+1} = \theta_t + c_t(Z_{t+1} - \theta_t)/(n+t+1)$. The choices $c_t = 1$ for all t, and $c_0 = 1$ with $c_t = 2$ for $t \ge 1$, define two frozen forecasters with identical initial laws and response maps. Both are exactly coherent with $\Gamma_t = c_t^2$, but for $H \ge 2$,

$$S_{n,H}^{(1)} = r_{n,H}^2, \qquad S_{n,H}^{(2)} - S_{n,H}^{(1)} = 3 \sum_{t=1}^{H-1} (n+t+1)^{-2} > 0. \tag{85}$$

Both processes are L^2-bounded since $\sup_H \mathbb{E}[\theta_H^2] \le 4r_n^2$. Thus current responses do not identify future covariance uniformly over frozen forecasters, although they may support useful prediction within a specified model and continuation regime.

D.3 Terminal Readout and Covariance Decomposition

Proposition D.2 (Terminal covariance under exact repair)**.** *Suppose exact mean repair is applied at every step and* $\sup_{t\geq 0}\mathbb{E}\|\theta_t\|^2 < \infty$*. Then* $\theta_t \to \theta_\infty^Q$ *almost surely and in* L^2*, with* $\mathbb{E}[\theta_\infty^Q \mid \mathcal{F}_t] = \theta_t$*. For every finite* H*,*

$$\begin{aligned}\mathrm{Cov}(\theta_\infty^Q \mid \mathcal{F}_0) &= \sum_{t=0}^{\infty} \frac{\mathbb{E}[\Gamma_t \mid \mathcal{F}_0]}{(n+t+1)^2} \\ &= S_{n,H} + \mathbb{E}[\mathrm{Cov}(\theta_\infty^Q \mid \mathcal{F}_H) \mid \mathcal{F}_0],\end{aligned} \tag{86}$$

and

$$\mathbb{E}[\mathrm{Cov}(\theta_\infty^Q \mid \mathcal{F}_H) \mid \mathcal{F}_0] = \sum_{t=H}^{\infty} \frac{\mathbb{E}[\Gamma_t \mid \mathcal{F}_0]}{(n+t+1)^2}. \tag{87}$$

Proof. Exact repair makes (θ_t) a martingale by Lemma 2.1. The L^2 martingale convergence theorem gives almost-sure and L^2 convergence (Doob, 1953). For fixed t, contractivity of conditional expectation yields

$$\theta_t = \mathbb{E}[\theta_j \mid \mathcal{F}_t] \longrightarrow \mathbb{E}[\theta_\infty^Q \mid \mathcal{F}_t] \quad \text{in } L^2 \qquad (j \to \infty). \tag{88}$$

Hence $\mathbb{E}[\theta_\infty^Q \mid \mathcal{F}_t] = \theta_t$.

For each fixed $v \in \mathbb{R}^d$, the orthogonality argument in Appendix B.6 gives

$$\begin{aligned}\sum_{t=0}^{H-1} \frac{\mathbb{E}[v^\top \Gamma_t v \mid \mathcal{F}_0]}{(n+t+1)^2} &= \mathbb{E}[(v^\top(\theta_H - \theta_0))^2 \mid \mathcal{F}_0] \\ &\longrightarrow \mathbb{E}[(v^\top(\theta_\infty^Q - \theta_0))^2 \mid \mathcal{F}_0] \quad \text{in } L^1.\end{aligned} \tag{89}$$

The left side is nondecreasing in H, so uniqueness of limits in probability identifies its almost-sure limit with $\mathrm{Var}(v^\top \theta_\infty^Q \mid \mathcal{F}_0)$. Taking v to be basis vectors and their pairwise sums proves the first equality in eq. (86). The trace series has finite expectation by L^2 boundedness, and positive semidefiniteness gives absolute convergence of each off-diagonal series.

Since $\mathbb{E}[\theta_\infty^Q \mid \mathcal{F}_H] = \theta_H$, conditional total covariance gives

$$\mathrm{Cov}(\theta_\infty^Q \mid \mathcal{F}_0) = \mathrm{Cov}(\theta_H \mid \mathcal{F}_0) + \mathbb{E}[\mathrm{Cov}(\theta_\infty^Q \mid \mathcal{F}_H) \mid \mathcal{F}_0]. \tag{90}$$

The finite-horizon orthogonality identity gives $\mathrm{Cov}(\theta_H \mid \mathcal{F}_0) = S_{n,H}$. Subtracting this finite sum proves eq. (87). □

Finite-horizon and terminal approximation. Proposition D.2 gives the readout-level martingale posterior interpretation in Section 4. For scalar readouts, the remaining variance in eq. (87) equals

$$\mathbb{E}[(\theta_\infty^Q - \theta_H)^2 \mid \mathcal{F}_0] = \mathbb{E}[\mathrm{Var}(\theta_\infty^Q \mid \mathcal{F}_H) \mid \mathcal{F}_0]. \tag{91}$$

If $S_{n,H_n} > 0$ and

$$\frac{\mathbb{E}[(\theta_\infty^Q - \theta_{H_n})^2 \mid \mathcal{F}_0]}{S_{n,H_n}} \xrightarrow{p} 0, \tag{92}$$

then conditional Markov's inequality gives, for every $\varepsilon > 0$,

$$\Pr\left\{\frac{|\theta_\infty^Q - \theta_{H_n}|}{\sqrt{S_{n,H_n}}} > \varepsilon\right\} \leq \mathbb{E}\left[\min\left\{1, \frac{\mathbb{E}[(\theta_\infty^Q - \theta_{H_n})^2 \mid \mathcal{F}_0]}{\varepsilon^2 S_{n,H_n}}\right\}\right] \longrightarrow 0. \tag{93}$$

The last convergence follows because the integrand is bounded and converges to zero in probability. Thus the finite-horizon CLT also applies to θ_∞^Q by Slutsky's theorem when its hypotheses hold. This relative tail condition is additional to L^2 convergence for each fixed n.

D.4 Covariance Decomposition under Approximate Repair

Fix $H < \infty$ and assume θ_0 and $\psi_t(Z_{t+1})$, $0 \le t < H$, are square integrable under the selected law. With the achieved moments $b_t = \mathbb{E}_{Q_t}[\psi_t]$ and $\Gamma_t = \mathrm{Cov}_{Q_t}(\psi_t)$, define

$$A_{n,H} = \sum_{t=0}^{H-1} \frac{b_t}{n+t+1}, \qquad M_{n,H} = \sum_{t=0}^{H-1} \frac{\psi_t(Z_{t+1}) - b_t}{n+t+1}. \tag{94}$$

Proposition D.3 (Finite-horizon decomposition with residual drift)**.** *The readout satisfies* $\theta_H - \theta_0 = A_{n,H} + M_{n,H}$*, with*

$$\mathbb{E}[M_{n,H} \mid \mathcal{F}_0] = 0, \qquad \mathrm{Cov}(M_{n,H} \mid \mathcal{F}_0) = S_{n,H}. \tag{95}$$

Consequently, $\mathbb{E}[\theta_H \mid \mathcal{F}_0] = \theta_0 + \mathbb{E}[A_{n,H} \mid \mathcal{F}_0]$ *and*

$$\begin{aligned}\mathrm{Cov}(\theta_H \mid \mathcal{F}_0) = S_{n,H} + \mathrm{Cov}(A_{n,H} \mid \mathcal{F}_0) \\ + \mathrm{Cov}(M_{n,H}, A_{n,H} \mid \mathcal{F}_0) + \mathrm{Cov}(A_{n,H}, M_{n,H} \mid \mathcal{F}_0).\end{aligned} \tag{96}$$

Moreover,

$$\begin{aligned}\|\mathrm{Cov}(\theta_H \mid \mathcal{F}_0) - S_{n,H}\|_{\mathrm{op}} \le \|\mathrm{Cov}(A_{n,H} \mid \mathcal{F}_0)\|_{\mathrm{op}} \\ + 2\sqrt{\|S_{n,H}\|_{\mathrm{op}} \|\mathrm{Cov}(A_{n,H} \mid \mathcal{F}_0)\|_{\mathrm{op}}}.\end{aligned} \tag{97}$$

Proof. Conditional Jensen's inequality makes b_t and the centered increments square integrable. Summing Lemma 2.1 gives $\theta_H - \theta_0 = A_{n,H} + M_{n,H}$. Each summand of $M_{n,H}$ has conditional mean zero and conditional covariance $\Gamma_t/(n+t+1)^2$. The orthogonality argument in Appendix B.6 gives its stated mean and covariance. Taking the conditional mean and expanding the covariance of $A_{n,H} + M_{n,H}$ proves eq. (96).

