THE PRICE OF NOT FORGETTING: A COVERING-BASED MEMORY-ERROR ANALYSIS FOR CONTINUAL TIME-SERIES FORECASTING UNDER REGIME DRIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

A forecaster running on a live stream faces an old dilemma: adapting to a new regime can erase what it knew about the old one. The usual escape is to freeze what has been learned and grow a fresh expert per regime - forgetting becomes impossible by construction, but the guarantee is bought with memory, and nothing says how much, or when to stop growing. We ask what can be certified about that bargain. The regimes a forecaster serves form a small metric space of distributions, and it is the intrinsic dimension of that space, not the raw channel count, that governs how error falls as experts are added. Covering arguments turn expert allocation into a budgeting problem with explicit upper certificates: for the regimes observed, and, under stated assumptions, across the whole condition space. On one narrow, fully specified class of drifts we also match the rate from below over a limited range of budgets - a conditional match, not universal optimality, and our slack terms are upper bounds rather than proved error floors. Three design rules follow: how large an adapter must be, when adding experts stops paying, and which regimes to keep in one pass. On five real datasets, freezing gives exactly zero measured forgetting - because old-regime parameters never move - yet does not uniformly out-learn methods that tolerate some. Low-rank adapters help only behind a bottleneck; a cross-channel drift statistic repairs a real blind spot in a controlled test but not on real streams; and the picture of fine-tuning sliding smoothly along a Wasserstein geodesic does not survive measurement. What we offer is a set of explicit certificates and their conditions - not a free lunch, but a priced one.

1 INTRODUCTION

Forecasting models on real streams are rarely trained once and left alone: electricity load, exchange rates, traffic volume and epidemiological counts all drift, so a deployed forecaster is continually re-exposed to new data and must decide how to update. Naive fine-tuning is the default and the source of a known failure: gradient steps that improve the new regime overwrite what made the old one accurate (catastrophic forgetting). This costs more than in classification - a redeployed forecaster is still expected to perform if the old regime recurs, as it routinely does in seasonal data.

The standard fix is to freeze parameters trained on past regimes and grow a new, dedicated component (an expert, an adapter, a prompt) per regime, so no update to new data touches old parameters. This eliminates parameter interference - nothing frozen can change - but the simplicity is also the problem: the strategy pays for forgetting-avoidance in memory and says nothing about how its error should scale with that memory or when growth should stop. Mixture-of-experts work shows freeze-and-grow works empirically, and parameter-efficient adaptation work shows small updates often suffice in practice; neither, to our knowledge, connects the two into a geometric memory-error certificate tied to the drift actually observed.

The consequences are concrete: a full new component is wasteful when regimes differ in only a few directions, one too small silently under-fits, and online it is unclear when a component is needed versus reusable. Answering both requires characterizing the geometry of drift, not treating each regime as an unrelated worst case.

We treat the regimes a forecaster serves as a small metric space inside the Wasserstein space of per-regime distributions and measure its intrinsic dimension rather than raw parameter count. This turns frozen-expert allocation into a covering-number problem and yields general upper certificates on

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.

observed regimes and, with stronger assumptions, over the full condition space. A lower construction matches the exponent only for one anchored translation/model-realizable class equipped with public condition nets and a q-bit codebook (Corollary 1), and only before the additive codebook-resolution offset η dominates. It does not establish universal minimax optimality over arbitrary maps into $\mathcal{P}_2(\mathbb{R}^d)$. The embedding and fit terms $B + C\bar{\epsilon}$, like η, are upper offsets in the corresponding certificates rather than proved irreducible floors.

Within a concrete shared-basis scheme, a naive frozen-expert component is typically as large as the whole model even when actual drift occupies only a handful of directions. We show how that scheme can be compressed when the backbone routes drift through a latent representation where it becomes low-rank - a condition we state and test directly, since a bare linear model with no such bottleneck can leave the same drift full-rank in parameter space even when it is low-rank in the data. The resulting full-expert/shared-basis ratio is scheme-specific, not the slack in a universal minimax comparison. Distinguishing data-space from parameter-space drift is itself one of our findings.

We evaluate on real multivariate data spanning a wide range of channel counts and drift characters, against fine-tuning and standard continual-learning baselines, and report candidly: frozen-and-adapt has exactly zero measured MSE forgetting because old-regime parameters remain frozen, but it does not universally out-learn methods that tolerate some - a genuine tradeoff, not a knockout win. We also test and refute the intuition that a fine-tuned model slides continuously toward the new regime in Wasserstein space; it falls off the path. Separately we address a blind spot in how drift is detected across correlated channels.

Our contributions are:

- **General frozen-expert upper certificates and a conditional rate match** (Theorems 1-3 and Corollary 1): covering controls the frozen-expert error, while the matching exponent is confined to the anchored translation/model-realizable class, a public net, a q-bit codebook, and the pre-η scale range.
- **A scheme-specific rank-adapter design principle with an explicit applicability condition** (Design Principle 4), sizing an adapter by drift rank in the representation it modifies; the evidence distinguishes data-space from parameter-space drift without turning the resulting compression ratio into a minimax claim.
- **Saturation and streaming certificates** (Theorems 5-6): the first identifies diminishing returns in an upper bound, while the second is an exact-W_2-oracle 8-approximation using $O(k)$ center records, not a finite-sample or total-memory guarantee.
- **An honest empirical account on five real datasets**: exactly zero measured MSE forgetting from frozen parameters, with a genuine, non-dominant learning tradeoff; a cross-channel statistic that repairs marginal blindness to lead-lag shift on a controlled construction but not on real streams; and a tested-and-refuted Wasserstein-geometric hypothesis.

2 Related Work

Continual learning against catastrophic forgetting. Regularization methods penalize movement of parameters important to past tasks, via Fisher information (Kirkpatrick et al., 2017) or a path-integral of parameter usefulness (Zenke et al., 2017); replay methods mix a buffer of past-regime examples into future training (Rolnick et al., 2019); expert-growing methods (Shazeer et al., 2017; Fedus et al., 2022) freeze existing experts and route new data to new ones, a mechanism whose forgetting and generalization behavior Li et al. (2025) analyze theoretically. None gives the covering-based frozen-expert certificates studied here. Our exponent match is deliberately narrower than a universal memory lower bound: it concerns an anchored translation/model-realizable class with public condition nets and a finite codebook, before codebook resolution dominates.

Time-series forecasting under distribution shift. A separate literature addresses non-stationarity *within* one window: RevIN (Kim et al., 2022) and Dish-TS (Fan et al., 2023) strip and re-insert per-window statistics, and non-stationary Transformers (Liu et al., 2022) separate a de-stationarizing statistic from the backbone. Long-horizon forecasting is well served by linear or patch-based biases (Zeng et al., 2023; Xu et al., 2024; Nie et al., 2023; Liu et al., 2024), with linear-expert mixtures (Ni et al., 2024) specializing to local trends. All correct non-stationarity inside one window; none addresses regimes seen sequentially under a memory budget.

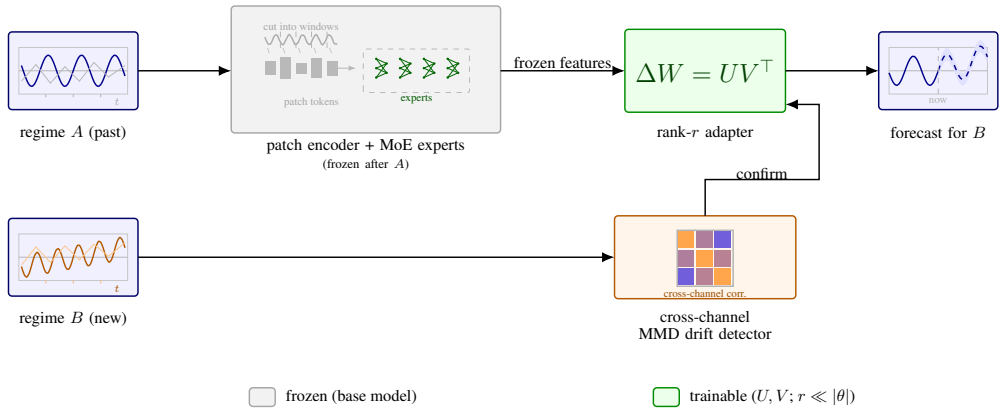


Figure 1: The frozen-and-adapt mechanism. A cross-channel drift detector identifies a new regime; the backbone remains frozen, and a regime-specific low-rank adapter is trained. The rank certificate applies only when the functional displacement is realized by the adapted representation.

Continual and streaming time-series forecasting. Regime-guided test-time adaptation (Kumar et al., 2026), attention-weighted replay (Besnard & Ragot, 2026), structural-knowledge-regularized forecasting (Pan et al., 2024), forecaster pooling under recurring drift (Zhan et al., 2025), and score-driven online updates (Urettini et al., 2026) all target better forecast accuracy or stability-plasticity balance under drift. Our complementary contribution is a set of explicit frozen-expert upper certificates, plus a conditional matching exponent for a model-realizable subclass rather than a universal memory necessity result.

Parameter-efficient adaptation. Low-rank adaptation (Hu et al., 2022) shows a frozen backbone plus a small trainable update often suffices - the mechanism our strategy uses. Missing is a rule for choosing rank from a property of the task rather than convention or a sweep: dynamic-rank methods PEARL (Bhat et al., 2026) and C-LoRA (Zhang et al., 2025) allocate rank empirically, with no geometric reason a rank should suffice. Design Principle 4 supplies such a rule, tied to a measurable drift rank and conditional on a compressible latent representation - a condition existing heuristics neither state nor test.

Drift detection. Spectral change-point methods (Hinder et al., 2022) detect drift from kernel-matrix block structure but, like marginal MMD, target the whole-window signal rather than cross-channel correlation shift - the gap Section 3.6 targets.

Theoretical tools. Our general upper results cover observed regimes or the condition space at a target resolution. The exponent-matching lower construction uses anchored translations together with a public net and finite codebook; it does not range over arbitrary $\mathcal{P}_2(\mathbb{R}^d)$-valued maps. Streaming eviction adapts incremental doubling for online k-center (Charikar et al., 2004) to an *exact* Wasserstein-distance oracle and stores $O(k)$ center records. Wasserstein empirical-measure rates (Fournier & Guillin, 2015) provide only illustrative sample scaling under their moment assumptions, not the streaming theorem's total-memory guarantee or a necessary replay-memory lower bound. The closest antecedent is a Wasserstein-geometric decomposition of out-of-distribution error into drift, distortion, and training-fit terms (Pham, 2026); Section 3 explains why our upper certificates instead use distance-clamping with constant one.

3 METHOD

We formulate forgetting-free continual forecasting as a finite-regime covering problem (Section 3.1), derive clamping certificates for observed regimes and for a fixed full condition map (Section 3.2), and give a minimax comparison on one common anchored translation class (Section 3.3). Sections 3.4-3.5 turn the certificates into rank, saturation, and streaming design rules; Section 3.6 supplies the drift statistic that triggers them. Figure 1 shows the mechanism, and Appendix B contains the proofs.

3.1 PROBLEM SETUP

A forecaster serves conditions $c_1, \dots, c_T \in \mathcal{C}$, with ground-truth forecast law $\mu_{c_j} = f(c_j)$ for $f : \mathcal{C} \to \mathcal{P}_2(\mathbb{R}^d)$ and integer $d \geq 1$. We work in the *condition-observable* setting: a condition descriptor is available at test time from a recent window. Distances between regimes are measured by W_2; $d_j := W_2(\mu_{c_{j-1}}, \mu_{c_j})$ is one-step drift. Let

$$\mathcal{M} := \{\mu_{c_1}, \dots, \mu_{c_T}\}, \qquad n := |\mathcal{M}|, \qquad \Delta := \operatorname{diam}_{W_2}(\mathcal{M}).$$

Thus n counts distinct forecast laws. Assumption 5's internal covering exponent D describes this finite metric space, whereas $D_{\mathcal{C}}$ in Assumption 4 describes the full condition space. We reserve the integer $M \geq 0$ for *expert-parameter memory*. If an expert stores $|\theta| \in \mathbb{N}$ scalars at an integer precision $b \in \mathbb{N}$ bits per scalar, then one full expert costs $q := |\theta| b \in \mathbb{N}$ bits. Routing keys, condition descriptors, and samples or sketches used to estimate W_2 are separate costs.

Given an expert θ_i trained at c_i, *clamping* serves $\hat{p}(\cdot \mid c_i; \theta_i)$ without evaluating the expert at a new condition. Its only modeling requirement is a uniform in-distribution certificate.

Assumption 1 (In-distribution fit)**.** For every trained regime eligible to be a center, $W_2(\hat{p}(\cdot \mid c_i; \theta_i), \mu_{c_i}) \leq C\epsilon_i$. We write $\bar{\epsilon} := \max_i \epsilon_i$ over those relevant trained regimes.

All routing below uses clamping. The triangle inequality then separates approximation of the selected center from geometric coverage of the target regime; Appendix B compares this certificate with conditioning.

3.2 CLAMPING UPPER CERTIFICATES

For a nonempty $S \subseteq \mathcal{M}$, define

$$r(S) := \max_{\mu \in \mathcal{M}} \min_{s \in S} W_2(\mu, s), \qquad r_k := \inf_{\substack{S \subseteq \mathcal{M} \\ 1 \leq |S| \leq k}} r(S), \quad k \geq 1.$$

The "at most k" definition is valid at every integer budget; in particular, $r_k = 0$ for $k \geq n$. We write $\operatorname{err}_{\mathcal{M}}(M)$ for the smallest worst-regime error achieved by an offline nearest-center clamping choice that stores at most $\lfloor M/q \rfloor$ eligible experts; this is a scheme-specific certificate, not a minimax risk over maps f.

