COMPLEXITY-ADAPTIVE STEIN TRANSPORT FOR PARALLEL TEMPERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sampling from complex distributions with parallel tempering can be limited by poor exchange between neighboring temperatures. Transport can improve these exchanges, but a fixed expressive map may be unnecessary when cross-temperature changes have simple structure. We introduce CAT-PT, which organizes transports into a hierarchy of structured velocity classes. CAT-PT fits a candidate in each class by minimizing a Stein residual that measures the annealing variation left unexplained by the corresponding motion. CAT-PT then tests increasingly expressive transports and stops as soon as one achieves a prescribed exchange-acceptance target. We show that the accumulated Stein residual directly controls finite-gap exchange rejection. In structured regimes, CAT-PT achieves round-trip performance governed by the residual variation after transport rather than by the ambient dimension. Under additional conditions, this residual scaling is order-optimal within the same transport class. This adaptivity can avoid the cost of a fixed high-capacity accelerator when a simple transport is sufficient. Numerical experiments validate the predicted communication and computational performance.

1 INTRODUCTION

Sampling from complex, multimodal distributions requires effective exploration: local Markov chains can remain trapped in a single mode and fail to visit other high-probability regions. Parallel tempering (PT) addresses this by running replicas along a temperature ladder, from the target distribution to progressively flatter distributions where transitions between modes are easier. PT periodically proposes exchanges between neighboring replicas and accepts them with a Metropolis rule. These exchanges allow states that explore different modes at high temperature to propagate back toward the target. The efficiency of this mechanism depends on overlap between adjacent distributions: poor overlap leads to low exchange acceptance and creates a communication bottleneck (Syed et al., 2022).

Transport-accelerated PT addresses this bottleneck by introducing a transport function that maps states between neighboring distributions before exchange. Recent approaches use expressive neural transports and can substantially improve communication (Zhang et al., 2026; Papamakarios et al., 2021). However, neighboring temperature levels may involve high-dimensional and complex distributions while differing only through a simple transformation, such as a global rescaling. In such settings, using a fixed high-capacity transport may introduce unnecessary modeling and computational cost. This motivates the central design question of our work: *can we design transport-accelerated PT that automatically identifies and uses the simplest transport complexity sufficient for effective cross-temperature communication?*

The key challenge is to determine whether a simple transport is already sufficient between neighboring temperatures. This cannot be inferred from the complexity of either distribution alone; what matters is the structure of their change. We therefore represent transport locally by a velocity field and ask how much of the distributional change it can explain. A Stein operator characterizes the infinitesimal density change induced by a given velocity, and the difference between this induced change and the actual change defines a Stein residual. A small residual means that the candidate motion captures most of the cross-temperature variation. This provides a principled way to adapt transport complexity to the structure of the change rather than to the ambient complexity of the distributions.

Based on this idea, we develop *Complexity-Adaptive Transport Parallel Tempering* (CAT-PT). CAT-PT organizes candidate transports into a nested hierarchy of increasingly expressive, cheaply integrable

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.

velocity classes. A calibration run of ordinary PT provides samples that are split into fitting and held-out validation blocks. Within each class, Stein regression uses the fitting block to identify the motion that best explains the local annealing change, and integrating this motion yields a finite invertible transport. The resulting candidates are then examined sequentially on the held-out block, starting from raw exchange and moving to richer transports only when needed. For each neighboring pair of annealing levels, CAT-PT selects the first candidate whose exchange performance is sufficient, so production uses the simplest certified kernel rather than a fixed high-capacity transport.

Our theory explains why this adaptive strategy improves PT communication. We first show that the accumulated Stein mismatch directly controls the rejection probability of the resulting finite transport. When most cross-temperature change is captured by a low-complexity transport, CAT-PT depends on a residual scale rather than the full ambient dimension. Under additional conditions, this yields a growing advantage over raw PT's fine-ladder limit. Moreover, under additional conditions, no transport in the same structured class can improve this fine-ladder scaling order. CAT-PT can also avoid the cost of an unnecessarily expressive transport. Numerical experiments support these predictions and demonstrate the communication and computational benefits of CAT-PT.

Overall, our main contributions are:

- We propose CAT-PT, a Stein-guided PT method that sequentially selects the simplest transport sufficient for each neighboring annealing pair.
- We connect Stein residual to finite-gap exchange performance and show that, in structured regimes, CAT-PT scales with residual rather than ambient complexity, with fine-ladder optimality of Stein fitting under aligned residuals.
- Numerical experiments validate the theory and show favorable communication and computational efficiency.

2 Related Work

Parallel tempering. Parallel tempering (PT) and replica-exchange methods improve exploration of multimodal distributions through exchanges across temperatures (Swendsen & Wang, 1986; Hukushima & Nemoto, 1996; Earl & Deem, 2005). Their efficiency depends on communication between neighboring distributions, and poor overlap can cause slow or torpid mixing (Woodard et al., 2009a;b). Prior work studies temperature schedules, optimized annealing paths, and non-reversible exchanges (Atchadé et al., 2011; Miasojedow et al., 2013; Vousden et al., 2016; Syed et al., 2021; 2022). CAT-PT instead adapts the complexity of the transport used between neighboring distributions.

Transport-accelerated sampling. Transport maps and normalizing flows provide flexible tools for simplifying sampling problems (Marzouk et al., 2017; Parno & Marzouk, 2018; Rezende & Mohamed, 2015; Dinh et al., 2016; Durkan et al., 2019; Papamakarios et al., 2021). Learned transports have also been combined with MCMC and annealed sampling (Hoffman et al., 2019; Arbel et al., 2021; Vargas et al., 2024). Neural transport accelerates PT by learning expressive maps that reduce cross-temperature mismatch (Zhang et al., 2026). CAT-PT addresses a different question: how much transport complexity is actually needed. It uses the Stein residual to quantify the variation left unexplained by structured motions and selects the simplest class that achieves sufficient exchange.

Stein methods. Stein operators underpin computable discrepancies and particle-transport methods (Liu et al., 2016; Liu & Wang, 2016; Liu, 2017; Detommaso et al., 2018). Stein transport uses continuity-equation geometry to move particles along tempered distributions (Nüsken, 2024). CAT-PT uses the same geometry to fit structured cross-temperature motions and measure the variation left unexplained by each transport class, enabling adaptive transport complexity.

3 Problem Formulation

We consider sampling from a target distribution π_1 through a smooth annealing path $\pi_t(\mathbf{x}) = \widetilde{\pi}_t(\mathbf{x})/Z_t, t \in [0,1]$, where π_0 is an easier reference distribution. Common examples include temperature annealing, $\widetilde{\pi}_t(\mathbf{x}) = \exp\{-\beta(t)U(\mathbf{x})\}$, and Bayesian likelihood tempering, $\widetilde{\pi}_t(\mathbf{x}) = p(\mathbf{x})p(\mathbf{y} \mid \mathbf{x})^{\beta(t)}$. We assume that $\widetilde{\pi}_t$, $\nabla_{\mathbf{x}} \log \widetilde{\pi}_t$, and $\partial_t \log \widetilde{\pi}_t$ can be evaluated.

3.1 PARALLEL TEMPERING AND TRANSPORTED EXCHANGE

PT introduces a smooth annealing path $\{\pi_t\}_{t\in[0,1]}$, interpolating from an easier reference distribution π_0 to the target π_1. The temperature ladder is obtained by discretizing this path at $0 = t_0 < t_1 < \cdots < t_K = 1$, with one Markov chain targeting each π_{t_k}.

During sampling, each replica first performs local MCMC updates at its own annealing level. PT then periodically proposes exchanges between neighboring replicas so that states discovered at easier levels can propagate toward the target. For adjacent levels $a = t_k$ and $b = t_{k+1}$, with $\mathbf{x} \sim \pi_a$ and $\mathbf{y} \sim \pi_b$, the raw exchange proposes $(\mathbf{x}, \mathbf{y}) \mapsto (\mathbf{y}, \mathbf{x})$ and is accepted with probability

$$\alpha_{a,b}^{\mathrm{raw}}(\mathbf{x}, \mathbf{y}) = 1 \wedge \frac{\widetilde{\pi}_a(\mathbf{y})\widetilde{\pi}_b(\mathbf{x})}{\widetilde{\pi}_a(\mathbf{x})\widetilde{\pi}_b(\mathbf{y})}.$$

When neighboring distributions overlap poorly, these exchange attempts are frequently rejected.

Transport-accelerated PT reduces this mismatch by an invertible map T that approximately maps π_a to π_b, it proposes $(\mathbf{x}, \mathbf{y}) \mapsto (T^{-1}(\mathbf{y}), T(\mathbf{x}))$. A Metropolis–Hastings correction, including the transport Jacobian, preserves exactness even when T is imperfect. Existing approaches can learn highly expressive transports, for example with neural flow models. However, this capacity can be unnecessary when neighboring distributions differ through simple structured changes.

3.2 STRUCTURED TRANSPORT THROUGH VELOCITY FIELDS

Transport capacity should reflect the structure of the change between neighboring levels, rather than the complexity of the distributions themselves. We represent such changes by *velocity fields*, where $\mathbf{v}(\mathbf{x})$ specifies how a state moves under an infinitesimal change of the annealing parameter.

We organize candidate motions into a nested hierarchy

$$\mathcal{V}_0 \subseteq \mathcal{V}_1 \subseteq \cdots \subseteq \mathcal{V}_M, \qquad \mathcal{V}_0 = \{\mathbf{0}\}, \tag{1}$$

with higher levels allowing richer deformations. For $m \geq 1$, we define the class $\mathcal{V}_m = \left\{\mathbf{v}_{m,\boldsymbol{\theta}} = \sum_{\ell=1}^{p_m} \theta_\ell \mathbf{v}_{m,\ell} : \boldsymbol{\theta} \in \mathbb{R}^{p_m}\right\}$, each basis field $\mathbf{v}_{m,\ell}$ represents an elementary motion, such as a translation, scaling, or blockwise affine deformation, while $\boldsymbol{\theta}$ determines how strongly these motions are combined. Thus, choosing $\boldsymbol{\theta}$ selects a particular motion within class $\mathcal{V}_m$; CAT-PT estimates these coefficients from data. A fitted velocity induces a finite transport through $d\mathbf{X}_u/du = \mathbf{v}_{m,\boldsymbol{\theta}}(\mathbf{X}_u)$. We restrict the hierarchy to fields whose flow, inverse, and Jacobian can be computed exactly or at low cost, so greater expressiveness does not require a high-capacity learned transport.

4 CAT-PT: A SIMPLE COMPLEXITY-ADAPTIVE SAMPLER

We propose *Complexity-Adaptive Transport Parallel Tempering* (CAT-PT) in this section. It first runs a calibration period of ordinary PT and splits the collected samples into fitting and validation blocks. Using the fitting samples, it constructs candidate transports from a nested hierarchy of velocity classes via Stein regression. It then evaluates these candidates sequentially on held-out samples, starting from raw exchange, and selects the first one that reaches a prescribed acceptance target. The selected kernel is then frozen and used throughout production.

4.1 PRE-PRODUCTION CALIBRATION

CAT-PT first uses a pre-production calibration period to construct and select the exchange kernel for each neighboring pair. Specifically, we run ordinary PT for B_{burn} iterations to reduce initialization bias and then collect B_{cal} further iterations. The collected samples are split into a fitting block D_{fit}, used to fit candidate transports, and a held-out block D_{val}, used to evaluate their exchange performance and select among them. A gap of g discarded iterations separates the blocks and is included in B_{cal}. The calibration budget should be large enough to support both transport fitting and held-out selection; the corresponding assumptions and sample-size conditions are given in Appendices A.3, C.1, and D.1.

4.2 STEIN-GUIDED VELOCITY FITTING

After calibration, we need to construct a candidate transport for each velocity class using D_{fit}. Ideally, this transport should capture as much as possible of the change from π_t to nearby distributions along the annealing path. Rather than fitting a finite map directly, we characterize this change locally through a velocity field and use Stein's identity to measure how well a candidate velocity explains it.

Let $g_t = \partial_t \log \widetilde{\pi}_t$ denote the directly evaluable change of the unnormalized density, and let $s_t = \partial_t \log \pi_t = g_t - \mathbb{E}_{\pi_t}[g_t]$ denote the corresponding normalized path score. For a velocity $\mathbf{v}$, define the Langevin–Stein operator $\mathcal{S}_t \mathbf{v} = \nabla \cdot \mathbf{v} + \mathbf{v}^\top \nabla \log \pi_t$. If particles move according to $\mathbf{v}_t$ while their distribution follows π_t, mass conservation gives $\partial_t \pi_t + \nabla \cdot (\pi_t \mathbf{v}_t) = 0$, which is equivalent to $s_t + \mathcal{S}_t \mathbf{v}_t = 0$. Thus, $-\mathcal{S}_t \mathbf{v}$ represents the log-density change induced by the motion, and $s_t + \mathcal{S}_t \mathbf{v}$ measures the part of the annealing change that it fails to explain.

For class $\mathcal{V}_m$, write $\mathbf{v}_{m,\boldsymbol{\theta}} = \sum_{\ell=1}^{p_m} \theta_\ell \mathbf{v}_{m,\ell}$. Since the Stein operator is linear, $\mathcal{S}_{t_j} \mathbf{v}_{m,\boldsymbol{\theta}} = \boldsymbol{\psi}_{j,m}^\top \boldsymbol{\theta}$, where $\psi_{j,m,\ell} = \mathcal{S}_{t_j} \mathbf{v}_{m,\ell}$. Hence the unexplained local change under this velocity is $s_{t_j} + \boldsymbol{\psi}_{j,m}^\top \boldsymbol{\theta}$. We learn a $\boldsymbol{\theta}$ that minimizes its squared $L^2(\pi_{t_j})$ magnitude. Although s_{t_j} contains the unknown normalizing-constant derivative, Stein features satisfy $\mathbb{E}_{\pi_{t_j}}[\boldsymbol{\psi}_{j,m}] = 0$. Therefore $\mathbb{E}_{\pi_{t_j}}[\big(s_{t_j} + \boldsymbol{\psi}_{j,m}^\top \boldsymbol{\theta}\big)^2] = \mathrm{Var}_{\pi_{t_j}}[g_{t_j} + \boldsymbol{\psi}_{j,m}^\top \boldsymbol{\theta}]$. Using the samples in D_{fit} from node t_j, we thus estimate

$$\widehat{\boldsymbol{\theta}}_{j,m} \in \arg\min_{\boldsymbol{\theta}} \widehat{\mathrm{Var}}\big[g_{t_j}(\mathbf{X}) + \boldsymbol{\psi}_{j,m}(\mathbf{X})^\top \boldsymbol{\theta}\big]. \tag{2}$$

This is a quadratic least-squares problem. Writing $\widehat{\mathbf{G}}_{j,m} = \widehat{\mathrm{Cov}}(\boldsymbol{\psi}_{j,m}, \boldsymbol{\psi}_{j,m})$ and $\widehat{\mathbf{b}}_{j,m} = \widehat{\mathrm{Cov}}(\boldsymbol{\psi}_{j,m}, g_{t_j})$, its minimum-norm solution is $\widehat{\boldsymbol{\theta}}_{j,m} = -\widehat{\mathbf{G}}_{j,m}^\dagger \widehat{\mathbf{b}}_{j,m}$. Thus each candidate can be fitted directly from sample covariances, without iterative optimization.

The fitted coefficients specify a local velocity at each annealing node. For edge $k = (t_k, t_{k+1})$, we average the endpoint fits as $\overline{\boldsymbol{\theta}}_{k,m} = (\widehat{\boldsymbol{\theta}}_{k,m} + \widehat{\boldsymbol{\theta}}_{k+1,m})/2$ and evolve $d\mathbf{X}_u/du = \mathbf{v}_{m,\overline{\boldsymbol{\theta}}_{k,m}}(\mathbf{X}_u)$ from $u = 0$ to $h_k = t_{k+1} - t_k$. The resulting map

$$\widehat{T}_{k,m}(\mathbf{x}) := \mathbf{X}_{h_k}, \qquad \mathbf{X}_0 = \mathbf{x}, \tag{3}$$

is the finite transport associated with class $\mathcal{V}_m$ on this edge. Appendix H.1 gives cheap exact flows, inverses, and Jacobians for the translation, diagonal-affine, and fixed-block affine classes.

4.3 EXCHANGE SELECTION AND FROZEN PRODUCTION

The fitted transport $\widehat{T}_{k,m}$ alone is not enough. To use it within PT while preserving the correct stationary distribution, we equip the proposal with a Metropolis–Hastings correction. For edge $k = (t_k, t_{k+1})$, candidate 0 is raw exchange, with $\widehat{T}_{k,0} = \mathrm{I}$. For $m \geq 1$, candidate m proposes $(\mathbf{x}, \mathbf{y}) \mapsto \big(\widehat{T}_{k,m}^{-1}(\mathbf{y}), \widehat{T}_{k,m}(\mathbf{x})\big)$. For a generic invertible map T, this proposal is accepted with probability $\alpha_{k,T}(\mathbf{x}, \mathbf{y}) = 1 \wedge R_{k,T}(\mathbf{x}, \mathbf{y})$, where

$$R_{k,T}(\mathbf{x}, \mathbf{y}) = \frac{\widetilde{\pi}_{t_k}(T^{-1}(\mathbf{y}))\widetilde{\pi}_{t_{k+1}}(T(\mathbf{x}))}{\widetilde{\pi}_{t_k}(\mathbf{x})\widetilde{\pi}_{t_{k+1}}(\mathbf{y})} \left|\det \nabla T^{-1}(\mathbf{y})\right| \left|\det \nabla T(\mathbf{x})\right|. \tag{4}$$

This correction makes every fitted candidate an exact exchange kernel. We denote its stationary acceptance by $A_{k,m} = \mathbb{E}_{\pi_{t_k} \otimes \pi_{t_{k+1}}}[\alpha_{k,m}]$ and rejection by $r_{k,m} = 1 - A_{k,m}$.

We use D_{val} to select the transport complexity on each edge. CAT-PT evaluates candidates sequentially, starting from raw exchange and moving to richer velocity classes only when needed. Let $\widehat{A}_{k,m}$ denote the average validation acceptance of candidate m. We select the first candidate satisfying

$$\widehat{A}_{k,m} - \omega \geq a_\star, \tag{5}$$

where $a_\star \in (0, 1)$ is the target acceptance and $\omega \geq 0$ is a validation margin. If no candidate satisfies the criterion, we choose the evaluated candidate with the largest $\widehat{A}_{k,m}$. The selected index $\widehat{c}_k$ and its transport are then frozen for production. Each edge uses raw exchange when $\widehat{c}_k = 0$ and the corresponding transported exchange otherwise. Production proceeds as in standard PT, with local MCMC updates followed by neighboring exchanges.

Algorithm 1 Complexity-adaptive transport parallel tempering (CAT-PT)

Require: Fixed ladder and local kernels; $\{\mathcal{V}_m\}_{m=1}^{M}$; $a_\star, \omega$; $B_{\mathrm{burn}}, B_{\mathrm{cal}}, B_{\mathrm{prod}}$; gap g.

Run ordinary PT for B_{burn} iterations, then collect B_{cal} calibration iterations and split them into D_{fit} and D_{val}, discarding g intervening iterations.
On every edge, evaluate candidate 0 (raw exchange, $\widehat{T}_{k,0} = \mathrm{I}$) on D_{val} using equation (4).
Set $\widehat{c}_k = 0$ for edges satisfying equation (5), and let $\mathcal{U}$ contain the unresolved edges.
for $m = 1, \ldots, M$ **do**
    if $\mathcal{U} = \varnothing$ **then**
        break
    end if
    Fit the level-m velocities at nodes incident to $\mathcal{U}$ using D_{fit} via equation (2).
    for each $k \in \mathcal{U}$ **do**
        Construct $\widehat{T}_{k,m}$ via equation (3) and evaluate its exchange acceptance on D_{val} using equation (4).
        if candidate m satisfies equation (5) **then**
            Set $\widehat{c}_k \leftarrow m$ and remove k from $\mathcal{U}$.
        end if
    end for
end for
For each remaining $k \in \mathcal{U}$, choose the evaluated candidate with the largest validation acceptance, including candidate 0. Freeze all selected indices and maps.
for $b = 1, \ldots, B_{\mathrm{prod}}$ **do**
    Apply the local MCMC updates to all replicas.
    Attempt neighboring exchanges using candidate $\widehat{c}_k$ on each scheduled edge, with acceptance given by equation (4).
end for
return production samples from the target replica.

Thus, CAT-PT fits increasingly expressive Stein transports using D_{fit}, selects the simplest one with sufficient exchange performance using D_{val}, and produces with the resulting frozen edgewise kernels. Algorithm 1 summarizes the procedure.

5 Theoretical Analysis

We now analyze why the Stein-guided selection used by CAT-PT improves cross-temperature communication. We first show that the Stein residual directly controls the rejection probability of the finite transported exchange. We then show that held-out selection transfers this edgewise guarantee to the frozen CAT-PT sampler. Finally, we characterize structured regimes where the resulting communication depends on the residual variation left after transport, rather than the ambient dimension.

Assumption 1 (Regular path and transports)**.** *The annealing path and velocity fields are sufficiently regular for the Stein identities below to hold and for each fitted velocity to generate a valid invertible transport. The calibration samples are sufficiently close to stationarity and satisfy the concentration conditions required for reliable acceptance estimation.*

Precise sufficient conditions are given in Appendix A.

5.1 Stein fitting controls finite exchange

The Stein regression in Section 4.2 fits a velocity by matching the *local* change of the annealing distribution. What matters for PT, however, is whether the resulting *finite* transport produces a high exchange acceptance between neighboring levels. We now connect these two quantities.

For a velocity $\mathbf{v}$, define its Stein residual $\rho_t^{\mathbf{v}} = s_t + \mathcal{S}_t \mathbf{v}$. Within class $\mathcal{V}_m$, its squared residual is $\mathcal{R}_m(t, \boldsymbol{\theta}) = \|\rho_t^{\mathbf{v}_{m,\boldsymbol{\theta}}}\|^2_{L^2(\pi_t)}$. The smallest residual achievable by the class is

$$\mathcal{E}_m(t) := \inf_{\boldsymbol{\theta}} \mathcal{R}_m(t, \boldsymbol{\theta}), \qquad \mathcal{L}_{k,m} := \int_{t_k}^{t_{k+1}} \sqrt{\mathcal{E}_m(t)}\, dt. \tag{6}$$

Here $\mathcal{E}_m(t)$ measures the annealing variation that cannot be explained by the best velocity in $\mathcal{V}_m$ at time t, while $\mathcal{L}_{k,m}$ accumulates this irreducible mismatch across edge k. Because the hierarchy is nested, $\mathcal{E}_{m+1}(t) \le \mathcal{E}_m(t)$. The implemented transport uses the estimated edge coefficient $\overline{\boldsymbol{\theta}}_{k,m}$ rather than the population optimum. We therefore define

$$\varepsilon^{\mathrm{fit}}_{k,m} := \int_{t_k}^{t_{k+1}} \sqrt{\mathcal{R}_m(t, \overline{\boldsymbol{\theta}}_{k,m}) - \mathcal{E}_m(t)}\, dt, \qquad \Delta_{k,m} := \mathcal{L}_{k,m} + \varepsilon^{\mathrm{fit}}_{k,m}. \tag{7}$$

The first term in $\Delta_{k,m}$ is the approximation error of the velocity class itself; the second captures the additional error from estimating the endpoint velocities from finite calibration data and holding their average constant across the edge. Thus $\Delta_{k,m}$ measures the total unexplained variation of the fitted transport over edge k. For a velocity curve $\mathbf{v} = \{\mathbf{v}_u\}_{u\in[s,t]}$, let $T^{\mathbf{v}}_{s\to t}$ denote the finite transport obtained by integrating the flow $d\mathbf{X}_u/du = \mathbf{v}_u(\mathbf{X}_u)$ from level s to level t. Let $r^{\mathrm{tr}}_{s,t}(T)$ denote the mean rejection probability of the Metropolis-corrected transported exchange between π_s and π_t when the two replicas are at stationarity.

Theorem 1 (Stein residual controls finite-gap exchange)**.** *For any admissible velocity curve* $\mathbf{v}$ *with exact flow* $T^{\mathbf{v}}_{s\to t}$, $r^{\mathrm{tr}}_{s,t}(T^{\mathbf{v}}_{s\to t}) \le \min\left\{1, \int_s^t \|s_u + \mathcal{S}_u \mathbf{v}_u\|_{L^2(\pi_u)}\, du\right\}$. *In particular, every fitted CAT-PT candidate satisfies* $1 - A_{k,m} \le \Delta_{k,m}$.

The theorem provides the key link between Stein fitting and the exchange kernel used by CAT-PT. If a class captures most of the annealing change ($\mathcal{L}_{k,m}$ is small) and its velocity is accurately fitted and approximated across the edge ($\varepsilon^{\mathrm{fit}}_{k,m}$ is small), then the corresponding finite transport must have high exchange acceptance.

5.2 Global communication guarantee

Theorem 1 controls the rejection probability of each candidate transport on a single edge. CAT-PT, however, selects one candidate per edge using D_{val} and then runs the resulting frozen kernels across the entire ladder. We next show how reliable held-out selection converts these edgewise guarantees into a bound on global communication.

We measure communication by the round-trip rate τ, defined as the aggregate number of completed reference–target–reference journeys per exchange sweep across all replica identities. Under deterministic even–odd (DEO) scheduling and the efficient-local-exploration (ELE) model, the rate:

$$\tau = \left(2 + 2\sum_{k=0}^{K-1} \frac{r_k}{1 - r_k}\right)^{-1}.$$

Appendix C.2 derives this identity and its normalization (Syed et al., 2022; Zhang et al., 2026).

Define the uniform validation event $\mathsf{V}(\omega) = \{\max_{k,c} |\widehat{A}_{k,c} - A_{k,c}| \le \omega\}$, where the maximum is over all edge–candidate pairs. On this event, the validation acceptance of every candidate is accurate within ω. Hence any candidate satisfying the selection rule $\widehat{A}_{k,m} - \omega \ge a_\star$ has true acceptance $A_{k,m} \ge a_\star$. The connection with Theorem 1 is direct: since $1 - A_{k,m} \le \Delta_{k,m}$, a small residual implies high true acceptance. In particular, $\Delta_{k,m} \le 1 - a_\star$ is sufficient for $A_{k,m} \ge a_\star$, while the stronger condition $\Delta_{k,m} \le 1 - a_\star - 2\omega$ guarantees that the candidate is certified by the validation rule on $\mathsf{V}(\omega)$. Appendix C.1 proves that $\mathsf{V}(\omega)$ holds with high probability.