For unit vectors u, v, conditional Cauchy–Schwarz gives

$$\begin{aligned}|u^\top \mathrm{Cov}(M_{n,H}, A_{n,H} \mid \mathcal{F}_0) v| &\le \sqrt{u^\top S_{n,H} u}\, \sqrt{v^\top \mathrm{Cov}(A_{n,H} \mid \mathcal{F}_0) v} \\ &\le \sqrt{\|S_{n,H}\|_{\mathrm{op}} \|\mathrm{Cov}(A_{n,H} \mid \mathcal{F}_0)\|_{\mathrm{op}}}.\end{aligned} \tag{98}$$

Taking the supremum over u, v bounds either cross-covariance in operator norm. The triangle inequality in eq. (96) proves eq. (97). □

Later drift may depend on earlier centered increments, so the cross terms need not vanish and residual drift need not increase endpoint covariance. Conditional Minkowski's inequality also gives

$$\left\{\mathbb{E}[\|A_{n,H}\|^2 \mid \mathcal{F}_0]\right\}^{1/2} \le \sum_{t=0}^{H-1} \frac{\left\{\mathbb{E}[\|b_t\|^2 \mid \mathcal{F}_0]\right\}^{1/2}}{n+t+1}. \tag{99}$$

Its squared left side bounds both $\|\mathbb{E}[A_{n,H} \mid \mathcal{F}_0]\|^2$ and $\|\mathrm{Cov}(A_{n,H} \mid \mathcal{F}_0)\|_{\mathrm{op}}$.

Gaussian approximation with residual drift. For scalar readouts and horizons $H_n \to \infty$, assume the preceding square-integrability conditions for each n and $S_{n,H_n} > 0$. Suppose eqs. (81) and (82) hold with $\psi_t(Z_{t+1})$ replaced by $\psi_t(Z_{t+1}) - b_t$. Then $M_{n,H_n}/\sqrt{S_{n,H_n}} \xrightarrow{d} \mathcal{N}(0,1)$ by the same martingale-array argument. If additionally

$$\frac{\mathbb{E}[A_{n,H_n}^2 \mid \mathcal{F}_0]}{S_{n,H_n}} \xrightarrow{p} 0, \tag{100}$$

then the conditional Markov argument above gives $A_{n,H_n}/\sqrt{S_{n,H_n}} \xrightarrow{p} 0$. Slutsky's theorem yields

$$\frac{\theta_{H_n} - \theta_0}{\sqrt{S_{n,H_n}}} = \frac{M_{n,H_n} + A_{n,H_n}}{\sqrt{S_{n,H_n}}} \xrightarrow{d} \mathcal{N}(0,1). \tag{101}$$

Relative consistency of the ACE estimate permits replacing S_{n,H_n} by $\widehat{S}_{n,H_n}$ as in Appendix D.2. A small drift covariance alone controls the variance discrepancy but not the centering error, and a small current residual b_0 does not control the accumulated drift.

E Covariance Target Selection

E.1 The Attainable Zero-Mean Covariance Set

Fix a context D_t and its finite candidate support. Let $\mathcal{Q}_t$ be the simplex of laws $Q \ll q_t$, or the fixed-marginal family $\mathcal{Q}_{t|G}$ in Appendix C.1. Define the attainable zero-mean covariance set by

$$\mathcal{K}_t = \left\{ \mathbb{E}_Q[\psi_t \psi_t^\top] : Q \in \mathcal{Q}_t,\ \mathbb{E}_Q[\psi_t] = 0 \right\}. \tag{102}$$

Every member of $\mathcal{K}_t$ is a centered covariance because its mean is zero. The set is empty if and only if the zero-mean constraint is infeasible in $\mathcal{Q}_t$. Otherwise, the zero-mean laws form a nonempty compact convex set, and their second-moment map is linear and continuous. Thus $\mathcal{K}_t$ is compact and convex, with $\mathcal{K}_t \subseteq \mathbb{S}_+^d$. Positive semidefiniteness alone does not imply attainability.

In the unrestricted case, $\Gamma \in \mathcal{K}_t$ if and only if $(0, \Gamma) \in \mathcal{C}_t'$. For fixed covariate marginals, the corresponding condition is $(0, \Gamma) \in \mathcal{C}_{t|G}'$. A finite full-support tilt additionally requires the augmented target to lie in the relative interior of the corresponding moment set. In particular, $\mathcal{K}_t$ can be nonempty even when zero mean is attainable only by boundary laws.

E.2 Covariance Projection and Interior Adjustment

We write $\Gamma_t^{\mathrm{ref}} = \mathrm{Cov}_{q_t}(\psi_t)$ for the base-response covariance used as a reference in Section 3. It is computed from the current responses and base probabilities before repair. For $\mathcal{K}_t \neq \varnothing$, define

$$\overline{\Gamma}_t = \operatorname*{arg\,min}_{C \in \mathcal{K}_t} \|C - \Gamma_t^{\mathrm{ref}}\|_{\mathrm{F}}^2. \tag{103}$$

Proposition E.1 (Covariance projection and full-support adjustment)**.** *If $\mathcal{K}_t \neq \varnothing$, the projection $\overline{\Gamma}_t$ exists and is unique. It equals Γ_t^{ref} exactly when $\Gamma_t^{\mathrm{ref}} \in \mathcal{K}_t$. Suppose additionally that a law $Q_t^0 \in \mathcal{Q}_t$ is positive on every retained atom and satisfies $\mathbb{E}_{Q_t^0}[\psi_t] = 0$. Writing $\Gamma_t^0 = \mathbb{E}_{Q_t^0}[\psi_t \psi_t^\top]$, for any $\rho \in (0,1)$ the target*

$$\Gamma_t^\star = (1-\rho)\overline{\Gamma}_t + \rho \Gamma_t^0 \tag{104}$$

admits a finite full-support KL-minimal mean-covariance repair in $\mathcal{Q}_t$.

Proof. The set $\mathcal{K}_t$ is compact and convex, and $C \mapsto \|C - \Gamma_t^{\mathrm{ref}}\|_{\mathrm{F}}^2$ is continuous and strictly convex. Hence eq. (103) has a unique minimizer. If $\Gamma_t^{\mathrm{ref}} \in \mathcal{K}_t$, it attains the minimum value zero. Conversely, equality of the projection and reference implies membership in $\mathcal{K}_t$.

Choose any zero-mean law $\overline{Q}_t \in \mathcal{Q}_t$ attaining $\overline{\Gamma}_t$. The mixture $(1-\rho)\overline{Q}_t + \rho Q_t^0$ lies in $\mathcal{Q}_t$, preserves zero mean, and is positive on every retained atom. Since both component means vanish, its covariance equals the target in eq. (104):

$$\mathrm{Cov}_{(1-\rho)\overline{Q}_t + \rho Q_t^0}(\psi_t) = (1-\rho)\overline{\Gamma}_t + \rho \Gamma_t^0 = \Gamma_t^\star. \tag{105}$$

The positive-weight characterization in Appendices B.3 and C.1 places $(0, \Gamma_t^\star)$ in the relative interior of the corresponding augmented moment set. The joint repair result therefore gives a unique KL minimizer with a finite full-support tilt. □

The mixture above certifies feasibility but need not be the KL minimizer. Projection first selects the closest attainable covariance, and KL minimization then selects a law at the chosen target. These are separate objectives, not a joint minimization of covariance discrepancy and KL. If $(0, \overline{\Gamma}_t)$ already lies in the relevant relative interior, no adjustment is needed and we take $\Gamma_t^\star = \overline{\Gamma}_t$.