Theorem 1 (Finite-regime clamping certificate)**.** *Under Assumption 1, clamping to the nearest member of any S gives*

$$\max_{j \leq T} W_2(\hat{p}_j, \mu_{c_j}) \leq r(S) + C\bar{\epsilon}.$$

If $\Delta > 0$ and Assumption 5 holds, then

$$r_k \leq \Delta k^{-1/D} \qquad \text{for every } k \geq \lceil 2^D \rceil,$$

while $r_k \leq \Delta$ for every $k \geq 1$. For an arbitrary integer expert-parameter budget $M \geq q$, let $k_M := \min\{n, \lfloor M/q \rfloor\}$. An offline choice of centers therefore satisfies

$$\operatorname{err}_{\mathcal{M}}(M) \leq r_{k_M} + C\bar{\epsilon} \leq C\bar{\epsilon} + \begin{cases} 0, & k_M = n, \\ \Delta k_M^{-1/D}, & \lceil 2^D \rceil \leq k_M < n, \\ \Delta, & 1 \leq k_M < \min\{n, \lceil 2^D \rceil\}. \end{cases} \tag{1}$$

Theorem 1 is a finite-instance certificate, not a minimax statement. Its last line also makes the finite-set endpoint explicit: once all n distinct regimes are retained, the covering term is exactly zero. The stated M excludes routing and center metadata.

Fix a full map f; let $r_{\mathcal{C}}(k)$ be the k-center covering radius of $\mathcal{C}$ and $\operatorname{err}_{\mathcal{C}}^{(f)}(M)$ the infimal worst-condition clamping error at budget M (Appendix E).

Theorem 2 (General per-instance condition-space certificate)**.** *Fix $f : \mathcal{C} \to \mathcal{P}_2(\mathbb{R}^d)$. Suppose experts satisfying Assumption 1 with residual at most $\bar{\epsilon}$ are available at the selected centers. For every finite nonempty $S \subseteq \mathcal{C}$, nearest-center clamping and Assumption 3 give*

$$\sup_{c \in \mathcal{C}} W_2\big(\hat{p}(\cdot \mid s(c); \theta_{s(c)}), \mu_c\big) \leq C\bar{\epsilon} + A \sup_{c \in \mathcal{C}} \min_{s \in S} d_{\mathcal{C}}(c, s) + B.$$

Consequently the infimal k-center certificate is $C\bar{\epsilon} + Ar_{\mathcal{C}}(k) + B$. Under Assumption 4, for every integer $k \geq 1$,

$$r_{\mathcal{C}}(k) \leq R \min\left\{1, \left(\frac{c_2}{k}\right)^{1/D_{\mathcal{C}}}\right\}.$$

Thus, for every integer $M \geq q$ and $k = \lfloor M/q \rfloor$,

$$\mathrm{err}_{\mathcal{C}}^{(f)}(M) \leq C\bar{\epsilon} + B + AR \min\left\{1, \left(\frac{c_2}{\lfloor M/q \rfloor}\right)^{1/D_{\mathcal{C}}}\right\}. \tag{2}$$

The minimum with one handles the entire integer range: the entropy inversion is used only when its candidate radius is at most R, and the diameter bound handles smaller k. The additive terms in Equation 2 are slack in this upper certificate; they are not lower bounds on achievable error. The result is deliberately per-instance because its fitted experts and constants can depend on f.

3.3 A SAME-CLASS MINIMAX COMPARISON

To obtain a finite, well-posed minimax problem we use the *anchored translation class* $\mathcal{T}_L^0$ of L-Lipschitz translations of a public base law, pinned at an anchor, and write $\mathfrak{R}_M^{\mathrm{det}}(\mathcal{T}_L^0)$ for the best worst-case d_∞ risk of a deterministic M-bit encoder/decoder (Appendix F). This oracle encoder is at least as informative as a state computed from a stream, so a lower bound for it also applies to continual algorithms.

Theorem 3 (Anchored-translation minimax lower bound)**.** *Under Assumption 4, every integer $M \geq 0$ satisfies*

$$\mathfrak{R}_M^{\mathrm{det}}(\mathcal{T}_L^0) \geq \frac{LR}{4}\left(\frac{c_1}{M+1}\right)^{1/D_{\mathcal{C}}}. \tag{3}$$

More generally, $\mathfrak{R}_M^{\mathrm{det}} \geq \varepsilon$ whenever $M < \log_2 P(\mathcal{T}_L^0, 2\varepsilon)$. If encoders and decoders are randomized and risk is measured in worst-case expected d_∞, then

$$\mathfrak{R}_{M,\mathrm{rand}}^{\mathrm{exp}}(\mathcal{T}_L^0) \geq \frac{LR}{8}\left(\frac{c_1}{M+1}\right)^{1/D_{\mathcal{C}}}.$$

The proof packs anchored Lipschitz translation bumps. It is a representation lower bound: allowing the encoder oracle access to the entire f avoids any unspoken stream-identifiability assumption.

The next statement supplies an upper construction on exactly the same class. Its extra representability premise is explicit.

Corollary 1 (Matching exponent before codebook resolution)**.** *Under Assumption 4, assume canonical condition nets with at most k points, chosen from the public metric space $\mathcal{C}$, are free shared problem data. Also assume a fixed public q-bit expert codebook $Q : \{0,1\}^q \to \mathcal{P}_2(\mathbb{R}^d)$ satisfies*

$$\sup_{\|u\| \leq LR} \min_{z \in \{0,1\}^q} W_2\big(Q(z), (\mathrm{id} + u)_{\#}\nu_0\big) \leq \eta.$$

Then for every integer $M \geq q$, with $k = \lfloor M/q \rfloor$,

$$\frac{LR}{4}\left(\frac{c_1}{M+1}\right)^{1/D_{\mathcal{C}}} \leq \mathfrak{R}_M^{\mathrm{det}}(\mathcal{T}_L^0) \leq \eta + Lc_2^{1/D_{\mathcal{C}}} R \lfloor M/q \rfloor^{-1/D_{\mathcal{C}}}. \tag{4}$$

Hence the two sides have exponent $-1/D_{\mathcal{C}}$ in the explicit pre-resolution range

$$\eta \leq Lc_2^{1/D_{\mathcal{C}}} R \lfloor M/q \rfloor^{-1/D_{\mathcal{C}}}.$$

Here η is an upper resolution offset *of the assumed codebook, not a lower error floor. Center locations are uncharged only because the canonical nets are public; learned routing metadata must otherwise be added to M.*

Figure 3 in Appendix G illustrates the same-class comparison in its pre-resolution range.

3.4 A RANK-ADAPTER DESIGN RULE

The often-quoted ratio $|\theta|b/[k_{\mathrm{drift}} \log_2(\Delta_{\mathrm{coef}}/\varepsilon_q)]$ is *not* the slack in Corollary 1. It is scheme-specific: if all displacements use a public shared k_{drift}-dimensional basis and each regime stores only quantized coefficients of range Δ_{coef} and precision ε_q, then the per-regime code costs approximately $k_{\mathrm{drift}} \log_2(\Delta_{\mathrm{coef}}/\varepsilon_q)$ bits, giving that ratio against a $|\theta|b$-bit full expert. Fresh LoRA factor pairs use a different accounting.

Appendix G defines the object a rigorous rank certificate applies to: over a weighted finite family of source-target pairs with selected optimal-transport maps, the displacements $T_\alpha - \mathrm{id}$ span a subspace of L^2, and the *weighted functional drift rank* $k_{\mathrm{drift}}(\tau)$ (Definition 1) is the smallest projection rank leaving at most a τ-fraction of their energy unexplained. Proposition 6 turns that tail into a weighted root-mean-square W_2 bound - and charges $r(d_{\mathrm{in}} + d_{\mathrm{out}})b$ bits per pair and layer *only* under the extra architectural assumption that the projected displacement is realized by a rank-r matrix update.

Design Principle 4 (Rank adapter)**.** If the adapted representation provides an explicit realization map from the leading functional displacement subspace to rank-r parameter updates, choose r from Definition 1 at the desired weighted error tolerance. Without such a realization, functional drift rank alone gives no parameter-memory guarantee.

Section 4.4 tests this proviso on a bottleneck backbone and on a bare linear forecaster. The latter is a useful negative control because low functional rank need not imply low matrix-update rank.

3.5 SATURATION AND STREAMING EVICTION

Write $e := C\bar{\epsilon}$ for the residual term in Theorem 1's certified upper bound.

Theorem 5 (Saturation of the certified upper bound)**.** *Suppose $\Delta > 0$ and $0 < e \le \Delta/2$, and define $k^\star := \min\{k \ge 1 : r_k \le e\}$. Then*

$$k^\star \le \min\left\{n, \left\lceil (\Delta/e)^D \right\rceil\right\}. \tag{5}$$

At $k \ge k^\star$, the certificate $r_k + e$ is at most $2e$; further center additions can improve this certificate by at most e. This is saturation of a proved upper certificate, not a claim that the true error has stopped improving or that e is a lower floor.

If $\Delta = 0$, every r_k is zero and one center suffices. If $\Delta > 0$ but $e = 0$, distinctness of the n metric points gives $k^\star = n$. The quantities n, Δ, D, and $\bar{\epsilon}$ are snapshot-dependent and can change as regimes arrive.

For the online case we use a complete incremental-doubling procedure (Charikar et al., 2004) with an exact W_2 oracle. Appendix I states its steps, together with the online k-center radius $r_t^\star(k)$ over the distinct regimes seen through time t and the realized clamping error $\mathrm{err}_{\mathrm{online},t}$ used below.

Theorem 6 (One-pass streaming certificate)**.** *With an exact W_2 metric oracle, the procedure above makes one pass over the stream, stores $O(k)$ center records, and at every time t has coverage radius at most $8r_t^\star(k)$. If every retained center has an expert satisfying Assumption 1, nearest-center clamping therefore gives*

$$\mathrm{err}_{\mathrm{online},t} \le 8r_t^\star(k) + C\bar{\epsilon}.$$

The $O(k)$ claim counts center records, not total bits. Retained expert parameters cost kq bits after warm-up; buffering all $k+1$ warm-up experts would cost $(k+1)q$, while selecting centers before instantiating the newest expert keeps the expert allocation at kq. Center descriptors and routing records are additional, and samples, sketches, or other state needed to implement the metric oracle are a separate memory term.

The proof maintains coverage within $4R_{\mathrm{alg}}$ and shows that every completed doubling phase has $R_{\mathrm{alg}} \le 2r_t^\star(k)$. It does not assert that an original evicted point remains within $2R_{\mathrm{alg}}$ after repeated merges. The policy spends memory on metric coverage rather than recency; after a center is removed, queries to its regime incur the surviving center's nonzero clamping distance.

3.6 DETECTING DRIFT ACROSS CORRELATED CHANNELS

Every mechanism above is triggered by a statistic deciding whether the current window is still consistent with an existing center. A detector based only on per-channel marginal MMD is blind to lead-lag changes that preserve each marginal but alter cross-channel dependence. We therefore append features from the window's pairwise correlation matrix to the MMD gate; Section 4.5 evaluates both its gains and its failure modes.

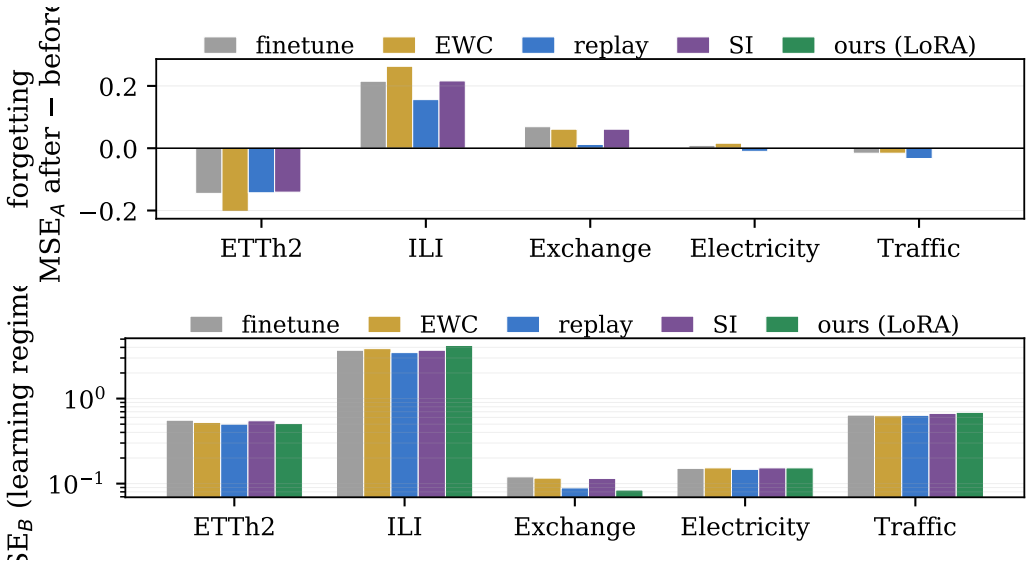


Figure 2: **Top:** Forgetting across five real datasets. Frozen-and-adapt is exactly zero everywhere; every baseline that keeps training shared parameters forgets by a dataset-dependent amount (occasionally negative, i.e. backward transfer, when B is easier). **Bottom:** MSE_B across the same datasets, log scale. Frozen-and-adapt is competitive but not uniformly best: wins Exchange, ties replay on ETTh2/Electricity, loses on ILI/Traffic.

4 Experiments

We organize experiments around quantities and mechanisms suggested by Section 3: zero measured MSE forgetting traded off against, not dominating, learning speed (Section 4.2); descriptive covering regressions on synthetic and real regime summaries (Section 4.3); a rank-adapter principle tested with and without a latent bottleneck (Section 4.4); the cross-channel drift statistic closing a controlled detection gap (Section 4.5); and, with the same candor as a positive result, a plausible Wasserstein-geometric account that does not survive measurement (Section 4.6). These measurements probe the theory's objects but do not certify its worst-scale covering assumptions or its restricted matching construction.

4.1 Setup

We use a patch-and-mixture-of-experts backbone with instance normalization on five public multivariate datasets spanning two orders of magnitude in channel count (7-862): ETTh2, illness surveillance (ILI), exchange rate, electricity and traffic. Each splits chronologically into past regime A (70%) and future B (30%); we train on A, continue on B per strategy, and report MSE_B and forgetting (increase in MSE_A). Baselines are fine-tuning, EWC, SI and replay (50% mixing), against frozen-and-adapt with a rank-16 projection adapter. Details and full tables are in Appendix M.