Proposition 1 (Held-out round-trip guarantee)**.** *For any frozen choices, CAT-PT preserves* $\bigotimes_{k=0}^{K} \pi_{t_k}$. *On* $\mathsf{V}(\omega)$*, if every selected candidate satisfies the selection criterion, then under DEO and ELE,*

$$\tau_{\text{CAT-PT}} \ge \left(2 + 2K\frac{1 - a_\star}{a_\star}\right)^{-1}. \tag{8}$$

Thus, Theorem 1 converts Stein residual into edgewise acceptance, while Proposition 1 converts sufficiently high acceptance across the ladder into a global round-trip guarantee.

5.3 STRUCTURED COMMUNICATION ADVANTAGE

The previous results show that a small Stein residual yields high exchange acceptance, which in turn improves communication across the ladder. We now ask when this leads to an advantage over raw PT.

Consider d-dimensional problems for which some level m_d leaves only residual variation of scale r_d, in the sense that $\mathcal{E}_{m_d}(t) \le C_0 r_d$ uniformly in t. The quantity r_d may be much smaller than d when most cross-temperature variation is captured by a simple structured motion.

Theorem 2 (Structured performance of CAT-PT)**.** *Fix $a_\star$ and $\omega < (1 - a_\star)/2$. Suppose that, for some level m_d, $\mathcal{E}_{m_d}(t) \le C_0 r_d$ uniformly in t, and that $\varepsilon^{\mathrm{fit}}_{k,m_d} \le C_1 h_k(1 + \sqrt{r_d})$ with dimension-independent C_0, C_1. Consider a uniform temperature ladder with $K_d = \lceil C(1 + \sqrt{r_d}) \rceil$ edges. For sufficiently large C, CAT-PT selects no level above m_d on $\mathsf{V}(\omega)$. Under DEO and ELE,*

$$\tau_{\text{CAT-PT}} \ge \left(2 + 2K_d \frac{1 - a_\star}{a_\star}\right)^{-1} = \Omega\big((1 + \sqrt{r_d})^{-1}\big)\,. \tag{9}$$

The theorem shows that CAT-PT communication is governed by the variation left unexplained after structured transport rather than directly by the ambient dimension. In particular, bounded r_d permits a constant-size ladder and a round-trip rate bounded away from zero as d grows. On this ladder, the sequential procedure does not need to know m_d in advance: once a sufficiently expressive level is reached, validation stops there or at an even simpler candidate.

For comparison, let $\tau^{\mathrm{fine}}_{\mathrm{raw}}$ denote the DEO/ELE round-trip rate of raw PT in the fine-ladder limit. Under the raw-score conditions of Appendix D.3, $\tau^{\mathrm{fine}}_{\mathrm{raw}} = O(d^{-1/2})$. Thus, even arbitrarily fine raw ladders retain an ambient-dimensional communication barrier, whereas CAT-PT depends on the residual scale r_d. Appendix D.1 discusses sufficient conditions for $\varepsilon^{\mathrm{fit}}_{k,m_d} \le C_1 h_k(1 + \sqrt{r_d})$ to hold. Appendix E gives concrete problem families satisfying the residual condition.

6 NUMERICAL EXPERIMENTS

We use synthetic experiments to examine two questions: whether CAT-PT can achieve strong temperature-space communication with low-complexity transports, whether its communication remains stable as ambient dimension grows while the relevant structured complexity is fixed. Full implementation details, benchmark definitions, and additional comparisons are given in Appendix J.

Baselines. The main text compares CAT-PT with Raw DEO-PT and NF-APT. *Raw DEO-PT* uses standard neighboring-state exchanges without transport and shares the same temperature ladder and local MCMC kernels as CAT-PT. It therefore isolates the effect of transported exchange.

NF-APT is the deterministic normalizing-flow variant of accelerated parallel tempering (Zhang et al., 2026). Our implementation fits a separate 20-layer RealNVP transport on each neighboring edge. NF-APT and CAT-PT both use deterministic transported exchanges with a Metropolis correction, but NF-APT deploys a fixed high-capacity neural map on every edge, whereas CAT-PT selects from a low-cost structured hierarchy.

Metrics. We report three complementary metrics that separate communication quality, computational cost, and downstream sampling efficiency. The number of completed round trips R measures how effectively replica identities move between the target and the easiest distribution under a fixed number of exchange sweeps. It is also the empirical quantity most directly corresponding to the round-trip rate analyzed in Section 5. Round trips per second, R/s, measures the same communication mechanism after accounting for production wall-clock cost. It therefore distinguishes a transport that produces more successful communication per sweep from one that achieves more communication per unit computation.

Finally, we report effective sample size per second (ESS/s) for prespecified scalar statistics of samples from the target replica. Whereas R and R/s describe movement through the temperature ladder, ESS/s measures the resulting efficiency for estimating target-distribution quantities. Improved temperature-space communication can increase ESS by reducing long-range autocorrelation, but the two need not move identically because ESS also depends on local mixing and on the statistic being

estimated. ESS is computed separately for each independent production run, summed across runs, and divided by production time (Vehtari et al., 2021). When several statistics are summarized by one number, we report the geometric mean of their ESS/s.

6.1 Main performance

We consider three complementary synthetic problems. The shifted Gaussian has an exactly representable low-complexity transport. Structured Many-Well combines multimodality with low-dimensional cross-temperature structure in a higher-dimensional state space. The 40-mode Gaussian mixture provides a more complex multimodal benchmark similar to those used in recent transport-accelerated PT experiments (Midgley et al., 2022; Zhang et al., 2026).

Shifted Gaussian. We consider the 10-dimensional path $\pi_t = \mathcal{N}\left((-1+2t)\mathbf{1},\, 0.1^2 I_{10}\right)$. The covariance is fixed while the mean moves linearly along the annealing path, so the exact cross-temperature motion is a translation. We use a 10-replica ladder.

Structured Many-Well. The target is 32-dimensional with four active double-well coordinates, giving $2^4 = 16$ target modes. The remaining coordinates are Gaussian nuisance directions. A global linear transformation induces structured variation along the annealing path, while the residual multimodal variation is confined to an $r_{\mathrm{str}} = 8$ subspace. The experiment therefore combines a multimodal target with cross-temperature variation that remains low-dimensional relative to the ambient state space.

40-mode Gaussian mixture. We use a 10-dimensional equal-weight mixture of 40 Gaussian components with component standard deviation 0.025. The component means vary in the first two coordinates, while the remaining eight coordinates have zero component means. A 10-dimensional standard Gaussian is connected to the target through a geometric annealing path. This benchmark provides a setting in which a more expressive transport can be useful because the cross-temperature change is not exactly represented by the simple structured motions used in the preceding examples.

Table 1: Main synthetic results. R denotes completed round trips under the fixed production budget. GM-ESS/s is the geometric mean of ESS/s over the benchmark-specific target statistics defined in Appendix J.

Target	Method	R	R/s	GM-ESS/s
Shifted Gaussian	Raw PT	0	0	**36,232**
	NF-APT	40,670	1,911	4,700
	CAT-PT	**49,995**	**13,189**	26,381
Many-Well	Raw PT	2,639	154.6	130.2
	NF-APT	6,984	198.0	161.4
	CAT-PT	**7,379**	**410.0**	**335.8**
40-mode GMM	Raw PT	703.1	563.7	782.8
	NF-APT	**10,720.4**	1,620.5	1,000.0
	CAT-PT	7,376.1	**4,485.2**	**2,761.9**

Table 1 reports absolute performance, while Figure 1 emphasizes the relative CAT-PT/NF-APT comparison. On the shifted Gaussian and Structured Many-Well targets, low-complexity structured transports are sufficient to match or improve NF-APT communication while providing higher computational efficiency. On the shifted Gaussian, CAT-PT recovers the low-complexity transport and substantially improves communication over Raw PT and NF-APT. The high Raw-PT ESS/s for local Gaussian moments illustrates that temperature-space communication and observable-specific estimation efficiency need not move together. On the 40-mode GMM, the more expressive NF-APT transport achieves higher communication per sweep, but CAT-PT remains substantially more efficient per unit wall-clock time. These results illustrate the capacity–cost tradeoff targeted by CAT-PT: additional transport capacity can improve exchange quality, but need not improve overall sampling efficiency when simpler cross-temperature structure is sufficient.

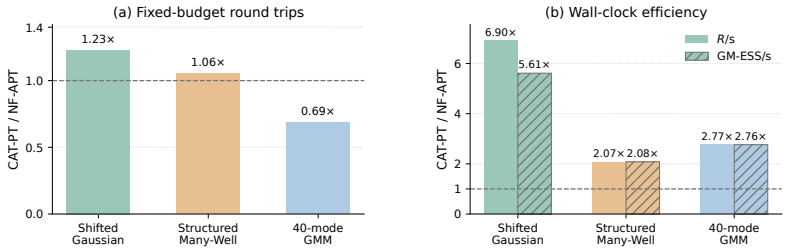

Figure 1: CAT-PT relative to NF-APT on the three main synthetic benchmarks. **Left:** fixed-budget round-trip ratio. **Right:** production round trips per second and GM-ESS/s ratio. Values above one favor CAT-PT; the dashed line marks parity.

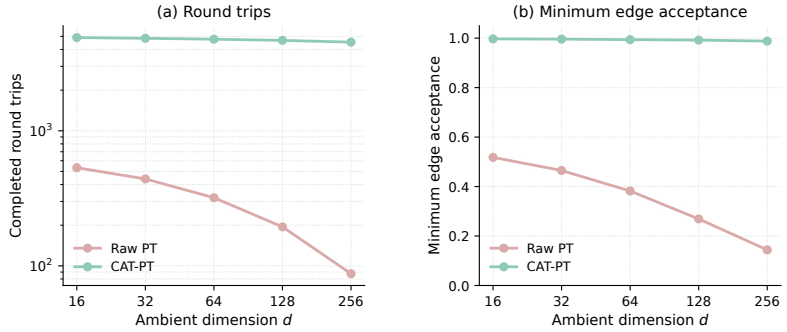

Figure 2: Scaling with ambient dimension at fixed structured rank $r_{\mathrm{str}} = 8$. **Left:** completed round trips under 10,000 exchange sweeps. **Right:** minimum edge acceptance.

6.2 Ambient dimension at fixed structured complexity

We next isolate the ambient-dimension dependence predicted by Theorem 2. We consider a family of zero-mean Gaussian paths in which all coordinates undergo a common scale change, while a fixed set of $r_{\mathrm{str}} = 8$ coordinates has an additional structured scale variation. CAT-PT is given the structured coordinate set but must estimate the corresponding transport coefficients from calibration samples. We vary $d \in \{16, 32, 64, 128, 256\}$ while holding $r_{\mathrm{str}} = 8$ fixed. All dimensions use the same 10-replica ladder and 10,000 exchange sweeps, so increasing d adds nuisance dimensions without increasing the dimension of the structured correction. The complete construction and calibration protocol are given in Appendix J.

CAT-PT remains nearly stable over the full range of ambient dimensions: its minimum edge acceptance stays between 0.988 and 0.997, and it retains 92.5% of its $d = 16$ round-trip count at $d = 256$. Raw PT deteriorates substantially over the same range, retaining only 16.4% of its initial round-trip count as its minimum edge acceptance falls from 0.518 to 0.144. Consequently, the CAT-PT/Raw-PT round-trip ratio grows from $9.19\times$ at $d = 16$ to $51.76\times$ at $d = 256$. This experiment tests a favorable structured regime consistent with Theorem 2: the representable cross-temperature variation has fixed structured complexity as ambient nuisance dimensions are added.

7 Conclusion

We introduced CAT-PT, a complexity-adaptive transport method for parallel tempering that matches transport capacity to the structure of cross-temperature variation. Its Stein residual provides a common quantity for fitting structured motions and analyzing exchange performance, while sequential selection avoids unnecessary transport complexity. Our theory shows that, in structured high-dimensional regimes, communication can scale with the residual variation left after transport rather than with the ambient dimension, with matching within-class guarantees under additional conditions. Experiments support these predictions and demonstrate improved communication and computational efficiency.

AI USE STATEMENT

In this work, we used generative AI tools to help develop conceptual frameworks, design and provide feedback on research methodology and experiments, and assist with method implementation. We also used generative AI tools to draft parts of the paper, summarize and analyze existing literature, and edit the manuscript to improve readability. We did not use generative AI tools for other tasks requiring disclosure. We reviewed all AI-assisted work. Implementation suggestions were tested through our experimental pipeline, and literature-based statements were verified against the cited sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce both the theoretical and experimental results. Appendices A–I state the technical assumptions and provide proofs of all theoretical results, including finite-sample fitting and validation conditions and computational-cost analyses. Appendix J documents the synthetic targets, transport hierarchy, baseline implementations, evaluation metrics, sampling protocols, and main experimental settings.

REFERENCES

Michael Arbel, Alex Matthews, and Arnaud Doucet. Annealed flow transport monte carlo. In *International Conference on Machine Learning*, pp. 318–330. PMLR, 2021.

Yves F Atchadé, Gareth O Roberts, and Jeffrey S Rosenthal. Towards optimal scaling of metropolis-coupled markov chain monte carlo. *Statistics and Computing*, 21(4):555–568, 2011.

Gianluca Detommaso, Tiangang Cui, Youssef Marzouk, Alessio Spantini, and Robert Scheichl. A stein variational newton method. *Advances in Neural Information Processing Systems*, 31, 2018.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

David J Earl and Michael W Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

Matthew Hoffman, Pavel Sountsov, Joshua V Dillon, Ian Langmore, Dustin Tran, and Srinivas Vasudevan. Neutra-lizing bad geometry in hamiltonian monte carlo using neural transport. *arXiv preprint arXiv:1903.03704*, 2019.

Koji Hukushima and Koji Nemoto. Exchange monte carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, 1996.

Qiang Liu. Stein variational gradient descent as gradient flow. *Advances in neural information processing systems*, 30, 2017.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pp. 276–284. PMLR, 2016.

Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. Sampling via measure transport: An introduction. *Springer Books*, pp. 785–825, 2017.

Błażej Miasojedow, Eric Moulines, and Matti Vihola. An adaptive parallel tempering algorithm. *Journal of Computational and Graphical Statistics*, 22(3):649–664, 2013.

Laurence Illing Midgley, Vincent Stimper, Gregor NC Simm, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Flow annealed importance sampling bootstrap. *arXiv preprint arXiv:2208.01893*, 2022.

Nikolas Nüsken. Stein transport for bayesian inference. *arXiv preprint arXiv:2409.01464*, 2024.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Matthew D Parno and Youssef M Marzouk. Transport map accelerated markov chain monte carlo. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):645–682, 2018.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Robert H Swendsen and Jian-Sheng Wang. Replica monte carlo simulation of spin-glasses. *Physical review letters*, 57(21):2607, 1986.

Saifuddin Syed, Vittorio Romaniello, Trevor Campbell, and Alexandre Bouchard-Côté. Parallel tempering on optimized paths. In *International Conference on Machine Learning*, pp. 10033–10042. PMLR, 2021.

Saifuddin Syed, Alexandre Bouchard-Côté, George Deligiannidis, and Arnaud Doucet. Non-reversible parallel tempering: A scalable highly parallel mcmc scheme. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(2):321–350, 2022.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Francisco Vargas, Shreyas Padhy, Denis Blessing, and Nikolas Nüsken. Transport meets variational inference: Controlled monte carlo diffusions. In *International Conference on Learning Representations*, volume 2024, pp. 55236–55278, 2024.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved r for assessing convergence of mcmc (with discussion). *Bayesian analysis*, 16(2):667–718, 2021.

WD Vousden, Will M Farr, and Ilya Mandel. Dynamic temperature selection for parallel tempering in markov chain monte carlo simulations. *Monthly Notices of the Royal Astronomical Society*, 455 (2):1919–1937, 2016.

Dawn Woodard, Scott Schmidler, and Mark Huber. Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electronic Journal of Probability*, 14:780–804, 2009a.

Dawn B Woodard, Scott C Schmidler, and Mark Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *Annals of Applied Probability*, 19:617–640, 2009b.

Leo Zhang, Peter Potaptchik, Jiajun He, Yuanqi Du, Arnaud Doucet, Francisco Vargas, Hai-Dang Dau, and Saifuddin Syed. Accelerated parallel tempering via neural transports. In *International Conference on Learning Representations*, volume 2026, pp. 90874–90907, 2026.

A TECHNICAL SETUP AND ASSUMPTIONS

This section collects the notation and assumptions shared by the main results. Proofs of the three main results follow in Appendices B–D.

A.1 ANNEALING DISTRIBUTIONS, EXCHANGE KERNELS, AND CALIBRATION DATA

Let p_t be the density of π_t and let $s_t(\mathbf{x}) = \partial_t \log p_t(\mathbf{x})$ be its normalized path score. Since $p_t = \widetilde{\pi}_t/Z_t$ and $g_t = \partial_t \log \widetilde{\pi}_t$,

$$s_t = g_t - \mathbb{E}_{\pi_t}[g_t], \qquad \mathbb{E}_{\pi_t}[s_t] = 0. \tag{10}$$

The centering term is the unknown derivative of $\log Z_t$; this is why the algorithm minimizes an empirical variance rather than an uncentered square.

For a regular vector field $\mathbf{v} : \mathbb{R}^d \to \mathbb{R}^d$, define the Langevin–Stein operator

$$\mathcal{S}_t\mathbf{v} := \nabla\cdot\mathbf{v} + \mathbf{v}^\top \nabla \log p_t. \tag{11}$$

Under the zero-flux condition stated below, $\mathbb{E}_{\pi_t}[\mathcal{S}_t\mathbf{v}] = 0$. Hence the Stein residual $s_t + \mathcal{S}_t\mathbf{v}$ is centered and can be compared directly with the centered path score.

For finite measures P, Q of equal mass, write $\|P - Q\|_{\mathrm{TV}} = \frac{1}{2}\int |dP - dQ|$. Total variation is unchanged by a measurable bijection. For a time-dependent velocity $\mathbf{v}_t$, let $\Phi^{\mathbf{v}}_{s,t}$ solve

$$\partial_t \Phi^{\mathbf{v}}_{s,t}(\mathbf{x}) = \mathbf{v}_t(\Phi^{\mathbf{v}}_{s,t}(\mathbf{x})), \qquad \Phi^{\mathbf{v}}_{s,s}(\mathbf{x}) = \mathbf{x}. \tag{12}$$

The flow property gives $\Phi_{r,t} = \Phi_{s,t} \circ \Phi_{r,s}$ and $(\Phi_{s,t})^{-1} = \Phi_{t,s}$.

For an invertible map T, write $G_T(\mathbf{x}, \mathbf{y}) = (T^{-1}(\mathbf{y}), T(\mathbf{x}))$ for the transported pair proposal. On edge k, let $\mu_k = \pi_{t_k} \otimes \pi_{t_{k+1}}$ and let $R_{k,T}$ be the ratio in equation (4); when the edge is fixed, we abbreviate it as R_T. Candidate zero uses the identity map; candidate $m \geq 1$ uses $\widehat{T}_{k,m}$. Its direct acceptance is $\alpha_{k,c} = \alpha_{k,\widehat{T}_{k,c}} = 1 \wedge R_{k,\widehat{T}_{k,c}}$, so $A_{k,c} = \mathbb{E}_{\mu_k}\alpha_{k,c}$ and $r_{k,c} = 1 - A_{k,c}$. In particular, $A_{k,0} = A^{\mathrm{raw}}_k$ and $A_{k,m} = A^{\mathrm{tr}}_{k,m}$ for $m \geq 1$. After selection, $A^{\text{CAT-PT}}_k = A_{k,\widehat{c}_k}$ is the acceptance of the single frozen kernel on that edge. The validation loss is $\ell_{k,c} = 1 - \alpha_{k,c}$, and $\widehat{A}_{k,c} = 1 - \widehat{\mathbb{E}}\ell_{k,c}$. Validation therefore estimates the acceptance of the exchange kernel that will be used during production. We also use the full-path residual length $\mathcal{L}_m := \int_0^1 \sqrt{\mathcal{E}_m(t)}\, dt = \sum_{k=0}^{K-1} \mathcal{L}_{k,m}$.

We write N_{fit} and N_{val} for the numbers of retained ensemble observations in the two blocks. Thus each fitting observation supplies one state at every node, and each validation observation supplies a simultaneous pair at every edge; these counts are not multiplied by the number of replicas. The effective sample sizes $N^{\mathrm{eff}}_{\mathrm{fit}}$ and $N^{\mathrm{eff}}_{\mathrm{val}}$ are those appearing in the stated concentration bounds and need not equal the retained iteration counts.

A.2 REGULARITY OF THE ANNEALING PATH AND TRANSPORT FLOWS

The following assumption gives a concrete set of sufficient conditions for the path and flow regularity required by Assumption 1. Conditions on the calibration samples are stated in Section A.3.

Assumption 2. *For every structured level used by the algorithm:*

1. *The density $p_t(\mathbf{x})$ is strictly positive and continuously differentiable in t and $\mathbf{x}$. The path score s_t belongs to $L^2(\pi_t)$, and differentiation and integration may be interchanged for the quantities used below.*

2. *Every basis field $\mathbf{v}_{m,\ell}$ is continuously differentiable and globally Lipschitz. On bounded coefficient sets, the fields, their spatial derivatives, divergences, and Stein features admit common envelopes integrable under the relevant π_t. The Stein features $\mathcal{S}_t\mathbf{v}_{m,\ell}$ belong to $L^2(\pi_t)$ for every $t \in [0, 1]$.*

3. *Every bounded measurable coefficient curve generates a global C^1 diffeomorphic flow, with*

$$\log|\det \nabla\Phi_{s,t}(\mathbf{x})| = \int_s^t \nabla\cdot\mathbf{v}_u(\Phi_{s,u}(\mathbf{x}))\, du. \tag{13}$$

4. *The boundary flux vanishes:* $\int \nabla \cdot \{p_t(\mathbf{x})\mathbf{v}_{m,\boldsymbol{\theta}}(\mathbf{x})\}\, d\mathbf{x} = 0$.

5. *For every admissible velocity curve, the co-moving square-root density introduced in Appendix B is absolutely continuous as an L^2-valued map.*

Conditions 1–2 ensure that the path score and Stein features are well defined, smooth in annealing time, and have the moments needed by the regression and subsequent differentiation arguments. They hold, for example, for smooth Gaussian paths with covariance uniformly bounded away from singularity, together with translation or affine velocity fields.

Condition 3 guarantees that a fitted velocity defines a valid finite transport: the flow exists globally, is invertible, and has a computable Jacobian. The translation, diagonal-affine, and block-affine classes used by CAT-PT satisfy these requirements directly; their explicit flows are given in Appendix H.1.

Condition 4 is the boundary condition required for Stein integration by parts, so that $\mathbb{E}_{\pi_t}[\mathcal{S}_t\mathbf{v}] = 0$. It holds, for example, when the target tails dominate the growth of the velocity field, as with Gaussian targets and affine fields, or under the corresponding no-flux condition on a bounded domain. Condition 5 provides the L^2 path regularity used in the finite-gap rejection argument; smooth Gaussian paths with bounded affine velocities satisfy it under the same moment-control conditions.

A.3 Calibration law and concentration assumptions

CAT-PT fits transports using D_{fit} and chooses among them using D_{val}. For the validation estimates to describe stationary exchange performance, the pre-production PT run must have sufficiently reduced the effect of initialization. The two blocks must also be sufficiently separated to control their dependence, and the validation block must contain enough information to estimate acceptance. The assumption below states these three requirements separately.

Let P_{ind} denote the law in which D_{fit} and D_{val} are independent stationary blocks of the ordinary-PT chain. Let $P^{(g)}_{\mathrm{stat}}$ denote the law of two stationary blocks from a single trajectory separated by a discarded gap of g iterations, and let Q_{cal} denote the actual retained-block law after B_{burn} iterations.

Assumption 3 (Calibration regularity)**.** *The calibration trajectory satisfies the following conditions.*

1. ***Burn-in accuracy.*** *The retained calibration blocks are close to their stationary counterparts:*
$$\|Q_{\mathrm{cal}} - P^{(g)}_{\mathrm{stat}}\|_{\mathrm{TV}} \le \eta_{\mathrm{burn}}.$$

2. ***Block dependence.*** *The dependence between the stationary fitting and validation blocks is controlled by*
$$\left\|P^{(g)}_{\mathrm{stat}} - P_{\mathrm{ind}}\right\|_{\mathrm{TV}} \le \beta_{\mathrm{mix}}(g). \tag{14}$$

3. ***Validation concentration.*** *Under P_{ind}, condition on the fitting block and all fitting randomness. For every bounded candidate loss f in the finite library and every $\delta \in (0,1)$,*
$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{Z}_i) - \mathbb{E}f(\mathbf{Z})\right| > C_{\mathrm{val}}\|f\|_\infty \sqrt{\frac{\log(2/\delta)}{N^{\mathrm{eff}}_{\mathrm{val}}}}\right) \le \delta. \tag{15}$$

The first condition controls initialization error. If the complete replica ensemble after burn-in is within η_{burn} of $\Pi = \bigotimes_{j=0}^{K} \pi_{t_j}$ in total variation, the joint law of all subsequently retained samples is within the same distance of its stationary counterpart. This follows by applying the same PT transitions and sample retention rule to the two initial laws; total variation cannot increase under this operation. Any quantitative convergence bound for the ordinary-PT chain can be used to choose B_{burn}. For example, if $\|\mathrm{Law}(X_n) - \Pi\|_{\mathrm{TV}} \le C_{\mathrm{init}}\rho^n$ for some $\rho < 1$, then it is sufficient to choose B_{burn} such that $C_{\mathrm{init}}\rho^{B_{\mathrm{burn}}} \le \eta_{\mathrm{burn}}$, where X_n denotes the complete ensemble. This specifies burn-in through a convergence bound for the preliminary chain, separately from the number of samples retained for fitting and validation.