If $\mathcal{K}_t = \varnothing$, covariance projection cannot restore mean feasibility, so we use the regularized repair in Section 3.2. If zero mean is feasible only on the boundary, changing the covariance target alone cannot yield full support on the original candidates. One may instead use the boundary formulation, an explicitly restricted support, or approximate repair.

The reference Γ_t^{ref}, projection $\overline{\Gamma}_t$, selected target $\Gamma_t^\star$, and achieved covariance Γ_t are distinct. Thus, the projection discrepancy $\|\overline{\Gamma}_t - \Gamma_t^{\mathrm{ref}}\|_{\mathrm{F}}$ and additional adjustment $\|\Gamma_t^\star - \overline{\Gamma}_t\|_{\mathrm{F}}$ should be reported separately. Exact mean-covariance repair gives $\Gamma_t = \Gamma_t^\star$, while approximate repair satisfies eq. (67).

E.3 Scalar Covariance Bands

For scalar responses with fixed covariate marginals, the attainable mean interval is $[m_{t,-}, m_{t,+}]$,

$$\text{where} \qquad m_{t,-} = \sum_{a=1}^{A} g_{t,a} \min_{y \in \mathcal{Y}_{t,a}} \psi_t(u_a, y), \qquad m_{t,+} = \sum_{a=1}^{A} g_{t,a} \max_{y \in \mathcal{Y}_{t,a}} \psi_t(u_a, y). \tag{106}$$

Thus $\mathcal{K}_t \neq \varnothing$ if and only if $m_{t,-} \leq 0 \leq m_{t,+}$. A strictly positive zero-mean law exists exactly when $m_{t,-} < 0 < m_{t,+}$ or $m_{t,-} = m_{t,+} = 0$. Both statements follow from eq. (50) and its relative-interior characterization. When nonempty, $\mathcal{K}_t$ is a closed interval $[\Gamma_{t,-}, \Gamma_{t,+}]$.

Proposition E.2 (Scalar attainable covariance band)**.** *If the zero-mean constraint is feasible, the attainable covariance endpoints are*

$$\Gamma_{t,-} = \sup_{\mu \in \mathbb{R}} \sum_{a=1}^{A} g_{t,a} \min_{y \in \mathcal{Y}_{t,a}} \{\psi_t(u_a, y)^2 - \mu \psi_t(u_a, y)\}, \tag{107}$$

$$\Gamma_{t,+} = \inf_{\mu \in \mathbb{R}} \sum_{a=1}^{A} g_{t,a} \max_{y \in \mathcal{Y}_{t,a}} \{\psi_t(u_a, y)^2 - \mu \psi_t(u_a, y)\}. \tag{108}$$

Both endpoints and every covariance between them are attained. The unrestricted case is recovered with one covariate of mass one.

Proof. The extrema of $\mathbb{E}_Q[\psi_t^2]$ over $Q \in \mathcal{Q}_{t|G}$ with $\mathbb{E}_Q[\psi_t] = 0$ are feasible bounded linear programs. For the minimum, introducing a multiplier $\mu \in \mathbb{R}$ for the mean constraint gives

$$\min_{Q \in \mathcal{Q}_{t|G}} \{\mathbb{E}_Q[\psi_t^2] - \mu \mathbb{E}_Q[\psi_t]\} = \sum_{a=1}^{A} g_{t,a} \min_{y \in \mathcal{Y}_{t,a}} \{\psi_t(u_a, y)^2 - \mu \psi_t(u_a, y)\}. \tag{109}$$

For fixed μ, each conditional simplex minimum is attained by a point mass. By strong duality for finite-dimensional linear programs, taking the supremum over μ gives eq. (107). Applying the same argument to $-\mathbb{E}_Q[\psi_t^2]$ and replacing μ by $-\mu$ gives eq. (108). Strictly positive feasibility is not required for these finite linear programs.

Compactness of the feasible set gives laws attaining both endpoints. Their convex mixtures preserve zero mean and the covariate marginal, while the second moment varies linearly between the endpoints. Thus the full interval $[\Gamma_{t,-}, \Gamma_{t,+}]$ is attained. □

Scalar projection and interior targets. Projection onto the interval is

$$\overline{\Gamma}_t = \min\{\max\{\Gamma_t^{\text{ref}}, \Gamma_{t,-}\}, \Gamma_{t,+}\}. \tag{110}$$

If $W_t = \Gamma_{t,+} - \Gamma_{t,-} > 0$ and a strictly positive zero-mean law exists, then for $0 < \rho < \frac{1}{2}$ an interior target is

$$\Gamma_t^\star = \min\left\{\max\{\Gamma_t^{\text{ref}}, \Gamma_{t,-} + \rho W_t\}, \Gamma_{t,+} - \rho W_t\right\}. \tag{111}$$

This target admits a finite full-support mean-covariance tilt. To see this, take a positive zero-mean law Q_t^0 with covariance Γ_t^0. For any $\gamma \in (\Gamma_{t,-}, \Gamma_{t,+})$ and sufficiently small $\varepsilon \in (0, 1)$,

$$\gamma' = \frac{\gamma - \varepsilon \Gamma_t^0}{1 - \varepsilon} \in [\Gamma_{t,-}, \Gamma_{t,+}]. \tag{112}$$

Mixing a zero-mean law attaining γ' with Q_t^0 gives a positive admissible law with covariance γ. The relative-interior characterization therefore applies to the target in eq. (111).

A singleton band need not imply zero variation. For example, unrestricted responses $\{-c, c\}$ with $c > 0$ force covariance c^2 under zero mean, so the second-moment constraint is redundant. When a positive zero-mean law exists, a singleton covariance band needs no interior adjustment. Otherwise, it still provides no full-support guarantee.

The reference and target-selection rule are evaluated at each visited context. Preserving the current response scale does not require future covariances to equal Γ_0.

F TabPFN Experimental Details and Additional Results

We provide the protocols and additional results for Section 5, followed by a computational cost analysis. The breadth, loss-defined, nonlinear synthetic, and endpoint-distribution studies all keep TabPFN frozen but use distinct evaluation panels.

F.1 Fixed-Query Continuations and Repair Diagnostics

Readout and candidate law. The breadth experiment uses 30 classification datasets, comprising 17 binary and 13 multiclass tasks with $K \in \{2, \dots, 10\}$. For a context D_0 of size $n = 50$, we fix a query $x^\star$ and its initially most probable class $c^\star$, and track $\theta_t = \mathsf{F}(D_t, x^\star)(\{c^\star\})$. Both $x^\star$ and $c^\star$ remain fixed even if the most probable class later changes. The original horizon-conditioned study uses $H \in \{8, 16, 32, 64\}$, and the enlarged held-out evaluation uses $H = 32$.

At a visited context, let u_a be a retained candidate covariate with prescribed weight $g_{t,a}$. The base joint law and exact response are

$$\begin{aligned} q_t(u_a, y) &= g_{t,a}\mathsf{F}(D_t, u_a)(y), \\ \psi_t(u_a, y) &= (n + t + 1)\{\Theta(D_t \oplus (u_a, y)) - \Theta(D_t)\}. \end{aligned} \tag{113}$$

Conditional repair preserves the current covariate weights and reweights the label law, as in Appendix C.1. Responses and repair are recomputed after each imputation. For randomized candidate pools, all achieved moments are conditioned on the realized pool according to Appendix C.1.