4.2 Zero forgetting is real, and it is not a free lunch

Figure 2 (top) shows forgetting is zero everywhere for frozen-and-adapt because the parameters used by regime A remain frozen - unsurprising by construction; the informative question is cost. This is an empirical MSE identity for our implementation, not a prediction of Theorem 1, whose certificate is stated in W_2.

Figure 2 (bottom) shows the tradeoff we set out to characterize, not resolve: no strategy wins both axes. Replay achieves the best or near-best MSE_B on four of five datasets by paying small non-zero forgetting; frozen-and-adapt pays zero forgetting and is competitive but loses on ILI (a small window where unregularized adaptation overfits and replay's mixture regularizes, which our adapter lacks) and Traffic (largest drift in our set, where rank-16 is under-capacity, consistent with Section 4.4). Forgetting-free is achievable; achievable and dominant are different claims.

4.3 DESCRIPTIVE SCALING DIAGNOSTICS

Synthetic regime sequences of known intrinsic dimension check that the fitting pipeline recovers the planted power-law slope (Figure 6, left, in Appendix M). On four real datasets extended to $T = 30$ regimes, the descriptive slope $\hat{D}$ is low (≈ 0.8-2.3), suggesting that these sampled Gaussian regime summaries occupy a far lower-dimensional set than the raw channel count. It is nevertheless a finite-sample average, not a certificate of the scale-uniform exponent required by Theorem 1, and it does not test Corollary 1's anchored translation/model-realizable class, public net, or codebook assumptions. Because no certified $C\bar{\epsilon}$ was estimated, the displayed $k^{\star}_{\mathrm{est}}$ values use the illustrative offsets reported in Appendix M; they demonstrate the proxy calculation but are not empirical evidence that the true diminishing-returns threshold has or has not been reached.

4.4 WHEN DOES A LOW-RANK ADAPTER ACTUALLY WORK?

Design Principle 4 proposes that $r \approx k_{\mathrm{drift}}$ can suffice for a particular adapter/shared-basis scheme *only* when drift routes through a compressible latent space; it is not a claim about minimax slack. We test both sides by sweeping rank on each. On the MoE backbone the sweep first appeared to refute the principle: at the main benchmark's 30-epoch adapter schedule, MSE_B is flat from $r{=}1$ to $r{=}64$ and a freeze-only baseline beats every adapted variant. Decoupling adapter epochs from backbone epochs explains why (Appendix M.7): the adapter over-trains on B, worsening with epochs, which both inflates MSE_B and washes out any rank dependence. At a one-epoch budget the adapter beats freeze-only on both datasets and $r{=}4$ beats $r{=}16$ - with $k_{\mathrm{drift}} = 4$ measured independently (Section 4.3) - the first evidence for the rank rule on this backbone, though from a single seed and two ranks. It also means our main tables report an over-trained adapter and understate frozen-and-adapt. On the failure side, a bare linear forecaster swept directly on output weights (Figure 7a, Appendix M) never closes the gap to fine-tuning even at rank 256, consistent with the proviso: parameter-space $\mathrm{rank}(W_B - W_A) = 469$, two orders of magnitude above that pair's data-space k_{drift}. Low-rank in data space and in parameter space are different statements; only a latent bottleneck licenses moving between them.

4.5 CROSS-CHANNEL DRIFT DETECTION

Figure 7b (Appendix M) confirms the gap motivating Section 3.6: on drift preserving each channel's marginal statistics while changing lead-lag structure, marginal MMD is nearly identical within- and across-regime (near-blind), while the cross-channel feature separates the cases clearly. This statistic drives the confirmation gate in Figure 1. But it is a capability demonstration under controlled ground truth, and measuring detection on the *real* streams against their $70/30$ cut, with block lag matched between classes, neither statistic beats chance usefully (AUROC 0.24-0.64; Appendix N.1) - largely because that cut is a chronological convenience, not an annotated event. The automatic trigger is thus validated in principle, not on real data: an open gap, not a result.

4.6 A SELECTED-GEODESIC HYPOTHESIS, AND WHERE IT FAILS

For each old/new pair we selected an optimal Wasserstein coupling and conjectured that the fine-tuned forecast remains close to its associated geodesic. On two real trajectories, the off-path residual instead exceeds the along-path term by 2-$5\times$ (Appendix J). This rejects the small-residual hypothesis for the selected geodesics on those pairs, not for every geodesic when optimal couplings are nonunique. It is also not a theorem about empirical MSE forgetting, and the two quantities should not be conflated.

5 LIMITATIONS

The theory has a deliberately limited scope. The general frozen-expert results are upper certificates, not universal memory necessities. The matching exponent is proved only for the anchored translation/model-realizable class with public condition nets and a q-bit codebook, and only before the codebook-resolution offset η dominates; it does not establish minimax optimality over arbitrary $\mathcal{P}_2(\mathbb{R}^d)$-valued maps. The additive terms $B + C\bar{\epsilon}$ and η are upper-bound offsets, not proved floors. Every result also assumes the regime label c is observable at test time - appropriate for regime-based forecasting, not latent task identity - and the bounds require the stated uniform residual control.

The streaming result assumes exact W_2 access and proves an 8-approximation with $O(k)$ center records; it neither counts expert, router, sample, or sketch bits nor proves robustness to estimated

distances. We never run doubling-and-merge far enough to force an eviction, and no experiment revisits an evicted regime. Likewise, the fitted $\hat{D}$, diameter, and $k^{\star}_{\mathrm{est}}$ from $T = 30$ regimes are descriptive finite-sample summaries rather than certificates of the all-scale covering hypothesis. Corollary 3 bounds W_2 error, not empirical MSE, and its path-independent envelope needs a residual bound uniform across the compared paths; real sequences are not collinear enough to provide the matched-endpoint test. The two-dataset clamping/conditioning MSE comparison is descriptive only and cannot validate Corollary 2's ordering of upper certificates.

Empirically, five gaps matter. (i) On Electricity and Traffic, quadratic-cost heads were disabled and a seasonal expert ran at reduced batch size to avoid out-of-memory failures, on a subsample with fewer epochs, so we trust those numbers less. (ii) Replay's occasional MSE_B edge buys a stored regime-A sample that frozen-and-adapt forgoes; it is not an equalized-budget win. Appendix L's Wasserstein replay scaling is illustrative under its moment assumptions, not a necessary replay-memory lower bound, and its crossover is unswept on real data. (iii) Our certificates are in W_2 while we report MSE, a mismatch we do not reconcile - though the Gaussian surrogate moves the fitted descriptive exponent by under 1.5% against exact optimal transport (Appendix M.6). Zero MSE forgetting follows from frozen parameters in this implementation, not from the W_2 theorem. (iv) The adapter trains for a fixed 30 epochs with no validation-based stopping, which over-trains it (Appendix M.7): the main tables therefore *understate* frozen-and-adapt, and the rank principle is supported only at a one-epoch budget on two datasets and one seed. Main tables are otherwise single-seed, replicated across three seeds for the low-channel datasets only (Appendix M.5). (v) We do not benchmark the dynamic-rank or structural-knowledge continual-learning methods of Section 2, and the cross-channel detector, measured on real streams, does not beat chance against the $70/30$ cut once block lag is matched (Appendix N.1) - so its automatic-trigger role rests on the synthetic construction alone.

6 Conclusion

This paper asked what can be certified about the memory-error behavior of frozen experts in continual forecasting, rather than treating "freeze and grow" as a black-box fix. Treating served regimes as a metric space yields covering-based upper certificates. A matching exponent is obtained only for an anchored translation/model-realizable class with public condition nets and a q-bit codebook, before the codebook-resolution offset dominates; it is not universal minimax optimality over arbitrary Wasserstein-valued maps, and the additive offsets are not proved floors. Within a particular shared-basis construction, drift rank gives a conditional adapter-sizing rule rather than a minimax-gap closure. The online result is likewise precise but idealized: an exact-W_2-oracle 8-approximation using $O(k)$ center records. On five real datasets, frozen parameters produce exactly zero measured MSE forgetting while learning competitively but not dominantly; on two trajectories, proximity to a selected Wasserstein geodesic is rejected. The next steps are the ones our measurements forced: validation-based adapter stopping, finite-sample routing with complete memory accounting, eviction at a scale that merges, and annotated change points on which a detector can be scored.

AI use statement

In this work, generative AI tools were used to (i) create or modify visual figures based on our experimental results, (ii) edit English grammar, phrasing, and vocabulary, and (iii) assist in producing Lean formalizations for checking selected mathematical equations. The tools were not used to generate data, formulate the research questions or theoretical claims, design or run experiments, analyze or interpret experimental results, or make scientific decisions. They were also not used to replace the authors' judgment about the scientific content.

We reviewed all AI-assisted work. The experimental data and results were checked against the original runs and stored result files; AI-assisted figures were checked against those results; and the Lean formalizations were compiled and checked by the authors to ensure that they corresponded to the equations stated in the paper. AI-assisted language edits were manually reviewed for meaning and accuracy. The authors take responsibility for the final content of this work, including all text, claims, figures, and formalizations produced with AI assistance.

REFERENCES

Quentin Besnard and Nicolas Ragot. Attention-based experience replay framework for continual learning of agnostic time series forecasting models. In *CAp & RFIAP*, Montpellier, France, 2026.

Prashant Shivaram Bhat, Shakib Yazdani, Elahe Arani, and Bahram Zonooz. Parameter efficient continual learning with dynamic low-rank adaptation. *Transactions on Machine Learning Research (TMLR)*, 2026.

Moses Charikar, Chandra Chekuri, Tomas Feder, and Rajeev Motwani. Incremental clustering and dynamic information retrieval. *SIAM Journal on Computing*, 33(6):1417–1440, 2004.

Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. Dish-ts: A general paradigm for alleviating distribution shift in time series forecasting. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 7522–7529, 2023.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

Fabian Hinder, André Artelt, Valerie Vaquet, and Barbara Hammer. Precise change point detection using spectral drift detection. *arXiv preprint arXiv:2205.06507*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jangho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations (ICLR)*, 2022.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Indar Kumar, Akanksha Tiwari, Sai Krishna Jasti, and Ankit Hemant Lade. Rg-tta: Regime-guided meta-control for test-time adaptation in streaming time series. *arXiv preprint arXiv:2603.27814*, 2026.

Hongbo Li, Sen Lin, Lingjie Duan, Yingbin Liang, and Ness B. Shroff. Theory on mixture-of-experts in continual learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

Ronghao Ni, Zinan Lin, Shuaiqi Wang, and Giulia Fanti. Mixture-of-linear-experts for long-term time series forecasting. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Zijie Pan, Yushan Jiang, Dongjin Song, Sahil Garg, Kashif Rasul, Anderson Schneider, and Yuriy Nevmyvaka. Structural knowledge informed continual multivariate time series forecasting. *arXiv preprint arXiv:2402.12722*, 2024.

Long H. C. Pham. Decomposing out-of-distribution error in conditional flow matching via wasserstein geometry. In *International Conference on Machine Learning (ICML)*, 2026.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Edoardo Urettini, Daniele Atzeni, Ioanna-Yvonni Tsaknaki, and Antonio Carta. Online continual learning for time series: a natural score-driven approach. *arXiv preprint arXiv:2601.12931*, 2026.

Zhijian Xu, Ailing Zeng, and Qiang Xu. Fits: Modeling time series with 10k parameters. In *International Conference on Learning Representations (ICLR)*, 2024.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning (ICML)*, pp. 3987–3995, 2017.

Tianxiang Zhan, Ming Jin, Yuanpeng He, Yuxuan Liang, and Shirui Pan. Continuous evolution pool: Taming recurring concept drift in online time series forecasting. *arXiv preprint arXiv:2506.14790*, 2025.

Xin Zhang, Liang Bai, Xian Yang, and Jiye Liang. C-lora: Continual low-rank adaptation for pre-trained models. *arXiv preprint arXiv:2502.17920*, 2025.

A ASSUMPTIONS AND NOTATION

Table 1 collects notation used throughout the appendix; it repeats and extends the notation introduced in Section 3.1.

Table 1: Notation.

Symbol	Meaning
$(\mathcal{P}_2(\mathbb{R}^d), W_2)$	Wasserstein-2 space; d is the *ambient* dimension of forecast targets
$\mathcal{C}, d_{\mathcal{C}}$	condition (regime) space and its metric
$f : \mathcal{C} \to \mathcal{P}_2$, $c \mapsto \mu_c$	conditional ground truth
$\hat{p}(\cdot \mid c; \theta)$	distribution generated by model θ conditioned on c
$\mathcal{M} = \{\mu_{c_1}, \dots, \mu_{c_T}\}$	the set of regimes actually visited, inside $\mathcal{P}_2$
D	internal covering exponent of $\mathcal{M}$ in W_2 (Assumption 5) – *intrinsic*, distinct from d
$d_j := W_2(\mu_{c_{j-1}}, \mu_{c_j})$	one-step drift
ϵ_j	flow/forecast training residual on regime j; C the fit constant of Assumption 1

Throughout, $C \geq 0$ and every training residual $\epsilon_i \geq 0$. Whenever a theorem uses $\bar{\epsilon}$, it assumes the fit bound uniformly over all eligible centers and sets $\bar{\epsilon} := \sup_i \epsilon_i < \infty$. Define the unit-time Grönwall factor

$$\Gamma_x := \begin{cases} (e^{L_x} - 1)/L_x, & L_x > 0, \\ 1, & L_x = 0. \end{cases}$$

Assumption 2 (Flow realization and uniform Lipschitz stability)**.** There is a common initial law $\lambda_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and a common horizon $t \in [0, 1]$. For every admissible θ and c, the ODEs $\dot{X}_t = v_\theta(t, X_t, c)$ and $\dot{X}_t = v^\star(t, X_t, c)$ have measurable flow maps $\Phi_t^{\theta,c}$ and $\Phi_t^{\star,c}$ for λ_0-almost every initial point, with endpoint laws

$$\hat{p}(\cdot \mid c; \theta) = (\Phi_1^{\theta,c})_{\#}\lambda_0, \qquad \mu_c = (\Phi_1^{\star,c})_{\#}\lambda_0.$$

Uniformly in t, θ, c, both fields are L_x-Lipschitz in x; v_θ is L_c-Lipschitz in c, and $v^\star$ is $L_c^\star$-Lipschitz in c, for $L_x, L_c, L_c^\star \geq 0$. The fields have sufficient measurability and growth for these flows and endpoint laws to exist in $\mathcal{P}_2(\mathbb{R}^d)$.