The second condition controls the fact that D_{fit} and D_{val} come from the same trajectory. Under stationarity, a β-mixing chain satisfies equation (14), with $\beta_{\mathrm{mix}}(g)$ decreasing as the gap g grows: the two retained blocks are functions of a past and a future segment of the chain, whose dependence

is bounded by $\beta_{\mathrm{mix}}(g)$. Here independence refers to the two blocks. Observations within either block may still be correlated.

The third condition controls the error of an acceptance estimate for a fixed fitted candidate. Here $\mathbf{Z}_i$ is a retained neighboring-state pair, $\mathbb{E}f(\mathbf{Z})$ is its stationary expected loss, and $N_{\mathrm{val}}^{\mathrm{eff}}$ describes the information available in the N validation observations through the displayed inequality. Under P_{ind}, a candidate constructed only from D_{fit} is independent of these validation observations, so its loss can be treated as fixed. For iid observations, equation (15) follows from Hoeffding's inequality with $N_{\mathrm{val}}^{\mathrm{eff}} = N$ and $C_{\mathrm{val}} = 1/\sqrt{2}$. For stationary Markov samples, suitable spectral-gap or mixing bounds give concentration of this form with a dependence-adjusted effective sample size.

These conditions control three different sources of validation uncertainty: initialization, dependence between blocks, and finite-sample averaging within the validation block. Proposition 3 combines them into a simultaneous error bound for the candidate acceptances. Appendix D.1 gives the separate sample and conditioning requirements for estimating the velocities. The invariant distribution of the frozen production kernels is established in Appendix C.2.

B Proof of Theorem 1: exchange rejection

Theorem 1 connects the quantity minimized by Stein regression to the rejection probability of the finite exchange map actually used by CAT-PT. We first state the population residual decomposition and the co-moving score identity. Together, these identify the unexplained change after transport. We then prove the finite-gap rejection bound and apply it to the fitted edge velocity. The proofs of the two supporting results follow the main argument.

B.1 Population Stein regression

Stein regression chooses a velocity whose induced density change is close to the change required by the annealing path. The following result separates the error caused by the chosen class from the error caused by its coefficients. At fixed t and level m, write $\boldsymbol{\psi}_m = (\mathcal{S}_t\mathbf{v}_{m,1}, \dots, \mathcal{S}_t\mathbf{v}_{m,p_m})^\top$, $\mathbf{G}_m = \mathbb{E}[\boldsymbol{\psi}_m\boldsymbol{\psi}_m^\top]$, and $\mathbf{b}_m = \mathbb{E}[\boldsymbol{\psi}_m s_t]$. The squared residual is $\mathcal{R}_m(t,\boldsymbol{\theta}) = \mathbb{E}_{\pi_t}[(s_t + \boldsymbol{\psi}_m^\top\boldsymbol{\theta})^2]$. Recall its minimum and the accumulated residual length:

$$\mathcal{E}_m(t) := \inf_{\boldsymbol{\theta}} \mathcal{R}_m(t,\boldsymbol{\theta}), \qquad \mathcal{L}_m := \int_0^1 \sqrt{\mathcal{E}_m(t)}\, dt = \sum_{k=0}^{K-1} \mathcal{L}_{k,m}. \tag{16}$$

Proposition 2 (Population Stein fit and residual decomposition)**.** *The minimum-norm population coefficient is $\boldsymbol{\theta}_m^\star = -\mathbf{G}_m^\dagger\mathbf{b}_m$, and*

$$\mathcal{E}_m(t) = \mathbb{E}[s_t^2] - \mathbf{b}_m^\top\mathbf{G}_m^\dagger\mathbf{b}_m, \tag{17}$$

$$\mathcal{R}_m(t,\boldsymbol{\theta}) = \mathcal{E}_m(t) + (\boldsymbol{\theta} - \boldsymbol{\theta}_m^\star)^\top\mathbf{G}_m(\boldsymbol{\theta} - \boldsymbol{\theta}_m^\star). \tag{18}$$

If $\mathcal{V}_m \subseteq \mathcal{V}_{m+1}$, then $\mathcal{E}_{m+1}(t) \le \mathcal{E}_m(t)$ and $\mathcal{L}_{m+1} \le \mathcal{L}_m$.

Equation (18) separates two limitations of a fitted transport. The class leaves an irreducible squared residual $\mathcal{E}_m(t)$, and a coefficient different from the population optimum adds the nonnegative quadratic term. A richer class can reduce the first term; estimating its coefficients accurately controls the second. We next relate their accumulated effect to exchange rejection. The variance identity and regression calculation are proved after the finite-gap argument below.

B.2 Bounding exchange rejection by the Stein residual

A small regression error is useful for PT only if it leads to a map with high exchange acceptance. To compare the map with the desired distribution, we express each π_t in coordinates that move with the same velocity. This removes the change explained by the transport and makes the remaining distributional mismatch easier to measure.

Set $F_t = \Phi_{0,t}$ and define the co-moving law

$$q_t = (F_t^{-1})_{\#}\pi_t. \tag{19}$$

Its density is $q_t(\mathbf{z}) = p_t(F_t(\mathbf{z}))|\det \nabla F_t(\mathbf{z})|$. Here q_t denotes both the transformed law and its density. If the velocity accounts for the entire annealing change, undoing its flow should make this law independent of t. The following identity quantifies the change that remains when the fit is imperfect.

Lemma 1 (Exact co-moving score identity)**.** *The normalized path score of q_t is*

$$\partial_t \log q_t(\mathbf{z}) = \{s_t + \mathcal{S}_t \mathbf{v}_t\}(F_t(\mathbf{z})). \tag{20}$$

Consequently,

$$\mathbb{E}_{q_t}\left[(\partial_t \log q_t)^2\right] = \mathbb{E}_{\pi_t}\left[(s_t + \mathcal{S}_t \mathbf{v}_t)^2\right]. \tag{21}$$

The above lemma gives the residual a direct probabilistic meaning: it is the path score after the transport has been undone. A zero residual makes the transformed distribution constant. For a nonzero residual, its integrated size bounds how far the transformed endpoint distributions can differ. This is the connection needed to control the Metropolis rejection.

We use the identity to prove Theorem 1; its derivative calculation is given afterward.

Proof of Theorem 1. We compare exchange in the original coordinates with exchange in the transformed distributions q_s and q_t. For $s < t$, the flow property gives $T^{\mathbf{v}}_{s\to t} = F_t \circ F_s^{-1}$. Define the pair-coordinate bijection $H_{s,t}(\mathbf{u}, \mathbf{w}) = (F_s(\mathbf{u}), F_t(\mathbf{w}))$. Then

$$\begin{aligned} G_{T^{\mathbf{v}}_{s\to t}} H_{s,t}(\mathbf{u}, \mathbf{w}) &= G_{T^{\mathbf{v}}_{s\to t}}(F_s(\mathbf{u}), F_t(\mathbf{w})) \\ &= \left((T^{\mathbf{v}}_{s\to t})^{-1} F_t(\mathbf{w}), T^{\mathbf{v}}_{s\to t} F_s(\mathbf{u})\right) \\ &= (F_s(\mathbf{w}), F_t(\mathbf{u})) = H_{s,t}(\mathbf{w}, \mathbf{u}). \end{aligned}$$

Thus transported exchange under $\pi_s \otimes \pi_t$ is conjugate to the raw swap under $q_s \otimes q_t$.

The rejection of this swap can be expressed as a distance between its stationary pair law and the law obtained by proposing a swap. The following calculation gives that identity for any proposal that is its own inverse. Let P be a finite measure and G a measurable involution such that P and $G_{\#}P$ have the same mass. The stationary rejection mass of the Metropolis proposal $\mathbf{z} \mapsto G\mathbf{z}$ is

$$r_G(P) = \|P - G_{\#}P\|_{\mathrm{TV}}. \tag{22}$$

For a probability measure this is the stationary rejection probability.

Let $\lambda = P + G_{\#}P$ and write $p = dP/d\lambda$ and $p_G = dG_{\#}P/d\lambda$. The Metropolis acceptance ratio is p_G/p on $\{p > 0\}$, so the accepted mass at state $\mathbf{z}$ is $p \min\{1, p_G/p\} = \min\{p, p_G\}$. The rejected mass is therefore

$$\int \{p - \min(p, p_G)\}\, d\lambda = \int (p - p_G)_+\, d\lambda.$$

Because P and $G_{\#}P$ have equal mass, $\int (p - p_G) d\lambda = 0$. Hence the positive and negative parts of $p - p_G$ have equal integrals, and

$$\int (p - p_G)_+ d\lambda = \frac{1}{2} \int |p - p_G| d\lambda = \|P - G_{\#}P\|_{\mathrm{TV}}.$$

Apply equation (22) after the bijection $H_{s,t}$. Total variation is invariant under bijections, so

$$r^{\mathrm{tr}}_{s,t}(T^{\mathbf{v}}_{s\to t}) = \|q_s \otimes q_t - q_t \otimes q_s\|_{\mathrm{TV}}. \tag{23}$$

The nonlinear transport map has now disappeared from the rejection expression; it enters only through the co-moving path q_t.

It remains to bound the distance between the transformed endpoints by the change along their path. Square-root densities are useful here because their L^2 derivative is proportional to the path score. We give the calculation for a general density path, then apply it to q_t. Let $\{Q_t\}$ have densities q_t such that $t \mapsto \sqrt{q_t}$ is absolutely continuous in L^2. If $u_t = \partial_t \log q_t$ and $I_Q(t) = \mathbb{E}_{Q_t}[u_t^2]$, then

$$\|Q_s - Q_t\|_{\mathrm{TV}} \le \frac{1}{2} \int_s^t \sqrt{I_Q(u)}\, du. \tag{24}$$

Where $q_t > 0$,

$$\partial_t \sqrt{q_t} = \frac{\partial_t q_t}{2\sqrt{q_t}} = \frac{1}{2} u_t \sqrt{q_t}.$$

Absolute continuity in L^2 and the fundamental theorem of calculus give

$$\begin{aligned} \|\sqrt{q_t} - \sqrt{q_s}\|_2 &\le \int_s^t \|\partial_u \sqrt{q_u}\|_2 \, du \\ &= \frac{1}{2} \int_s^t \left\{ \int u_u^2 q_u \right\}^{1/2} du = \frac{1}{2} \int_s^t \sqrt{I_Q(u)} \, du. \end{aligned}$$

Next factor $q_s - q_t = (\sqrt{q_s} - \sqrt{q_t})(\sqrt{q_s} + \sqrt{q_t})$. Cauchy–Schwarz gives

$$\begin{aligned} \|Q_s - Q_t\|_{\mathrm{TV}} &= \frac{1}{2} \int |\sqrt{q_s} - \sqrt{q_t}|(\sqrt{q_s} + \sqrt{q_t}) \\ &\le \frac{1}{2} \|\sqrt{q_s} - \sqrt{q_t}\|_2 \|\sqrt{q_s} + \sqrt{q_t}\|_2. \end{aligned}$$

Since each square-root density has L^2 norm one, the triangle inequality gives $\|\sqrt{q_s} + \sqrt{q_t}\|_2 \le 2$. Combining the displays proves the claim.

For the swap between the two endpoint distributions, this gives

$$\|Q_s \otimes Q_t - Q_t \otimes Q_s\|_{\mathrm{TV}} \le \int_s^t \sqrt{I_Q(u)} \, du. \tag{25}$$

Insert the intermediate product $Q_t \otimes Q_t$ and use the triangle inequality:

$$\begin{aligned} &\|Q_s \otimes Q_t - Q_t \otimes Q_s\|_{\mathrm{TV}} \\ &\quad \le \|Q_s \otimes Q_t - Q_t \otimes Q_t\|_{\mathrm{TV}} + \|Q_t \otimes Q_t - Q_t \otimes Q_s\|_{\mathrm{TV}}. \end{aligned}$$

For a common probability factor R, $\|P \otimes R - Q \otimes R\|_{\mathrm{TV}} = \|P - Q\|_{\mathrm{TV}}$; this follows directly by integrating the absolute density difference over the common factor. Each term above therefore equals $\|Q_s - Q_t\|_{\mathrm{TV}}$. Apply equation (24); the factor two cancels its factor one half.

Applying this bound to the transformed laws, equations (23) and (25) give

$$r_{s,t}^{\mathrm{tr}}(T_{s \to t}^{\mathbf{v}}) \le \int_s^t \left\{ \mathbb{E}_{q_u}[(\partial_u \log q_u)^2] \right\}^{1/2} du.$$

By Lemma 1, the integrand is exactly $\|s_u + \mathcal{S}_u \mathbf{v}_u\|_{L^2(\pi_u)}$. This proves the integral bound in Theorem 1; taking the minimum with one uses only the fact that rejection is a probability.

For the fitted edge map, use the constant edge coefficient $\overline{\boldsymbol{\theta}}_{k,m}$ on $[t_k, t_{k+1}]$. The general inequality gives

$$r_k^{\mathrm{tr}}(\widehat{T}_{k,m}) \le \int_{t_k}^{t_{k+1}} \sqrt{\mathcal{R}_m(t, \overline{\boldsymbol{\theta}}_{k,m})} \, dt.$$

Write $\mathcal{R}_m = \mathcal{E}_m + (\mathcal{R}_m - \mathcal{E}_m)$. Both terms are nonnegative, and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. Therefore

$$\begin{aligned} r_k^{\mathrm{tr}}(\widehat{T}_{k,m}) &\le \int_{t_k}^{t_{k+1}} \sqrt{\mathcal{E}_m(t)} \, dt \\ &\quad + \int_{t_k}^{t_{k+1}} \sqrt{\mathcal{R}_m(t, \overline{\boldsymbol{\theta}}_{k,m}) - \mathcal{E}_m(t)} \, dt \\ &= \mathcal{L}_{k,m} + \varepsilon_{k,m}^{\mathrm{fit}} = \Delta_{k,m}. \end{aligned}$$

□

We now verify the two supporting results. The first calculation identifies the density change in moving coordinates; the second gives the population regression and its residual decomposition.

Proof of Lemma 1. The density q_t changes through both p_t and the volume change of F_t. We first compute that volume change. The flow Jacobian satisfies

$$\partial_t \log|\det \nabla_{\mathbf{x}} \Phi_{s,t}(\mathbf{x})| = \nabla\cdot\mathbf{v}_t(\Phi_{s,t}(\mathbf{x})). \tag{26}$$

Let $\mathbf{J}_{s,t}(\mathbf{x}) = \nabla_{\mathbf{x}}\Phi_{s,t}(\mathbf{x})$. Differentiate the flow equation with respect to the initial state. The chain rule gives the variational equation

$$\partial_t \mathbf{J}_{s,t}(\mathbf{x}) = \nabla\mathbf{v}_t(\Phi_{s,t}(\mathbf{x}))\mathbf{J}_{s,t}(\mathbf{x}).$$

The flow is a diffeomorphism, so $\mathbf{J}_{s,t}$ is invertible. Jacobi's formula yields

$$\begin{aligned}\partial_t \log|\det \mathbf{J}_{s,t}| &= \operatorname{tr}\big(\mathbf{J}_{s,t}^{-1}\partial_t \mathbf{J}_{s,t}\big)\\ &= \operatorname{tr}\big(\mathbf{J}_{s,t}^{-1}\nabla\mathbf{v}_t(\Phi_{s,t})\mathbf{J}_{s,t}\big)\,.\end{aligned}$$

By cyclicity of the trace, the last expression is $\operatorname{tr}\nabla\mathbf{v}_t(\Phi_{s,t}) = \nabla\cdot\mathbf{v}_t(\Phi_{s,t})$, proving equation (26).

Differentiate $\log q_t(\mathbf{z})$ at a fixed co-moving coordinate $\mathbf{z}$. The density term gives, by the chain rule,

$$\begin{aligned}\partial_t \log p_t(F_t(\mathbf{z})) &= \partial_t \log p_t(\mathbf{x})|_{\mathbf{x}=F_t(\mathbf{z})} + \nabla \log p_t(F_t(\mathbf{z}))^\top \partial_t F_t(\mathbf{z})\\ &= s_t(F_t(\mathbf{z})) + \nabla \log p_t(F_t(\mathbf{z}))^\top \mathbf{v}_t(F_t(\mathbf{z})).\end{aligned}$$

Equation (26) gives

$$\partial_t \log|\det \nabla F_t(\mathbf{z})| = \nabla\cdot\mathbf{v}_t(F_t(\mathbf{z})).$$

Adding the two derivatives produces $s_t + \nabla\cdot\mathbf{v}_t + \mathbf{v}_t^\top\nabla\log p_t$, which is exactly equation (20). If $\mathbf{Z}\sim q_t$, then $F_t(\mathbf{Z})\sim\pi_t$ by definition of the pullback. Squaring equation (20) and changing variables therefore gives equation (21). □

Proof of Proposition 2. We first justify the residual used in the population objective. Suppose particles evolve according to $\dot{\mathbf{X}}_t = \mathbf{v}_t(\mathbf{X}_t)$ and have density p_t. For every smooth compactly supported test function f, the chain rule gives

$$\frac{d}{dt}\mathbb{E}[f(\mathbf{X}_t)] = \mathbb{E}[\nabla f(\mathbf{X}_t)^\top \mathbf{v}_t(\mathbf{X}_t)]. \tag{27}$$

Using the density of $\mathbf{X}_t$ and integrating by parts,

$$\begin{aligned}\mathbb{E}[\nabla f(\mathbf{X}_t)^\top \mathbf{v}_t(\mathbf{X}_t)] &= \int \nabla f(\mathbf{x})^\top \mathbf{v}_t(\mathbf{x}) p_t(\mathbf{x})\,d\mathbf{x}\\ &= -\int f(\mathbf{x})\nabla\cdot\{p_t(\mathbf{x})\mathbf{v}_t(\mathbf{x})\}\,d\mathbf{x}.\end{aligned} \tag{28}$$

On the other hand, $d\mathbb{E}[f(\mathbf{X}_t)]/dt = \int f\,\partial_t p_t$. Equality for every test function yields the continuity equation

$$\partial_t p_t + \nabla\cdot(p_t\mathbf{v}_t) = 0. \tag{29}$$

Because $p_t > 0$, divide by p_t and use $\partial_t p_t/p_t = s_t$ and $\nabla\cdot(p_t\mathbf{v}_t)/p_t = \nabla\cdot\mathbf{v}_t + \mathbf{v}_t^\top\nabla\log p_t$ to obtain

$$s_t + \mathcal{S}_t\mathbf{v}_t = 0. \tag{30}$$

Thus $-\mathcal{S}_t\mathbf{v}$ is the log-density change induced by moving particles with velocity $\mathbf{v}$. The desired path score minus this induced change is $s_t + \mathcal{S}_t\mathbf{v}$, the unexplained annealing change.

At fixed t and level m, let $\boldsymbol{\psi}_m = (\mathcal{S}_t\mathbf{v}_{m,1},\dots,\mathcal{S}_t\mathbf{v}_{m,p_m})^\top$. Since the zero-flux identity gives $\mathbb{E}\boldsymbol{\psi}_m = \mathbf{0}$ and $g_t = s_t + \mathbb{E}g_t$, centering the regression variable gives

$$g_t(\mathbf{X}) + \boldsymbol{\psi}_m(\mathbf{X})^\top\boldsymbol{\theta} - \mathbb{E}\{g_t(\mathbf{X}) + \boldsymbol{\psi}_m(\mathbf{X})^\top\boldsymbol{\theta}\} = s_t(\mathbf{X}) + \boldsymbol{\psi}_m(\mathbf{X})^\top\boldsymbol{\theta}.$$

Consequently,

$$\operatorname{Var}_{\pi_t}\big[g_t(\mathbf{X}) + \boldsymbol{\psi}_m(\mathbf{X})^\top\boldsymbol{\theta}\big] = \mathbb{E}_{\pi_t}\big[\{s_t(\mathbf{X}) + \boldsymbol{\psi}_m(\mathbf{X})^\top\boldsymbol{\theta}\}^2\big] = \mathcal{R}_m(t,\boldsymbol{\theta}). \tag{31}$$

This identity justifies the empirical objective in Algorithm 1: it estimates the normalized Stein residual without evaluating $\partial_t \log Z_t$.

Fix t, m and abbreviate $I_t = \mathbb{E}[s_t^2]$, $\mathbf{G} = \mathbb{E}[\boldsymbol{\psi}_m \boldsymbol{\psi}_m^\top]$, and $\mathbf{b} = \mathbb{E}[\boldsymbol{\psi}_m s_t]$. Expanding the square gives

$$\mathcal{R}_m(t, \boldsymbol{\theta}) = I_t + 2\boldsymbol{\theta}^\top \mathbf{b} + \boldsymbol{\theta}^\top \mathbf{G} \boldsymbol{\theta}. \tag{32}$$

We next check that redundant features do not prevent the normal equations from having a solution. Let $\mathbf{c} \in \ker(\mathbf{G})$. Since $\mathbf{G}$ is positive semidefinite,

$$0 = \mathbf{c}^\top \mathbf{G} \mathbf{c} = \mathbb{E}[(\mathbf{c}^\top \boldsymbol{\psi}_m)^2].$$

A nonnegative random variable has zero expectation only if it vanishes almost surely, so $\mathbf{c}^\top \boldsymbol{\psi}_m = 0$ almost surely. Therefore

$$\mathbf{c}^\top \mathbf{b} = \mathbb{E}[(\mathbf{c}^\top \boldsymbol{\psi}_m) s_t] = 0.$$

Thus $\mathbf{b}$ is orthogonal to $\ker(\mathbf{G})$. For a symmetric matrix, $\ker(\mathbf{G})^\perp = \mathrm{range}(\mathbf{G})$, proving the claim. Consequently, the normal equation $\mathbf{G}\boldsymbol{\theta} = -\mathbf{b}$ is solvable. Among all of its solutions, the Moore–Penrose solution $\boldsymbol{\theta}_m^\star = -\mathbf{G}^\dagger \mathbf{b}$ has minimum Euclidean norm. Write $\boldsymbol{\theta} = \boldsymbol{\theta}_m^\star + \mathbf{u}$ in equation (32). Since $\mathbf{G}\boldsymbol{\theta}_m^\star = -\mathbf{b}$,

$$\begin{aligned} \mathcal{R}_m(t, \boldsymbol{\theta}) &= I_t + 2(\boldsymbol{\theta}_m^\star + \mathbf{u})^\top \mathbf{b} + (\boldsymbol{\theta}_m^\star + \mathbf{u})^\top \mathbf{G} (\boldsymbol{\theta}_m^\star + \mathbf{u}) \\ &= I_t + 2(\boldsymbol{\theta}_m^\star)^\top \mathbf{b} + (\boldsymbol{\theta}_m^\star)^\top \mathbf{G} \boldsymbol{\theta}_m^\star + \mathbf{u}^\top \mathbf{G} \mathbf{u}. \end{aligned}$$

The terms linear in $\mathbf{u}$ cancel because $\mathbf{b} + \mathbf{G}\boldsymbol{\theta}_m^\star = 0$. Moreover, $(\boldsymbol{\theta}_m^\star)^\top \mathbf{G} \boldsymbol{\theta}_m^\star = -(\boldsymbol{\theta}_m^\star)^\top \mathbf{b} = \mathbf{b}^\top \mathbf{G}^\dagger \mathbf{b}$. Substitution gives both equation (17) and equation (18).

If $\mathcal{V}_m \subseteq \mathcal{V}_{m+1}$, every level-$m$ velocity is feasible at level $m+1$. Taking an infimum over the larger set cannot increase the risk, so $\mathcal{E}_{m+1}(t) \le \mathcal{E}_m(t)$ pointwise. Taking square roots and integrating over t gives $\mathcal{L}_{m+1} \le \mathcal{L}_m$. □

The bound applies to the exact finite map generated by the velocity, at any edge width covered by the regularity conditions. For a fitted CAT-PT candidate, its two contributions are the class residual $\mathcal{L}_{k,m}$ and the fitting and interpolation error $\varepsilon_{k,m}^{\mathrm{fit}}$. Making their sum small is sufficient for high exchange acceptance. Selection must then identify a candidate with adequate acceptance from the held-out data; that statistical step is analyzed next.

C Proof of Proposition 1: global communication

This section proves Proposition 1. Two features of selection are important: A candidate that passes the held-out rule must really meet the acceptance target, and a sufficiently good candidate must pass when the search reaches it. We state these selection properties and the DEO round-trip identity first, then combine them to prove the main proposition.

C.1 Accuracy of held-out acceptance and complexity selection

CAT-PT uses the same held-out block D_{val} throughout sequential selection. To justify this adaptive reuse, we require the validation acceptance to be accurate uniformly over all edge–candidate pairs that the algorithm could encounter. Then any candidate that passes the validation threshold has high population acceptance, while a genuinely good candidate will not be missed because of the stopping rule.

Conditional on D_{fit} and the fitting randomness, the fitting procedure defines a candidate map for every edge and complexity level, including levels that may never be reached in the actual sequential search. We refer to this finite collection as the *potential candidates*. Under P_{ind}, these candidates are fixed before D_{val} is examined, so a uniform concentration bound over the collection remains valid for whichever candidates the algorithm eventually evaluates.

For candidate (k, c), let $\ell_{k,c} = 1 - \alpha_{k,c} \in [0, 1]$ denote its rejection loss, whose population and held-out means are $1 - A_{k,c}$ and $1 - \widehat{A}_{k,c}$, respectively. The following result controls these estimation errors simultaneously and gives the consequences for sequential stopping and fallback selection.

Proposition 3 (Uniform validation and sequential selection)**.** *Suppose each of the K edges has at most J potential candidates, with $J = M + 1$ for the structured hierarchy including raw exchange. Under Assumption 3, the event*

$$\mathsf{V}(\omega) := \left\{ \max_{k,c} |\widehat{A}_{k,c} - A_{k,c}| \le \omega \right\} \tag{33}$$

has probability at least $1 - \delta$ under P_{ind} for

$$\omega = C_{\mathrm{val}} \sqrt{\frac{\log(2KJ/\delta)}{N_{\mathrm{val}}^{\mathrm{eff}}}}. \tag{34}$$

Under Q_{cal}, its probability is at least $1 - \delta - \beta_{\mathrm{mix}}(g) - \eta_{\mathrm{burn}}$. On this event, for every edge:

1. *A candidate passing equation* (5) *has $A_{k,c} \ge a_\star$.*
2. *If $A_{k,c^\star} \ge a_\star + 2\omega$, the search stops at or before $c^\star$ and the selected candidate passes equation* (5)*.*
3. *Whether the search stops early or uses the empirical-best fallback,*

$$A_{k,\widehat{c}_k} \ge \max_{c \in \mathcal{C}_k^{\mathrm{eval}}} A_{k,c} - 2\omega \ge A_{k,0} - 2\omega. \tag{35}$$

The proposition gives a guarantee for whichever candidate is selected, although the choice depends on the validation values. On $\mathsf{V}(\omega)$, an early stopping decision selects a kernel with stationary acceptance at least $a_\star$. If some level has acceptance at least $a_\star + 2\omega$, the algorithm stops there or earlier. The two ω terms allow for underestimating that candidate and for the additional margin subtracted by the selection rule. These are the selection properties used in Theorem 2.