Executed mean-covariance repair. The breadth pipeline performs joint mean-covariance repair, using $\Gamma_t^{\mathrm{ref}} = \mathrm{Cov}_{q_t}(\psi_t)$ when it is compatible with zero mean. Otherwise, when zero mean is feasible, it selects a target from the attainable covariance band using the projection procedure in Appendix E.2. Steps without a certified exact repair use the approximate fallback. Consequently, the reported exact-repair rate describes the executed joint repair, not a separate mean-only continuation. The achieved covariance $\Gamma_t = \mathrm{Cov}_{Q_t}(\psi_t)$ is always computed under the law actually used for sampling, including at fallback steps.

Evaluation units and residuals. The fresh breadth audit contains 1,440 initial evaluation states and 64 selected-law paths per state, with depths $t = 0, \dots, 31$. Thus the feasibility statistics pool approximately 2.95 million visited path-step evaluations, rather than counting only initial states. Reference-preserving joint feasibility additionally requires the unmodified reference covariance, whereas the target-adaptive rate allows projection when needed. On exact steps, the reported residuals are $|\mathbb{E}_{Q_t}[\psi_t]|$ and the relative covariance error $|\Gamma_t/\Gamma_t^\star - 1|$ for positive targets.

Table 6: Additional diagnostics for the fresh audit. Feasibility and exact-solve rates use visited path steps. Residual summaries use certified exact joint-repair steps, and the covariance residual is relative to the selected target.

Diagnostic	Value
Mean feasibility	93.6%
Joint repair preserving Γ_t^{ref}	71.6%
Certified exact joint repair, projection allowed	93.5%
Mean residual, median / maximum	1.4×10^{-16} / 9.8×10^{-9}
Relative covariance residual, median	4.4×10^{-16}
Relative covariance residual, mean	2.4×10^{-14}
Relative covariance residual, 99th percentile / maximum	3.4×10^{-13} / 2.8×10^{-9}
$D_{\mathrm{KL}}[Q_t \| q_t]$, median	0.214

F.2 Loss-Defined Targets and Predictive-Probability Comparisons

Matched loss-defined target. For the TabMGP-aligned comparison, the readout is the unpenalized logistic-loss minimizer on the completed empirical distribution, $\theta_t = \arg\min_\vartheta \mathbb{E}_{F_t}[\ell(Z, \vartheta)]$. The baseline law q_t and repaired law Q_t use the same functional and empirical covariate law, with

Table 7: Matched predictive-probability comparison at $H = 32$ with 32 paths and 32 states per dataset. Logistic projection and Direct q_t use the same base-law paths, while Direct Q_t uses the repaired continuation. Rate, Width, and Error use the definitions in the text. The nonlinear row is the macro average over the five nonlinear tasks.

	Logistic projection			Direct q_t			Direct Q_t		
Dataset	Rate	Width	Error	Rate	Width	Error	Rate	Width	Error
XOR-2D	.031	.16	.34	.625	.22	.07	.719	.29	.07
XOR-20D	.312	.40	.40	.438	.41	.30	.344	.38	.35
Four-Gaussian XOR	.031	.14	.29	.594	.17	.08	.562	.22	.07
Checkerboard	.000	.18	.46	.375	.27	.24	.281	.30	.31
Two Spirals	.062	.22	.38	.250	.29	.20	.344	.31	.23
Linear-logistic	.719	.25	.08	.812	.28	.06	.906	.30	.07
Nonlinear	.087	.22	.37	.456	.27	.18	.450	.30	.21

the logistic fit defining the readout after each augmentation. The real-data results in Table 2 follow the TabMGP benchmark (Ng et al., 2026), using blood (OpenML 1464), banknote (OpenML 1462), phoneme (OpenML 1489), skin (OpenML 1502), mozilla (OpenML 1046), rice (UCI 545), telescope (UCI 159), and sepsis (OpenML 827), originated from OpenML and UCI repositories. We use 500 imputations and 100 paths per repetition. We retain the baseline's joint coverage statistic and the covariance trace of the induced parameter distribution. Drift is the endpoint displacement of the path average from the initial readout, normalized by endpoint spread and summarized over repetitions, rather than the average absolute one-step residual.

Figure 2 shows the eight-dataset trajectory panels, with repaired path averages closer to the initial readout while retaining variation across paths. It illustrates centering, not terminal convergence.

Predictive-probability readouts. The nonlinear comparison instead evaluates the probability at a fixed query $x^\star$, with the known probability $p^\star(Y = 1 \mid x^\star)$ as the evaluation target. For a given path, the logistic readout fits the unpenalized logistic model on the completed dataset and evaluates it at $x^\star$. The direct readout evaluates $\mathsf{F}(D_H, x^\star)(\{1\})$ on the same completed dataset. The logistic projection and direct probability readouts therefore share continuation paths, isolating the readout choice. The repaired direct arm instead uses Q_t while retaining the same probability readout.

For each evaluation state, the interval is given by the empirical 2.5% and 97.5% quantiles of the endpoint readouts across paths. Rate is the fraction of evaluation states whose interval contains $p^\star$. Width is the median interval width over states. Error is the median over states of $|\operatorname{median}_r \theta_H^{(r)} - p^\star|$. Thus Error measures absolute error in probability units, not classification error or a distance between entire distributions.

The completed matched comparison in Table 7 instead uses all 32 states per dataset, 32 paths, and $H = 32$, with common random numbers between the base and repaired continuations. We keep the two panels separate because their horizons, path budgets, and evaluation states differ.

The short-horizon comparison retains the readout distinction, but repair does not uniformly improve coverage, width, or error relative to $p^\star$. These ground-truth metrics measure a different property from conditional coherence under the selected law.