Assumption 3 (Coarse embedding)**.** There are $a, A > 0$ and $b_{\mathrm{emb}}, B \geq 0$ such that, for all $c, c' \in \mathcal{C}$,

$$a\, d_{\mathcal{C}}(c, c') - b_{\mathrm{emb}} \leq W_2(\mu_c, \mu_{c'}) \leq A\, d_{\mathcal{C}}(c, c') + B.$$

For a nonempty metric space X, let $N_X^{\mathrm{int}}(r)$ be the least number of closed radius-r balls, centered in X, that cover X. A *strict r-packing* has pairwise distances $> r$, and $P_X(r)$ is the supremum of its cardinalities. Every maximal strict r-packing is an internal r-net, hence $N_X^{\mathrm{int}}(r) \leq P_X(r)$.

Assumption 4 (Power-law entropy of $\mathcal{C}$)**.** The space $\mathcal{C}$ is nonempty and totally bounded, $R := \mathrm{diam}(\mathcal{C}) \in (0, \infty)$, and there are $D_{\mathcal{C}}, c_1, c_2 > 0$ such that

$$c_1(R/r)^{D_{\mathcal{C}}} \leq N_{\mathcal{C}}^{\mathrm{int}}(r) \leq c_2(R/r)^{D_{\mathcal{C}}}, \qquad 0 < r \leq R.$$

The excluded singleton case $R = 0$ is trivial for both covering and the anchored lower-bound class.

Assumption 5 (Internal covering exponent of $\mathcal{M}$)**.** Let $n := |\mathcal{M}|$ and $\Delta := \mathrm{diam}_{W_2}(\mathcal{M})$. In the nondegenerate case $n \geq 2$ and $\Delta > 0$, there is $D > 0$ such that

$$N_{\mathcal{M}}^{\mathrm{int}}(r) \leq (\Delta/r)^D, \qquad 0 < r \leq \Delta/2.$$

If $\Delta = 0$, then $\mathcal{M}$, as a set of laws, is a singleton and one center has radius zero.

Assumption 4 supplies the scale-uniform constants needed in Appendices F and E; no separate packing regularity is required. Assumption 5's half-diameter restriction is essential: three mutually equidistant points need three internal balls at every radius below the diameter, although $(\Delta/r)^D \to 1$ as $r \uparrow \Delta$. Thus its integer-budget inversion is used only for $k \geq \lceil 2^D \rceil$.

Assumption 1's C is an opaque fit constant and is not identified with e^{L_x} or Γ_x. The theorem errors are in Wasserstein distance units. Empirical MSE forgetting is a different quantity and requires a separate loss-comparison argument.

Theorem 1 concerns the finite observed set, whereas Theorem 3 concerns all of $\mathcal{C}$; those two results do not form a minimax sandwich. The only like-for-like comparison below restricts both sides to the anchored translation class $\mathcal{T}_L^0$ and assumes a public expert codebook with uniform resolution offset η. The generic additive quantities B and $C\bar{\epsilon}$ are upper-certificate slack, never lower error floors.

B WHY CLAMPING, NOT CONDITIONING

Consider an expert θ_i trained on regime c_i, queried to serve regime c_j. Two test-time policies are available.

Proposition 1 (Clamping bound)**.** *Under Assumption 1 alone,*

$$W_2\big(\hat{p}(\cdot \mid c_i;\theta_i), \mu_{c_j}\big) \;\le\; C\epsilon_i + W_2(\mu_{c_i}, \mu_{c_j}).$$

Proof. Triangle inequality through μ_{c_i}: $W_2(\hat{p}(\cdot|c_i;\theta_i), \mu_{c_j}) \le W_2(\hat{p}(\cdot|c_i;\theta_i), \mu_{c_i}) + W_2(\mu_{c_i}, \mu_{c_j})$, and the first term is $\le C\epsilon_i$ by Assumption 1. □

Proposition 2 (Conditioning bound)**.** *Under Assumptions 1–3, for the conditioning policy $\hat{p}(\cdot \mid \bar{c};\theta_i)$ evaluated at any $\bar{c}$,*

$$W_2\big(\hat{p}(\cdot \mid \bar{c};\theta_i), \mu_{\bar{c}}\big) \le K_\star W_2(\mu_{c_i}, \mu_{\bar{c}}) + \delta_\star + C\epsilon_i, \quad K_\star = 1 + \frac{\Gamma_x L_c}{a}, \quad \delta_\star = \frac{\Gamma_x L_c b_{\mathrm{emb}}}{a}.$$

Proof. Couple both model flows by the same $X_0 \sim \lambda_0$. If $z(t) := \|\Phi_t^{\theta,\bar{c}}(X_0) - \Phi_t^{\theta,c_i}(X_0)\|$, then $z'(t) \le L_x z(t) + L_c d_{\mathcal{C}}(c_i, \bar{c})$ and $z(0) = 0$. Grönwall gives $z(1) \le \Gamma_x L_c d_{\mathcal{C}}(c_i, \bar{c})$, hence the same bound on the Wasserstein distance between the two endpoint laws. Insert these endpoint laws and μ_{c_i} in the triangle inequality, use Assumption 1, and then use $d_{\mathcal{C}}(c_i, \bar{c}) \le [W_2(\mu_{c_i}, \mu_{\bar{c}}) + b_{\mathrm{emb}}]/a$. □

Corollary 2 (Clamping dominates conditioning)**.** *For the same i and $\bar{c}$, the conditioning* certificate *exceeds the clamping certificate by*

$$\frac{\Gamma_x L_c}{a}\big(W_2(\mu_{c_i}, \mu_{\bar{c}}) + b_{\mathrm{emb}}\big) \ge 0.$$

The inequality is strict exactly when $L_c > 0$ and $W_2(\mu_{c_i}, \mu_{\bar{c}}) + b_{\mathrm{emb}} > 0$. This compares the two proved upper certificates; it does not order their realized errors.

This is a negative result about what is *provable* under our stated assumptions, not a claim that conditioning is useless in practice; a model whose true conditional map f has more structure than Assumptions 2–3 capture could benefit from conditioning, and Section 4.5's companion measurement (clamping vs. conditioning, Appendix M) is precisely designed to detect that case empirically. All theorems in Section 3 use clamping and therefore carry constant 1, not $K_\star$.

C COMPARISON TO A DIRECT APPLICATION OF THE WASSERSTEIN OOD DECOMPOSITION

Pham (2026) decomposes out-of-distribution error for conditional flow matching into a drift term, a distortion term, and a training-fit term. On the common unit horizon and with the endpoint realization of Assumption 2, its corresponding condition-sensitivity factor is $K_{\mathrm{T1}} = \Gamma_x(L_c + L_c^\star)/a$. The comparison below is only between upper certificates.

Proposition 3 (Structural floor on K_{T1})**.** *Under Assumptions 2 and 3, assume $d_{\min} > 0$ and $L_c^\star > 0$. For any pair with $d_{\mathcal{C}}(c, c') \ge d_{\min}$,*

$$K_{\mathrm{T1}} \;\ge\; \Big(1 - \tfrac{b_{\mathrm{emb}}}{a\, d_{\min}}\Big)\Big(1 + \tfrac{L_c}{L_c^\star}\Big).$$

In particular, if $b_{\mathrm{emb}} = 0$ and $L_c \ge L_c^\star$, then $K_{\mathrm{T1}} \ge 2$.

Proof. Coupling the two true flows through the common $X_0 \sim \lambda_0$ gives $W_2(\mu_c, \mu_{c'}) \le \Gamma_x L_c^\star d_{\mathcal{C}}(c, c')$. Combining this with Assumption 3's lower half and dividing by $a d_{\mathcal{C}}(c, c') > 0$ yields $\Gamma_x L_c^\star/a \ge 1 - b_{\mathrm{emb}}/(a d_{\mathcal{C}}(c, c')) \ge 1 - b_{\mathrm{emb}}/(a d_{\min})$. Multiply by $1 + L_c/L_c^\star$. □

Proposition 4 (No first-order improvement over the triangle bound)**.** *Under Proposition 3's hypotheses, write $\delta_{\mathrm{T1}} := \Gamma_x(L_c + L_c^\star) b_{\mathrm{emb}}/a$. Then*

$$K_{\mathrm{T1}} - K_\star = \frac{\Gamma_x L_c^\star}{a} - 1 \ge -\frac{b_{\mathrm{emb}}}{a d_{\min}}, \qquad \delta_{\mathrm{T1}} - \delta_\star = \frac{\Gamma_x L_c^\star b_{\mathrm{emb}}}{a} \ge 0.$$

Let $W := W_2(\mu_c, \mu_{c'})$ and $\gamma := \Gamma_x L_c^\star / a$. The same flow and embedding inequalities give $W \le a\gamma d_{\mathcal{C}}(c, c') \le \gamma(W + b_{\mathrm{emb}})$, and hence the complete OOD certificate is no smaller than the conditioning certificate:

$$\{K_{\mathrm{T1}} W + \delta_{\mathrm{T1}}\} - \{K_\star W + \delta_\star\} = \gamma(W + b_{\mathrm{emb}}) - W \ge 0.$$

The componentwise additive inequality is strict when $b_{\mathrm{emb}} > 0$. No claim that either multiplicative constant is at least one is made when $b_{\mathrm{emb}} > 0$. Moreover, for any real certificates z_i and weights $w_i \ge 0$ with $\sum_i w_i = 1$, the exact direction is $\min_i z_i \le \sum_i w_i z_i$; thus the best componentwise certificate is no larger than their convex average.

We take Propositions 3–4 as a preliminary structural finding, not a rebuttal of Pham (2026) outside our setting: the comparison assumes the decomposition's training-fit term bounds Assumption 1-type residual for *every* training regime (its definition via a $\max_i$), matching the source's construction; a decomposition defined differently (e.g. only at a weighted mixture) would need a separate comparison.

D Proof of Theorem 1 (upper bound)

Let $k \ge 1$ and let $S \subseteq \mathcal{M}$ be nonempty with $|S| \le k$. Assumption 1 supplies the uniform residual $\epsilon_i \le \bar{\epsilon}$ over every eligible center. Nearest-center clamping and Proposition 1 give

$$\max_{j \le T} W_2(\hat{p}_j, \mu_{c_j}) \le r(S) + C\bar{\epsilon}.$$

Taking an infimum over S (rather than silently assuming an optimizer in a more general space) gives the first assertion with

$$r_k := \inf_{\substack{\varnothing \ne S \subseteq \mathcal{M} \\ |S| \le k}} r(S).$$

Here the infimum is attained because $\mathcal{M}$ is finite, but the notation remains valid in the full-space analogue.

The covering inversion has three cases. If $\Delta = 0$ or $k \ge n$, then $r_k = 0$. If $\Delta > 0$, $k < n$, and $k \ge \lceil 2^D \rceil$, put $r = \Delta k^{-1/D} \le \Delta/2$. Assumption 5 supplies an internal cover using at most $(\Delta/r)^D = k$ centers, hence $r_k \le r$. If $1 \le k < \min\{n, \lceil 2^D \rceil\}$, only $r_k \le \Delta$ follows from the stated assumption.

For the expert-parameter budget used in Theorem 1, let $q := |\theta| b > 0$ be the bits per full expert and

$$k_M := \min\{n, \lfloor M/q \rfloor\}.$$

When $k_M \ge 1$, substitute this integer budget in the preceding display; there is no equality $k = M/q$ unless M/q is an integer. In particular, in the power-law range $\lceil 2^D \rceil \le k_M < n$,

$$\mathrm{err}_{\mathcal{M}}(M) \le \Delta k_M^{-1/D} + C\bar{\epsilon}.$$

Routing metadata and stored samples are outside the theorem's expert-parameter budget and must be added for total-memory accounting. □

The centers-in-$\mathcal{M}$ convention is essential. Unrestricted covering centers generally require conversion back to data centers and can enlarge the radius. Likewise D here is the stated internal covering exponent, not a doubling dimension; a doubling-only hypothesis yields a bound with an additional constant.

E Proof of Theorem 2 (upper bound over the condition space)

Notation used in Theorem 2. For a fixed full map f, define the condition-space radius

$$r_{\mathcal{C}}(k) := \inf_{\substack{S \subseteq \mathcal{C} \\ 1 \le |S| \le k}} \sup_{c \in \mathcal{C}} \min_{s \in S} d_{\mathcal{C}}(c, s).$$

We analogously write $\mathrm{err}_{\mathcal{C}}^{(f)}(M)$ for the infimal worst-condition error of this nearest-center clamping scheme for the fixed map f, among choices using at most $\lfloor M/q \rfloor$ eligible experts.

Let $S \subseteq \mathcal{C}$ be a nonempty finite set with $|S| \le k$, and assume the residual bound $\epsilon_i \le \bar{\epsilon}$ uniformly for every $c_i \in S$. For each c, choose a nearest center $i(c)$ (finite S makes the minimum attainable). Proposition 1 gives

$$W_2\big(\hat{p}(\cdot \mid c_{i(c)}; \theta_{i(c)}), \mu_c\big) \;\le\; C\epsilon_{i(c)} + W_2\big(\mu_{c_{i(c)}}, \mu_c\big).$$

The upper half of Assumption 3 and a supremum over c yield $C\bar{\epsilon} + Ar_{\mathcal{C}}(S) + B$.