If none of the candidates passes, the algorithm instead uses the best observed acceptance. Equation (35) shows that this fallback loses at most 2ω relative to the best evaluated candidate. The same comparison holds after early stopping and includes raw exchange. The conclusion concerns stationary acceptance on the edge; global communication is analyzed in the next subsection. It requires no monotonicity of fitted acceptance as the class becomes richer.

The probability bound also specifies the validation requirement. For a desired positive radius ω, it is enough to have $N_{\mathrm{val}}^{\mathrm{eff}} \ge C_{\mathrm{val}}^2 \log(2KJ/\delta)/\omega^2$. More candidate levels increase this requirement only through the logarithm. The effective sample size accounts for correlation within the validation block, whereas $\beta_{\mathrm{mix}}(g)$ and η_{burn} account for dependence between blocks and initialization. A large retained sample count alone does not control these latter two errors.

C.2 Invariance and round-trip rate of the selected sampler

Good acceptance on individual edges should allow states to travel between the reference and target distributions. Proposition 1 quantifies this through the round-trip rate and shows that the frozen CAT-PT kernels preserve the product target. We first characterize the round-trip rate under the DEO exchange schedule, and then combine this expression with the acceptance guarantees from the previous subsection.

A replica identity moves with its state whenever an exchange is accepted, including when the state is transformed by a transport map. Let R_B denote the total number of completed $0 \to K \to 0$ journeys of all $K + 1$ replica identities during B exchange sweeps. Under deterministic even–odd (DEO) scheduling, one sweep attempts either the even or the odd neighboring edges, with the two parities alternating. We define

$$\tau := \lim_{B \to \infty} \frac{\mathbb{E} R_B}{B}. \tag{36}$$

A full even–odd cycle contains two sweeps, so rates reported per cycle are twice equation (36).

The stationary acceptance probabilities of the individual edges do not by themselves determine τ, because successive exchange outcomes can remain correlated through slow local exploration. To isolate the effect of cross-temperature communication, we adopt the efficient-local-exploration (ELE) model used in DEO analyses (Syed et al., 2022; Zhang et al., 2026).

Definition 1 (Efficient local exploration model)**.** *For a tagged replica, conditional on its current edge and travel direction, the next exchange attempt has the stationary acceptance probability of that edge and is independent of its previous exchange history.*

Independent refreshment from each π_{t_j} between exchange sweeps is a sufficient realization of Definition 1. Under this model, the motion of a tagged replica is determined only by its current level, travel direction, and the edge acceptance probabilities.

To compute its travel time, write $a_k = 1 - r_k > 0$. A tagged replica has lifted state $(i, +)$ or $(i, -)$: at $(i, +)$ it attempts edge i upward, while at $(i, -)$ it attempts edge $i - 1$ downward. A successful exchange moves the replica by one level and preserves its direction; a rejection keeps it at the same level and reverses its direction. At the endpoints, the inactive outward step is $(0, -) \to (0, +)$ or $(K, +) \to (K, -)$ and takes one sweep.

Lemma 2 (DEO round-trip identity)**.** *Under the preceding model, define* $\mathcal{I}(\mathbf{r}) = \sum_{k=0}^{K-1} r_k/(1 - r_k)$. *Then*

$$\mathbb{E}T_{\mathrm{cycle}} = 2(K + 1)\{1 + \mathcal{I}(\mathbf{r})\}, \qquad \tau = \{2 + 2\mathcal{I}(\mathbf{r})\}^{-1}. \tag{37}$$

Here T_{cycle} *is a tagged identity's return time from* $(0, -)$ *through* $(K, +)$ *and back to* $(0, -)$.

Rejection-free DEO has aggregate rate $1/2$: each identity completes a journey every $2(K+1)$ sweeps, and there are $K + 1$ identities. Rejections increase this journey time through the sum of edgewise rejection odds $\mathcal{I}(\mathbf{r})$. A poorly accepting edge contributes a large term to this sum, so its effect is felt by journeys across the whole ladder. The sweep cost associated with a given ladder is analyzed in Appendix H.3.

Proposition 1 follows once the selected transitions are shown to preserve the product target and their stationary acceptances are bounded below. The invariance calculation applies to every frozen map. The acceptance bound additionally uses $\mathsf{V}(\omega)$, and the round-trip conclusion uses DEO and ELE.

Proof of Proposition 1. Fix the maps and indices obtained after selection. We first derive the round-trip guarantee from the two preceding results, then verify invariance for the frozen kernels.

For the communication claim, work on $\mathsf{V}(\omega)$ and suppose every selected candidate passes equation (5). Proposition 3(1) then gives $A_{k,\widehat{c}_k} \geq a_\star$ on every edge. Since $a \mapsto (1 - a)/a$ decreases for $a > 0$,

$$\frac{1 - A_{k,\widehat{c}_k}}{A_{k,\widehat{c}_k}} \leq \frac{1 - a_\star}{a_\star}.$$

Under DEO and ELE, Lemma 2 applies to these frozen edge acceptances. Summing over K edges and substituting into equation (37) yields

$$\tau_{\text{CAT-PT}} \geq \left(2 + 2K\frac{1 - a_\star}{a_\star}\right)^{-1},$$

as claimed under DEO and ELE.

It remains to verify invariance. This part does not require successful validation or the ELE assumption.

Fix an edge with endpoints $a < b$ and write $\mu = \pi_a \otimes \pi_b$ and $m(\mathbf{x}, \mathbf{y}) = p_a(\mathbf{x})p_b(\mathbf{y})$ for its density. Let T be a C^1 diffeomorphism and define $G_T(\mathbf{x}, \mathbf{y}) = (T^{-1}(\mathbf{y}), T(\mathbf{x}))$. Direct substitution gives

$$G_T(G_T(\mathbf{x}, \mathbf{y})) = (T^{-1}(T(\mathbf{x})), T(T^{-1}(\mathbf{y}))) = (\mathbf{x}, \mathbf{y}).$$

The absolute Jacobian determinant of this pair map is

$$J_{G_T}(\mathbf{x}, \mathbf{y}) = |\det \nabla T^{-1}(\mathbf{y})|\,|\det \nabla T(\mathbf{x})|. \tag{38}$$

Indeed, its derivative has the two off-diagonal blocks ∇T^{-1} and ∇T, and exchanging the two block columns changes only the sign of the determinant.

Since $G_T^{-1} = G_T$, the change-of-variables formula yields

$$\frac{d(G_T)_{\#}\mu}{d\mu}(\mathbf{z}) = \frac{m(G_T\mathbf{z})J_{G_T}(\mathbf{z})}{m(\mathbf{z})} = R_T(\mathbf{z}). \tag{39}$$

Each normalizing constant occurs once in the numerator and once in the denominator; hence it cancels, leaving exactly the unnormalized-density formula used in the algorithm. Differentiating $G_T \circ G_T = \mathrm{Id}$ also gives $J_{G_T}(G_T\mathbf{z})J_{G_T}(\mathbf{z}) = 1$, and consequently

$$R_T(G_T\mathbf{z})R_T(\mathbf{z}) = 1. \tag{40}$$

The forward and reverse ratios are reciprocal. The Metropolis rule therefore assigns equal stationary probability flow to an accepted move and its reverse, as verified below.

For this map, the pair transition is

$$K_T(\mathbf{z}, d\mathbf{z}') = \alpha_T(\mathbf{z})\delta_{G_T\mathbf{z}}(d\mathbf{z}') + \{1 - \alpha_T(\mathbf{z})\}\delta_{\mathbf{z}}(d\mathbf{z}'), \quad \alpha_T = 1 \wedge R_T. \tag{41}$$

To establish detailed balance, we compare the probability of an accepted move with that of its reverse.

Let $\lambda = \mu + (G_T)_{\#}\mu$ and write $f = d\mu/d\lambda$ and $f_G = d(G_T)_{\#}\mu/d\lambda$. The measure λ is invariant under G_T because applying G_T swaps its two summands. Thus $f_G = f \circ G_T$ and $f = f_G \circ G_T$, up to null sets. On the support of μ, $\alpha_T = \min\{1, f_G/f\}$, so

$$f\alpha_T = \min\{f, f_G\}.$$

For any bounded measurable function H on the pair of old and new states,

$$\begin{aligned}\int \mu(d\mathbf{z})\alpha_T(\mathbf{z})H(\mathbf{z}, G_T\mathbf{z}) &= \int \min\{f, f_G\}(\mathbf{z})H(\mathbf{z}, G_T\mathbf{z})\lambda(d\mathbf{z}) \\ &= \int \min\{f, f_G\}(\mathbf{z})H(G_T\mathbf{z}, \mathbf{z})\lambda(d\mathbf{z}).\end{aligned}$$

The second equality changes variables by G_T and uses the invariance of both λ and $\min\{f, f_G\}$. It proves symmetry of the accepted flow. The remaining transition stays at $\mathbf{z}$, so its joint measure is supported on the diagonal and is also symmetric. This proves detailed balance.

The rejection identity equation (22) applies to the same involution and identifies its stationary rejection. Finally, for $T = \mathrm{Id}$ the two Jacobians are one and $G_T(\mathbf{x}, \mathbf{y}) = (\mathbf{y}, \mathbf{x})$, so its ratio and proposal are precisely those of raw PT. This also identifies its stationary rejection:

$$r_T = \|\mu - (G_T)_{\#}\mu\|_{\mathrm{TV}}, \qquad A_T = 1 - r_T. \tag{42}$$

If $T_{\#}\pi_a = \pi_b$, then $(T^{-1})_{\#}\pi_b = \pi_a$ and the transformed pair again has law $\pi_a \otimes \pi_b$. The distance in equation (42) is then zero, so exact transport has acceptance one. For an approximate map, the same correction still preserves the pair law, with the mismatch appearing in rejection.

We next lift the pair update to the full ensemble $\Pi = \bigotimes_{j=0}^{K} \pi_{t_j}$. A local update at node j preserves π_{t_j} and leaves the other coordinates unchanged. Integrating first over that coordinate shows that it preserves Π. The pair argument above likewise shows that an exchange on edge k preserves $\pi_{t_k} \otimes \pi_{t_{k+1}}$; integrating over those two coordinates establishes invariance of its ensemble transition.

If P and Q both preserve Π, then for every measurable set B,

$$\int \Pi(d\mathbf{x})(PQ)(\mathbf{x}, B) = \int (\Pi P)(d\mathbf{y})Q(\mathbf{y}, B) = (\Pi Q)(B) = \Pi(B).$$

Repeated composition therefore preserves Π, including the alternating even–odd schedule. A state-independent mixture preserves it as well, since $\Pi \sum_i w_i P_i = \sum_i w_i \Pi = \Pi$ whenever $w_i \geq 0$ and $\sum_i w_i = 1$. This proves invariance for every realized frozen choice. □

Once the selected maps are fixed, every production transition preserves the product target $\Pi = \bigotimes_{k=0}^{K} \pi_{t_k}$. Calibration therefore affects the acceptance guarantees, but not the invariant distribution of the frozen sampler.

When some edges use the fallback rather than reaching the common target $a_\star$, the individual validation estimates still give a round-trip bound. Let $\underline{A}_k = \max\{0, \widehat{A}_{k,\widehat{c}_k} - \omega\}$. On $\mathsf{V}(\omega)$, $A_{k,\widehat{c}_k} \geq \underline{A}_k$, and hence

$$\tau_{\text{CAT-PT}} \geq \left(2 + 2\sum_{k=0}^{K-1} \frac{1 - \underline{A}_k}{\underline{A}_k}\right)^{-1}, \tag{43}$$

with the right-hand side defined as zero if any $\underline{A}_k = 0$. Thus the selected sampler remains quantitatively controlled even when a common acceptance target is not reached on every edge.

We finish by proving the uniform validation statement and the DEO identity used above. The former controls the data-dependent choice of kernel; the latter computes the journey time for its fixed edge acceptances.

Proof of Proposition 3. We first control validation under P_{ind}, conditional on the fitting block and its randomness. Every loss in the library is then a fixed function of a validation pair. For one edge–candidate pair, apply equation (15) with failure probability $\delta/(KJ)$. Since $\|\ell_{k,c}\|_\infty \le 1$, its error radius is at most

$$C_{\mathrm{val}}\sqrt{\frac{\log(2KJ/\delta)}{N_{\mathrm{val}}^{\mathrm{eff}}}} = \omega.$$

The estimation error of acceptance is the negative of the loss estimation error. A union bound over at most KJ pairs therefore gives

$$P_{\mathrm{ind}}\left(\max_{k,c}|\widehat{A}_{k,c} - A_{k,c}| > \omega \,\middle|\, D_{\mathrm{fit}},\ \text{fitting randomness}\right) \le KJ\frac{\delta}{KJ} = \delta.$$

Integrating over the fitting block and randomness gives the same bound without conditioning. Dependence between different candidate estimates is allowed: the union bound requires only their individual error bounds.

To pass from this idealized law to the actual calibration trajectory, apply the total-variation bounds of Assumption 3. For any event E determined by the two retained blocks,

$$\begin{aligned} Q_{\mathrm{cal}}(\mathsf{E}) &\ge P_{\mathrm{stat}}^{(g)}(\mathsf{E}) - \|Q_{\mathrm{cal}} - P_{\mathrm{stat}}^{(g)}\|_{\mathrm{TV}} \\ &\ge P_{\mathrm{ind}}(\mathsf{E}) - \|P_{\mathrm{stat}}^{(g)} - P_{\mathrm{ind}}\|_{\mathrm{TV}} - \eta_{\mathrm{burn}}. \end{aligned}$$

Consequently,

$$Q_{\mathrm{cal}}(\mathsf{E}) \ge P_{\mathrm{ind}}(\mathsf{E}) - \beta_{\mathrm{mix}}(g) - \eta_{\mathrm{burn}}. \tag{44}$$

Taking $\mathsf{E} = \mathsf{V}(\omega)$ proves the stated probability. When the fit uses auxiliary randomness independent of the trajectory, the same argument applies after adjoining that randomness to each law.

It remains to examine what the algorithm does on $\mathsf{V}(\omega)$. The two inequalities

$$A_{k,c} \ge \widehat{A}_{k,c} - \omega, \qquad \widehat{A}_{k,c} \ge A_{k,c} - \omega$$

hold simultaneously over the library. The first shows that any candidate passing the selection rule has $A_{k,c} \ge \widehat{A}_{k,c} - \omega \ge a_\star$, proving item 1. For a candidate with $A_{k,c^\star} \ge a_\star + 2\omega$, the second gives

$$\widehat{A}_{k,c^\star} - \omega \ge A_{k,c^\star} - 2\omega \ge a_\star.$$

Thus the candidate passes when it is evaluated. Either the search has already stopped at a simpler candidate, or it stops at $c^\star$. In both cases the chosen candidate passes the rule, proving item 2.

For item 3, first suppose no candidate passes. Let $c^\star$ maximize population acceptance over $\mathcal{C}_k^{\mathrm{eval}}$, and let $\widehat{c}_k$ be the empirical-best fallback. Then

$$A_{k,\widehat{c}_k} \ge \widehat{A}_{k,\widehat{c}_k} - \omega \ge \widehat{A}_{k,c^\star} - \omega \ge A_{k,c^\star} - 2\omega.$$

The middle inequality is the defining property of the fallback. Now suppose the search stops early at $\widehat{c}_k$. Each earlier evaluated candidate $c < \widehat{c}_k$ failed the rule, so

$$\widehat{A}_{k,c} - \omega < a_\star, \qquad A_{k,c} \le \widehat{A}_{k,c} + \omega < a_\star + 2\omega.$$

By item 1, $A_{k,\widehat{c}_k} \ge a_\star$. It is therefore within 2ω of every earlier candidate, and hence of the best candidate in the evaluated prefix. Raw exchange belongs to this prefix, giving the last inequality in equation (35). Resolving empirical ties toward smaller levels leaves the argument unchanged. □

Proof of Lemma 2. Let u_i^+ and u_i^- be the expected times to first hit $(K,+)$, starting from $(i,+)$ and $(i,-)$, respectively. Set $u_K^+ = 0$. For $0 \le i < K$, first-step decomposition gives

$$u_i^+ = 1 + a_i u_{i+1}^+ + r_i u_i^-.$$

For $1 \le i \le K$, the downward counterpart is $u_i^- = 1 + a_{i-1} u_{i-1}^- + r_{i-1} u_i^+$, and the reflecting lower boundary has $u_0^- = 1 + u_0^+$. Put $d_i = u_i^- - u_i^+$. Rearranging the upward equation,

$$u_i^+ - u_{i+1}^+ = \frac{1 + r_i d_i}{a_i}.$$

Insert this expression into the downward equation at $i+1$:

$$\begin{aligned} d_{i+1} &= 1 + a_i(u_i^- - u_{i+1}^+) \\ &= 1 + a_i \left(d_i + \frac{1 + r_i d_i}{a_i} \right) = 2 + d_i. \end{aligned}$$

Since $d_0 = 1$, induction yields $d_i = 2i + 1$. Therefore

$$u_i^+ - u_{i+1}^+ = \frac{1 + (2i+1) r_i}{a_i} = 1 + 2(i+1)\frac{r_i}{a_i}.$$

Summing from $i = 0$ to $K - 1$ and including the outward boundary step gives

$$u_0^- = K + 1 + 2 \sum_{i=0}^{K-1} (i+1)\frac{r_i}{a_i}.$$

For the return journey, reflect the ladder by sending level i to $K - i$ and reversing the travel direction. The reflected edge j has acceptance a_{K-1-j}. Applying the preceding upward calculation to this reflected ladder gives the expected time from $(K,+)$ to $(0,-)$ as

$$K + 1 + 2 \sum_{j=0}^{K-1} (j+1)\frac{r_{K-1-j}}{a_{K-1-j}} = K + 1 + 2 \sum_{i=0}^{K-1} (K-i)\frac{r_i}{a_i}.$$

The strong Markov property at $(K,+)$ allows the outward and return times to be added. Their sum is

$$2(K+1) + 2 \sum_{i=0}^{K-1} (K+1)\frac{r_i}{a_i} = 2(K+1)\{1 + \mathcal{I}(\mathbf{r})\}.$$

For each identity the completed-cycle process is regenerative, with long-run rate $1/\mathbb{E}T_{\mathrm{cycle}}$. Summing over the $K+1$ identities gives the second formula. The identities need not be mutually independent for this sum of rates. Initial incomplete journeys contribute only a bounded counting error and disappear after division by B. □

D STRUCTURED PERFORMANCE OF CAT-PT AND COMPARISON WITH RAW PT

This section develops the structured-performance result for CAT-PT and the raw-PT comparison stated after Theorem 2. First, we give sufficient finite-sample conditions for the fitting-error premise used by the theorem. We then prove the finite-ladder CAT-PT guarantee itself. Finally, we analyze raw PT when the temperature ladder is made arbitrarily fine and derive the high-dimensional benchmark used for comparison.

D.1 FINITE-SAMPLE CONTROL OF THE STEIN FITTING ERROR

The fitting premise of Theorem 2 is the uniform bound

$$\varepsilon_{k,m_d}^{\mathrm{fit}} \le C_1 h_k (1 + \sqrt{r_d}).$$

We give one set of sufficient conditions for establishing this premise. Assumption 4 supplies nodewise concentration; control of population coefficient variation then yields an edgewise bound. These

conditions are not hypotheses of Theorem 2 itself, which applies whenever its fitting-error bound holds.

There are two sources of fitting error. The nodewise coefficients are estimated from finitely many observations in D_{fit}, and the average of the two endpoint estimates is held constant across an edge even when the population optimum varies with temperature. We first control estimation at the nodes and then convert endpoint errors and coefficient variation into the integrated residual used by the theorem.

We begin at a single node and level. Abbreviate the feature covariance and cross-covariance as $\mathbf{G}$ and $\mathbf{b}$. The population coefficient is $\boldsymbol{\theta}^\star = -\mathbf{G}^{-1}\mathbf{b}$ when $\mathbf{G} \succeq \kappa \mathbf{I}_p$ for $\kappa > 0$.

Coefficient errors affect the unexplained density change through the quadratic form defined by $\mathbf{G}$. The following bounded-feature regime permits stable moment estimation and an explicit coefficient-error bound. Its constants describe both the size of an observation and the conditioning of the regression.

Assumption 4 (A sufficient bounded-feature fitting regime)**.** *At a fixed node and level, the fitting observations are iid from π_t, $\|\boldsymbol{\psi}_m(\mathbf{X})\|_2 \le B$, $|g_t(\mathbf{X}) - \mathbb{E}g_t(\mathbf{X})| \le S$, $\mathbf{G}_m \succeq \kappa \mathbf{I}_p$, and $\|\boldsymbol{\theta}_m^\star\|_2 \le \Theta$.*

The bounds B and S limit the contribution of one sample to the feature covariance and cross-covariance. The eigenvalue condition says that distinct coefficient directions produce distinguishable density changes; when κ is small, moment-estimation errors can cause a large coefficient error. The bound Θ controls this amplification at the population solution. These conditions apply, for example, when the features and path score are bounded on the support and the feature covariance is nonsingular.

The proposition retains all of these constants, so its sample requirement depends on the target and class as well as the parameter count p. For unbounded features, the same coefficient-error calculation can be combined with appropriate moment-concentration bounds. The scalar Gaussian regression in Appendix F.3 supplies a worked example with unbounded observations.

Proposition 4 (Coefficient and residual error of Stein regression)**.** *Under Assumption 4, there is a universal constant C such that, for every $\delta \in (0,1)$, if*

$$N_{\mathrm{fit}} \ge C \frac{B^4}{\kappa^2}\{p + \log(1/\delta)\}, \tag{45}$$

then with probability at least $1-\delta$,

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \le \frac{C}{\kappa}(BS + B^2\Theta)\sqrt{\frac{p + \log(1/\delta)}{N_{\mathrm{fit}}}}, \tag{46}$$

$$\mathcal{R}(t, \widehat{\boldsymbol{\theta}}) - \mathcal{E}_m(t) \le \frac{C\lambda_{\max}(\mathbf{G})}{\kappa^2}(BS + B^2\Theta)^2 \frac{p + \log(1/\delta)}{N_{\mathrm{fit}}}. \tag{47}$$

Proof. We first relate moment errors to coefficient error. This algebraic step explains which empirical quantities must be controlled. Work first on $\|\widehat{\mathbf{G}} - \mathbf{G}\|_{\mathrm{op}} \le \kappa/2$; we will obtain this event from the sample-size condition below. The population and empirical normal equations are $\mathbf{G}\boldsymbol{\theta}^\star = -\mathbf{b}$ and $\widehat{\mathbf{G}}\widehat{\boldsymbol{\theta}} = -\widehat{\mathbf{b}}$. Add and subtract $\widehat{\mathbf{G}}\boldsymbol{\theta}^\star$ to obtain

$$\begin{aligned}\widehat{\mathbf{G}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) &= -\widehat{\mathbf{b}} - \widehat{\mathbf{G}}\boldsymbol{\theta}^\star \\ &= -(\widehat{\mathbf{b}} - \mathbf{b}) - (\widehat{\mathbf{G}} - \mathbf{G})\boldsymbol{\theta}^\star.\end{aligned}$$

Weyl's eigenvalue inequality gives

$$\lambda_{\min}(\widehat{\mathbf{G}}) \ge \lambda_{\min}(\mathbf{G}) - \|\widehat{\mathbf{G}} - \mathbf{G}\|_{\mathrm{op}} \ge \kappa/2.$$

Hence $\widehat{\mathbf{G}}$ is invertible and $\|\widehat{\mathbf{G}}^{-1}\|_{\mathrm{op}} \le 2/\kappa$. Multiply the normal-equation difference by $\widehat{\mathbf{G}}^{-1}$ and apply the triangle and operator-norm inequalities to obtain

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \le \frac{2}{\kappa}\left(\|\widehat{\mathbf{b}} - \mathbf{b}\|_2 + \|\widehat{\mathbf{G}} - \mathbf{G}\|_{\mathrm{op}}\|\boldsymbol{\theta}^\star\|_2\right). \tag{48}$$

We now show that the fitting block makes both terms on the right small. Write $N = N_{\mathrm{fit}}$, $\overline{\boldsymbol{\psi}} = N^{-1}\sum_i \boldsymbol{\psi}_i$, and $\overline{g} = N^{-1}\sum_i g_i$. Since the population Stein features are centered, the empirical

covariance and cross-covariance can be written as

$$\widehat{\mathbf{G}} = N^{-1}\sum_i \boldsymbol{\psi}_i\boldsymbol{\psi}_i^\top - \overline{\boldsymbol{\psi}}\,\overline{\boldsymbol{\psi}}^\top,$$

$$\widehat{\mathbf{b}} = N^{-1}\sum_i (\boldsymbol{\psi}_i - \overline{\boldsymbol{\psi}})(g_i - \overline{g}).$$

For the Gram matrix, decompose

$$\widehat{\mathbf{G}} - \mathbf{G} = \left\{ N^{-1}\sum_i (\boldsymbol{\psi}_i\boldsymbol{\psi}_i^\top - \mathbb{E}\boldsymbol{\psi}\boldsymbol{\psi}^\top) \right\} - \overline{\boldsymbol{\psi}}\,\overline{\boldsymbol{\psi}}^\top.$$

We bound these fluctuations uniformly over feature directions by a finite covering of the unit sphere. Let $\mathcal{N}$ be a $1/4$-net of the unit sphere in $\mathbb{R}^p$, with $|\mathcal{N}| \le 9^p$. For any symmetric matrix $\mathbf{H}$, $\|\mathbf{H}\|_{\mathrm{op}} \le 2\max_{\mathbf{u}\in\mathcal{N}} |\mathbf{u}^\top\mathbf{H}\mathbf{u}|$; for any vector $\mathbf{z}$, $\|\mathbf{z}\|_2 \le 2\max_{\mathbf{u}\in\mathcal{N}} |\mathbf{u}^\top\mathbf{z}|$. To see the first inequality, approximate a unit vector maximizing the quadratic form by a net point. The difference of the two quadratic forms is at most $2(1/4)\|\mathbf{H}\|_{\mathrm{op}}$ and can be moved to the left; the vector bound follows by the same approximation.