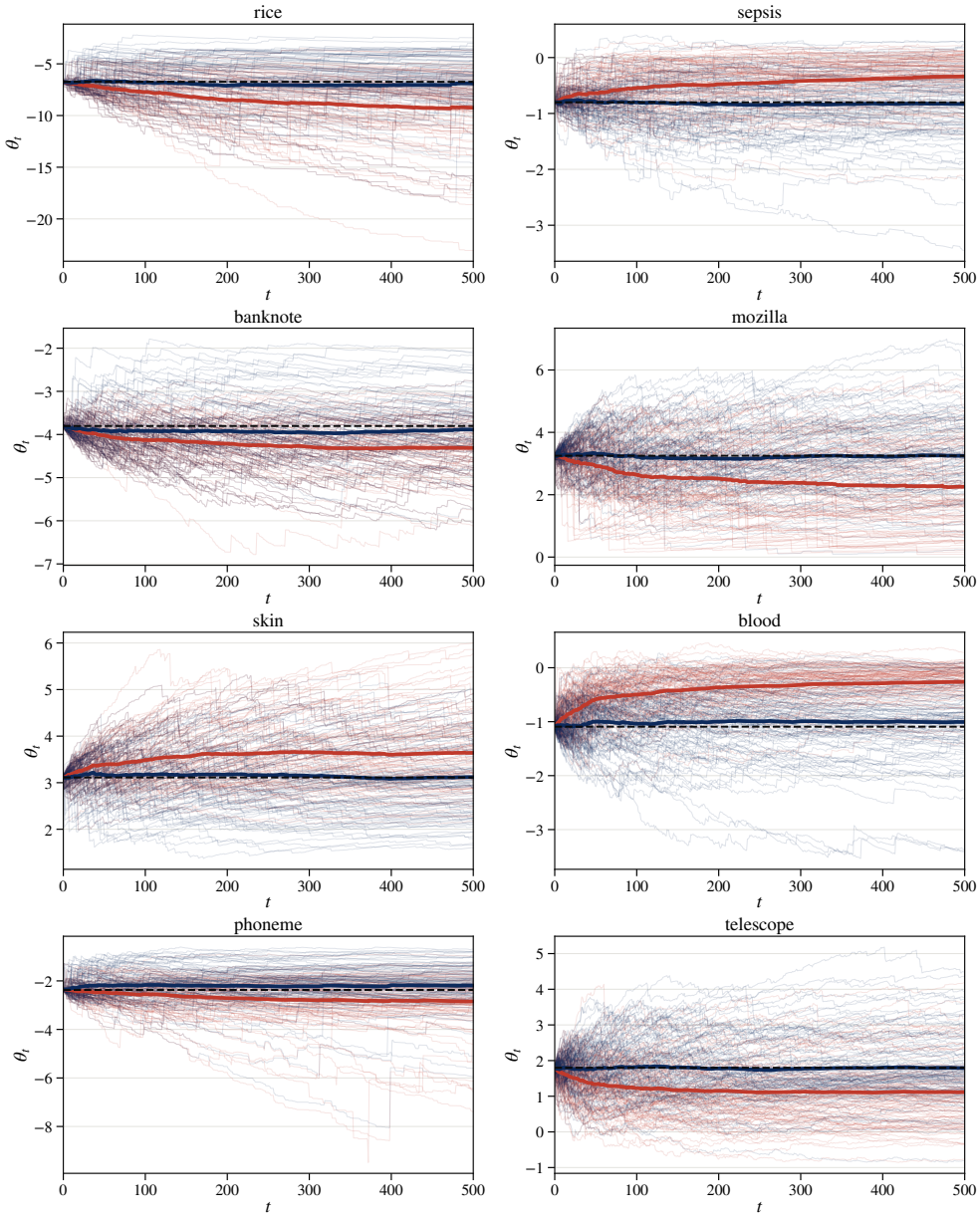

Figure 2: Predictive-resampling trajectories for the eight real-data loss-defined comparisons over 500 imputations. Red and blue denote the base and repaired continuations, respectively. Thin curves show individual paths, thick curves show their averages, and the dashed horizontal line marks the initial readout. The panels show the plotted evaluation, with a separate vertical scale for each dataset.

F.3 ACE Supervision, Baselines, and Held-Out Evaluation

Centered-covariance supervision. For each initial state and horizon, independent repaired paths supply a Monte Carlo estimate of the centered covariance in eq. (16):

$$S_{n,H}^{\mathrm{MC}} = \frac{1}{P}\sum_{r=1}^{P}\sum_{t=0}^{H-1}\frac{\Gamma_t^{(r)}}{(n+t+1)^2}. \tag{114}$$

The achieved $\Gamma_t^{(r)}$ is recomputed at every visited context under its selected law. For a fixed path budget, averaging conditionally independent paths estimates $S_{n,H}$ without substituting the current covariance for future covariances. Approximate steps are included through their centered achieved covariance, so this label is not automatically the endpoint variance. For $\Gamma_0 > 0$, the regression target is $\log\{S_{n,H}^{\mathrm{MC}}/(r_{n,H}^2\Gamma_0)\}$, and inference reconstructs $\widehat{S}_{n,H} = r_{n,H}^2\Gamma_0\exp(\widehat{\log\kappa}_{n,H})$.

The ACE input consists of the current prediction, candidate responses and probabilities, repair summaries, and horizon information. TabPFN stays frozen, and no future endpoint, future covariance, or future fallback frequency is supplied as an inference input. The horizon-conditioned predictor is trained on stored rollout targets, then frozen before the fresh evaluation expansion.

Evaluation panel and metrics. Combined-64 joins 480 historical held-out states with 1,440 fresh states from the same 30 datasets, giving 1,920 evaluation states, with 64 states per dataset. Fresh states are not used to refit or select the frozen ACE predictor. This is generalization to new states from the studied datasets, not an unseen-dataset evaluation. Independently simulated endpoint paths estimate $V_{n,32} = \mathrm{Var}(\theta_{32} \mid \mathcal{F}_0)$, denoted V_{32} in Table 8. Spearman correlation is computed within each dataset against V_{32}, and high-instability AUROC identifies its top quartile. Macro summaries weight datasets equally. These ranking metrics assess prediction of future endpoint instability, rather than directly measuring log-covariance regression error or validating a Gaussian limit.

Baselines. Local uses $r_{n,H}^2\Gamma_0$. H0 predicts log spread from the 192-dimensional hidden representation entering TabPFN's prediction head at the fixed query, and H1 combines that representation with the local baseline. The hidden representation is extracted without changing the forecaster, and the reported hidden-control scores average the three seeds. The $S_{n,32}$ column uses future Monte Carlo information and is a reference rather than an inference-time competitor.

Additional results. ACE exceeds Local on 24 of 30 datasets in Spearman correlation and 25 of 30 in AUROC, and is the best inference-time method among Local, H0, H1, and ACE on 17 and 20 datasets, respectively. The gains are heterogeneous rather than universal. On fresh states alone, ACE also improves over Local from 0.788 to 0.842 in Spearman and from 0.882 to 0.925 in AUROC.

F.4 Endpoint-Moment Gaussian Approximation

Target and learned approximation. The distributional experiment uses 384 fresh states, 32 from each of 12 classification datasets, with $n = 200$ and primary horizon $H = 128$. The query and initially selected class remain fixed. Under the selected continuation law, define $\mu_{n,H} = \mathbb{E}[\theta_H \mid \mathcal{F}_0]$ and $V_{n,H} = \mathrm{Var}(\theta_H \mid \mathcal{F}_0)$. By proposition D.3, residual drift can make these moments differ from $(\theta_0, S_{n,H})$. The endpoint extension therefore predicts both moments, rather than treating centered covariance alone as an endpoint distribution.

The selected model adapts a copy of the small ACE encoder and adds two output heads, while leaving TabPFN, the continuation rule, and the original ACE checkpoint unchanged. Its decoder is

$$\widehat{\mu}_{n,H} = \mathrm{sigmoid}\{\mathrm{logit}(\theta_0) + u_\phi\}, \qquad \widehat{V}_{n,H} = \widehat{S}_{n,H}\exp(v_\phi), \tag{115}$$

where u_ϕ and v_ϕ use current prediction, response and repair information, and n, H. The implementation uses numerical guards for probabilities and variances. Stored empirical endpoint means and sample variances with denominator $P-1$ provide supervision. The predictive approximation is $\widehat{P}_{n,H} = \mathcal{N}(\widehat{\mu}_{n,H}, \widehat{V}_{n,H})$. This extends moment prediction without introducing a new repair.

Table 8: Per-dataset prediction of future fixed-query instability at $H = 32$ on the Combined-64 held-out evaluation. Spearman is the within-dataset rank correlation with estimated endpoint variance V_{32}. AUROC identifies the top 25% of V_{32} within each dataset. H0 uses the initial hidden representation alone, and H1 also uses the local covariance baseline. $S_{n,32}$ is the Monte Carlo centered-covariance reference. Bold indicates the best result, excluding this reference.