For the covering inversion, let $k \ge 1$. If $k \ge c_2$, set $\rho_k := c_2^{1/D_c} R k^{-1/D_c} \le R$; then Assumption 4 gives $N_{\mathcal{C}}^{\mathrm{int}}(\rho_k) \le k$. If $1 \le k < c_2$, one arbitrary center has radius at most R. Therefore

$$r_{\mathcal{C}}(k) := \inf_{\substack{\varnothing \ne S \subseteq \mathcal{C} \\ |S| \le k}} r_{\mathcal{C}}(S) \le R \min\left\{1, \left(\frac{c_2}{k}\right)^{1/D_c}\right\}.$$

With $q := |\theta| b > 0$ and $k_M := \lfloor M/q \rfloor \ge 1$, the memory form uses k_M, not M/q:

$$\mathrm{err}_{\mathcal{C}}^{(f)}(M) \le C\bar{\epsilon} + B + AR \min\left\{1, \left(\frac{c_2}{k_M}\right)^{1/D_c}\right\}.$$

This conclusion requires uniform representability/fit for every center in the chosen net; per-instance fit on an already fixed finite set is not sufficient. □

The finite-set and full-space certificates have different scopes. Merely sharing the domain $\mathcal{C}$ with a lower bound is also not enough: a minimax comparison requires the same function class, information model, memory accounting, and a uniform representability premise. Appendix G states that restricted comparison explicitly.

F Proof of Theorem 3 (minimax lower bound)

The class, the information model, and the risk. To obtain a finite, well-posed minimax problem, fix $L > 0$, $c_0 \in \mathcal{C}$, and a public base law $\nu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, and consider the anchored translation class

$$\mathcal{T}_L^0 := \left\{ f_\tau(c) = (\mathrm{id} + \tau(c))_{\#}\nu_0 : \mathrm{Lip}(\tau) \le L,\ \tau(c_0) = 0,\ \tau : \mathcal{C} \to \mathbb{R}^d \right\}.$$

Anchoring implies $\|\tau(c)\| \le LR$ and removes arbitrary global translations. For functions g, h let $d_\infty(g,h) := \sup_{c \in \mathcal{C}} W_2(g(c), h(c))$. A deterministic M-bit representation consists of an encoder $E : \mathcal{T}_L^0 \to \{1, \dots, 2^M\}$ and a decoder D from a state and condition to a forecast law. Define

$$\mathfrak{R}_M^{\mathrm{det}}(\mathcal{T}_L^0) := \inf_{E,D} \sup_{f \in \mathcal{T}_L^0} d_\infty(D(E(f), \cdot), f).$$

We use the strict packing convention: $P(\mathcal{G}, \rho)$ is the largest family with every pair at distance strictly greater than ρ.

Fix an anchor $c_0 \in \mathcal{C}$, a base law $\nu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, and $L > 0$. The lower bound is for the anchored translation class

$$\mathcal{T}_L^0 := \{ f_\tau(c) = (\mathrm{id} + \tau(c))_{\#}\nu_0 : \tau(c_0) = 0,\ \mathrm{Lip}(\tau) \le L \}.$$

Anchoring makes the problem totally bounded at the relevant scales: it removes the arbitrary global translations that make the minimax risk over the unanchored Lipschitz class infinite. Equip this class with $d_\infty(f,g) := \sup_{c \in \mathcal{C}} W_2(f(c), g(c))$ and use the strict-packing convention from Appendix A.

Step 1 (anchored bump packing). Fix $0 < \varepsilon \le LR/4$ and put $s := 4\varepsilon/L \le R$. Let $Z = \{z_1, \dots, z_N\}$ be a maximal strict s-packing. It is an internal s-net, so

$$N \ge N_{\mathcal{C}}^{\mathrm{int}}(s) \ge c_1 (R/s)^{D_c} = c_1 (LR/(4\varepsilon))^{D_c}.$$

At most one closed ball $\bar{B}(z_i, s/2)$ contains c_0, because two such centers would be at distance at most s. Remove that center if it exists and relabel the remaining m centers. Then

$$m \ge c_1 (LR/(4\varepsilon))^{D_c} - 1.$$

Choose a unit vector $u \in \mathbb{R}^d$ and define

$$\phi_i(c) := L \max\{0, s/2 - d_{\mathcal{C}}(c, z_i)\}, \qquad \tau_\sigma(c) := u \sum_{i=1}^{m} \sigma_i \phi_i(c), \quad \sigma \in \{-1, 1\}^m.$$

The supports are disjoint and avoid c_0, so $\tau_\sigma(c_0) = 0$. Each map is L-Lipschitz. Indeed, within one support this follows from the reverse triangle inequality; between supports $i \neq j$,

$$\|\tau_\sigma(c) - \tau_\sigma(c')\| \leq L\{s - d(c, z_i) - d(c', z_j)\} < Ld(c, c'),$$

where strict separation gives $d(z_i, z_j) > s$; the support-to-complement case is the same one-cone estimate. If $\sigma \neq \sigma'$, then at a differing coordinate i,

$$\|\tau_\sigma(z_i) - \tau_{\sigma'}(z_i)\| = Ls = 4\varepsilon > 2\varepsilon.$$

Finally, translations satisfy

$$W_2\big((\mathrm{id} + v)_\# \nu_0, (\mathrm{id} + w)_\# \nu_0\big) = \|v - w\| :$$

the common translation coupling gives "$\leq$", and the difference of means gives "$\geq$". Thus the 2^m maps f_{τ_σ} form a strict 2ε-packing in d_∞.

Step 2 (deterministic state counting). An M-bit deterministic learner has at most 2^M final states and therefore at most 2^M decoded maps. If $m \geq M + 1$, the packing contains more maps than states, so two targets share a decoded map. Their strict 2ε separation and the triangle inequality imply that the learner has worst-case error $> \varepsilon$ on at least one. Every

$$0 < \varepsilon < \varepsilon_M := \frac{LR}{4}\left(\frac{c_1}{M+1}\right)^{1/D_\mathcal{C}}$$

has $c_1(LR/(4\varepsilon))^{D_\mathcal{C}} > M + 1$, and hence $m \geq c_1(LR/(4\varepsilon))^{D_\mathcal{C}} - 1 > M$. Because m is an integer, $m \geq M + 1$. Taking $\varepsilon \uparrow \varepsilon_M$ gives

$$\mathfrak{R}_M^{\mathrm{det}}(\mathcal{T}_L^0) := \inf_{E,D} \sup_{f \in \mathcal{T}_L^0} d_\infty(D(E(f), \cdot), f) \geq \varepsilon_M.$$

(Assumption 4 at $r = R$ implies $c_1 \leq 1$, so $\varepsilon_M \leq LR/4$.)

Step 3 (randomized expected risk). Put the uniform prior on the 2^m packed targets. For any deterministic M-bit learner, at most 2^M targets can have loss at most ε, because one decoded map cannot lie in two disjoint closed ε-balls. When $m \geq M + 1$, at least half the targets therefore have loss $> \varepsilon$, and the prior-average loss is $> \varepsilon/2$. Yao's minimax principle now gives

$$\inf_{\text{randomized alg.}} \sup_{f \in \mathcal{T}_L^0} \mathbb{E}\, d_\infty(\widehat{f}, f) \geq \frac{\varepsilon_M}{2}.$$

The expectation is over the algorithm's randomness. This factor $1/2$ is necessary for the stated Yao argument; conditioning on a random seed alone would not prove a worst-case expected lower bound. □

G THE RANK-ADAPTER GAP AND DERIVATION OF DESIGN PRINCIPLE 4

Here is the codebook upper construction used in Corollary 1. Let canonical k-point internal nets of the public metric space $\mathcal{C}$ be shared problem data, and let the public decoder $Q : \{0,1\}^q \to \mathcal{P}_2(\mathbb{R}^d)$ satisfy

$$\sup_{\|u\| \leq LR} \min_{z \in \{0,1\}^q} W_2\big(Q(z), (\mathrm{id} + u)_\# \nu_0\big) \leq \eta.$$

For $f_\tau \in \mathcal{T}_L^0$, anchoring gives $\|\tau(c_i)\| \leq LR$, so store at each net point c_i a codeword z_i attaining (or approaching arbitrarily closely) this resolution. Route c to its nearest net point and decode $Q(z_i)$. Translation invariance and Lipschitzness give

$$W_2(Q(z_i), f_\tau(c)) \leq \eta + \|\tau(c_i) - \tau(c)\| \leq \eta + Ld_\mathcal{C}(c_i, c).$$

With $k = \lfloor M/q \rfloor \geq 1$, Assumption 4's covering inversion and Theorem 3 therefore give the same-class sandwich

$$\frac{LR}{4}\left(\frac{c_1}{M+1}\right)^{1/D_\mathcal{C}} \leq \mathfrak{R}_M^{\mathrm{det}}(\mathcal{T}_L^0) \leq Lc_2^{1/D_\mathcal{C}} Rk^{-1/D_\mathcal{C}} + \eta.$$

The randomized expected lower envelope is one half of the displayed lower term. This is a genuine sandwich because the domain, target class, loss, state budget, and anchor agree. The right-hand

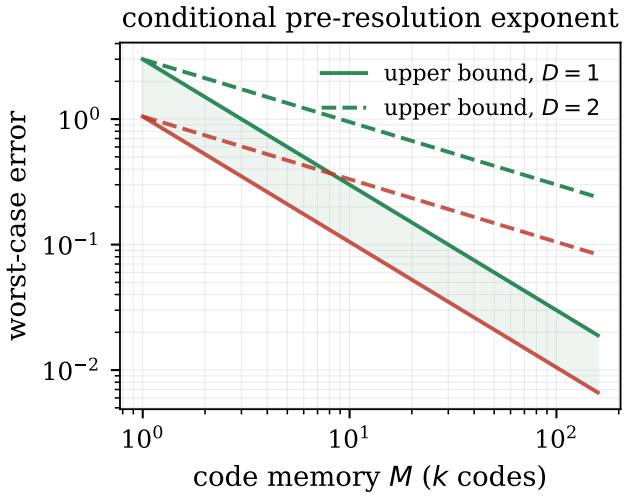


Figure 3: Corollary 1's same-class bounds have slope $-1/D_{\mathcal{C}}$ while the covering term dominates the codebook's upper resolution offset η. The shaded band represents constant-factor and per-center coding costs, not a universal minimax floor.

additive term is only the assumed codebook's upper resolution offset; it is not a lower error floor. If net locations or routing metadata are learned rather than public, their bits must be added to M.

The power-law term dominates that upper certificate exactly in the explicit range

$$\eta = 0 \quad \text{or} \quad 1 \le k \le c_2 \left(\frac{LR}{\eta}\right)^{D_{\mathcal{C}}}.$$

In this range the upper certificate is at most twice its covering term. Equivalently, for a target total error $t > \eta$, frozen-and-add needs at most

$$M_{\text{freeze+add}}(t) \le q \left\lceil c_2 \left(\frac{LR}{t-\eta}\right)^{D_{\mathcal{C}}} \right\rceil,$$

whereas the deterministic lower bound requires, up to the harmless one-state offset, $M_{\text{opt}}(t) \ge c_1(LR/(4t))^{D_{\mathcal{C}}} - 1$. Consequently the comparison carries the explicit factor $[t/(t-\eta)]^{D_{\mathcal{C}}}$; when $t \ge 2\eta$ it is bounded by $2^{D_{\mathcal{C}}}$. No bound here implies the scheme-specific adapter ratio below, and no Θ claim follows from a one-sided memory lower bound.

Proposition 5 (Shared-basis drift code)**.** *Suppose $\mu_c = (\mathrm{id} + \Delta_c)_{\#}\mu_0$ for a common reference μ_0. Fix an integer $r \ge 1$, a shared r-dimensional subspace $V \subset L^2(\mu_0)$, and an orthonormal basis of V. Assume that for every c there is a specified $v_c \in V$ with $\|\Delta_c - v_c\|_{L^2(\mu_0)} \le \delta$ whose coefficient vector lies in $[-\Delta_{\max}, \Delta_{\max}]^r$. Assume each basis function has a finite representation using exactly h scalars at b bits per scalar. For target coefficient error $\varepsilon_q > 0$, let*

$$b_q := \begin{cases} 0, & \Delta_{\max} = 0, \\ \max\left\{0, \left\lceil \log_2 \dfrac{\Delta_{\max}\sqrt{r}}{\varepsilon_q} \right\rceil\right\}, & \Delta_{\max} > 0, \end{cases}$$

Coordinatewise uniform midpoint quantization then has ℓ_2 coefficient error at most ε_q, so the induced transport error is at most $\delta + \varepsilon_q$. The one-off basis cost is rhb bits and the per-regime code costs rb_q bits.

Proof. Quantize each specified v_c coefficient using 2^{b_q} equal-width midpoint cells (using the zero vector when $b_q = 0$). Each coordinate error is at most $\Delta_{\max} 2^{-b_q} \le \varepsilon_q/\sqrt{r}$, so $\|v_c - \widehat{v}_c\|_{L^2(\mu_0)} \le \varepsilon_q$. Coupling the pushforwards by the same $X \sim \mu_0$ gives

$$W_2\big((\mathrm{id} + \widehat{v}_c)_{\#}\mu_0, \mu_c\big) \le \|\widehat{v}_c - \Delta_c\|_{L^2(\mu_0)} \le \varepsilon_q + \delta.$$

There are r basis functions of h stored scalars and r quantized coefficients per regime, giving the stated bit counts. □

Under these additional coding assumptions, $M_{\text{shared}} = rhb + krb_q$ and $M_{\text{freeze+add}} = kq$, so their large-$k$ ratio tends to $q/(rb_q)$ when $b_q > 0$. This is a comparison between two specified schemes, not a minimax gap. The rank r here is the dimension of a common family $\{\Delta_c\}$ in one reference coordinate system. It need not equal the numerical rank of pairwise optimal-transport displacements $\{T_{i\to j} - \mathrm{id}\}$ unless an additional compatibility or pullback relation identifies those two families.

Proposition 5 is not the accounting used by the implemented LoRA-style adapters. If $|\mathcal{I}| = k$ represents one fixed-source displacement per stored regime, one specified layer is adapted, and the positively realized matrix-update rank is $k_{\text{drift}} \geq 1$, then fresh factor pairs per regime give

$$\frac{M_{\text{freeze+add}}}{M_{\text{adapt}}} \xrightarrow[k\to\infty]{} \frac{|\theta|}{k_{\text{drift}}\,(d_{\text{in}} + d_{\text{out}})},$$

The factor-pair costs must instead be summed over pairs and layers in a more general implementation. Proposition 5's smaller scalar code applies only when its shared finite basis is stored and used explicitly.