For fixed $\mathbf{u} \in \mathcal{N}$, $(\mathbf{u}^\top\boldsymbol{\psi}_i)^2$ lies in $[0, B^2]$. Hoeffding's inequality and a union bound over the net therefore bound the uncentered Gram fluctuation by $CB^2\sqrt{H/N}$, where $H = p + \log(8/\delta)$. Applying the same argument to the bounded scalar $\mathbf{u}^\top\boldsymbol{\psi}_i \in [-B, B]$ gives $\|\overline{\boldsymbol{\psi}}\|_2 \le CB\sqrt{H/N}$. The centering correction has operator norm $\|\overline{\boldsymbol{\psi}}\|_2^2$, and hence

$$\|\widehat{\mathbf{G}} - \mathbf{G}\|_{\mathrm{op}} \le CB^2\left(\sqrt{H/N} + H/N\right). \tag{49}$$

Put $\zeta_i = g_i - \mathbb{E}g_i$, so $|\zeta_i| \le S$ and $\mathbf{b} = \mathbb{E}[\boldsymbol{\psi}\zeta]$. Expanding the empirical centering gives

$$\widehat{\mathbf{b}} - \mathbf{b} = \left\{ N^{-1}\sum_i \boldsymbol{\psi}_i\zeta_i - \mathbf{b} \right\} - \overline{\boldsymbol{\psi}}\,\overline{\zeta}.$$

For each net direction, the scalar $(\mathbf{u}^\top\boldsymbol{\psi}_i)\zeta_i$ is bounded in absolute value by BS. Hoeffding and the net bound control the first term by $CBS\sqrt{H/N}$. A scalar Hoeffding bound gives $|\overline{\zeta}| \le CS\sqrt{H/N}$. Multiplying this with the feature-mean bound controls the centering term by $CBSH/N$, and thus

$$\|\widehat{\mathbf{b}} - \mathbf{b}\|_2 \le CBS\left(\sqrt{H/N} + H/N\right). \tag{50}$$

Allocating failure probability to these finitely many bounds yields joint probability at least $1 - \delta$. Matrix concentration can sharpen the explicit dimension dependence when additional variance information is available (Tropp, 2012).

Since $\kappa \le \lambda_{\max}(\mathbf{G}) \le B^2$, the sample requirement equation (45) also ensures $N \ge CH$ after adjusting the universal constant. Thus the H/N terms are bounded by constant multiples of the square-root terms, and equation (49) is at most $\kappa/2$ after enlarging C. Equation (48), together with $\|\boldsymbol{\theta}^\star\|_2 \le \Theta$, then gives equation (46). Finally, the projection identity equation (18) gives

$$\mathcal{R}(t, \widehat{\boldsymbol{\theta}}) - \mathcal{R}(t, \boldsymbol{\theta}^\star) = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top\mathbf{G}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star). \tag{51}$$

Using $\mathbf{u}^\top\mathbf{G}\mathbf{u} \le \lambda_{\max}(\mathbf{G})\|\mathbf{u}\|_2^2$ and squaring the coefficient bound now gives equation (47). □

The deterministic perturbation bound equation (48) does not require independent observations. For a correlated fitting block, the same proof applies whenever the matrix, vector, and sample-mean concentration bounds above hold with N_{fit} replaced by a valid effective sample size $N_{\mathrm{fit}}^{\mathrm{eff}}$. This is a separate concentration requirement for fitting moments; concentration of the bounded acceptance losses used in validation does not by itself establish it.

Suppose these nodewise bounds hold under the independent stationary-block law P_{ind}, and the sample size meets the corresponding Gram-stability requirement at every node and level under consideration. Allocate failure probability $\delta/\{(K+1)M\}$ to each node–level pair. A union bound gives, with probability at least $1 - \delta$ under that law,

$$\max_{j,m} \|\widehat{\boldsymbol{\theta}}_{j,m} - \boldsymbol{\theta}_m^\star(t_j)\|_2 \le C_{\mathrm{fit}}\sqrt{\frac{p_{\max} + \log\{(K+1)M/\delta\}}{N_{\mathrm{fit}}^{\mathrm{eff}}}}. \tag{52}$$

A level-specific statement replaces $p_{\max}$ by p_m and uses the corresponding constant $C_{\mathrm{fit},m}$. Dependence between estimates at different nodes or levels does not affect this union bound. The transfer from P_{ind} to the actual calibration law is accounted for in Appendix D.2.

The constants retain the feature-moment and conditioning dependence in Proposition 4. Consequently, a small parameter count does not alone imply a dimension-independent fitting requirement. In a high-dimensional application, these constants and the edge-interpolation error below must be checked when verifying the stated bound on $\varepsilon^{\mathrm{fit}}_{k,m_d}$.

Accurate nodewise fits do not by themselves control a map used throughout an edge: even exact endpoint coefficients may differ from the optimum between them. The next result bounds this additional error when the population coefficient is Lipschitz in temperature. Its proof also gives sufficient moment conditions for that regularity. The endpoint coefficient is the average of the two nodewise fits, as in the algorithm.

Proposition 5 (Fitting error across an annealing edge)**.** *Suppose $\boldsymbol{\theta}^\star_m$ is $L_{\theta,m}$-Lipschitz, $\mathbf{G}_m(t) \preceq \overline{\kappa}_m \mathbf{I}$ on edge k, and both endpoint coefficient estimates have error at most e_m. Then*

$$\varepsilon^{\mathrm{fit}}_{k,m} \le h_k \sqrt{\overline{\kappa}_m} \left(e_m + \frac{1}{2} L_{\theta,m} h_k \right). \tag{53}$$

When the simultaneous fitting conditions for equation (52) *hold, combining the two bounds yields*

$$\varepsilon^{\mathrm{fit}}_{k,m} = O_{\mathbb{P}}\left(h_k \sqrt{\frac{p_m + \log(KM/\delta)}{N^{\mathrm{eff}}_{\mathrm{fit}}}} + h_k^2 \right) \tag{54}$$

up to level-dependent conditioning constants.

Proof. Let $\boldsymbol{\theta}^\star_k = \boldsymbol{\theta}^\star_m(t_k)$ and $\boldsymbol{\theta}^\star_{k+1} = \boldsymbol{\theta}^\star_m(t_{k+1})$. Add and subtract their average:

$$\begin{aligned} &\|\overline{\boldsymbol{\theta}}_{k,m} - \boldsymbol{\theta}^\star_m(t)\|_2 \\ &\quad \le \frac{1}{2}\|\widehat{\boldsymbol{\theta}}_{k,m} - \boldsymbol{\theta}^\star_k\|_2 + \frac{1}{2}\|\widehat{\boldsymbol{\theta}}_{k+1,m} - \boldsymbol{\theta}^\star_{k+1}\|_2 \\ &\qquad + \left\| \frac{1}{2}(\boldsymbol{\theta}^\star_k + \boldsymbol{\theta}^\star_{k+1}) - \boldsymbol{\theta}^\star_m(t) \right\|_2 . \end{aligned}$$

The first two terms sum to at most e_m. By convexity of the norm and the Lipschitz property, the last term is at most

$$\frac{L_{\theta,m}}{2}\{|t - t_k| + |t_{k+1} - t|\} = \frac{1}{2} L_{\theta,m} h_k.$$

Consequently, the edge coefficient satisfies

$$\|\overline{\boldsymbol{\theta}}_{k,m} - \boldsymbol{\theta}^\star_m(t)\|_2 \le e_m + \frac{1}{2} L_{\theta,m} h_k. \tag{55}$$

To translate this coefficient bound into residual length, use the geometry of the population regression. The Pythagorean identity equation (18) gives

$$\mathcal{R}_m(t, \overline{\boldsymbol{\theta}}_{k,m}) - \mathcal{E}_m(t) = (\overline{\boldsymbol{\theta}}_{k,m} - \boldsymbol{\theta}^\star_m(t))^\top \mathbf{G}_m(t) (\overline{\boldsymbol{\theta}}_{k,m} - \boldsymbol{\theta}^\star_m(t)).$$

The Gram upper bound implies that the square root of this quantity is at most

$$\sqrt{\overline{\kappa}_m}\, \|\overline{\boldsymbol{\theta}}_{k,m} - \boldsymbol{\theta}^\star_m(t)\|_2.$$

Apply equation (55); the resulting upper bound is constant over the edge. Integrating over an interval of length h_k gives equation (53). Substituting the level-specific version of equation (52) for e_m gives the first term in equation (54); the interpolation term is $h_k(L_{\theta,m} h_k/2) = O(h_k^2)$.

To check the Lipschitz premise from population moments, the following additional conditions are sufficient. Suppose $\mathbf{G}_m(t) \succeq \kappa \mathbf{I}$, $\|\mathbf{G}_m(t) - \mathbf{G}_m(s)\|_{\mathrm{op}} \le L_G |t - s|$, $\|\mathbf{b}_m(t) - \mathbf{b}_m(s)\|_2 \le L_b |t - s|$, and $\sup_t \|\boldsymbol{\theta}^\star_m(t)\|_2 \le \Theta$. These bounds on the moments imply a bound on the coefficients. Their two population normal equations are

$$\mathbf{G}_m(t)\boldsymbol{\theta}^\star_m(t) = -\mathbf{b}_m(t), \qquad \mathbf{G}_m(s)\boldsymbol{\theta}^\star_m(s) = -\mathbf{b}_m(s).$$

Add and subtract $\mathbf{G}_m(t)\boldsymbol{\theta}_m^\star(s)$ to obtain

$$\mathbf{G}_m(t)\{\boldsymbol{\theta}_m^\star(t) - \boldsymbol{\theta}_m^\star(s)\} = -\{\mathbf{b}_m(t) - \mathbf{b}_m(s)\} - \{\mathbf{G}_m(t) - \mathbf{G}_m(s)\}\boldsymbol{\theta}_m^\star(s).$$

Because $\mathbf{G}_m(t) \succeq \kappa\mathbf{I}$, $\|\mathbf{G}_m(t)^{-1}\|_{\mathrm{op}} \le 1/\kappa$. Multiplying by the inverse and applying the assumed Lipschitz and norm bounds yields

$$\|\boldsymbol{\theta}_m^\star(t) - \boldsymbol{\theta}_m^\star(s)\|_2 \le \frac{L_b|t-s| + L_G|t-s|\Theta}{\kappa},$$

or, equivalently,

$$\|\boldsymbol{\theta}_m^\star(t) - \boldsymbol{\theta}_m^\star(s)\|_2 \le \frac{L_b + L_G\Theta}{\kappa}|t-s|. \tag{56}$$

The Lipschitz hypotheses can be checked from moments along the path. For a smooth time-dependent integrand f_t,

$$\frac{d}{dt}\mathbb{E}_{\pi_t} f_t = \mathbb{E}_{\pi_t}[\partial_t f_t + s_t f_t],$$

which follows by differentiating $f_t p_t$ and using $\partial_t p_t = s_t p_t$. With $\mathbf{G}_m = \mathbb{E}[\boldsymbol{\psi}_m\boldsymbol{\psi}_m^\top]$ and $\mathbf{b}_m = \mathbb{E}[\boldsymbol{\psi}_m s_t]$, this yields

$$\dot{\mathbf{G}}_m = \mathbb{E}[\dot{\boldsymbol{\psi}}_m\boldsymbol{\psi}_m^\top + \boldsymbol{\psi}_m\dot{\boldsymbol{\psi}}_m^\top + s_t\boldsymbol{\psi}_m\boldsymbol{\psi}_m^\top],$$
$$\dot{\mathbf{b}}_m = \mathbb{E}[\dot{\boldsymbol{\psi}}_m s_t + \boldsymbol{\psi}_m\dot{s}_t + \boldsymbol{\psi}_m s_t^2].$$

Uniform bounds on the norms of these expectations imply the constants L_G, L_b in equation (56). Thus smooth target parameters and suitable moment envelopes, together with a nonsingular Gram matrix, give a quantitative bound on how rapidly the population fit changes with t. □

Equation (53) separates the two errors identified at the start of this subsection. The term proportional to $h_k e_m$ comes from estimating the endpoint velocities and decreases with fitting sample size. The term proportional to h_k^2 comes from using a constant edge coefficient and decreases with ladder spacing. The explicit constants show what must remain controlled for fitting to preserve the residual-dimensional performance in Theorem 2.

When the empirical features are nearly collinear, ridge regularization improves numerical conditioning. Its population bias can be included in the same edge calculation. For $\lambda > 0$, define $\boldsymbol{\theta}_\lambda = -(\mathbf{G} + \lambda\mathbf{I})^{-1}\mathbf{b}$. Since $\mathbf{b} = -\mathbf{G}\boldsymbol{\theta}^\star$,

$$\boldsymbol{\theta}_\lambda = (\mathbf{G} + \lambda\mathbf{I})^{-1}\mathbf{G}\boldsymbol{\theta}^\star.$$

Therefore

$$\boldsymbol{\theta}_\lambda - \boldsymbol{\theta}^\star = -\lambda(\mathbf{G} + \lambda\mathbf{I})^{-1}\boldsymbol{\theta}^\star.$$

The smallest eigenvalue of $\mathbf{G} + \lambda\mathbf{I}$ is at least $\kappa + \lambda$, so

$$\|\boldsymbol{\theta}_\lambda - \boldsymbol{\theta}^\star\|_2 \le \frac{\lambda}{\kappa + \lambda}\|\boldsymbol{\theta}^\star\|_2. \tag{57}$$

The resulting excess risk is at most $\lambda_{\max}(\mathbf{G})\|\boldsymbol{\theta}_\lambda - \boldsymbol{\theta}^\star\|_2^2$ by equation (18) and $\mathbf{u}^\top\mathbf{G}\mathbf{u} \le \lambda_{\max}(\mathbf{G})\|\mathbf{u}\|_2^2$. For an empirical ridge fit, its estimation error relative to $\boldsymbol{\theta}_\lambda$ must also be controlled. Adding that error to the bias in equation (57) gives an endpoint-error bound relative to $\boldsymbol{\theta}^\star$, which can be used as e_m in equation (53). The finite-gap argument still applies to the exact flow of the fitted regularized coefficient.

D.2 Proof of Theorem 2

We prove that a ladder at the residual scale makes CAT-PT stop by level m_d and attain the finite-ladder rate in Theorem 2. We use the structured-residual condition $\mathcal{E}_{m_d}(t) \le C_0 r_d$ specified before the theorem, together with its fitting-error premise. The proof uses the finite-gap rejection bound, without a local likelihood expansion. It suffices to show that the level-m_d candidate has stationary acceptance at least $a_\star + 2\omega$: Proposition 3 then guarantees, on $\mathsf{V}(\omega)$, that it passes the held-out rule when reached or that a simpler sufficient candidate has already been selected.

Proof of Theorem 2. First consider an arbitrary edge and class. Suppose a level m satisfies $\mathcal{E}_m(t) \le C_0 r$ on edge k and $\varepsilon^{\mathrm{fit}}_{k,m} \le e_{k,m}$. Then

$$\mathcal{L}_{k,m} = \int_{t_k}^{t_{k+1}} \sqrt{\mathcal{E}_m(t)}\, dt \le h_k \sqrt{C_0 r}.$$

Adding the fitting error and applying Theorem 1 gives

$$A_{k,m} \ge 1 - \Delta_{k,m}, \qquad \Delta_{k,m} \le h_k \sqrt{C_0 r} + e_{k,m}. \tag{58}$$

Thus $h_k\sqrt{C_0 r} + e_{k,m} \le 1 - a_\star - 2\omega$ yields $A_{k,m} \ge a_\star + 2\omega$. Proposition 3(2) then guarantees that the search stops no later than m on $\mathsf{V}(\omega)$.

Write $b_0 = 1 - a_\star - 2\omega > 0$ and choose

$$C \ge \max\left\{1, \frac{\sqrt{C_0} + C_1}{b_0}\right\}, \qquad K_d = \lceil C(1 + \sqrt{r_d})\rceil. \tag{59}$$

The constant C is independent of dimension because C_0, C_1, b_0 are. On the uniform grid, $h_k = 1/K_d$. The residual and fitting assumptions give

$$\begin{aligned}\Delta_{k,m_d} &\le h_k\sqrt{C_0 r_d} + C_1 h_k (1 + \sqrt{r_d}) \\ &\le \frac{(\sqrt{C_0} + C_1)(1 + \sqrt{r_d})}{K_d} \le b_0 = 1 - a_\star - 2\omega.\end{aligned}$$

Equation (58) therefore implies $A_{k,m_d} \ge a_\star + 2\omega$ on every edge.

We now account for the data-dependent choice of kernel. On $\mathsf{V}(\omega)$,

$$\widehat{A}_{k,m_d} - \omega \ge A_{k,m_d} - 2\omega \ge a_\star.$$

If the search reaches level m_d, it passes the selection rule there. Otherwise it has already stopped at an earlier passing candidate. Hence $\widehat{c}_k \le m_d$ on every edge, and no edge reaches the empirical-best fallback. For whichever level was selected, the same uniform event gives

$$A_{k,\widehat{c}_k} \ge \widehat{A}_{k,\widehat{c}_k} - \omega \ge a_\star.$$

These are the two selection conclusions in Proposition 3: the useful level bounds the search depth, and uniform validation controls the population acceptance of an earlier choice.

Applying Proposition 1 now yields

$$\tau_{\text{CAT-PT}} \ge \left(2 + 2K_d \frac{1 - a_\star}{a_\star}\right)^{-1}.$$

Since $K_d \le (C+1)(1 + \sqrt{r_d})$, the denominator is at most

$$\left\{2 + 2(C+1)\frac{1 - a_\star}{a_\star}\right\}(1 + \sqrt{r_d}).$$

Taking its reciprocal proves $\tau_{\text{CAT-PT}} = \Omega((1+\sqrt{r_d})^{-1})$ with a dimension-independent lower constant.

□

The result concerns the data-fitted, selected sampler: the useful class supplies a candidate, and simultaneous validation protects whichever earlier candidate the algorithm might choose. Different edges may therefore stop at different levels. The theorem identifies a sufficient ladder size in terms of r_d; the algorithm still receives a fixed ladder and adapts only the exchange class. A nonuniform ladder also works when each edge satisfies equation (60).

The sufficient fitting conditions of Appendix D.1 provide one concrete way to meet this acceptance margin. By Proposition 5, endpoint errors of at most e_m give the sufficient requirement

$$h_k\sqrt{C_0 r} + h_k\sqrt{\kappa_m}\, e_m + \tfrac{1}{2}\sqrt{\kappa_m} L_{\theta,m} h_k^2 \le b, \qquad b := 1 - a_\star - 2\omega > 0. \tag{60}$$

The left side adds the class residual, the estimation error, and the error from averaging endpoint coefficients. The right side is the rejection allowance left after accounting for the acceptance target and validation uncertainty. This makes the roles of ladder spacing and calibration size visible in a single inequality.

For example, allocate $b/3$ to each term. The class and interpolation requirements become

$$h_k\sqrt{C_0 r} \le b/3, \qquad \sqrt{\overline{\kappa}_m} L_{\theta,m} h_k^2 \le 2b/3.$$

If $e_m \le C_{\mathrm{fit},m}\sqrt{H_m/N_{\mathrm{fit}}^{\mathrm{eff}}}$, where $H_m = p_m + \log((K+1)M/\delta_{\mathrm{fit}})$, the estimation requirement holds when

$$N_{\mathrm{fit}}^{\mathrm{eff}} \ge \frac{9h_k^2\overline{\kappa}_m C_{\mathrm{fit},m}^2 H_m}{b^2}, \tag{61}$$

together with the relevant Gram-stability requirement: this is equation (45) in the iid bounded-feature regime, or its counterpart under valid dependent-sample moment concentration. More data can reduce the second term of equation (60); the first and third are controlled by the represented structure and the edge width.

It remains to record the probability of the event on which the fitting and validation guarantees hold simultaneously. Let F be the event that $\varepsilon_{k,m_d}^{\mathrm{fit}} \le C_1 h_k(1+\sqrt{r_d})$ on every edge. The proof applies on $\mathsf{F}\cap\mathsf{V}(\omega)$. To quantify its probability, suppose an applicable fitting analysis gives $P_{\mathrm{ind}}(\mathsf{F}) \ge 1-\delta_{\mathrm{fit}}$, and choose the validation radius with failure probability δ_{val}. The moment, interpolation, and sample-size conditions above provide one way to establish this fitting probability.

A union bound under P_{ind} followed by the calibration-law transfer equation (44) gives

$$Q_{\mathrm{cal}}(\mathsf{F}\cap\mathsf{V}(\omega)) \ge 1-\delta_{\mathrm{fit}}-\delta_{\mathrm{val}}-\beta_{\mathrm{mix}}(g)-\eta_{\mathrm{burn}}. \tag{62}$$

The fitting and validation events need not be independent. Their failure probabilities are first combined under the same stationary-block law, and the block-dependence and burn-in terms are then charged once in transferring the joint event to Q_{cal}. For a prescribed $\omega_0 > 0$, the requirement $\omega \le \omega_0$ follows from

$$N_{\mathrm{val}}^{\mathrm{eff}} \ge C_{\mathrm{val}}^2 \log\{2K(M+1)/\delta_{\mathrm{val}}\}/\omega_0^2.$$

Appendices E.2 and F.3 verify fitting accuracy for structured Gaussian paths, including examples with unbounded features that do not satisfy Assumption 4.

D.3 High-dimensional performance of raw PT with arbitrarily fine ladders

We now derive the raw-PT benchmark. The purpose is to determine whether the ambient-dimensional slowdown of raw PT can be removed simply by refining the temperature ladder. We first establish the limiting round-trip rate for a general transported exchange as the ladder mesh tends to zero. We then specialize this result to raw PT and show that, under suitable score conditions, its limiting DEO/ELE rate remains $O(d^{-1/2})$. The local-expansion assumption used here is additional to the conditions of Theorem 2; the mesh tends to zero at each fixed dimension before the dependence of the limiting rate on d is examined.

D.3.1 Fine-ladder limit for general transported exchange

Fix t and an admissible short-time velocity, with flow $T_{t\to t+h}^{\mathbf{v}}$. Express the destination distribution in the source coordinates:

$$Q_{t,h}^{\mathbf{v}} := (T_{t\to t+h}^{\mathbf{v}})_{\#}^{-1}\pi_{t+h}, \qquad D_{t,h}^{\mathbf{v}}(\mathbf{x}) := \log\frac{dQ_{t,h}^{\mathbf{v}}}{d\pi_t}(\mathbf{x}). \tag{63}$$

The two laws agree at $h=0$. Their first-order log-density difference is $\rho_t^{\mathbf{v}} = s_t + \mathcal{S}_t\mathbf{v}_t$, by the transformed-score identity in Lemma 1. To convert this derivative into an expansion of rejection, the approximation must remain accurate after averaging over source states. The following condition bounds that average remainder.

Assumption 5 (Uniform local likelihood expansion)**.** *For every admissible velocity in the compact coefficient class under consideration and all sufficiently small $h > 0$ with $t + h \le 1$,*

$$e^{D_{t,h}^{\mathbf{v}}(\mathbf{x})} = 1 + h\rho_t^{\mathbf{v}}(\mathbf{x}) + R_{t,h}^{\mathbf{v}}(\mathbf{x}), \qquad \sup_{t,\mathbf{v}} \mathbb{E}_{\pi_t}|R_{t,h}^{\mathbf{v}}| \le Ch^2. \tag{64}$$

Moreover, $t \mapsto \rho_t^{\mathbf{v}}$ and the first-order coefficient defined below are continuous, uniformly over the stated class.

The assumption bounds the average error in replacing the short-time density ratio by its first-order term. It holds, for example, when the log ratio has a uniform second-order expansion in $L^3(\pi_t)$ and an integrable exponential envelope. Smooth Gaussian parameter paths with uniformly nonsingular covariance and sufficiently regular affine coefficient curves satisfy these conditions on compact coefficient sets for small h. The remainder constant may depend on the fixed dimension. The bound is used below while taking the mesh to zero; finite-ladder comparisons are developed from endpoint distributions in Appendix F.

For independent $\mathbf{X}, \mathbf{X}' \sim \pi_t$, define $\lambda_t(\mathbf{v}) = \frac{1}{2}\mathbb{E}|\rho_t^{\mathbf{v}}(\mathbf{X}) - \rho_t^{\mathbf{v}}(\mathbf{X}')|$. The result below shows that this mean absolute difference is exactly the first-order rejection per unit edge width. We use its integral to express the limiting round-trip rate:

$$\Lambda(\mathbf{v}) := \int_0^1 \lambda_t(\mathbf{v}_t)\, dt. \tag{65}$$

Let $\mathcal{P}_K = \{0 = t_0 < \cdots < t_K = 1\}$ have mesh $\|\mathcal{P}_K\| = \max_k h_k$, and use the flow of $\mathbf{v}_t$ on each edge.

Theorem 3 (Local rejection and the fine-ladder limit)**.** *Under Assumption 5,*

$$r_{t,t+h}^{\mathrm{tr}}(T_{t\to t+h}^{\mathbf{v}}) = h\lambda_t(\mathbf{v}) + O(h^2), \qquad \lambda_t(\mathbf{v}) = \frac{1}{2}\mathbb{E}\left|\rho_t^{\mathbf{v}}(\mathbf{X}) - \rho_t^{\mathbf{v}}(\mathbf{X}')\right|, \tag{66}$$

uniformly over the stated class. Let $\omega_\lambda(\epsilon) = \sup_{|u-t|\le\epsilon} |\lambda_u(\mathbf{v}_u) - \lambda_t(\mathbf{v}_t)|$. As $\|\mathcal{P}_K\| \to 0$, under DEO and ELE,

$$\sum_{k=0}^{K-1} \frac{r_k^{\mathrm{tr}}}{1 - r_k^{\mathrm{tr}}} = \Lambda(\mathbf{v}) + O(\|\mathcal{P}_K\|) + O(\omega_\lambda(\|\mathcal{P}_K\|)), \tag{67}$$

$$\tau_{\mathrm{DEO}}^{\mathrm{tr}}(\mathcal{P}_K) \longrightarrow \frac{1}{2 + 2\Lambda(\mathbf{v})}. \tag{68}$$

Moreover,

$$\Lambda(\mathbf{v}) \le \frac{1}{\sqrt{2}} \int_0^1 \|s_t + \mathcal{S}_t \mathbf{v}_t\|_2\, dt. \tag{69}$$

For an admissible population level-m Stein curve satisfying the same conditions, writing $\Lambda_m := \Lambda(\boldsymbol{\theta}_m^\star)$ gives $\Lambda_m \le \mathcal{L}_m/\sqrt{2}$.