		Spearman with V_{32}					High-instability AUROC				
Dataset	K	Local	H0	H1	ACE	$S_{n,32}$	Local	H0	H1	ACE	$S_{n,32}$
analcatdata-authorship	4	**.893**	.818	.874	.890	.928	**.957**	.908	.943	.945	.971
analcatdata-dmft	6	**.689**	.552	.624	.680	.725	.780	.768	.747	**.865**	.927
balance-scale	3	.936	.905	.936	**.959**	.972	.951	.917	**.971**	.958	.988
banknote-authentication	2	.722	**.749**	.746	.733	.751	.930	**.947**	.935	.923	.918
blood-transfusion	2	**.680**	.515	.623	.545	.540	.891	.892	**.924**	.835	.866
breast-w	2	.894	.935	**.936**	.911	.944	.990	**.991**	.988	.983	.983
car	4	.892	.757	.925	**.955**	.972	.956	.809	.953	**.973**	.982
climate-model-simulation-crashes	2	.788	**.917**	.916	.908	.913	.889	.938	.945	**.949**	.961
cmc	3	**.734**	.467	.647	.731	.874	.839	.656	.781	**.870**	.871
credit-approval	2	.826	.872	.861	**.888**	.932	.874	.879	.872	**.902**	.934
credit-g	2	.610	.717	.678	**.803**	.867	.796	.872	.832	**.931**	.961
cylinder-bands	2	.671	.408	.593	**.703**	.789	.882	.737	.844	**.919**	.945
diabetes	2	.856	.782	.820	**.873**	.942	.902	.906	.910	**.917**	.961
dresses-sales	2	.514	.580	.483	**.715**	.906	.750	.771	.699	**.918**	.974
eucalyptus	5	.815	.274	.865	**.897**	.941	.910	.573	.936	**.953**	.984
ilpd	2	**.760**	.597	.654	.687	.808	.895	.827	.866	**.900**	.918
kc2	2	.814	**.885**	.872	.870	.889	.953	.953	**.966**	.960	.970
mfeat-fourier	10	.927	.630	.915	**.953**	.973	.932	.754	.926	**.973**	.988
mfeat-karhunen	10	.813	.687	.830	**.838**	.920	.848	.771	**.870**	.854	.941
mfeat-morphological	10	.900	.587	.852	**.942**	.960	.947	.682	.865	**.970**	.973
mfeat-zernike	10	.819	.462	.883	**.894**	.933	.897	.660	**.923**	.919	.978
MiceProtein	8	.704	.556	.636	**.855**	.904	.831	.691	.785	**.922**	.922
pc1	2	**.876**	.850	.870	.874	.902	**.947**	.900	.901	.939	.958
pc3	2	.784	**.933**	.897	.920	.927	.908	.973	.953	**.980**	.984
pc4	2	.907	.948	.952	**.954**	.973	.949	.997	.986	**.999**	.999
qsar-biodeg	2	.857	.879	**.907**	.900	.947	.868	.922	**.938**	.935	.974
steel-plates-fault	7	.739	.402	.722	**.858**	.947	.892	.574	.815	**.934**	.974
tic-tac-toe	2	.596	.675	.645	**.749**	.814	.757	.840	.832	**.893**	.928
vehicle	4	.794	.716	.697	**.903**	.913	.902	.759	.827	**.938**	.934
wdbc	2	.875	**.923**	.900	.918	.902	.952	.971	.962	**.977**	.970
Best / 30		6	5	2	**17**	–	2	2	6	**20**	–
Macro mean		.789	.699	.792	**.844**	.890	.892	.828	.890	**.931**	.955
Median		.813	.717	.856	**.881**	.917	.900	.856	.917	**.934**	.965

Independent references and scoring. Each fresh state has 32 reference paths and 128 separate evaluation paths. Reference paths determine the rollout Gaussian's mean and variance or its empirical distribution, while evaluation paths provide the outcomes used for scoring every method. A preselected 48-state dense panel uses 64 reference and 256 evaluation paths. The model and primary optimization seed are selected on development data before the fresh evaluation, and no fresh evaluation state is used for fitting or selection. Scores first average over evaluation endpoints within a state, then over states within each dataset, and finally equally over datasets. The rollout references are finite Monte Carlo estimates, not exact oracles.

For a predictive CDF F and evaluation endpoint y, CRPS is $\int_{\mathbb{R}}(F(z) - \mathbf{1}\{y \le z\})^2 \, \mathrm{d}z$. For its nominal 90% interval $[l, u]$, the interval score is

$$\mathrm{IS90}(l, u; y) = (u - l) + 20(l - y)\mathbf{1}\{y < l\} + 20(y - u)\mathbf{1}\{y > u\}. \tag{116}$$

Lower CRPS and IS90 are better. Coverage is the fraction of independent evaluation endpoints inside $[l, u]$ and is compared with 0.9, not maximized. This coverage concerns future readouts under the selected law, unlike coverage of $p^\star$ in the nonlinear synthetic comparison.

Table 9: Additional methods for the endpoint-distribution experiment at $n = 200$, $H = 128$. All entries use the same 384 fresh states and independent evaluation endpoints. The selected endpoint model adapts an encoder copy, whereas the frozen-encoder control trains only its new heads.

Approximation	CRPS ↓	IS90 ↓	Cover90
Centered ACE Gaussian	.0603	.558	.931
Global endpoint correction	.0514	.465	.924
Frozen ACE encoder with endpoint heads	.0458	.393	.825
Adapted ACE encoder with endpoint heads	.0441	.369	.841
Direct endpoint-moment regression	.0446	.378	.824
32-path centered-covariance Gaussian	.0600	.540	.954
8-path endpoint Gaussian	.0416	.348	.841
32-path endpoint Gaussian	.0391	.302	.903
8-path empirical distribution	.0426	.389	.711
32-path empirical distribution	.0388	.297	.853

Table 10: Per-dataset Gaussian distributional results at $n = 200$ and $H = 128$. Each dataset contributes 32 fresh evaluation states, with 128 independent evaluation endpoints per state. Centered denotes $\mathcal{N}(\theta_0, \widehat{S}_{n,H})$, Endpoint denotes the learned $\mathcal{N}(\widehat{\mu}_{n,H}, \widehat{V}_{n,H})$, and MC32 uses moments from 32 independent reference paths. CRPS and IS90 are lower-is-better scores, whereas Cover90 is compared with the nominal value 0.9.

	CRPS			Cover90			IS90		
Dataset	Centered	Endpoint	MC32	Centered	Endpoint	MC32	Centered	Endpoint	MC32
analcatdata_dmft	0.0404	0.0254	0.0200	0.997	0.929	0.895	0.445	0.214	0.167
balance-scale	0.0455	0.0379	0.0352	0.918	0.883	0.940	0.406	0.318	0.276
blood-transfusion	0.0727	0.0441	0.0344	0.980	0.904	0.881	0.657	0.338	0.263
breast-w	0.0436	0.0356	0.0286	0.782	0.737	0.919	0.426	0.389	0.243
car	0.0701	0.0439	0.0421	0.820	0.794	0.917	0.543	0.364	0.320
cmc	0.0732	0.0606	0.0561	0.990	0.803	0.900	0.682	0.471	0.410
credit-approval	0.0437	0.0392	0.0363	0.937	0.809	0.907	0.401	0.335	0.295
credit-g	0.0690	0.0570	0.0540	0.988	0.806	0.900	0.638	0.445	0.400
diabetes	0.0594	0.0409	0.0372	0.961	0.890	0.888	0.560	0.322	0.282
pc3	0.0680	0.0472	0.0362	0.886	0.804	0.896	0.582	0.446	0.273
pc4	0.0443	0.0313	0.0290	0.924	0.873	0.916	0.456	0.288	0.243
tic-tac-toe	0.0942	0.0658	0.0606	0.994	0.854	0.880	0.895	0.497	0.453
Macro mean	0.0603	0.0441	0.0391	0.931	0.841	0.903	0.558	0.369	0.302

Distributional results and moment ablations. The learned endpoint Gaussian improves dataset-level mean CRPS over the centered ACE Gaussian on all 12 datasets. Its paired macro difference is -0.0163, with a 95% dataset-bootstrap interval of $[-0.0208, -0.0118]$. Paired differences use the unrounded scores. Relative to the improvement from centered ACE to the 32-path endpoint Gaussian, it recovers 77% of the CRPS gap, with interval $[70\%, 83\%]$. This is a recovery fraction in scoring performance, not a fraction of a probability distribution recovered.

Mean-only correction $\mathcal{N}(\widehat{\mu}_{n,H}, \widehat{S}_{n,H})$ gives CRPS 0.0513, and variance-only correction $\mathcal{N}(\theta_0, \widehat{V}_{n,H})$ gives 0.0572, compared with 0.0441 when both moments are corrected. Both contributions therefore matter for this finite-horizon approximation. The endpoint model remains above the 32-path Gaussian's CRPS 0.0391, and its nominal 90% intervals cover 84.1% of evaluation endpoints. The prespecified noninferiority criterion against that reference was not met.