G.1 Weighted functional drift rank and its certificate

We now define the functional quantity to which a rigorous rank certificate applies. Fix a finite set $\mathcal{I}$ of ordered source–target pairs and weights $w_\alpha > 0$, $\sum_{\alpha\in\mathcal{I}} w_\alpha = 1$. Assume every pair $\alpha = (i, j)$ admits a selected deterministic optimal-transport map T_α from μ_{c_i} to μ_{c_j}. For a reference measure μ, choose measurable versions for which $g_\alpha := T_\alpha - \mathrm{id}$ belongs to $\mathcal{H} := L^2(\mu; \mathbb{R}^d)$, and define the finite-family synthesis operator

$$\mathsf{D} : \ell_2(\mathcal{I}) \to \mathcal{H}, \qquad \mathsf{D}e_\alpha := \sqrt{w_\alpha}\, g_\alpha.$$

Definition 1 (Weighted functional drift rank)**.** Let $\sigma_1 \geq \sigma_2 \geq \cdots$ be the singular values of D, let P_r project onto its first r left singular functions, and set

$$\delta_r^2 := \sum_{\alpha\in\mathcal{I}} w_\alpha \|(I - P_r)g_\alpha\|_{\mathcal{H}}^2 = \sum_{s>r} \sigma_s^2.$$

For $\tau \in (0, 1)$, define $k_{\text{drift}}(\tau) := \min\{r \in \mathbb{Z}_{\geq 0} : \delta_r^2 \leq \tau \sum_s \sigma_s^2\}$. In particular, $k_{\text{drift}}(\tau) = 0$ when the total singular energy is zero. This is a weighted finite-family projection rank. Its value depends on the reference measure μ, the selected optimal maps and their chosen μ-almost-everywhere versions, the pair set, the weights, and the tolerance; all of these are fixed parts of the definition.

Proposition 6 (Functional rank certificate and conditional adapter cost)**.** *Assume the optimal maps above exist and every source law obeys $\mu_{c_i} \ll \mu$ with $\|\mathrm{d}\mu_{c_i}/\mathrm{d}\mu\|_\infty \leq \Lambda$. For $\alpha = (i, j)$ define $\widetilde{\mu}_\alpha^{(r)} := (\mathrm{id} + P_r g_\alpha)_{\#}\mu_{c_i}$. Then*

$$\left(\sum_{\alpha\in\mathcal{I}} w_\alpha W_2^2(\widetilde{\mu}_\alpha^{(r)}, \mu_{c_j})\right)^{1/2} \leq \sqrt{\Lambda}\,\delta_r, \tag{6}$$

and, for each pair, $W_2(\widetilde{\mu}_\alpha^{(r)}, \mu_{c_j}) \leq \sqrt{\Lambda/w_\alpha}\,\delta_r$. Only under the additional architectural assumption that $P_r g_\alpha$ is realized by a rank-at-most-r matrix update $U_\alpha V_\alpha^\top$ in a specified $d_{\text{out}} \times d_{\text{in}}$ layer does storing fresh factors cost $r(d_{\text{in}} + d_{\text{out}})b$ bits per pair and layer.

Equation 6 is a weighted root-mean-square functional W_2 certificate. A common-base pushforward representation alone does not imply $\Lambda = 1$; that constant disappears only when the reference/source relation explicitly gives it (or after an isometric pullback with the corresponding proof). Likewise, $k_{\text{drift}}(\tau)$ is the smallest rank for the stated weighted $L^2(\mu)$ tail criterion, not the smallest LoRA rank or the exact smallest rank under pointwise W_2 error.

Proof of Proposition 6. Let $\mathcal{I}$ be the finite pair index set from Definition 1, with $w_\alpha > 0$ and $\sum_\alpha w_\alpha = 1$. Existence of a deterministic optimal map T_α is an explicit hypothesis; it is not automatic for arbitrary source laws. Choose representatives $g_\alpha := T_\alpha - \mathrm{id}$ in $\mathcal{H} = L^2(\mu; \mathbb{R}^d)$ and define the finite-rank Hilbert–Schmidt synthesis operator

$$\mathsf{D} : \ell_2(\mathcal{I}) \to \mathcal{H}, \qquad \mathsf{D}e_\alpha = \sqrt{w_\alpha}\, g_\alpha.$$

If P_r projects onto its leading r left singular functions, Hilbert-space Eckart–Young gives

$$\delta_r^2 = \sum_{s>r} \sigma_s^2 = \sum_{\alpha\in\mathcal{I}} w_\alpha \|(I - P_r)g_\alpha\|_{L^2(\mu)}^2.$$

For $\alpha = (i, j)$, the joint law of $\big(T_\alpha(X), X + P_r g_\alpha(X)\big)$ with $X \sim \mu_{c_i}$ couples μ_{c_j} and $\widetilde{\mu}_\alpha^{(r)}$. Therefore

$$W_2^2(\widetilde{\mu}_\alpha^{(r)}, \mu_{c_j}) \le \int \|(I - P_r)g_\alpha\|^2 \, d\mu_{c_i} \le \Lambda \|(I - P_r)g_\alpha\|_{L^2(\mu)}^2.$$

Multiplying by w_α, summing, and taking square roots proves Equation 6. Keeping one summand and using $w_\alpha \|(I - P_r)g_\alpha\|^2 \le \delta_r^2$ gives the individual factor $\sqrt{\Lambda / w_\alpha}$.

A shared base-pushforward description does not imply $\Lambda = 1$. For probability measures, the condition $d\mu_{c_i}/d\mu \le 1$ actually forces $\mu_{c_i} = \mu$; alternatively one may pull every displacement back to a common base by specified maps and prove the corresponding norm identity there. Finally, $P_r g_\alpha$ has *functional* rank r. The cost $r(d_{\text{in}} + d_{\text{out}})b$ follows only after assuming a specified layer realizes it as a matrix update $U_\alpha V_\alpha^\top$; no linear-algebra implication from functional rank to LoRA rank is being asserted. □

H Proof of Theorem 5 (saturation of the bound)

Write $e := C\bar{\epsilon}$. First suppose $\Delta > 0$ and $0 < e \le \Delta/2$, and put

$$K := \lceil (\Delta/e)^D \rceil .$$

Then $K \ge \lceil 2^D \rceil$. If $K < n$, Theorem 1's power-law case gives $r_K \le \Delta K^{-1/D} \le e$, hence $k^\star \le K$. If $K \ge n$, the finite-set endpoint gives $r_n = 0 \le e$, hence $k^\star \le n$. These two cases prove

$$k^\star \le \min\{n, K\}.$$

For every $k \ge k^\star$, monotonicity gives $r_k + e \le 2e$; adding further centers can reduce this *certificate* by at most e. It does not make the certificate constant, prove that true error stops improving, or make e a lower error floor.

The degenerate cases are separate. If $\Delta = 0$, $\mathcal{M}$ is a singleton and $r_1 = 0$. If $\Delta > 0$ and $e = 0$, the finiteness and distinctness of the n metric points imply $r_k > 0$ for $k < n$ and $r_n = 0$, so $k^\star = n$. If $\Delta/2 < e < \Delta$, Assumption 5 gives no power-law inversion at that scale; the safe conclusion is only $k^\star \le n$. If $e \ge \Delta$, one center already has radius at most e.

The empirical quantity $k^\star_{\text{est}} = \lceil (\Delta/e)^{\hat{D}} \rceil$ in Table 7 substitutes an estimated exponent into the nondegenerate formula. It is a continuous-relaxation diagnostic, not the exact $k^\star$; the exact finite-set threshold never exceeds n.

I Proof of Theorem 6 (streaming eviction)

The procedure, and the online quantities it is measured against. Fix $k \ge 1$ and ignore duplicate metric points. Buffer the first $k+1$ distinct regimes (or retain all regimes exactly if fewer ever arrive). Set the working radius R_{alg} to half the closest-pair distance in that buffer, and replace the buffer by any maximal subset whose pairwise distances are strictly greater than $2R_{\text{alg}}$. Thereafter, discard a new point if it lies within $2R_{\text{alg}}$ of a current center; otherwise add it. Whenever this makes $|S| > k$, repeatedly set $R_{\text{alg}} \leftarrow 2R_{\text{alg}}$ and replace S by a maximal $2R_{\text{alg}}$-separated subset until $|S| \le k$.

For the distinct regimes seen through time t, let

$$r_t^\star(k) := \inf_{\substack{S \subseteq \mathcal{M}_t \\ 1 \le |S| \le k}} \max_{\mu \in \mathcal{M}_t} \min_{s \in S} W_2(\mu, s).$$

For the current center set S_t, let $\hat{p}_t(\mu)$ denote its nearest-center clamped forecast for $\mu \in \mathcal{M}_t$, and define

$$\text{err}_{\text{online},t} := \max_{\mu \in \mathcal{M}_t} W_2\big(\hat{p}_t(\mu), \mu\big).$$

The theorem assumes an *exact* W_2 oracle. Ignore duplicate metric points. If the stream contains at most k distinct regimes, retain all and the claim is immediate. Otherwise buffer the first $k+1$ distinct points, let $d_{\min} > 0$ be their closest-pair distance, initialize $R_{\text{alg}} := d_{\min}/2$, and replace the buffer by a maximal strict $2R_{\text{alg}}$-separated subset. It has at most k elements because a closest pair cannot both survive. Thereafter, a point within $2R_{\text{alg}}$ of S is covered without insertion; a farther point is inserted. If $|S| > k$, repeatedly double R_{alg} and replace S by a maximal strict $2R_{\text{alg}}$-separated subset until $|S| \le k$.

We prove two invariants. First, every point seen so far lies at distance at most $4R_{\mathrm{alg}}$ from a current center. It holds after initialization because maximality gives distance at most $2R_{\mathrm{alg}}$. A point rejected between doublings is also within $2R_{\mathrm{alg}}$. During a doubling from R to $R' = 2R$, every removed old center is within $2R'$ of a surviving one. A historical point previously within $4R$ of that old center is therefore within

$$4R + 2R' = 4R' = 4R_{\mathrm{alg}}$$

so repeated merges accumulate to at most $4R_{\mathrm{alg}}$, not $2R_{\mathrm{alg}}$.

Second, after initialization and after every completed doubling phase,

$$R_{\mathrm{alg}} < 2r_t^\star(k).$$

At initialization, the $k+1$ buffered points are pairwise at distance at least $d_{\min} = 2R_{\mathrm{alg}}$. No k balls of radius strictly below $d_{\min}/2$ can cover them, so $r_t^\star(k) \geq R_{\mathrm{alg}}$ and the displayed strict inequality follows. For a later phase, immediately before its final doubling the center set contains more than k historical points pairwise farther than $2R_{\mathrm{old}}$. If $r_t^\star(k) \leq R_{\mathrm{old}}$, assigning these witnesses to k optimal radius-R_{old} balls would put two in one ball and make their distance at most $2R_{\mathrm{old}}$, a contradiction. Thus $r_t^\star(k) > R_{\mathrm{old}}$, and the final scale satisfies $R_{\mathrm{alg}} = 2R_{\mathrm{old}} < 2r_t^\star(k)$. This same packing witness remains among the demands at every later time, so the inequality persists between phases even though the set of permissible internal centers expands and $r_t^\star(k)$ need not be monotone.

Combining the invariants gives coverage radius at most $4R_{\mathrm{alg}} < 8r_t^\star(k)$, hence in particular at most $8r_t^\star(k)$. Nearest-center clamping and the uniform residual yield

$$\mathrm{err}_{\mathrm{online},t} \leq 8r_t^\star(k) + C\bar{\epsilon}.$$

The $O(k)$ statement counts center records only (and a transient $k+1$-point initialization buffer), not bits for retained experts, router metadata, condition descriptors, or samples/sketches needed to implement the exact metric oracle. Once a historical center is removed, its error is bounded by the coverage invariant but generally is not zero.

J A SELECTED-GEODESIC DIAGNOSTIC AND ITS TESTED HYPOTHESIS

We record this construction in full because Section 4.6 tests proximity to a selected geodesic. The decomposition below is always valid for that selection; small off-geodesic error is a separate empirical hypothesis.

Fix an optimal coupling $\pi_j \in \Pi(\mu_{c_{j-1}}, \mu_{c_j})$, whose existence holds in $\mathcal{P}_2(\mathbb{R}^d)$, and fix the associated constant-speed geodesic

$$\mu^{(t)} := ((x,y) \mapsto (1-t)x + ty)_{\#}\pi_j, \qquad t \in [0,1].$$

This construction does not assume that an optimal Monge map exists; when geodesics are nonunique, all quantities below are relative to the one fixed geodesic. For $q_j := \hat{p}(\cdot \mid c_{j-1}; \theta_j)$, define

$$t_j^\star \in \arg\min_{t\in[0,1]} W_2(q_j, \mu^{(t)}), \quad \hat{\rho}_j := 1 - t_j^\star, \quad \eta_j := \min_{t\in[0,1]} W_2(q_j, \mu^{(t)}),$$

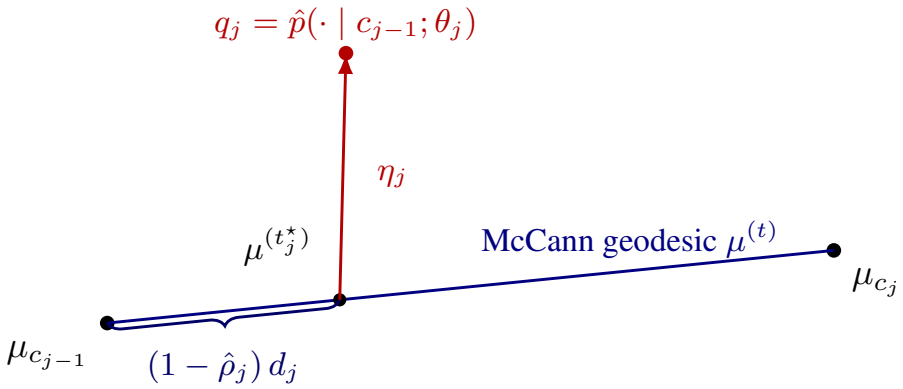
where existence of a minimizer follows from continuity on the compact interval. Define the Wasserstein old-regime error

$$\mathrm{Forget}_j^{W_2} := W_2(q_j, \mu_{c_{j-1}}).$$

Proposition 7 (Forgetting decomposition)**.** $\mathrm{Forget}_j^{W_2} \leq \eta_j + (1-\hat{\rho}_j)d_j$.