Proof. We first express rejection in source coordinates, so that its leading term can be read from the density-ratio expansion. In source coordinates, transported exchange between π_t and π_{t+h} is an ordinary swap between π_t and $Q_{t,h}^{\mathbf{v}}$. Under $\pi_t \otimes \pi_t$, the two product measures involved in this swap have Radon–Nikodym densities $e^{D_{t,h}^{\mathbf{v}}(\mathbf{X}')}$ and $e^{D_{t,h}^{\mathbf{v}}(\mathbf{X})}$, respectively. Equation (22) therefore gives the exact identity

$$r_{t,t+h}^{\mathrm{tr}} = \frac{1}{2}\mathbb{E}\left|e^{D_{t,h}^{\mathbf{v}}(\mathbf{X})} - e^{D_{t,h}^{\mathbf{v}}(\mathbf{X}')}\right|. \tag{70}$$

Insert the expansion from Assumption 5. With $\rho = \rho_t^{\mathbf{v}}$ and $R = R_{t,h}^{\mathbf{v}}$, the expression inside the absolute value becomes

$$h\{\rho(\mathbf{X}) - \rho(\mathbf{X}')\} + R(\mathbf{X}) - R(\mathbf{X}').$$

The elementary inequality $||a+b| - |a|| \le |b|$ yields

$$\left| r_{t,t+h}^{\mathrm{tr}} - \frac{h}{2}\mathbb{E}|\rho(\mathbf{X}) - \rho(\mathbf{X}')| \right| \le \frac{1}{2}\mathbb{E}|R(\mathbf{X}) - R(\mathbf{X}')| \le \frac{1}{2}\mathbb{E}\{|R(\mathbf{X})| + |R(\mathbf{X}')|\} = \mathbb{E}|R(\mathbf{X})| \le Ch^2.$$

This is exactly equation (66); uniformity follows from the uniform remainder bound.

The residual is centered because both the normalized path score and the Stein features have mean zero. Cauchy–Schwarz therefore bounds the local coefficient by its second moment:

$$\begin{aligned}\lambda_t(\mathbf{v}) &\le \frac{1}{2}\left\{\mathbb{E}\big(\rho_t^{\mathbf{v}}(\mathbf{X}) - \rho_t^{\mathbf{v}}(\mathbf{X}')\big)^2\right\}^{1/2} \\ &= \frac{1}{\sqrt{2}}\|s_t + \mathcal{S}_t\mathbf{v}_t\|_{L^2(\pi_t)}.\end{aligned} \tag{71}$$

We now accumulate the local expansion over the ladder. Apply equation (66) on edge k:

$$r_k^{\mathrm{tr}} = h_k\lambda_{t_k}(\mathbf{v}_{t_k}) + O(h_k^2),$$

where the remainder is uniform. Summing the remainder gives

$$\sum_{k=0}^{K-1} h_k^2 \le \left(\max_k h_k\right)\sum_{k=0}^{K-1} h_k = \|\mathcal{P}_K\|.$$

For the first-order terms, integrate the difference from the left endpoint on each edge:

$$\left|\sum_k h_k\lambda_{t_k}(\mathbf{v}_{t_k}) - \Lambda(\mathbf{v})\right| \le \sum_k \int_{t_k}^{t_{k+1}} |\lambda_{t_k}(\mathbf{v}_{t_k}) - \lambda_u(\mathbf{v}_u)|\,du \le \omega_\lambda(\|\mathcal{P}_K\|).$$

The modulus tends to zero by uniform continuity on the compact time interval. Therefore

$$\sum_k r_k^{\mathrm{tr}} = \Lambda(\mathbf{v}) + O(\|\mathcal{P}_K\|) + O(\omega_\lambda(\|\mathcal{P}_K\|)). \tag{72}$$

The same local expansion and boundedness of the coefficient imply $\max_k r_k^{\mathrm{tr}} = O(\|\mathcal{P}_K\|)$. For a sufficiently fine ladder, all rejections are bounded away from one, and

$$\sum_k\left\{\frac{r_k^{\mathrm{tr}}}{1-r_k^{\mathrm{tr}}} - r_k^{\mathrm{tr}}\right\} = \sum_k \frac{(r_k^{\mathrm{tr}})^2}{1-r_k^{\mathrm{tr}}} \le C\max_k r_k^{\mathrm{tr}}\sum_k r_k^{\mathrm{tr}} = O(\|\mathcal{P}_K\|).$$

Combining this with equation (72) proves equation (67). Substitution into the DEO identity equation (37) and continuity of $x \mapsto (2+2x)^{-1}$ give equation (68).

Finally, integrate equation (71) over t to obtain equation (69). Choosing the pointwise Stein minimizer makes the integral on the right equal to $\mathcal{L}_m$, which proves the last claim. □

We write $\tau_{\mathrm{fine}}^{\mathrm{tr}}(\mathbf{v}) = (2+2\Lambda(\mathbf{v}))^{-1}$ for this limiting rate. If $t \mapsto \lambda_t(\mathbf{v}_t)$ is Lipschitz, then $\omega_\lambda(\epsilon) = O(\epsilon)$ and the full remainder is $O(\|\mathcal{P}_K\|)$. This limit concerns the specified velocity curve; the finite-ladder CAT-PT theorem instead controls the estimated edgewise maps through $\Delta_{k,m}$.

D.3.2 Consequence for raw PT in high dimensions

Raw PT is the zero-velocity case, with local coefficient

$$\lambda_0^{(d)}(t) := \frac{1}{2}\mathbb{E}|s_t^{(d)}(\mathbf{X}) - s_t^{(d)}(\mathbf{X}')|, \tag{73}$$

and integrated coefficient $\Lambda_0^{(d)} = \int_0^1 \lambda_0^{(d)}(t)\,dt$.

A large score variance produces a large raw barrier when that variance is visible in typical score differences. Concentrating the variance on very rare values can make the mean absolute difference small in comparison. The following fourth-moment bound supplies a quantitative condition that rules out this behavior and will also be used for within-class lower bounds.

Lemma 3 (Bounding mean absolute differences by variance)**.** *Let Z be centered with $\mathbb{E}Z^2 = 1$ and $\mathbb{E}Z^4 \le B$, and let Z' be an independent copy. Then*

$$\frac{1}{\sqrt{B+3}} \le \frac{1}{2}\mathbb{E}|Z - Z'| \le \frac{1}{\sqrt{2}}. \tag{74}$$

Consequently, a centered residual ρ with $\mathbb{E}\rho^4 \le B(\mathbb{E}\rho^2)^2$ satisfies $\frac{1}{2}\mathbb{E}|\rho - \rho'| \ge \|\rho\|_2/\sqrt{B+3}$.

Proof. Put $W = Z - Z'$. Independence and centering imply $\mathbb{E}W^2 = 2$ and $\mathbb{E}W^4 = 2\mathbb{E}Z^4 + 6 \le 2B + 6$. Hölder's inequality, with exponents $3/2$ and 3, gives

$$\mathbb{E}W^2 = \mathbb{E}\big[|W|^{2/3}|W|^{4/3}\big] \le (\mathbb{E}|W|)^{2/3}(\mathbb{E}|W|^4)^{1/3}.$$

Raising to the power $3/2$ and rearranging yields

$$\mathbb{E}|W| \ge \frac{(\mathbb{E}W^2)^{3/2}}{(\mathbb{E}W^4)^{1/2}} \ge \frac{2^{3/2}}{\sqrt{2B+6}}.$$

Dividing by two gives the lower bound in equation (74). The upper bound follows from Cauchy–Schwarz, $\mathbb{E}|W| \le \sqrt{\mathbb{E}W^2} = \sqrt{2}$. For nonzero ρ, apply the result to $Z = \rho/\|\rho\|_2$; a zero residual satisfies the conclusion trivially. □

For Gaussian standardized scores the fourth moment is three. For the standardized χ^2_d score of Appendix E.2, it is $3 + 12/d \le 15$. Both give a dimension-independent lower comparison constant. Independent centered score components with uniformly bounded standardized fourth moments retain such a bound after summation: their fourth moment expands into the sum of component fourth moments and six times the pairwise variance products. This covers additive scores whose variance grows with the number of changing coordinates.

We now give the score conditions needed for the comparison following Theorem 2. At each fixed dimension, assume Assumption 5 for the zero velocity. To express the raw and structured scales together, consider the regime

$$\mathbb{E}[(s_t^{(d)})^2] \ge c_0 d, \qquad \lambda_0^{(d)}(t) \ge c_G \|s_t^{(d)}\|_2, \qquad \mathcal{E}_{m_d}^{(d)}(t) \le C_0 r_d, \tag{75}$$

uniformly in t, where $0 \le r_d \le d$ and the positive constants are independent of d, t. Only the first two inequalities are needed for the raw-PT benchmark. The first gives its score-variance scale, and the second ensures that this scale contributes to average rejection. Lemma 3 verifies the second inequality when the standardized raw-score fourth moments are uniformly bounded. The third inequality is the separate structured-residual premise used by CAT-PT.

For the population-flow comparison also recorded below, additionally assume that the pointwise Stein minimizer curve is admissible and satisfies the hypotheses of Theorem 3. This requirement is needed to interpret its barrier as a fine-ladder rate; it is not needed for the raw bound or for the fitted finite-ladder comparison.

The first two conditions imply

$$\lambda_0^{(d)}(t) \ge c_G \|s_t^{(d)}\|_2 = c_G \sqrt{\mathbb{E}[(s_t^{(d)})^2]} \ge c_G \sqrt{c_0 d}.$$

Integrating over an interval of length one bounds $\Lambda_0^{(d)}$ from below. For the population Stein flow, equation (69) gives

$$\Lambda_{m_d}^{(d)} \le \frac{1}{\sqrt{2}} \int_0^1 \sqrt{\mathcal{E}_{m_d}^{(d)}(t)}\, dt \le \frac{1}{\sqrt{2}} \int_0^1 \sqrt{C_0 r_d}\, dt = \sqrt{C_0 r_d/2}.$$

Together,

$$\Lambda_0^{(d)} \ge c_G \sqrt{c_0 d}, \qquad \Lambda_{m_d}^{(d)} \le \sqrt{C_0 r_d/2}. \tag{76}$$

Substituting the raw bound into the limiting rate yields

$$\tau_{\mathrm{raw}}^{\mathrm{fine}}(d) \le \frac{1}{2 + 2c_G\sqrt{c_0 d}} = O(d^{-1/2}). \tag{77}$$

This proves the raw-PT benchmark used in the paragraph following Theorem 2. It does not require any fitting assumption. The corresponding comparison between population-flow fine-ladder rates is

$$\frac{\tau_{\mathrm{fine}}^{\mathrm{tr}}(\boldsymbol{\theta}_{m_d}^{\star})}{\tau_{\mathrm{raw}}^{\mathrm{fine}}} \ge \frac{2 + 2c_G\sqrt{c_0 d}}{2 + \sqrt{2C_0 r_d}}. \tag{78}$$

Indeed, the structured upper bound on Λ gives a lower bound on its rate, while the raw lower bound on Λ gives an upper bound on the raw rate. Their ratio diverges when $r_d = o(d)$ and is $\Omega(\sqrt{d/r_d})$ when $r_d \to \infty$.

For the implemented sampler, combine equation (77) directly with Theorem 2, rather than using the population-flow ratio above. On the fitting and validation event, a dimension-independent constant $c_{\mathrm{CAT}} > 0$ satisfies

$$\tau_{\text{CAT-PT}} \geq \frac{c_{\mathrm{CAT}}}{1+\sqrt{r_d}}.$$

Under the raw-score conditions above, it follows that

$$\frac{\tau_{\text{CAT-PT}}}{\tau_{\mathrm{raw}}^{\mathrm{fine}}} \geq c_{\mathrm{CAT}} \frac{2+2c_G\sqrt{c_0 d}}{1+\sqrt{r_d}} \geq c\frac{\sqrt{d}}{1+\sqrt{r_d}}$$

for a dimension-independent $c > 0$. When $r_d \to \infty$ and $r_d = o(d)$, this gives the $\Omega(\sqrt{d/r_d})$ improvement stated after the theorem. For bounded r_d, CAT-PT's rate stays bounded away from zero, while the raw benchmark is $O(d^{-1/2})$. The comparison holds on the same fitting and validation event, with probability controlled by equation (62) whenever its premises hold.

Both rates use the same DEO/ELE convention. CAT-PT uses its fitted finite ladder, whereas the raw-PT benchmark is the fine-ladder limit, not the best rate over all finite raw ladders. Under the score conditions above, refining the raw ladder leaves a limiting accumulated rejection-odds barrier of at least order $\sqrt{d}$. Structured transport instead reduces the change that exchanges must accommodate. Appendix E gives paths with small residual scale, and Appendix F develops within-class optimality and finite-ladder lower bounds under additional assumptions.

E Structured examples with small Stein residuals

Theorem 2 is useful when a simple velocity class leaves much less annealing variation than raw exchange faces. This section gives concrete conditions under which that happens. In a conditional location–scale family, a structured motion can account for changes in a large block of coordinates, leaving only a smaller marginal distribution to vary. In a Gaussian scale family, one scalar velocity accounts for the entire d-dimensional change. These examples explain when the theorem's residual scale r_d can be small and how that scale is related to the chosen velocity class.

E.1 Conditional distributions with low-dimensional residual change

A large set of coordinates need not contribute to the residual if their changes can be accounted for by a coordinate transformation in the supplied class. For example, subtracting a changing conditional mean and dividing by a changing scale may leave a conditional distribution whose shape is constant. We first describe this situation generally, then give the location–scale formulas and allow approximation error.

Let $(\mathbf{e}, \mathbf{z})$ denote coordinates in which the conditional law of $\mathbf{e}$ given $\mathbf{z}$ is independent of t. The vector $\mathbf{e}$ may be high-dimensional and dependent on $\mathbf{z}$. Write

$$\varpi_t(\mathbf{e}, \mathbf{z}) = \kappa(\mathbf{e} \mid \mathbf{z})\rho_t(\mathbf{z}), \qquad \pi_t = (\Gamma_t)_{\#}\varpi_t, \tag{79}$$

where κ is independent of t and Γ_t is a differentiable bijection. There are two sources of change in π_t: the coordinate map Γ_t moves points, and the marginal ρ_t changes their relative weights. A velocity that follows Γ_t accounts for the first source. At the current position $\mathbf{x}$, this velocity is

$$\mathbf{v}_t(\mathbf{x}) = \partial_t \Gamma_t(\Gamma_t^{-1}(\mathbf{x})). \tag{80}$$

The inverse recovers the coordinates $(\mathbf{e}, \mathbf{z})$ of a current state, and the derivative describes how that state moves when these coordinates are held fixed. The next result shows that this motion removes all annealing change associated with Γ_t. Only the change in the marginal ρ_t remains in the Stein residual.

Proposition 6 (Residual reduction by a conditional coordinate transformation)**.** *Assume that differentiation under the integral sign and Stein integration by parts are valid for the quantities below. Then*

$$s_t^{\pi}(\mathbf{x}) + \mathcal{S}_t^{\pi}\mathbf{v}_t(\mathbf{x}) = s_t^{\rho}(\mathbf{z}), \qquad \mathbf{x} = \Gamma_t(\mathbf{e}, \mathbf{z}). \tag{81}$$

Moreover, the flow of $\mathbf{v}_t$ from s to t is $\Gamma_t \circ \Gamma_s^{-1}$. If this velocity belongs to $\mathcal{V}_m$, then

$$\mathcal{E}_m(t) \leq \mathbb{E}_{\rho_t}[(s_t^{\rho}(\mathbf{Z}))^2]. \tag{82}$$

In particular, an upper bound Cr on this marginal score variance, uniform in t, gives $\mathcal{L}_m \leq \sqrt{Cr}$.

Proof. Let f be smooth and compactly supported. By the pushforward representation,

$$\mathbb{E}_{\pi_t} f(\mathbf{X}) = \int f(\Gamma_t(\mathbf{e}, \mathbf{z}))\kappa(\mathbf{e} \mid \mathbf{z})\rho_t(\mathbf{z})\, d\mathbf{e}\, d\mathbf{z}.$$

Differentiate this identity with respect to t. The derivative of the first factor is

$$\nabla f(\Gamma_t(\mathbf{e}, \mathbf{z}))^\top \partial_t \Gamma_t(\mathbf{e}, \mathbf{z}).$$

After pushing forward, this contributes the velocity term in the display below. The derivative of ρ_t contributes the residual-score term. Therefore

$$\frac{d}{dt}\mathbb{E}_{\pi_t} f = \mathbb{E}_{\pi_t}[\nabla f(\mathbf{X})^\top \mathbf{v}_t(\mathbf{X})] + \mathbb{E}_{\pi_t}[f(\mathbf{X}) s_t^\rho(\mathbf{Z})]. \tag{83}$$

Stein integration by parts gives

$$\mathbb{E}_{\pi_t}[\nabla f(\mathbf{X})^\top \mathbf{v}_t(\mathbf{X})] = -\mathbb{E}_{\pi_t}[f(\mathbf{X}) \mathcal{S}_t^\pi \mathbf{v}_t(\mathbf{X})].$$

On the other hand, direct differentiation of the normalized density gives $d\mathbb{E}_{\pi_t} f/dt = \mathbb{E}_{\pi_t}[f s_t^\pi]$. Substituting these two identities into equation (83) yields

$$\mathbb{E}_{\pi_t}[f(\mathbf{X})\{s_t^\pi(\mathbf{X}) + \mathcal{S}_t^\pi \mathbf{v}_t(\mathbf{X}) - s_t^\rho(\mathbf{Z})\}] = 0$$

for every such test function f. Hence the term in braces is zero π_t-almost surely, proving equation (81).

For the flow statement, fix canonical coordinates $(\mathbf{e}, \mathbf{z})$ and set $\mathbf{x}_u = \Gamma_u(\mathbf{e}, \mathbf{z})$. By equation (80), $\dot{\mathbf{x}}_u = \mathbf{v}_u(\mathbf{x}_u)$. Starting at $\mathbf{x}_s = \Gamma_s(\mathbf{e}, \mathbf{z})$, uniqueness of the ODE solution therefore gives $\Phi_{s,t}(\mathbf{x}_s) = \Gamma_t(\mathbf{e}, \mathbf{z}) = \Gamma_t \circ \Gamma_s^{-1}(\mathbf{x}_s)$.

The velocity in equation (80) is feasible in the minimization defining $\mathcal{E}_m(t)$. Its residual is $s_t^\rho(\mathbf{Z})$ by equation (81). Integrating its squared residual over π_t is therefore the same as integrating $(s_t^\rho)^2$ over the marginal ρ_t, proving equation (82). A uniform bound Cr then gives

$$\mathcal{L}_m = \int_0^1 \sqrt{\mathcal{E}_m(t)}\, dt \le \int_0^1 \sqrt{Cr}\, dt = \sqrt{Cr}.$$

□

When the velocity of Γ_t belongs to the chosen class, the residual bound involves only the marginal ρ_t, regardless of the dimension of $\mathbf{e}$. If this marginal is also constant, the structural velocity has zero residual. Otherwise, its score variance measures the remaining annealing change. A low-dimensional marginal gives a small residual scale when its score moments are controlled; dimension alone does not bound how rapidly the marginal can change.

To relate this variance to dimension, suppose $\rho_t(\mathbf{z}) = \prod_{i=1}^r \rho_{t,i}(z_i)$ and each coordinate score has second moment at most C. Since these scores are centered and independent,

$$\mathbb{E}_{\rho_t}[(s_t^\rho)^2] = \sum_{i=1}^r \mathbb{E}[(\partial_t \log \rho_{t,i}(Z_i))^2] \le Cr.$$

For dependent coordinates, the same conclusion follows when the sum of score-component covariances is $O(r)$. Thus an r-dimensional residual marginal gives the desired scale when its per-coordinate change is controlled. The factorization explains a small residual in the supplied class; CAT-PT fits that class in the original coordinates using the available scores.

For a conditional location–scale family, the coordinate transformation can be written explicitly. This makes it possible to check whether the velocity needed for the residual reduction is actually present in the hierarchy. Let $\mathbf{z} \in \mathbb{R}^r$ and $\mathbf{u} \in \mathbb{R}^{d-r}$, and suppose

$$\pi_t(\mathbf{z}, \mathbf{u}) = \zeta_t(\mathbf{z}) |\det \mathbf{L}_t(\mathbf{z})|^{-1} \varrho_{\mathbf{z}}\big(\mathbf{L}_t(\mathbf{z})^{-1}\{\mathbf{u} - \mathbf{m}_t(\mathbf{z})\}\big), \tag{84}$$

where $\mathbf{L}_t(\mathbf{z})$ is invertible and the standardized conditional density $\varrho_{\mathbf{z}}$ is fixed in t. Here the conditional shape of $\mathbf{u}$ stays the same: only its location and scale change. Set $\Gamma_t(\mathbf{e}, \mathbf{z}) = (\mathbf{z}, \mathbf{m}_t(\mathbf{z}) + \mathbf{L}_t(\mathbf{z})\mathbf{e})$. For fixed $\mathbf{z}$, the change of variables $\mathbf{u} = \mathbf{m}_t(\mathbf{z}) + \mathbf{L}_t(\mathbf{z})\mathbf{e}$ has Jacobian $|\det \mathbf{L}_t(\mathbf{z})|$. Substitution into equation (84) shows that $(\mathbf{e}, \mathbf{z})$ has density $\varrho_{\mathbf{z}}(\mathbf{e})\zeta_t(\mathbf{z})$. The conditional density of $\mathbf{e}$ is now independent of time, as required in equation (79).

To move from s to t, recover the standardized coordinate at the source, $\mathbf{e} = \mathbf{L}_s(\mathbf{z})^{-1}\{\mathbf{u} - \mathbf{m}_s(\mathbf{z})\}$, and apply the destination map to the same coordinate. This gives

$$T_{s\to t}(\mathbf{z}, \mathbf{u}) = \big(\mathbf{z}, \mathbf{m}_t(\mathbf{z}) + \mathbf{L}_t(\mathbf{z})\mathbf{L}_s(\mathbf{z})^{-1}\{\mathbf{u} - \mathbf{m}_s(\mathbf{z})\}\big). \tag{85}$$

Differentiating the map while holding $(\mathbf{e}, \mathbf{z})$ fixed gives

$$\mathbf{v}_t(\mathbf{z}, \mathbf{u}) = \Big(\mathbf{0}, \dot{\mathbf{m}}_t(\mathbf{z}) + \dot{\mathbf{L}}_t(\mathbf{z})\mathbf{L}_t(\mathbf{z})^{-1}\{\mathbf{u} - \mathbf{m}_t(\mathbf{z})\}\Big).$$

Whenever this field belongs to $\mathcal{V}_m$, Proposition 6 gives

$$\mathcal{E}_m(t) \le \mathbb{E}_{\zeta_t}[(\partial_t \log \zeta_t(\mathbf{Z}))^2]. \tag{86}$$

Thus changes in the large conditional block are removed by its location–scale motion, leaving only the marginal change in $\mathbf{z}$.

The reduction depends on class membership, which can now be checked from the displayed field. If its coefficients are independent of $\mathbf{z}$ and have global, diagonal, or fixed-block form, the field belongs to the affine families in Appendix H.1. Coefficients depending on $\mathbf{z}$ require the corresponding conditional fields in the class and the stated flow regularity. Thus the useful structure is a motion that the hierarchy can represent, not just a partition into small and large coordinate blocks.

A practical hierarchy may capture the conditional motion only approximately. We can quantify how much of the residual reduction is retained in that case. Suppose an ideal structural motion $\mathbf{v}_t^{\mathrm{str}}$ is approximated by $\overline{\mathbf{v}}_t \in \mathcal{V}_m$ with

$$\|\mathcal{S}_t(\overline{\mathbf{v}}_t - \mathbf{v}_t^{\mathrm{str}})\|_{L^2(\pi_t)} \le \eta_t. \tag{87}$$

This measures the difference in the density changes induced by the two velocities. Since $\overline{\mathbf{v}}_t$ is feasible, $\sqrt{\mathcal{E}_m(t)} \le \|s_t + \mathcal{S}_t\overline{\mathbf{v}}_t\|_2$. Adding and subtracting the structural velocity inside the norm gives

$$\begin{aligned}\sqrt{\mathcal{E}_m(t)} &\le \|s_t + \mathcal{S}_t\mathbf{v}_t^{\mathrm{str}}\|_2 + \|\mathcal{S}_t(\overline{\mathbf{v}}_t - \mathbf{v}_t^{\mathrm{str}})\|_2 \\ &\le \|s_t + \mathcal{S}_t\mathbf{v}_t^{\mathrm{str}}\|_2 + \eta_t.\end{aligned}$$

Integrating yields

$$\mathcal{L}_m \le \int_0^1 \|s_t + \mathcal{S}_t\mathbf{v}_t^{\mathrm{str}}\|_2\, dt + \int_0^1 \eta_t\, dt. \tag{88}$$

The ideal motion's residual and the error in representing its density change contribute additively. Approximate location–scale or block structure can therefore remain useful when the second integral is small. Restricting the integrals to edge k gives a bound on $\mathcal{L}_{k,m}$; adding the fitting error from Appendix D.1 then gives the rejection bound of Theorem 1. The same calculation separates imperfect problem structure from imperfect estimation of a useful motion.

E.2 Gaussian scale example: exact transport with one parameter

This example verifies both the small-residual condition and its benefit for the fitted sampler. All d coordinates change scale, but a single coefficient describes their common motion. Stein regression recovers that coefficient exactly from nondegenerate fitting samples, giving acceptance one, while raw PT's fine-ladder round-trip rate decreases as $d^{-1/2}$. Consider

$$\pi_t^{(d)} = \mathcal{N}(\mathbf{0}, e^{2\gamma t}\mathbf{\Sigma}_d), \qquad \gamma \ne 0, \qquad \mathbf{\Sigma}_d \succ 0, \tag{89}$$

and let the first structured class be generated by $\mathbf{v}_1(\mathbf{x}) = \mathbf{x}$.