Table 10 reports the per-dataset scores and coverage, and Table 11 reports the secondary horizons. The Gaussian family is motivated by the finite-horizon analysis, but this experiment does not establish its CLT assumptions or terminal martingale-posterior recovery. It evaluates a fast finite-horizon approximation under the selected continuation law.

Table 11: Secondary horizon results on the same 384 fresh states. Entries are dataset-macro CRPS point estimates, and $H = 128$ remains the primary horizon. Comparisons are made within each horizon because the endpoint distributions change with H.

H	Centered ACE	Amortized endpoint	32-path endpoint
8	.0179	.0177	.0175
16	.0241	.0234	.0231
32	.0316	.0297	.0289
64	.0425	.0369	.0350
128	.0603	.0441	.0391

F.5 Computational Cost

Forecaster evaluations. Let M be the number of covariates probed at a context, K the number of labels, and P the number of repaired rollout paths of length H. Candidate responses require at most MK augmented readouts, in addition to current predictions and the finite-dimensional repair. Once computed, candidate moments or reweighting require no additional forecaster calls. For fixed-query experiments, one augmented readout is one frozen TabPFN evaluation at the fixed query after appending a candidate pair.

A current probe-and-repair computation therefore uses $O(MK)$ augmented evaluations. Explicit estimation of $S_{n,H}$ uses $O(PHMK)$, whereas ACE inference uses $O(MK)$ plus the amortizer. The same rollouts also provide endpoint samples. For N training states, offline supervision requires $O(NPHMK)$ evaluations, with shorter-horizon labels reusable from path prefixes. These counts exclude independent evaluation paths.

Table 12: Forecaster-evaluation counts for one initial state. M counts probed covariates, K labels, P paths, and H imputation steps. Counts exclude amortizer training, candidate selection, and numerical moment solves. Parallel probing reduces latency but not the number of evaluations.

Operation	Augmented evaluations	Future sequential depth
Current probing and repair	$O(MK)$	0
ACE or endpoint-moment inference	$O(MK)$	0
P repaired paths to horizon H	$O(PHMK)$	H

Context growth and repair overhead. Let $C_{\mathrm{F}}(m)$ bound an augmented readout at context size m, $C_{\mathrm{R},t}$ the remaining per-state cost of current predictions, candidate selection, and repair at depth t, and C_{ACE} the amortizer cost. With bounded M and K,

$$C_{\mathrm{rollout}} = O\left(P \sum_{t=0}^{H-1} \{MK\, C_{\mathrm{F}}(n+t+1) + C_{\mathrm{R},t}\}\right), \tag{117}$$

$$C_{\mathrm{inference}} = O(MK\, C_{\mathrm{F}}(n+1) + C_{\mathrm{R},0} + C_{\mathrm{ACE}})\,. \tag{118}$$

These expressions retain context-growth cost rather than treating all forecaster calls as equal. If J atoms are retained and p is the identifiable moment dimension, each dense Newton iteration costs $O(Jp^2 + p^3)$. This is a per-iteration solver cost, not an iteration-count guarantee. For a scalar mean-covariance target, $p \le 2$, while covariate-subset selection and covariance projection are included in $C_{\mathrm{R},t}$. Sharing the initial response table across paths reduces common first-step work but does not change the $O(PHMK)$ scaling for $H \ge 2$.

Measured inference cost. In the NVIDIA A100 setting, the endpoint approximation takes 0.74 seconds per evaluation instance, including current probing and repair. The 8-path and 32-path references take approximately 500 and $1{,}693$ seconds at $n = 200, H = 128$. These are implementation-specific inference measurements, not hardware-independent speedups or equal-accuracy comparisons. Rollout supervision and training are offline and excluded from the 0.74 seconds. ACE avoids future sequential rollouts at inference, but not current probing or offline rollout supervision.

G LLM Experimental Details and Additional Results

We provide the protocols, additional results, and computational costs for repaired-law uncertainty and native-law directional prediction. These experiments use the same local response interface but have distinct prediction targets.

G.1 Models, Benchmarks, and Continuation Protocol

Models and benchmarks. Following the model setting of Kim et al. (2026), we use frozen Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and OLMo-3-7B-Instruct (Olmo et al., 2025). The clean benchmark comparison uses CommonsenseQA (CSQA; Talmor et al., 2019), AI2-ARC-Challenge (ARC-C; Clark et al., 2018), and MMLU-Pro (Wang et al., 2024). Earlier diagnostic panels also include TinyMMLU (Polo et al., 2024). We distinguish these panels from the clean comparison rather than combining their results. The forecaster weights remain fixed throughout probing, continuation, and amortizer training.

Reasoning blocks and the readout. A continuation step appends one sampled reasoning block to the current prefix D_t. Following work on intermediate answers from partial reasoning traces (Yang et al., 2025; Wang et al., 2026), a separate answer branch supplies the stop-now distribution π_t by normalizing the answer-option logits. The answer branch measures the readout and is not appended to the reasoning continuation. At the initial prefix, we fix $a^\star = \arg\max_a \pi_0(a)$ and track $\theta_t = \pi_t(a^\star)$, even if another option becomes most probable later.

Candidate law and repair. At each visited prefix, we draw a pool of 16 reasoning blocks from the native continuation law and evaluate the readout after each candidate augmentation. The base candidate law q_t is the empirical law on this pool. The selected law Q_t reweights the candidate probabilities to enforce scalar mean repair when an exact tilt is available. Otherwise, the selected process retains its base-pool fallback, and the achieved residual drift is included in the uncertainty calculation. For the empirical pool, this fallback is uniform over sampled candidate entries. This fallback is not an exact or KL-regularized repair. New pools are drawn along the continuation, and the responses and selected law are recomputed at each prefix. All local moments condition on the realized candidate pool as in Appendix C.1.

Two continuation targets. Repaired continuations generate the scalar centered-variation target S_H. Separate native continuations generate the full future mean $\mu_H^P = \mathbb{E}_P[\pi_H \mid D_0]$ used for directional prediction. In particular, μ_H^P and S_H are not the mean and variance of a common distribution. Independent continuation streams are used for Monte Carlo references and evaluation, and learned predictors receive only current-prefix information and current candidate responses at inference.

G.2 Finite-Horizon Reasoning Uncertainty

Centered variation and evaluation. We use the reasoning-block clock, with the initial prefix indexed by $t = 0$, and write the exact block response as $\psi_t(z) = (t+1)\{\Theta(D_t \oplus z) - \theta_t\}$. The clock counts reasoning steps, not the observed sample size in the asymptotic theory. Let $b_t = \mathbb{E}_{Q_t}[\psi_t]$ and $\Gamma_t = \mathrm{Cov}_{Q_t}(\psi_t)$ be the achieved moments, including at fallback steps. The accumulated residual drift and centered fluctuation are

$$A_H = \sum_{t=0}^{H-1} \frac{b_t}{t+1}, \qquad M_H = \theta_H - \theta_0 - A_H, \qquad S_H = \mathbb{E}_Q\left[\sum_{t=0}^{H-1} \frac{\Gamma_t}{(t+1)^2} \,\middle|\, D_0\right]. \tag{119}$$

Here the expectation also averages over candidate-pool draws in the selected continuation process. Conditional orthogonality gives $\mathbb{E}_Q[M_H^2 \mid D_0] = S_H$. The endpoint variance can additionally contain residual-drift and cross-covariance terms, as detailed in Proposition D.3. We score forecasts using $(\widehat{S}_H - M_H^2)^2$, averaged over independent evaluation continuations. The target is the centered second moment, not the uncentered displacement $(\theta_H - \theta_0)^2$.