Proof. Triangle inequality through $\mu^{(t_j^\star)}$, using the geodesic property $W_2(\mu^{(t)}, \mu^{(0)}) = t \cdot d_j$. □

This proposition does *not* bound the empirical forgetting reported as "MSE after minus MSE before." Such a conclusion would require an additional inequality relating changes in forecast risk to W_2 and appropriate loss/moment assumptions. Relative to the fixed geodesic selected for each of the two tested trajectories, η_j exceeds $(1-\hat{\rho}_j)d_j$ by a factor of 2.2–4.6. Thus those measurements reject the small-η_j hypothesis for the selected paths; they do not rule out every optimal geodesic when the geodesic is nonunique.



prediction: finetuned model slides along the geodesic (η_j small)
measured (E1): $\eta_j \gg (1-\hat{\rho}_j)d_j$ – model falls off the geodesic instead

Figure 4: The selected McCann geodesic. $\mu^{(t)}$ interpolates optimally between the old and new regime; $\hat{\rho}_j = 1 - t_j^\star$ measures position along this path, and η_j is distance from it. The tested hypothesis, not the decomposition itself, posits that η_j is small.

Table 2: Measured decomposition terms for the selected geodesic, in empirical W_2 distance units. $\eta_j \gg (1 - \hat{\rho}_j)d_j$ on both tested trajectories.

Dataset	η_j	$(1 - \hat{\rho}_j)d_j$	ratio
Exchange	0.869	0.189	4.6×
ETTh2	0.873	0.405	2.16×

K PATH-INDEPENDENCE

Corollary 3 (Path-independence of the bound under clamping)**.** *For a single model fine-tuned sequentially and clamped at its most recent regime c_T, evaluated against the original regime c_1,*

$$W_2\big(\hat{p}(\cdot \mid c_T; \theta_T), \mu_{c_1}\big) \le C\epsilon_T + W_2(\mu_{c_T}, \mu_{c_1}),$$

where the geometric term depends on net displacement rather than $\sum_j d_j$. For a family of paths with the same endpoints, a common path-independent envelope exists only if one additionally has a path-independent residual bound $\epsilon_T \le \bar{\epsilon}$, in which case the right-hand side is at most $C\bar{\epsilon} + W_2(\mu_{c_T}, \mu_{c_1})$.

Proof. Direct application of Assumption 1 at θ_T plus the triangle inequality through μ_{c_T}. □

The per-path quantity ϵ_T can depend on the entire optimization trajectory, so equal endpoints alone do not make the original per-path certificates identical. Even under a uniform path-independent envelope, the result does not imply that realized error or empirical MSE forgetting is path-independent; it says only that this particular Wasserstein certificate contains no explicit cumulative-drift term. A controlled test must fix both endpoints, vary the intervening path, and separately measure the final residual.

L MEMORY ACCOUNTING ACROSS THREE RESOURCE TYPES

Table 3 separates parameter bits, stored-data bits, and the Wasserstein quantity for which this appendix provides a certificate. It does not relabel empirical MSE forgetting as a Wasserstein distance.

For scale only, consider the stronger, scheme-specific model in which a replay buffer represents each regime solely by its empirical law $\hat{\mu}_m$, the target is $\mathbb{E}W_2(\hat{\mu}_m, \mu) \le \varepsilon$, samples are raw d-vectors, and

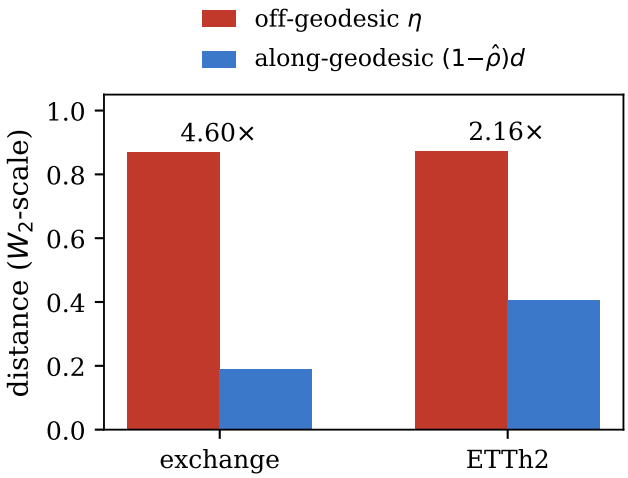


Figure 5: Off-path deviation η_j versus along-path distance $(1-\hat{\rho}_j)d_j$ for the selected geodesic on two real fine-tuning trajectories. The measurements reject the small-off-path hypothesis for these selections; Proposition 7 itself remains a valid inequality.

Table 3: Resource accounting. Both memory columns are in bits; the last column is an old-regime W_2 error certificate in target-distance units, not empirical MSE forgetting. Here m is samples stored per regime, each represented by d scalars at b bits, and a fresh rank-r factor pair has $r(d_{\text{in}}+d_{\text{out}})$ scalars. Router, center-descriptor, and metric-oracle sketch memory is additional.

Strategy	Parameter memory (bits)	Data memory (bits)	Old-regime W_2 certificate
Fine-tuning	$\lvert\theta\rvert b$	0	$\le \eta_j + (1-\hat{\rho}_j)d_j$
Replay	$\lvert\theta\rvert b$	$m\,d\,b$ per regime	no bound stated
Frozen-and-add	$k\,\lvert\theta\rvert b$	0	$\le C\bar{\epsilon}$ while dedicated
Frozen-and-adapt (D.P. 4)	$\lvert\theta\rvert b + k\,r(d_{\text{in}}+d_{\text{out}})\,b$	0	$\le C\bar{\epsilon}$ while dedicated

the regime laws have uniformly bounded support (sufficiently strong uniform moments also suffice). A Fournier–Guillin upper bound (Fournier & Guillin, 2015), followed by Jensen's inequality, yields the illustrative sufficient scalings

$$m \gtrsim \begin{cases} \varepsilon^{-d}, & d > 4, \\ \varepsilon^{-4}\{\log(1/\varepsilon)\}^2, & d = 4, \\ \varepsilon^{-4}, & d < 4, \end{cases}$$

up to distribution-dependent constants. This is neither a universal asymptotic equivalence nor a necessary lower bound for replay-trained predictors: optimization, compression, parametric structure, and task loss can change the comparison. Under the normalized $d > 4$ illustration only, raw replay storage is of order $db\varepsilon^{-d}$ bits per regime, so $|\theta| < d\varepsilon^{-d}$ merely marks where a $|\theta|b$-bit expert is smaller than that particular *sufficient* buffer budget.

Table 4: Illustrative sufficient raw-replay storage and equal-size expert count at $\varepsilon = 0.1$, $d > 4$, with hidden constants normalized to one. These are not necessary replay lower bounds.

Ambient dim. d	Sufficient replay bits/regime	Equal-size $\lvert\theta\rvert$
6	1.9×10^8	6.0×10^6
8	2.6×10^{10}	8.0×10^8
16	5.1×10^{18}	1.6×10^{17}

M Experimental Details

M.1 Datasets and Preprocessing

We use five public multivariate time-series forecasting benchmarks: ETTh2 (7 channels), national illness surveillance / ILI (7 channels), Exchange Rate (8 channels), Electricity (321 channels), and Traffic (862 channels). Each series is split chronologically 70%/30% into regime A (past, used for

initial training) and regime B (future, used for the continual-adaptation phase); this is a natural, unlabeled regime boundary rather than an injected one. Inputs are instance-normalized (RevIN-style) per window.

M.2 Model and training

The backbone is a patch-embedding, mixture-of-experts forecaster (patch length 24, stride 6, model width 128). Continual-adaptation training uses 30 epochs (10 for Electricity/Traffic, batch size reduced to 4/2 respectively to fit memory – Section 5), learning rate 8×10^{-4}, seed 42. EWC uses $\lambda = 1000$; SI uses $\lambda = 0.1$; replay mixes 50% regime-A batches into regime-B training. Our rank-adapter variant uses rank $r = 16$ on the shared projection layer; a full-encoder LoRA variant (rank 16 across all 120 linear layers of the patch-and-expert backbone) is reported alongside it in Table 5. The drift-confirmation gate's bootstrap-vigilance percentile is set to 90 (reduced from a default of 99, which we found overly conservative – it rejected genuine regime-B confirmations).

M.3 Full numeric results

Table 5 gives MSE_B for all eight methods (four baselines, a dense-head frozen expert, an MLP-head frozen expert, our projection-rank-adapter, and the full-encoder-adapter variant) on all five datasets; Table 6 gives the corresponding forgetting values. The dense- and MLP-head frozen-expert variants are provided for the ablation referenced in Section 4.4: an expert with its own small head (rather than a low-rank update inside the shared backbone) under-performs both fine-tuning and the rank-adapter variant, because a separate small head has strictly less capacity than a low-rank modification of the full shared backbone – capacity, not linearity versus nonlinearity, is the operative factor (a nonlinear bottleneck head does not close the gap either).

Table 5: $\mathrm{MSE}_B \downarrow$, all eight methods, all five datasets. '–' marks the quadratic-cost dense/nonlinear head disabled for high-channel datasets (Section 5).

Dataset	#ch.	finetune	EWC	replay	SI	dense-head	MLP-head	**rank-adapter**	full-adapter
ETTh2	7	0.556	0.524	**0.500**	0.548	0.552	0.559	0.509	0.519
ILI	7	3.697	3.864	**3.470**	3.697	4.027	4.386	4.209	3.966
Exchange	8	0.120	0.116	0.089	0.115	0.274	0.400	**0.084**	0.085
Electricity	321	0.151	0.153	**0.147**	0.153	–	–	0.153	0.150
Traffic	862	0.639	**0.628**	0.635	0.669	–	–	0.688	0.669

Table 6: Forgetting$\downarrow$ (MSE_A after $-$ before), all methods, all datasets.

Dataset	finetune	EWC	replay	SI	**rank-adapter** / full-adapter
ETTh2	-0.145	-0.203	-0.143	-0.141	**0.000**
ILI	$+0.215$	$+0.263$	$+0.156$	$+0.216$	**0.000**
Exchange	$+0.069$	$+0.061$	$+0.012$	$+0.061$	**0.000**
Electricity	$+0.009$	$+0.016$	-0.010	$+0.002$	**0.000**
Traffic	-0.016	-0.016	-0.033	$+0.001$	**0.000**

M.4 Estimated covering exponent, drift rank, and saturation (30 regimes)

Table 7 gives the per-dataset quantities underlying Figure 6 (real-data panel), computed from $T = 30$ z-scored, channel-averaged regimes per dataset. Each regime is summarized by the per-window channel-mean feature vector of its lookback windows and fit as a Gaussian; W_2 is then the closed-form Bures/Gelbrich distance between Gaussian summaries, not transport on the full raw empirical laws. The reported k_{drift} is the number of singular values of the pairwise regime-mean-shift matrix needed to explain 95% of its energy. Thus it is an experimental proxy, not Definition 1's weighted family of assumed optimal-map displacements unless an additional identification is justified. The fitted $\hat{D}$ is the regression slope of $\log N(r)$ on $\log(1/r)$ over the finite observed scale window; it is neither the scale-uniform exponent D in Assumption 5 nor the full-space $D_{\mathcal{C}}$ in Assumption 4. We did not estimate a certified Wasserstein residual or the constant C. Therefore the explicitly reported e_{plug} values are illustrative scenario offsets used to make the proxy calculation reproducible, not measured values of $C\bar{\epsilon}$. These measurements do not test Corollary 1's minimax premise.

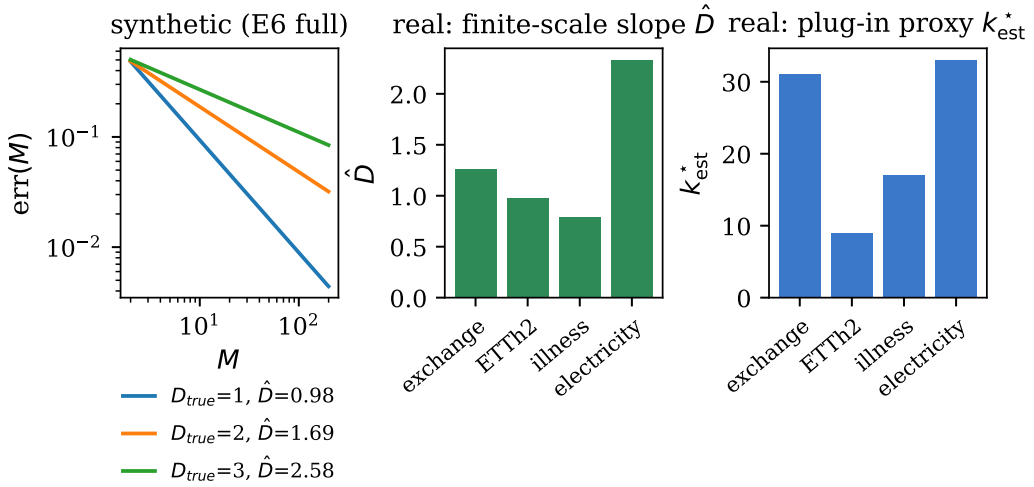


Figure 6: Left: on synthetic sequences of known intrinsic dimension D_{true}, fitting $\log \text{err}(M)$ vs. $\log M$ recovers $\hat{D} \approx D_{\text{true}}$, $R^2 > 0.99$. Middle/right: on four real datasets ($T \geq 30$ regimes), the finite-scale regression gives a descriptive slope $\hat{D}$ and a scenario proxy $k^{\star}_{\text{est}}$ computed from explicitly stated illustrative offsets (Appendix M); neither is a theorem certificate.

Table 7: Finite-scale slope $\hat{D}$, proxy drift rank k_{drift}, empirical diameter Δ, stated illustrative offset e_{plug}, and scenario proxy $k^{\star}_{\text{est}} = \lceil (\Delta/e_{\text{plug}})^{\hat{D}} \rceil$ ($T = 30$). This is not the exact $k^{\star} \leq n$, and e_{plug} is not a measured residual or lower floor.