Proposition 7 (Exact scalar transport and Gaussian raw-PT scaling)**.** *For equation* (89)*, the population Stein coefficient is γ and $\mathcal{E}_1(t) = 0$. The unregularized empirical regression recovers the same coefficient whenever its feature has nonzero sample variance. The resulting flow is $T_{s\to t}(\mathbf{x}) = e^{\gamma(t-s)}\mathbf{x}$ and has acceptance one on every finite edge.*

Maintaining a fixed raw acceptance $a \in (0,1)$ requires and can be achieved with $\Theta(\sqrt{d})$ ladder intervals. Under DEO and ELE, $\tau_{\mathrm{raw}}^{\mathrm{fine}} = \Theta(d^{-1/2})$. On a one-edge ladder, if the endpoint fits are nondegenerate and $\omega \le (1 - a_\star)/2$, CAT-PT selects the scaling candidate and has rate $1/2$ for all sufficiently large d on $\mathsf{V}(\omega)$.

Proof. We first compute the Stein fit. The normalized log density is

$$\log p_t(\mathbf{x}) = C_d - \gamma t d - \tfrac{1}{2} e^{-2\gamma t} \mathbf{x}^\top \mathbf{\Sigma}_d^{-1} \mathbf{x},$$

where C_d is independent of $t, \mathbf{x}$. Differentiation gives

$$s_t(\mathbf{x}) = \gamma \{ e^{-2\gamma t} \mathbf{x}^\top \mathbf{\Sigma}_d^{-1} \mathbf{x} - d \}. \tag{90}$$

The scaling field has divergence d and $\mathbf{v}_1(\mathbf{x})^\top \nabla \log p_t(\mathbf{x}) = -e^{-2\gamma t} \mathbf{x}^\top \mathbf{\Sigma}_d^{-1} \mathbf{x}$. Hence

$$\mathcal{S}_t \mathbf{v}_1(\mathbf{x}) = d - e^{-2\gamma t} \mathbf{x}^\top \mathbf{\Sigma}_d^{-1} \mathbf{x}. \tag{91}$$

Equations (90)–(91) show that $s_t + \gamma \mathcal{S}_t \mathbf{v}_1 = 0$. Thus the coefficient γ leaves zero population residual.

The empirical objective has the same minimizer because $g_t + \gamma \mathcal{S}_t \mathbf{v}_1$ is constant across samples. Its sample variance is therefore zero. Nonzero sample variance of the Stein feature makes the quadratic objective strictly convex in the scalar coefficient, so this minimizer is unique. The coefficient is constant across nodes, and averaging endpoint fits keeps it equal to γ. Solving $\dot{\mathbf{X}}_u = \gamma \mathbf{X}_u$ gives

$$\mathrm{Flow}_1(h, \gamma)(\mathbf{x}) = e^{\gamma h} \mathbf{x}. \tag{92}$$

This maps $\pi_t^{(d)}$ exactly to $\pi_{t+h}^{(d)}$, so its direct-exchange acceptance is one.

We next calculate the overlap available to raw exchange. Whitening by $\mathbf{\Sigma}_d^{-1/2}$ leaves both Hellinger affinity and swap acceptance unchanged. For one coordinate with standard deviations $\sigma_t = e^{\gamma t}$ and $\sigma_{t+h} = e^{\gamma(t+h)}$, the affinity is

$$\left(\frac{2\sigma_t \sigma_{t+h}}{\sigma_t^2 + \sigma_{t+h}^2} \right)^{1/2}.$$

The whitened coordinates are independent, so their affinities multiply. Using $2e^{\gamma h}/(1 + e^{2\gamma h}) = \mathrm{sech}(\gamma h)$,

$$\int \sqrt{d\pi_t^{(d)} d\pi_{t+h}^{(d)}} = \{ \mathrm{sech}(|\gamma| h) \}^{d/2}. \tag{93}$$

Raw-swap acceptance is the overlap of $P = \pi_t \otimes \pi_{t+h}$ and $Q = \pi_{t+h} \otimes \pi_t$. Since $\min\{a, b\} \le \sqrt{ab}$, this overlap is at most their Hellinger affinity. The product structure makes that affinity the square of equation (93), yielding

$$A_{t,t+h}^{\mathrm{raw}} \le \{ \mathrm{sech}(|\gamma| h) \}^d. \tag{94}$$

If $A_{t,t+h}^{\mathrm{raw}} \ge a$, then $d \log \cosh(|\gamma| h) \le -\log a$. For fixed γ, the inequality $\log \cosh(|\gamma| h) \ge c_\gamma h^2$ holds near zero. The displayed acceptance condition forces h into that neighborhood as d grows and then gives $h \le C_a / \sqrt{d}$. Covering $[0, 1]$ requires $\Omega(\sqrt{d})$ intervals. Conversely, $\mathbb{E}_{\pi_t} s_t^2 = 2\gamma^2 d$, so Theorem 1 applied with zero velocity gives $r_{t,t+h}^{\mathrm{raw}} \le |\gamma| h \sqrt{2d}$. A uniform spacing $h \le (1 - a)/(|\gamma| \sqrt{2d})$ therefore attains acceptance at least a on every edge with $O(\sqrt{d})$ intervals.

The round-trip comparison follows by evaluating the raw local barrier. Let Q_d, Q_d' be independent χ_d^2 variables. In whitened coordinates, $s_t(\mathbf{X}) = \gamma(Q_d - d)$, and hence

$$\lambda_0^{(d)}(t) = \frac{|\gamma|}{2} \mathbb{E} |Q_d - Q_d'|. \tag{95}$$

The central limit theorem gives $(Q_d - Q_d')/(2\sqrt{d}) \Rightarrow \mathcal{N}(0, 1)$. These variables have uniformly bounded second moments, so their absolute values are uniformly integrable and their first absolute moments converge. Therefore

$$\lambda_0^{(d)}(t) = |\gamma| \sqrt{\frac{2d}{\pi}} \{ 1 + o(1) \}. \tag{96}$$

This barrier is constant in t. Theorem 3 then yields

$$\tau_{\mathrm{raw}}^{\mathrm{fine}}(d) = \frac{1 + o(1)}{2 + 2|\gamma| \sqrt{2d/\pi}} = \Theta(d^{-1/2}). \tag{97}$$

Finally consider the selected kernel on a one-edge ladder. For sufficiently large d, equation (94) gives $A^{\mathrm{raw}} < a_\star$. On $\mathsf{V}(\omega)$, $\widehat{A}^{\mathrm{raw}} - \omega \le A^{\mathrm{raw}} < a_\star$, so raw exchange fails the selection rule. The scaling candidate has held-out acceptance exactly one and satisfies the rule when $\omega \le (1 - a_\star)/2$. It is therefore selected. The one-edge DEO/ELE identity is $\tau = A/2$, giving

$$\tau_{\text{CAT-PT}}(d) = \frac{1}{2} \quad \text{for all sufficiently large } d \text{ on } \mathsf{V}(\omega). \tag{98}$$

□

This example realizes the full construction with one parameter. The Stein fit recovers the common scale change, endpoint averaging preserves that coefficient, and its exact flow removes the finite-edge mismatch. The empirical fitting identity is algebraic, so fitting observations with nonzero feature variance recover γ even before invoking a sampling concentration bound. The validation event is used to compare the raw estimate with its population acceptance and determine the selected level.

The example separates three quantities: the state has dimension d, the raw score variance is proportional to d, and the useful transport has one parameter and zero residual. In the conditional families above, the same principle removes only the representable part of the change. Theorem 2 gives a rate based on what remains. Whether another velocity in the same class can improve that rate is the optimality question addressed in Appendix F.

F Optimality of Stein fitting and achievable round-trip rates

Theorem 2 gives a rate that CAT-PT can achieve when a structured class leaves residual scale r_d. Here we ask whether a different transport in the same class can do better. We first give conditions under which the population Stein velocity maximizes the fine-ladder round-trip rate. We then bound the rate attainable by any velocity in that class, obtaining the same $(1 + \sqrt{r_d})^{-1}$ order under two-sided residual and moment conditions. Finally, a Gaussian location family gives matching bounds for the fitted CAT-PT sampler on a finite ladder.

F.1 Optimality of Stein fitting in the fine-ladder limit

Stein regression minimizes a squared residual, while exchange rejection depends on the mean absolute difference of two residual values. Minimizing the former need not minimize the latter. We first give a condition under which the two objectives agree, so that Stein fitting optimizes the actual limiting round-trip rate. We then verify the projection and path-length properties that hold without this condition.

Recall that the first-order rejection coefficient is $\lambda_t(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}|\rho_{t,\boldsymbol{\theta}}(\mathbf{X}) - \rho_{t,\boldsymbol{\theta}}(\mathbf{X}')|$ from equation (66), where $\rho_{t,\boldsymbol{\theta}} = s_t + \boldsymbol{\psi}_m^\top \boldsymbol{\theta}$. The following assumption makes this coefficient proportional to the square-root Stein risk: changing the fitted coefficient changes the scale of the residual, but not its standardized distribution.

Assumption 6 (Shape-aligned residual family)**.** *At each t and level m, there exists a centered unit-variance scalar random variable Z_t such that, for every $\boldsymbol{\theta}$ with $\mathcal{R}_m(t, \boldsymbol{\theta}) > 0$,*

$$s_t(\mathbf{X}) + \boldsymbol{\psi}_m(\mathbf{X})^\top \boldsymbol{\theta} \overset{d}{=} \sqrt{\mathcal{R}_m(t, \boldsymbol{\theta})}\, Z_t, \qquad \mathbf{X} \sim \pi_t. \tag{99}$$

Theorem 4 (Fine-ladder communication optimality of Stein projection)**.** *Under Assumption 6,*

$$\lambda_t(\boldsymbol{\theta}) = c_t \sqrt{\mathcal{R}_m(t, \boldsymbol{\theta})}, \qquad c_t := \frac{1}{2}\mathbb{E}|Z_t - Z_t'|, \tag{100}$$

where Z_t' is an independent copy. Therefore the Stein coefficient $\boldsymbol{\theta}_m^\star(t)$ minimizes the exact local communication barrier over $\mathcal{V}_m$. If the pointwise minimizer curve is admissible, its flow minimizes the global barrier Λ. Under Assumption 5 and DEO/ELE, it therefore maximizes the limiting round-trip rate among admissible velocity curves in $\mathcal{V}_m$. For Gaussian Z_t, $c_t = 1/\sqrt{\pi}$.

Under the common-shape assumption, every coefficient changes local rejection through the same scale factor that appears in squared-residual fitting. Stein projection therefore minimizes first-order

rejection and maximizes the limiting round-trip rate within the class. This is a population result for the fine-ladder limit; fitting and finite-edge errors are analyzed in Appendix D.

The variable Z_t represents this common standardized distribution. Its shape can vary with t; at a fixed t, all coefficients share that shape and differ only through the factor $\sqrt{\mathcal{R}_m(t,\boldsymbol{\theta})}$. A sufficient example is an elliptically distributed centered vector $(s_t(\mathbf{X}), \boldsymbol{\psi}_m(\mathbf{X})^\top)^\top$ with finite covariance, including the joint Gaussian case. Each residual is a linear combination of this vector and therefore has the same standardized shape. This makes the relation between residual magnitude and exchange rejection independent of which coefficient is used.

Proof. By Assumption 6, independent residuals at $\mathbf{X}$ and $\mathbf{X}'$ have the joint representation

$$\rho_{t,\boldsymbol{\theta}}(\mathbf{X}) \overset{d}{=} \sqrt{\mathcal{R}_m(t,\boldsymbol{\theta})}Z_t, \qquad \rho_{t,\boldsymbol{\theta}}(\mathbf{X}') \overset{d}{=} \sqrt{\mathcal{R}_m(t,\boldsymbol{\theta})}Z_t',$$

with Z_t, Z_t' independent. Substituting this representation into the definition of λ_t gives

$$\lambda_t(\boldsymbol{\theta}) = \frac{1}{2}\sqrt{\mathcal{R}_m(t,\boldsymbol{\theta})}\mathbb{E}|Z_t - Z_t'|,$$

which is equation (100). If the risk is zero, the residual vanishes almost surely, so the same identity holds trivially.

The factor c_t does not depend on $\boldsymbol{\theta}$. Therefore minimizing λ_t is equivalent to minimizing $\mathcal{R}_m(t,\boldsymbol{\theta})$. Proposition 2 shows that the latter is minimized by $\boldsymbol{\theta}_m^\star(t)$. If the minimizer curve is admissible, this pointwise inequality integrates to $\Lambda(\boldsymbol{\theta}_m^\star) \le \Lambda(\boldsymbol{\theta})$ for every admissible curve in the class. The fine-ladder round-trip formula equation (68) is strictly decreasing in Λ, so the same curve maximizes the limiting round-trip rate.

If $Z_t \sim \mathcal{N}(0,1)$, then $Z_t - Z_t' \sim \mathcal{N}(0,2)$ and $\mathbb{E}|Z_t - Z_t'| = 2/\sqrt{\pi}$. Hence $c_t = 1/\sqrt{\pi}$. □

Thus, under shape alignment, population Stein fitting is exactly aligned with communication: it minimizes the first-order rejection barrier and therefore maximizes the fine-ladder round-trip rate within the class. Beyond this aligned setting, the Stein objective still retains several useful structural properties that help explain what the fitting procedure is optimizing.

Properties beyond shape alignment. Without shape alignment, the population fit still minimizes the residual-length upper bound from Theorem 1. This quantity depends on the available motions rather than their choice of basis, and is unchanged by a smooth increasing reparameterization of annealing time. We verify these properties next; they describe the fitting objective, not an additional assumption for communication optimality.

At a fixed t, the density-change functions represented by class $\mathcal{V}_m$ form its Stein image $\mathcal{A}_m(t) = \overline{\{\mathcal{S}_t\mathbf{v} : \mathbf{v} \in \mathcal{V}_m\}} \subset L^2(\pi_t)$. Let $\mathbf{P}_m(t)$ be the orthogonal projection onto this subspace. The regression seeks the element of $\mathcal{A}_m(t)$ closest to $-s_t$:

$$\mathcal{E}_m(t) = \inf_{a \in \mathcal{A}_m(t)} \|s_t + a\|_2^2.$$

That element is $-\mathbf{P}_m(t)s_t$, so the remaining score is $(\mathbf{I} - \mathbf{P}_m(t))s_t$ and

$$\mathcal{E}_m(t) = \|(\mathbf{I} - \mathbf{P}_m(t))s_t\|_{L^2(\pi_t)}^2. \tag{101}$$

This describes the residual through the motions available in the class. An invertible change of basis changes coefficient coordinates but leaves the Stein image, fitted density change, $\mathcal{E}_m$, and $\mathcal{L}_m$ unchanged.

For nested images, adding a richer class explains an additional orthogonal component of s_t. In fact,

$$(\mathbf{I} - \mathbf{P}_m)s_t = (\mathbf{I} - \mathbf{P}_{m+1})s_t + (\mathbf{P}_{m+1} - \mathbf{P}_m)s_t,$$

where the two terms on the right are orthogonal. The second is also orthogonal to $\mathcal{A}_m(t)$. Taking squared norms gives

$$\mathcal{E}_m(t) - \mathcal{E}_{m+1}(t) = \|\{\mathbf{P}_{m+1}(t) - \mathbf{P}_m(t)\}s_t\|_2^2. \tag{102}$$

The reduction in residual therefore measures exactly how much of the score lies in the newly available directions. A larger class improves the best Stein fit when those directions align with variation along the path.

Projection minimizes not only the residual at each temperature but also the integrated residual appearing in the rejection upper bound. For every admissible coefficient curve $\boldsymbol{\theta}(\cdot)$, $\mathcal{R}_m(t, \boldsymbol{\theta}(t)) \geq \mathcal{E}_m(t)$. Taking square roots and integrating yields

$$\int_0^1 \sqrt{\mathcal{R}_m(t, \boldsymbol{\theta}(t))}\, dt \geq \mathcal{L}_m. \tag{103}$$

If $\boldsymbol{\theta}^\star_m(t)$ is admissible, it attains the pointwise minimum almost everywhere and hence equality in this bound. Its flow therefore gives the smallest residual-length upper bound from Theorem 1 within $\mathcal{V}_m$.

This integrated residual describes the path and the chosen motion, rather than the speed used to traverse them. To see this, reparameterize annealing time by a monotone change of variable. If $t = \tau(u)$ is an increasing C^1 bijection and $\widetilde{s}_u = \dot{\tau}(u) s_{\tau(u)}$, $\widetilde{\mathbf{v}}_u = \dot{\tau}(u)\mathbf{v}_{\tau(u)}$, then linearity of the Stein operator gives

$$\widetilde{s}_u + \mathcal{S}_{\tau(u)}\widetilde{\mathbf{v}}_u = \dot{\tau}(u)\{s_{\tau(u)} + \mathcal{S}_{\tau(u)}\mathbf{v}_{\tau(u)}\}.$$

Changing variables $t = \tau(u)$ therefore yields

$$\int_0^1 \|\widetilde{s}_u + \mathcal{S}_{\tau(u)}\widetilde{\mathbf{v}}_u\|_2\, du = \int_0^1 \|s_t + \mathcal{S}_t\mathbf{v}_t\|_2\, dt. \tag{104}$$

A change of annealing speed rescales both the desired density change and the velocity by the same factor, leaving the integrated mismatch unchanged.

For general residual distributions, projection still minimizes the residual-length upper bound in equation (103). To establish the best attainable *scaling order*, however, exact agreement of standardized distributions is stronger than necessary. Uniform upper and lower comparisons between the two objectives suffice, as shown next.

F.2 Upper bounds on round-trip rates within a transport class

To show that the rate in Theorem 2 has the correct order, we need an upper bound on the communication attainable by competing transports. We give two such bounds. The first applies to fine-ladder rates when every velocity leaves residual variation of at least order r_d and its residual moments control rejection. The second applies directly to finite ladders when all maps leave a changing statistic of the state unchanged.

For the first bound, a small-residual upper bound is not enough: the class must also be unable to reduce the residual below that order. The following assumption states this two-sided requirement and the moment comparison needed to relate it to rejection.

Assumption 7 (Matching residual complexity and nondegeneracy)**.** *For levels m_d and scales $r_d \geq 1$, constants $0 < c_E \leq C_E < \infty$ independent of d, t satisfy*

$$c_E r_d \leq \mathcal{E}^{(d)}_{m_d}(t) \leq C_E r_d. \tag{105}$$

For every velocity in the comparison class, its centered residual obeys

$$\lambda_t(\mathbf{v}) \geq c_Z \|s^{(d)}_t + \mathcal{S}^{(d)}_t \mathbf{v}\|_{L^2(\pi^{(d)}_t)} \tag{106}$$

with $c_Z > 0$ independent of $d, t, \mathbf{v}$. The pointwise population minimizer curve is admissible.

The lower bound in equation (105) says that no velocity in the class reduces the squared residual below a constant times r_d; the upper bound says that the population Stein fit reaches this scale. Equation (106) prevents the residual variance from being concentrated in rare values that have little effect on average rejection. By Lemma 3, it holds when the standardized fourth moments are uniformly bounded over all competing velocities. Common residual shape with a uniform positive comparison constant is another sufficient condition. The Gaussian translation family below verifies these requirements through explicit residual calculations.

Theorem 5 (Optimal order of the within-class fine-ladder rate)**.** *Under Assumptions 5 and 7, and under DEO and ELE, every admissible level-m_d velocity curve satisfies*

$$\Lambda(\mathbf{v}^{(d)}) \geq c_Z\sqrt{c_E r_d}, \qquad \tau^{\mathrm{tr}}_{\mathrm{fine}}(\mathbf{v}^{(d)}) \leq \frac{1}{2 + 2c_Z\sqrt{c_E r_d}}. \tag{107}$$

The population Stein curve satisfies

$$\Lambda(\boldsymbol{\theta}^\star_{m_d}) \leq \sqrt{C_E r_d/2}, \qquad \tau^{\mathrm{tr}}_{\mathrm{fine}}(\boldsymbol{\theta}^\star_{m_d}) \geq \frac{1}{2 + \sqrt{2C_E r_d}}. \tag{108}$$

Thus the best fine-ladder DEO/ELE rate within the class is $\Theta((1+\sqrt{r_d})^{-1})$, achieved in order by Stein projection. Under shape alignment, Theorem 4 further makes that curve an exact maximizer of the limiting rate.

The upper bound for every competing velocity and the lower bound for Stein projection have the same dependence on r_d. Under these assumptions, choosing another velocity in the class can improve constants but not the fine-ladder scaling order. The comparison fixes the transport class and uses DEO/ELE rates.

Proof. Fix a competing curve $\mathbf{v}^{(d)}_t = \mathbf{v}_{m_d,\boldsymbol{\theta}_d(t)}$. Pointwise minimization and the lower energy bound give

$$\|s^{(d)}_t + \mathcal{S}^{(d)}_t \mathbf{v}^{(d)}_t\|_2^2 = \mathcal{R}^{(d)}_{m_d}(t, \boldsymbol{\theta}_d(t)) \geq \mathcal{E}^{(d)}_{m_d}(t) \geq c_E r_d.$$

Equation (106) therefore gives $\lambda_t(\mathbf{v}^{(d)}_t) \geq c_Z\sqrt{c_E r_d}$. Integrating over an interval of length one proves the first inequality in equation (107). The rate formula from Theorem 3 is decreasing in Λ and gives the second inequality.

For the population Stein curve, its squared residual is exactly $\mathcal{E}^{(d)}_{m_d}(t) \leq C_E r_d$. The universal upper bound equation (71) yields

$$\Lambda(\boldsymbol{\theta}^\star_{m_d}) \leq \frac{1}{\sqrt{2}} \int_0^1 \sqrt{\mathcal{E}^{(d)}_{m_d}(t)}\, dt \leq \sqrt{C_E r_d/2}.$$

Substitution in the same rate formula proves equation (108). The upper bound for every curve and the lower bound for one admissible curve sandwich the supremum between constant multiples of $(1+\sqrt{r_d})^{-1}$. Here each competing curve has its own fine-ladder limiting rate, and the supremum is taken over those rates. Exact optimality under shape alignment is Theorem 4. □

A finite ladder permits a more direct argument when the class leaves part of the state unchanged. Suppose every allowed map preserves a statistic $H(\mathbf{x})$. Even after transforming other coordinates, an exchange must swap the two values of H between temperatures. If their temperature marginals differ, that mismatch alone causes rejection. Write $\rho_t = H_\# \pi_t$ for the distribution of the unchanged statistic at time t. The next result bounds rejection by this marginal mismatch and gives an explicit rate bound when the marginal is Gaussian.

Theorem 6 (Finite-ladder rate bound for transports preserving a statistic)**.** *For an invertible map T satisfying $H(T\mathbf{x}) = H(\mathbf{x})$,*

$$r^{\mathrm{tr}}_{s,t}(T) \geq \|\rho_s \otimes \rho_t - \rho_t \otimes \rho_s\|_{\mathrm{TV}}. \tag{109}$$

If $\pi_t(\mathbf{w},\mathbf{z}) = \nu_t(\mathbf{w})\rho_t(\mathbf{z})$ and $T(\mathbf{w},\mathbf{z}) = (S\mathbf{w},\mathbf{z})$ with $S_\#\nu_s = \nu_t$, equality holds for $H(\mathbf{w},\mathbf{z}) = \mathbf{z}$.

If $H_\#\pi_t = \mathcal{N}(t\mathbf{b}, \mathbf{I}_r)$ and all allowed maps preserve H, then for any finite ladder and any frozen allowed maps, including identity maps, the DEO/ELE rate satisfies

$$\tau \leq \frac{1}{2 + 2\|\mathbf{b}\|_2/\sqrt{\pi}}. \tag{110}$$

In particular, $\|\mathbf{b}\|_2 \asymp \sqrt{r}$ implies $\tau = O((1+\sqrt{r})^{-1})$ uniformly over the ladder and map choices.

Proof. We first compare the full pair distributions through the statistic H. The assumption also implies $H(T^{-1}\mathbf{y}) = H(\mathbf{y})$, by applying it to $\mathbf{x} = T^{-1}\mathbf{y}$. Define $\overline{H}(\mathbf{x},\mathbf{y}) = (H(\mathbf{x}), H(\mathbf{y}))$. Under

$P = \pi_s \otimes \pi_t$, its law is $\rho_s \otimes \rho_t$; under $(G_T)_{\#}P$, its law is $\rho_t \otimes \rho_s$. For any measurable map h, total variation contracts:

$$\|h_{\#}P - h_{\#}Q\|_{\mathrm{TV}} = \sup_B |P(h^{-1}(B)) - Q(h^{-1}(B))| \le \sup_A |P(A) - Q(A)|.$$

Apply this with $h = \overline{H}$ and use equation (42) to prove the lower bound.

For equality in the product case, the transformed $\mathbf{w}$ pair again has law $\nu_s \otimes \nu_t$, independently of the reversed $\mathbf{z}$ pair. The two full pair laws therefore have an identical nuisance factor. Tensoring two measures by the same probability measure preserves their TV distance: integrating the absolute difference of their densities removes that common factor. The full distance is exactly the marginal distance.

To obtain the Gaussian rate bound, let $\rho_t = \mathcal{N}(t\mathbf{b}, \mathbf{I}_r)$ and abbreviate $b = \|\mathbf{b}\|_2$. We calculate the raw-swap acceptance $a_b(h)$ of this marginal across a gap h. Let Φ and ϕ be the standard normal distribution and density functions, respectively.

Take independent $\mathbf{X} \sim \rho_s$ and $\mathbf{Y} \sim \rho_{s+h}$. The log raw-swap ratio is

$$L = h\mathbf{b}^\top(\mathbf{X} - \mathbf{Y}).$$

Its mean is $-h^2b^2$ and its variance is $2h^2b^2$. For $bh > 0$, write $L = -q^2/2 + qZ$ with $q = \sqrt{2}bh$ and $Z \sim \mathcal{N}(0,1)$. Then

$$\begin{aligned}\mathbb{E}\min\{1, e^L\} &= \mathbb{P}(Z \ge q/2) + \int_{-\infty}^{q/2} e^{-q^2/2+qz}\phi(z)\,dz \\ &= \Phi(-q/2) + \int_{-\infty}^{q/2} \phi(z-q)\,dz = 2\Phi(-q/2).\end{aligned}$$

Here ϕ is the standard normal density, and the second equality completes the square in its exponent. The result also holds at $bh = 0$ by continuity.