Local and amortized estimators. The local baseline uses the current candidate covariance with the horizon factor $r_H^2 = \sum_{h=1}^{H} h^{-2}$, with its scale calibrated on development data. ACE learns

the finite-horizon correction from the current prediction and the same 16-block response information, using rollout-derived S_H targets. The comparison therefore holds the current probing budget fixed. Table 13 gives the scores used in the main comparison. ACE improves over the local estimator in both model families, indicating that the candidate responses are useful beyond their current covariance alone.

Table 13: Finite-horizon uncertainty prediction using one pool of 16 candidate reasoning blocks. Scores are squared second-moment errors against independent centered fluctuations, with lower values indicating better forecasts. The $H = 4$ ACE prediction transfers the $H = 2$ predictor using the horizon rescaling described below.

Model	H	Calibrated local	ACE
Llama-3.1-8B	2	0.01383	**0.01334**
	4	0.01699	**0.01630**
OLMo-3-7B	2	0.01688	**0.01638**
	4	0.01736	**0.01659**

Transfer to a longer horizon. The transferred prediction uses $\widehat{S}_4 = (r_4^2/r_2^2)\widehat{S}_2$, where

$$\frac{r_4^2}{r_2^2} = \frac{1 + 1/4 + 1/9 + 1/16}{1 + 1/4} = \frac{41}{36}. \tag{120}$$

This reuses the learned correction without fitting a new $H = 4$ head. It is an empirical transfer rule, not an assumption that every state's true future covariance scales by this factor. A separate held-out precision evaluation uses 1,656 initial states for Llama and 120 for OLMo. Table 14 gives its paired one-pool and two-pool comparisons, separately from the absolute-score panel in Table 13.

Table 14: Additional $H = 4$ precision evaluation. Entries are the ACE score minus the calibrated-local score, followed by the reported 95% confidence interval. Each pool contains 16 candidate blocks, and each comparison uses the same pool budget for both estimators. The evaluation contains 1,656 Llama states and 120 OLMo states.

Model	One pool	Two pools
Llama-3.1-8B	-6.65×10^{-4}	-2.99×10^{-4}
95% CI	$[-8.16, -5.20] \times 10^{-4}$	$[-4.25, -1.77] \times 10^{-4}$
OLMo-3-7B	-6.6×10^{-4}	-7.3×10^{-4}
95% CI	$[-11.3, -2.1] \times 10^{-4}$	$[-12.5, -2.5] \times 10^{-4}$

Moment-identity diagnostics. The independent-completion diagnostics give ratios of realized centered second moment to Monte Carlo spread of $0.97/1.00$ for Llama/OLMo at $H = 2$ and $1.02/0.94$ at $H = 4$. The reported contributions of steps 3 and 4 to the longer-horizon spread are $16\%/20\%$. These diagnostics concern the centered-variation identity and nontrivial variation beyond the first two blocks. They do not by themselves establish a Gaussian endpoint law or the assumptions of the terminal martingale CLT.

G.3 Directional Prediction of Native Future Answers

Prediction target and controls. For each candidate block z_j, the signed-response predictor retains the full vector change $\pi_0^{+z_j} - \pi_0$ rather than only the change in the fixed option $a^\star$. It predicts $\mu_H^P = \mathbb{E}_P[\pi_H \mid D_0]$, the future stop-now answer distribution averaged over native continuations. The native Monte Carlo reference averages endpoint answer probabilities, rather than sampled answer indicators. This experiment does not change the native continuation law or reinterpret its future mean as a repaired-law quantity.

We compare with temperature scaling (Guo et al., 2017; Xie et al., 2024) using two distinct controls. The adaptive-temperature predictor shares the signed predictor's inputs and trunk but can only rescale the current logits by a positive scalar. The gold-fit temperature is a global scalar fitted to

gold-label negative log likelihood on development data. Both preserve the current option ranking, whereas the signed-response predictor permits option-specific corrections. Future-distribution fidelity is compared with adaptive temperature, and accuracy and calibrated negative log likelihood are compared with gold-fit temperature.

Metrics and benchmark results. For a predicted answer distribution $\widehat{\pi}_H$ and an independent Monte Carlo estimate $\overline{\pi}_H^P$ of the native future mean, soft cross-entropy is

$$\text{Soft-CE}(\overline{\pi}_H^P, \widehat{\pi}_H) = -\sum_a \overline{\pi}_H^P(a) \log \widehat{\pi}_H(a). \tag{121}$$

Accuracy instead compares the predicted top option with the gold answer, and categorical NLL is the negative log probability assigned to that answer. Thus fidelity to future reasoning and correctness are evaluated against different targets. Table 15 reports the differences separately for each benchmark.

Table 15: Directional prediction of the native future answer distribution from 16 current candidate blocks. Each entry is signed-response prediction minus its comparator. Soft-CE uses adaptive temperature as the comparator, whereas accuracy and calibrated NLL use gold-fit temperature. Negative Soft-CE/NLL and positive accuracy differences favor signed responses. Accuracy differences are in percentage points.

	Llama-3.1-8B			OLMo-3-7B		
Benchmark	ΔSoft-CE	ΔAcc.	ΔNLL	ΔSoft-CE	ΔAcc.	ΔNLL
CSQA	-0.052	$+1.4$	-0.074	-0.103	$+1.6$	-0.076
ARC-C	-0.038	$+3.4$	-0.077	-0.093	$+7.9$	-0.096
MMLU-Pro	-0.131	$+4.5$	-0.110	-0.163	$+3.5$	-0.099

Signed responses improve future-mean fidelity on each benchmark in both families. The accuracy improvements cannot be produced by positive temperature rescaling of the same logits, which leaves their ranking unchanged. A magnitude-only control performs worse in both families, indicating that the direction of a candidate-induced change matters in addition to its size. Adding question, option, or candidate semantic embeddings did not improve over the signed-response representation in the reported ablations.

G.4 Computational Cost

Work per prediction. Let M be the number of candidate blocks, H the horizon, and R the number of reference paths. Current inference generates M blocks and evaluates their M augmented stop-now readouts, in addition to the initial readout. The uncertainty predictor adds one repair solve and an amortizer evaluation. The signed-response predictor uses the same response table without future repaired paths. For block length bounded by B, generation uses $O(MB)$ autoregressive decoding steps. Computing scalar local moments or the full response table requires $O(M)$ or $O(MK)$ arithmetic operations beyond the forecaster calls, where K is the number of answer options.

A repaired reference rebuilds the candidate pool at every visited prefix, requiring up to $O(RHM)$ block generations and readouts and $O(RH)$ repair solves. A native reference instead requires $O(RH)$ block generations and $O(R)$ endpoint readouts. For N training states, repaired-law supervision requires $O(NRHM)$ candidate generations and readouts, plus amortizer training. These counts distinguish the costs of the two reference laws and the offline cost retained by amortization.

Context length and latency. Each forecaster call depends on prefix length and generated tokens, so the preceding orders are work counts rather than context-independent FLOPs or wall-clock time. Candidates at a common prefix can run in parallel, but each reference path has up to H sequentially dependent steps. Early termination can reduce the realized work. Table 16 reports the measured latency in NVIDIA A100 GPU.

Both configurations generate 16 blocks, but their sequential structure and answer-readout work differ. Native rollouts remain more accurate than the signed-response predictor in the reported comparison, so lower latency does not imply equivalent prediction quality. The local probe instead supplies

Table 16: Measured inference latency in the matched A100 setting. The comparison is between the shared current probe and eight native two-block rollouts, not the repaired reference that rebuilds candidate pools at each step.

Computation	Llama-3.1-8B	OLMo-3-7B
Current 16-block probe	8.96 s	9.04 s
Eight native two-block rollouts	16.47 s	17.15 s

estimates of future centered variation or native answer movement without the corresponding future sequential rollouts at each new evaluation instance.