Dataset	#ch.	$\hat{D}$	k_{drift}	Δ	e_{plug}	$k^{\star}_{\text{est}}$
Exchange	8	1.26	4	8.30	0.544	31
ETTh2	7	0.98	4	6.90	0.734	9
Illness	7	0.79	1	5.35	0.149	17
Electricity	321	2.33	11	16.99	3.790	33

M.5 Seed variability

Tables 5–6 are single-seed (42). To say which gaps are real we reran the full eight-method benchmark on three *additional* seeds (1–3) for the three low-channel datasets; Table 8 reports mean $\pm$ standard deviation over those three runs. Forgetting was exactly 0.0000 for every frozen-and-adapt variant in all nine runs, so the zero-forgetting claim is not seed-dependent.

The variability is small relative to the effects the paper claims, and it supports the qualitative reading of Section 4.2 rather than overturning it. On ETTh2 the projection rank-adapter and replay are separated by 0.0004 against standard deviations of 0.003–0.008: a tie, exactly as claimed, and one that no single-seed table could have established. On Exchange the rank-adapter's advantage over replay (0.0826 vs. 0.0864) is several standard deviations wide and therefore real. On ILI replay's advantage is likewise unambiguous. The one caveat is that the ILI rank-adapter has the largest spread in the table (± 0.22), so its *margin* of loss there is less well determined than its direction.

M.6 Sensitivity of the geometry estimates

Two estimation shortcuts underlie Table 7: the Gaussian/Bures surrogate for W_2, and the $\tau = 0.05$ energy cutoff defining k_{drift}. We quantify both rather than assert they are harmless.

Gaussian surrogate versus exact optimal transport. We recomputed every pairwise regime distance with exact discrete OT (network-simplex, uniform empirical measures, ≤ 300 windows per regime) and refit all downstream quantities. Table 9 reports the comparison: pairwise distances

Table 8: $\mathrm{MSE}_B \downarrow$, mean $\pm$ s.d. over three seeds (1–3), in fixed method order; best per dataset in bold. Frozen-and-adapt variants had exactly zero forgetting in every run.

Method	ETTh2	ILI	Exchange
finetune	0.555 ± 0.023	3.788 ± 0.091	0.117 ± 0.008
EWC	0.543 ± 0.018	3.873 ± 0.060	0.117 ± 0.005
replay	0.513 ± 0.008	$\mathbf{3.448 \pm 0.026}$	0.086 ± 0.003
SI	0.538 ± 0.021	3.786 ± 0.095	0.111 ± 0.003
dense-head (ours)	0.556 ± 0.010	4.058 ± 0.030	0.274 ± 0.002
MLP-head	0.563 ± 0.017	4.591 ± 0.135	0.384 ± 0.020
rank-adapter	$\mathbf{0.512 \pm 0.003}$	4.211 ± 0.221	$\mathbf{0.083 \pm 0.001}$
full-adapter	0.520 ± 0.008	4.069 ± 0.077	0.084 ± 0.001

agree almost exactly (Pearson and Spearman correlation ≥ 0.999, median relative error $\leq 1.3\%$), and the fitted exponent moves by under 1.5% on three of four datasets. ETTh2 is the outlier at 8.4%, which is informative: the covering-curve regression amplifies small distance perturbations more than the distances themselves move, so $\hat{D}$ should be read as accurate to roughly one decimal, not three. Exact OT yields a *systematically slightly larger* $\hat{D}$ (and hence $k^\star_{\mathrm{est}}$) on all four datasets, so the Gaussian surrogate is mildly optimistic, never conservative – but not by enough to change any conclusion drawn from Table 7.

Table 9: Gaussian/Bures surrogate versus exact OT, $T = 30$ regimes. r_{P} is the Pearson correlation between the two pairwise-distance matrices and "rel. err" their median relative difference.

Dataset	$\hat{D}$ (Gauss)	$\hat{D}$ (exact)	diff.	r_{P}	rel. err	$k^\star_{\mathrm{est}}$ (G)	$k^\star_{\mathrm{est}}$ (E)
Exchange	1.264	1.283	1.4%	1.0000	0.2%	31	33
ETTh2	0.984	1.074	8.4%	0.9996	0.9%	9	11
Illness	0.791	0.797	0.8%	1.0000	0.0%	17	18
Electricity	2.334	2.357	1.0%	0.9994	1.3%	33	35

Drift-rank tolerance. Definition 1 makes the cutoff an explicit error tolerance τ, so we sweep it (Table 10). On the low-channel datasets k_{drift} is stable across a twentyfold change in τ (2–6 for Exchange, 3–6 for ETTh2, 1–2 for Illness), so the reported $k_{\mathrm{drift}} = 4$ is not an artifact of the cutoff. Electricity is the exception, moving $4 \to 21$: with 321 channels the displacement spectrum has a long tail and no clear numerical rank, which is precisely the regime where Design Principle 4's rank-sizing advice is least actionable. We report this rather than quote a single number for it.

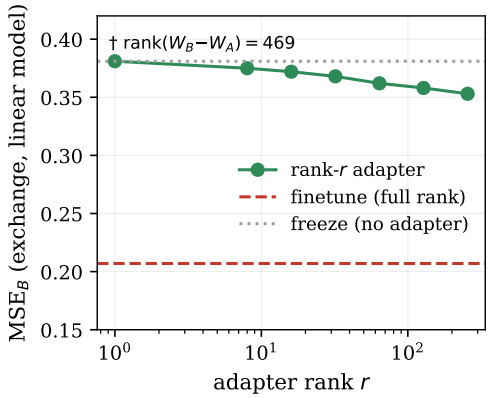
Table 10: $k_{\mathrm{drift}}(\tau)$ versus tolerance τ (tail displacement energy). The paper reports $\tau = 0.05$.

Dataset	τ=0.20	τ=0.10	τ=0.05	τ=0.01
Exchange	2	3	4	6
ETTh2	3	3	4	6
Illness	1	1	1	2
Electricity	4	7	11	21

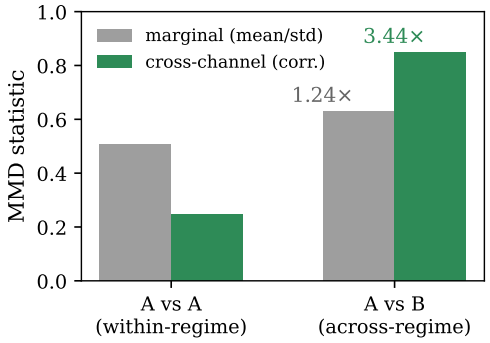
M.7 Rank sweep on the MoE backbone, and an over-training confound

Design Principle 4's positive side – that $r \approx k_{\mathrm{drift}}$ suffices on a backbone with a latent bottleneck – was previously untested: the only sweep in the paper is on a bare linear model, i.e. the failure case. Sweeping $r \in \{1, \ldots, 64\}$ on the MoE backbone at the main benchmark's 30-epoch adapter schedule produced an essentially flat curve from r=1 to r=64 across a $64\times$ range of adapter parameters; the r=4 versus r=16 spread at that schedule is 0.0056 on ETTh2 and 0.0022 on Exchange (Table 11), several times smaller than the swing the adapter-epoch axis induces and, more awkwardly, a *freeze-only* baseline – the regime-A model serving B with no adaptation at all – that beat every adapted variant. Since freeze-only also has exactly zero forgetting, it would dominate the proposed method outright.

That conclusion turns out to be an artifact of the adapter schedule, not a property of the method. Table 11 decouples adapter training length from backbone training length. MSE_B degrades with adapter epochs on both datasets – strictly monotonically on ETTh2, and monotonically on Exchange up to a reversal of at most 0.002 in the last two columns – and at a single epoch the adapter beats freeze-only on both; by three epochs it is already worse. The 30-epoch schedule used for the main ta-



(a) Adapter rank sweep on a *linear* forecaster (no latent bottleneck), Exchange. Even rank 256 does not reach fine-tuning's error, since parameter-space drift rank ($\mathrm{rank}(W_B - W_A) = 469$) exceeds every rank tested.



(b) MMD between within-regime (A vs. A) and across-regime (A vs. B), marginal vs. cross-channel feature, on a synthetic lead–lag drift where only cross-channel structure changes. Marginal barely separates the two cases; cross-channel does.

Figure 7: Two controlled ablations referenced in Sections 4.4 and 4.5.

bles therefore over-trains the adapter on regime B, which simultaneously inflates MSE_B and flattens the rank dependence – an over-trained rank-1 adapter and an over-trained rank-64 adapter converge to similarly poor solutions, which is why the 30-epoch sweep looked rank-independent.

Table 11: $\mathrm{MSE}_B \downarrow$ versus adapter rank and adapter training epochs (backbone trained 30 epochs on A throughout; seed 42). "freeze" attaches no adapter. Best per column in bold.

	adapter epochs					
Setting	1	3	5	10	30	freeze / finetune
ETTh2, r=4	**0.4711**	0.5155	0.5170	0.5222	0.5228	0.4871 / 0.5477
ETTh2, r=16	0.4760	0.5178	0.5186	0.5213	0.5284	
Exchange, r=4	**0.0770**	0.0814	0.0848	0.0852	0.0846	0.0792 / 0.1111
Exchange, r=16	0.0786	0.0818	0.0861	0.0844	0.0868	

At the one-epoch budget the rank ordering also matches the principle for the first time: r=4 beats r=16 on both datasets, and $k_{\mathrm{drift}}(\tau{=}0.05) = 4$ for both (Table 7). We state the strength of this evidence carefully. It is a single seed, two datasets, and two ranks, and the freeze baseline itself moved by 0.0065 (1.3%) between two identically-seeded runs of the same configuration on different Kaggle sessions, so GPU non-determinism sets a noise floor of roughly 1% here. The r=4 advantage over freeze-only is about 3% on ETTh2 and 3% on Exchange – above that floor, but established with $n = 1$, not replicated. What we take as solid is the degradation with adapter epochs, which is large, consistent in direction across both datasets and both ranks, and sufficient on its own to explain the flat 30-epoch sweep.

The practical consequence is that the main tables (Tables 5–8) report an over-trained adapter and so *understate* the frozen-and-adapt variants, and that a validation-based stopping rule – which this codebase does not implement for the adapter – is a prerequisite for the rank rule to be testable at all. We report the numbers as run rather than re-tuning and re-reporting the benchmark, and flag the schedule as the first thing to fix (Section 5).

M.8 Rank sweep on a bare linear forecaster

Table 12 gives the full numeric sweep underlying Figure 7a (Exchange dataset, linear forecaster with no latent bottleneck). The measured parameter-space drift rank is $\mathrm{rank}(W_B - W_A) = 469$.

Table 12: Rank sweep, linear forecaster, Exchange. Fine-tune (full rank) reaches 0.207; freeze-only (rank 0) is 0.381.

Adapter rank r	1	8	16	32	64	128	256
MSE_B	0.381	0.375	0.372	0.368	0.362	0.358	0.353

M.9 CLAMPING VERSUS CONDITIONING

On the multi-expert setting of Section 3, clamping gives $\mathrm{MSE} = 0.237$ versus conditioning's 0.435 on Exchange, and 0.550 versus 0.547 on ETTh2 (a difference within measurement noise). These realized MSE values are descriptive; they neither confirm nor refute Corollary 2, which compares Wasserstein upper certificates rather than realized errors.

N CROSS-CHANNEL DRIFT DETECTOR: CONSTRUCTION

The marginal drift feature is the per-channel $(\mathrm{mean}, \mathrm{std})$ vector of the current window, compared to the same vector computed on each frozen center's training window via MMD with an RBF kernel. The cross-channel feature is the upper triangle of the current window's pairwise Pearson correlation matrix across channels, flattened to a vector and compared the same way. Both features feed the same bootstrap-vigilance confirmation gate (Appendix M); a regime change is confirmed when either statistic exceeds its calibrated percentile threshold, so the cross-channel feature adds a detection channel rather than replacing the marginal one.

N.1 DETECTION METRICS ON REAL STREAMS (A NEGATIVE RESULT)

Section 4.5's evidence is an MMD ratio on a synthetic lead–lag construction. That demonstrates the cross-channel statistic *can* separate a drift the marginal statistic is blind to, but it is not a detection evaluation, so we ran one on the real streams and report what it found, which is unfavourable.

We take the reference block immediately before each dataset's $70/30$ past/future cut and, for each lag L, score one negative (a block L windows *earlier*, inside A) and one positive (a block L windows *later*, inside B) against it. Matching the lag is essential: without it, negatives are drawn from the far past of A while positives sit just after the cut, so MMD tracks elapsed time rather than regime membership and AUROC falls below 0.5 for that reason alone. Under the matched-lag control (Table 13), *neither* statistic separates the two classes meaningfully – AUROC ranges from 0.24 to 0.64 around the 0.5 chance line, and recall at a 10% false-alarm operating point is at or below 0.2 everywhere.

Table 13: Matched-lag detection metrics at a 10% false-alarm operating point. n is the number of matched negative/positive block pairs. ILI's $n = 12$ is too small to support any conclusion and is shown only for completeness.

	AUROC		precision		recall	
Dataset (n)	marg.	cross	marg.	cross	marg.	cross
ETTh2 (320)	0.456	0.444	0.333	0.319	0.050	0.047
Exchange (136)	0.236	0.398	0.000	0.417	0.000	0.074
Electricity (487)	0.568	0.538	0.647	0.667	0.185	0.201
ILI (12)	0.639	0.583	0.600	0.000	0.250	0.000

We read this as a statement about the *benchmark*, not only about the detector. The $70/30$ split these datasets are evaluated on is a chronological convenience boundary, not an event: distribution distance grows smoothly with elapsed time, and once that trend is controlled for, there is little left at the cut for any statistic to find. Two consequences we accept. First, the synthetic lead–lag construction is the only place we demonstrate the cross-channel feature's advantage, and it is a capability demonstration under controlled ground truth, not evidence of real-stream detection performance. Second, the automatic-trigger story of Figure 1 is validated in principle but *not* on these real streams: the gate would fire on this data close to chance, so a deployment would be relying on the confirmation percentile rather than on discriminative power. Validating the trigger properly needs streams with annotated change points, which these forecasting benchmarks do not provide.