For the odds bound, define the Gaussian upper-tail hazard $H_0(u) = \phi(u)/\Phi(-u)$ for $u \ge 0$. Since

$$\phi(u) = \int_u^\infty z\phi(z)\,dz > u\int_u^\infty \phi(z)\,dz,$$

we have $H_0(u) > u$ and $H_0'(u) = H_0(u)\{H_0(u) - u\} > 0$. Thus $H_0(u) \ge H_0(0) = \sqrt{2/\pi}$. Integrating the logarithmic derivative of $2\Phi(-u)$ from zero to u gives

$$-\log(2\Phi(-u)) = \int_0^u H_0(v)\,dv \ge \sqrt{2/\pi}\,u.$$

It follows that $1/(2\Phi(-u)) - 1 \ge e^{\sqrt{2/\pi}u} - 1 \ge \sqrt{2/\pi}u$. With $u = bh/\sqrt{2}$, the two conclusions are

$$a_b(h) = 2\Phi\left(-\frac{bh}{\sqrt{2}}\right), \qquad \frac{1 - a_b(h)}{a_b(h)} \ge \frac{bh}{\sqrt{\pi}}. \tag{111}$$

The marginal comparison in equation (109) gives $A_k \le a_b(h_k)$ for every full-state transport. Since $a \mapsto (1-a)/a$ decreases,

$$\frac{1 - A_k}{A_k} \ge \frac{1 - a_b(h_k)}{a_b(h_k)} \ge \frac{bh_k}{\sqrt{\pi}}.$$

Summing over the ladder and using $\sum_k h_k = 1$ gives $\mathcal{I}(\mathbf{r}) \ge b/\sqrt{\pi}$. Substitution into equation (37) proves equation (110). The calculation holds for each realized set of maps, including maps obtained by fitting and selection. □

For a differentiable statistic H, preservation follows when every field in the class satisfies

$$\nabla H(\mathbf{x})^\top \mathbf{v}(\mathbf{x}) = 0.$$

Along a flow, the chain rule then gives $dH(\mathbf{X}_u)/du = 0$. A hierarchy that moves a block $\mathbf{w}$ while leaving $\mathbf{z}$ fixed satisfies this condition with $H(\mathbf{w}, \mathbf{z}) = \mathbf{z}$. Raw exchange corresponds to the identity

map and preserves the statistic as well. The bound consequently applies when different edges select different allowed maps or choose raw exchange.

For the Gaussian marginal, the unavoidable rejection odds are proportional to edge width. Adding temperatures splits this contribution among more edges, but its sum is still at least $b/\sqrt{\pi}$. The rate bound therefore holds for any finite ladder, including one chosen to favor the competing maps. We next give a family where CAT-PT attains this order by transporting the other coordinates, and verify that finite-sample fitting preserves the result.

F.3 Gaussian location example: optimal finite-ladder performance

We now exhibit a fitted CAT-PT sampler whose finite-ladder rate matches the upper bound for its transport class. The target has two independent Gaussian blocks with changing means. The hierarchy can translate the larger block but leaves the r-dimensional block fixed, so only the latter imposes an unavoidable communication cost. We compute the residual, compare upper and lower rate bounds, and then check the fitting sample requirement explicitly. Let $q = d - r \geq 1$, $r \geq 1$, and fix nonzero constants γ, η. Define

$$\pi_t^{(d)}(\mathbf{w}, \mathbf{z}) = \mathcal{N}(\gamma t \mathbf{1}_q, \mathbf{I}_q)(\mathbf{w}) \mathcal{N}(\eta t \mathbf{1}_r, \mathbf{I}_r)(\mathbf{z}). \tag{112}$$

Suppose every field up to level m_d has zero $\mathbf{z}$ component and conditional zero flux in $\mathbf{w}$, and the class contains $\mathbf{v}^\star(\mathbf{w}, \mathbf{z}) = (\gamma \mathbf{1}_q, \mathbf{0})$. Translations or affine fields restricted to a predefined $\mathbf{w}$ block have these properties. The Gaussian factors supply the required moments and integration-by-parts identities.

Corollary 1 (Optimal finite-ladder rate in the Gaussian location family)**.** *For the hierarchy and path above, take a uniform ladder with $K = \lceil C(1 + \sqrt{r}) \rceil$ and fitting accuracy as in Theorem 2. For sufficiently large C, on the fitting and validation event CAT-PT selects no level above m_d and*

$$\tau_{\text{CAT-PT}} = \Theta((1 + \sqrt{r})^{-1}) \tag{113}$$

under DEO and ELE. No finite-ladder scheme using maps from these classes can improve this order. If $q \asymp d$ and $r = o(d)$, raw PT's fine-ladder rate is $\Theta(d^{-1/2})$.

The lower and upper bounds concern the same fitted sampler: the useful motion removes the changing mean of the $\mathbf{w}$ block, while every allowed map leaves the changing $\mathbf{z}$ marginal untouched. The proof checks this residual decomposition and then applies the finite-ladder comparison.

Proof. The normalized score separates into

$$s_t = s_t^w + s_t^z, \quad s_t^w = \gamma \mathbf{1}_q^\top (\mathbf{w} - \gamma t \mathbf{1}_q), \quad s_t^z = \eta \mathbf{1}_r^\top (\mathbf{z} - \eta t \mathbf{1}_r).$$

For an allowed field $\mathbf{v} = (\mathbf{v}_w, \mathbf{0})$, conditional Stein integration by parts gives $\mathbb{E}[s_t^w + \mathcal{S}_t \mathbf{v} \mid \mathbf{Z}] = 0$. Because s_t^z is determined by $\mathbf{Z}$, the cross term vanishes:

$$\mathbb{E}\big[s_t^z (s_t^w + \mathcal{S}_t \mathbf{v})\big] = \mathbb{E}\big[s_t^z \mathbb{E}[s_t^w + \mathcal{S}_t \mathbf{v} \mid \mathbf{Z}]\big] = 0.$$

Consequently,

$$\begin{aligned} \mathbb{E}(s_t + \mathcal{S}_t \mathbf{v})^2 &= \mathbb{E}(s_t^z)^2 + \mathbb{E}(s_t^w + \mathcal{S}_t \mathbf{v})^2 \\ &\geq \eta^2 r. \end{aligned}$$

The field $\mathbf{v}^\star$ cancels s_t^w exactly and attains equality. Its residual and flow are therefore

$$\mathcal{E}_{m_d}(t) = \eta^2 r, \qquad T^\star_{s \to t}(\mathbf{w}, \mathbf{z}) = (\mathbf{w} + \gamma(t - s)\mathbf{1}_q, \mathbf{z}). \tag{114}$$

The lower bound on $\mathcal{E}_{m_d}$ comes from the changing marginal left untouched by every allowed field. For translation-only classes, all residuals are linear combinations of Gaussian coordinates, so the common shape condition also holds exactly.

Equation (114) verifies the energy condition of Theorem 2; its fitting and validation assumptions then give the lower bound for $\tau_{\text{CAT-PT}}$ and selection by level m_d. Each selected map preserves $H(\mathbf{w}, \mathbf{z}) = \mathbf{z}$ because all fields up to this level do, and candidate zero is identity. Apply Theorem 6 with $\mathbf{b} = \eta \mathbf{1}_r$ to get the matching upper bound. The upper and lower bounds therefore apply to the same finite selected sampler.

For raw PT, the score is Gaussian with variance $\gamma^2 q + \eta^2 r$. The difference of two independent scores is Gaussian with twice this variance, so its local barrier is $\sqrt{\gamma^2 q + \eta^2 r}/\sqrt{\pi}$. It does not depend on t. The fine-ladder formula therefore gives

$$\tau_{\mathrm{raw}}^{\mathrm{fine}} = \left(2 + \frac{2}{\sqrt{\pi}}\sqrt{\gamma^2 q + \eta^2 r}\right)^{-1} = \Theta(d^{-1/2})$$

when $q \asymp d$. The same Gaussian calculation also lower-bounds the raw rejection odds on every finite ladder by the full-dimensional barrier. □

The corollary assumes an accurate fitted map. For the smallest class containing the useful motion, $\mathbf{v}_\theta = (\theta \mathbf{1}_q, \mathbf{0})$, this requirement can be verified without an ambient-dimensional sample bound. Let $J = K + 1$ and suppose each node has $N \geq 2$ independent fitting observations, with $N - 1 \geq 16 \log(2J/\delta_{\mathrm{fit}})$. Then, with probability at least $1 - \delta_{\mathrm{fit}}$, simultaneously over all edges,

$$\varepsilon_{k,1}^{\mathrm{fit}} \leq 2|\eta| h_k \sqrt{r} \sqrt{\frac{\log(4J/\delta_{\mathrm{fit}})}{N-1}}. \tag{115}$$

This has the form required by Theorem 2, with a constant independent of d. The factor $\sqrt{q}$ in residual error is balanced by the $q^{-1/2}$ factor in coefficient estimation. Thus fitting the common motion of a large block does not require a separate coefficient estimate for each coordinate.

To verify this bound, at a fixed node, define

$$U = q^{-1/2} \mathbf{1}_q^\top (\mathbf{W} - \gamma t \mathbf{1}_q), \qquad V = r^{-1/2} \mathbf{1}_r^\top (\mathbf{Z} - \eta t \mathbf{1}_r).$$

Then U, V are independent standard normals. The Stein feature is $\psi = -\sqrt{q}\, U$, while g_t differs from $\gamma\sqrt{q}\, U + \eta\sqrt{r}\, V$ by a constant. Centering removes this constant. For $N \geq 2$ independent fitting observations, the empirical covariance and cross-covariance are

$$\widehat{G} = \frac{q}{N} \sum_{i=1}^{N} (U_i - \overline{U})^2,$$

$$\widehat{b} = -\frac{\gamma q}{N} \sum_{i=1}^{N} (U_i - \overline{U})^2 - \frac{\eta\sqrt{qr}}{N} \sum_{i=1}^{N} (U_i - \overline{U}) V_i.$$

The term involving $\overline{V}$ vanishes because $\sum_i (U_i - \overline{U}) = 0$. Substituting into $\widehat{\theta} = -\widehat{b}/\widehat{G}$ gives

$$\widehat{\theta} - \gamma = \eta \sqrt{r/q}\, \frac{\sum_{i=1}^{N} (U_i - \overline{U}) V_i}{\sum_{i=1}^{N} (U_i - \overline{U})^2}. \tag{116}$$

The denominator is positive almost surely. The numerator is the empirical correlation between the transported block and the residual block; it is the only source of coefficient error in this model.

Let $S_U = \sum_i (U_i - \overline{U})^2$. Conditional on the U_i, the numerator in equation (116) is a linear combination of independent standard normal V_i with mean zero and variance S_U. Thus, for $s > 0$,

$$\mathbb{P}\Big(|\widehat{\theta} - \gamma| > |\eta| \sqrt{r/q} \sqrt{2s/S_U} \,\Big|\, U_1, \ldots, U_N\Big) \leq 2e^{-s}.$$

Also $S_U \sim \chi^2_{N-1}$. Writing $n = N - 1$, Markov's inequality applied to $e^{-S_U/2}$ gives

$$\mathbb{P}(S_U \leq n/2) \leq e^{n/4} \mathbb{E} e^{-S_U/2} = e^{n/4} 2^{-n/2} \leq e^{-n/16}.$$

This controls the possibility that the empirical feature variance is too small to estimate the coefficient stably. On $S_U \geq n/2$, the conditional normal bound has threshold at most $2|\eta|\sqrt{r/q}\sqrt{s/n}$. Combining the two tail bounds therefore controls both the numerator fluctuation and the random denominator.

Let $J = K + 1$ be the number of fitted nodes. If $n \geq 16 \log(2J/\delta_{\mathrm{fit}})$, choose $s = \log(4J/\delta_{\mathrm{fit}})$. The denominator failure probability at each node is at most $\delta_{\mathrm{fit}}/(2J)$, and the conditional normal failure

probability is at most the same amount. A union bound over nodes gives, with probability at least $1 - \delta_{\mathrm{fit}}$,

$$\max_j \sqrt{q}\,|\widehat{\theta}_j - \gamma| \le 2|\eta|\sqrt{r}\sqrt{\frac{\log(4J/\delta_{\mathrm{fit}})}{N-1}}. \tag{117}$$

The union bound allows dependence between the nodewise estimates; the single-node calculation uses the stated independent fitting observations.

The population coefficient is constant across t, so averaging endpoint fits adds no interpolation error. Since the scalar Gram matrix is q, the projection identity gives the excess squared residual $q(\overline{\theta}_{k,1} - \gamma)^2$ at every point on the edge. Its integrated square root is $h_k\sqrt{q}\,|\overline{\theta}_{k,1} - \gamma|$. The endpoint average is bounded by the larger endpoint error in equation (117), which gives equation (115).

For independent fitting observations, a fixed error margin therefore requires only logarithmic growth in the number of nodes and confidence level in this scalar class. The validation requirement has the same logarithmic dependence on library size. These are sample requirements once suitable observations are available; the cost of obtaining them from preliminary PT also depends on the concentration, block-dependence, and initialization conditions. Together, the calculations show that the finite selected sampler attains the within-class rate bound whenever those calibration and fitting conditions hold. The accuracy with which selection identifies the first sufficient class is considered next.

G Accuracy and cost of complexity selection

Theorem 2 shows that CAT-PT stops no later than a transport level that achieves the required exchange performance. It does not, however, guarantee that the selected level is the simplest one whose true acceptance already meets the target $a_\star$.

This section studies that stronger selection property. We first show that, when the first sufficient candidate has enough margin above $a_\star$, the held-out sequential rule identifies this candidate exactly. We then quantify the computational cost of searching through the preceding levels and compare it with an oracle that knows the correct level in advance. Finally, we characterize the number of validation samples needed to resolve a given acceptance margin.

G.1 Minimum sufficient level and oracle search cost

The intended benefit of sequential selection is to avoid unnecessary transport complexity. We therefore compare its choice with the first *fitted candidate* whose stationary acceptance reaches the target. Fix one edge and order its candidates as $0, 1, \dots, M$, with raw exchange at zero. Define

$$m^\dagger := \min\{m : A_m \ge a_\star\}, \tag{118}$$

with $m^\dagger = \infty$ if no level reaches the target. The index is determined by the candidate acceptances after fitting, rather than by the best map that could hypothetically be found in each class. Exact recovery requires the held-out rule to reject earlier insufficient candidates and to pass at $m^\dagger$.

An oracle that knows $m^\dagger$ fits and evaluates that level directly. CAT-PT also evaluates the preceding levels. Let F_m^{screen} be the cost of fitting, constructing, and validating a positive level m from scratch on the edge. These costs also provide upper budgets for incremental screening with reused statistics. We account for the shared initial raw-validation cost separately. When successive screening costs grow geometrically, the sum of preceding costs is controlled by the cost of the last level.

Theorem 7 (Minimum sufficient complexity and oracle screening cost)**.** *Suppose $m^\dagger < \infty$ and $A_{m^\dagger} \ge a_\star + 2\omega$. On $\mathsf{V}(\omega)$, CAT-PT selects $\widehat{m} = m^\dagger$. If also $m^\dagger \ge 1$ and $F_{m+1}^{\mathrm{screen}} \ge qF_m^{\mathrm{screen}}$ for some $q > 1$, then*

$$\sum_{m=1}^{\widehat{m}} F_m^{\mathrm{screen}} \le \frac{q}{q-1} F_{m^\dagger}^{\mathrm{screen}}. \tag{119}$$

For $m^\dagger = 0$, both procedures incur zero positive-level screening cost.

Proof. By the definition of $m^\dagger$, each earlier level $m < m^\dagger$ has $A_m < a_\star$. On the uniform event,

$$\widehat{A}_m - \omega \le A_m < a_\star,$$

so none of these levels passes the selection rule. At $m^\dagger$,

$$\widehat{A}_{m^\dagger} - \omega \geq A_{m^\dagger} - 2\omega \geq a_\star.$$

Thus the sequential search reaches $m^\dagger$ and stops there, proving $\widehat{m} = m^\dagger$.

We next sum the screening costs of the levels that were visited. For $m \leq m^\dagger$, repeated use of the geometric growth condition gives

$$F_m^{\mathrm{screen}} \leq q^{m-m^\dagger} F_{m^\dagger}^{\mathrm{screen}}.$$

We have just shown that $\widehat{m} = m^\dagger$ on the stated event. Therefore

$$\begin{aligned}
\sum_{m=1}^{\widehat{m}} F_m^{\mathrm{screen}} &\leq F_{m^\dagger}^{\mathrm{screen}} \sum_{m=1}^{m^\dagger} q^{m-m^\dagger} \\
&= F_{m^\dagger}^{\mathrm{screen}} \sum_{j=0}^{m^\dagger-1} q^{-j} \\
&\leq F_{m^\dagger}^{\mathrm{screen}} \sum_{j=0}^{\infty} q^{-j} = \frac{q}{q-1} F_{m^\dagger}^{\mathrm{screen}}.
\end{aligned}$$

When $m^\dagger = 0$, the first validation step selects raw exchange, so neither procedure fits a positive-level map. □

The result identifies the minimum sufficient level among the fitted candidates. Uniform validation prevents an earlier candidate with true acceptance below $a_\star$ from passing. The extra 2ω at $m^\dagger$ ensures that this first sufficient candidate does pass. Thus exact recovery uses the same error bound as Proposition 3, with a margin required specifically at the first sufficient level. It does not require acceptance to increase monotonically along the hierarchy.

The oracle comparison quantifies the price of not knowing that level. Geometric growth makes the sum of earlier screening costs comparable to the final cost. This can occur when regression dominates and basis sizes grow geometrically, but the hypothesis is on the full screening costs, including validation. Reusing feature moments and nodewise fits further reduces actual work relative to these from-scratch budgets, as described in Appendix H.1.

An expected-cost bound also applies when the required level varies with the fitting data. Let $M^\star$ be the first level with $A_m \geq a_\star + 2\omega$, setting $M^\star = \infty$ when none exists. Use deterministic screening budgets F_m^{screen}. Suppose the uniform validation event fails with probability at most δ and total screening cost on failure is bounded by $C_{\max}$. On the success event, reaching level m implies $M^\star \geq m$: otherwise the search would already have stopped at an earlier margin-sufficient level. Hence

$$\mathbb{E}C_{\mathrm{screen}}^{\text{CAT-PT}} \leq \sum_{m=1}^{M} F_m^{\mathrm{screen}} \mathbb{P}(M^\star \geq m) + \delta C_{\max}. \tag{120}$$

To see this directly, write the positive-level cost as a sum of level costs multiplied by their visit indicators. On the success event each visit indicator is bounded by $\mathbf{1}\{M^\star \geq m\}$; on failure the total is bounded by $C_{\max}$. Taking expectations yields equation (120). This expression shows how often richer classes need to be reached, as well as how much each such class costs.

G.2 Sample complexity of acceptance-based selection

A candidate close to the acceptance target is harder to distinguish from an inadequate one. The upper bound in equation (34) requires an effective sample size proportional to γ^{-2} to resolve an acceptance difference of size γ, up to confidence and library-size factors. We ask whether this inverse-square dependence can be improved.

The following lower bound answers the question for general iid observations in $[0, 1]$. It considers a single acceptance comparison: deciding whether the mean equals $a_\star$ or is larger by γ. A procedure valid for all such bounded observations must also work for Bernoulli observations, which provide the testing example used in the proof.

Theorem 8 (Validation samples required to resolve an acceptance margin)**.** *Fix $\eta \in (0, 1/2)$ and $\delta \in (0, 1/4)$, let $a_\star \in [\eta, 1-\eta]$, and let $0 < \gamma \le \eta/2$. Any procedure based on n iid observations in $[0, 1]$ that distinguishes*

$$H_0 : A = a_\star \qquad \textit{from} \qquad H_1 : A = a_\star + \gamma \tag{121}$$

with both error probabilities at most δ, uniformly over observation distributions in the two hypotheses, must satisfy

$$n \ge c_\eta \gamma^{-2} \log \frac{1}{4\delta} \tag{122}$$

for a constant $c_\eta > 0$ depending only on η. Consequently, $\omega = O(\sqrt{\log(KM/\delta)/N_{\mathrm{val}}})$ is rate-optimal in the acceptance margin, up to the logarithmic cost of simultaneous validation.

Proof. It suffices to prove the claim on the Bernoulli subclass, because any procedure valid for all observations in $[0, 1]$ must also work for Bernoulli observations. Let $P_0 = \mathrm{Bernoulli}(a_\star)$ and $P_1 = \mathrm{Bernoulli}(a_\star + \gamma)$. For Bernoulli parameters p, q in a compact subinterval of $(0, 1)$,

$$\mathrm{KL}(\mathrm{Bern}(p) \| \mathrm{Bern}(q)) \le \frac{(p-q)^2}{q(1-q)}.$$

Here $q = a_\star + \gamma$ lies in $[\eta, 1 - \eta/2]$ by the stated range of $a_\star$ and γ. Therefore $q(1-q)$ is bounded below by a positive constant depending only on η, and

$$\mathrm{KL}(P_0 \| P_1) \le C_\eta \gamma^2. \tag{123}$$

For any test ϕ, the Bretagnolle–Huber inequality gives

$$P_0^n(\phi = 1) + P_1^n(\phi = 0) \ge \frac{1}{2} \exp\{-n \,\mathrm{KL}(P_0 \| P_1)\}.$$

If both errors are at most δ, their sum is at most 2δ. Combining this fact with equation (123) yields

$$2\delta \ge \frac{1}{2} e^{-nC_\eta \gamma^2}.$$

Taking logarithms and rearranging gives $n \ge C_\eta^{-1} \gamma^{-2} \log(1/(4\delta))$, which is equation (122) with $c_\eta = C_\eta^{-1}$. □

The theorem shows that the inverse-square dependence on the acceptance gap is unavoidable in this bounded-observation model. For the Bernoulli pair used in the proof, each observation carries only order-γ^2 information about the difference in means, so reliable discrimination requires order-γ^{-2} observations at fixed confidence. This matches the margin dependence of the held-out upper bound. The lower bound concerns generic acceptance observations; it does not assert the same difficulty for every particular target or transport family.

The additional factor $\log(KJ)$ in equation (34) comes from controlling a whole finite library at once. The testing result addresses the one-comparison margin cost; the library factor in CAT-PT follows from simultaneous validation. For a correlated calibration trajectory, the effective sample size and the initialization and block-separation terms further determine how many retained iterations are needed. Once these statistical requirements are met, the computational cost is governed by the classes actually visited and deployed, which we analyze next.

H Computational cost of CAT-PT

Round trips per sweep measure communication, but the work required for a sweep also matters. CAT-PT can save work by stopping before expensive transport classes are fitted and by using only the selected map during production. This section quantifies both costs for the structured hierarchy and compares them with a fixed expressive transport at a common guaranteed acceptance target. We also account separately for calibration, local MCMC, and target evaluations, so the transport-cost bounds can be related to end-to-end runtime.

H.1 Fitting and evaluating structured transports

A simple map can be inexpensive to apply without being inexpensive to fit. We therefore calculate these costs separately. For the affine classes used here, the velocity fitted on an edge has an explicit flow. Its matrices and Jacobian can be precomputed, after which each exchange applies the stored forward and inverse maps. Fitting instead requires empirical Stein moments and a linear solve whose size is the number of velocity parameters. We write $\mathrm{Flow}(h, \cdot)$ for the time-h solution map of the indicated autonomous velocity field.

For a translation field $\mathbf{v}(\mathbf{x}) = \mathbf{b}$, integration gives

$$\mathrm{Flow}(h, \mathbf{b})(\mathbf{x}) = \mathbf{x} + h\mathbf{b}, \qquad \log|\det \nabla \mathrm{Flow}| = 0. \tag{124}$$

For $v_i(\mathbf{x}) = a_i x_i + b_i$, multiplying the scalar ODE by $e^{-a_i u}$ and integrating from 0 to h gives

$$e^{-ha_i} X_i(h) - x_i = \int_0^h e^{-ua_i} b_i \, du.$$

Consequently, with $\varphi_1(z) = (e^z - 1)/z$ and $\varphi_1(0) = 1$,

$$[\mathrm{Flow}(h, \mathbf{a}, \mathbf{b})(\mathbf{x})]_i = e^{ha_i} x_i + h\varphi_1(ha_i) b_i, \qquad \log|\det \nabla \mathrm{Flow}| = h \sum_i a_i. \tag{125}$$

The continuous definition at zero covers pure translation coordinates without an ill-conditioned division. In numerical code, $\mathrm{expm1}$ or its series expansion evaluates φ_1 near zero.

For a block-affine field $\mathbf{v}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$, the integrating-factor argument gives

$$\mathrm{Flow}(h, \mathbf{A}, \mathbf{b})(\mathbf{x}) = e^{h\mathbf{A}}\mathbf{x} + \int_0^h e^{(h-u)\mathbf{A}}\mathbf{b} \, du, \qquad \log|\det \nabla \mathrm{Flow}| = h \operatorname{tr} \mathbf{A}. \tag{126}$$

The integral is well defined even if $\mathbf{A}$ is singular. It can be computed by an augmented matrix exponential or the matrix φ_1 function, rather than by explicitly inverting $\mathbf{A}$. For all these autonomous flows the inverse is the same formula with h replaced by $-h$. If the maximum block size is a fixed q_0, exponentials are formed once at cost $O(dq_0^2)$ and each subsequent map or inverse evaluation costs $O(dq_0)$. Both are linear in d when q_0 is fixed. For affine maps the forward and inverse log Jacobians are constants of opposite sign; their sum in the exchange ratio cancels, although keeping both terms in a generic implementation makes the formula applicable to non-affine maps as well.

With fixed block size, applying the stored map takes work proportional to d. The covariance regression has a different dependence: its cost grows with the number p of fitted Stein features. The hierarchy can therefore save fitting work even when several levels have map-evaluation costs that are all linear in the state dimension.

The nested hierarchy permits reuse of these fitting quantities. Suppose its classes are prefixes of an ordered basis and the largest reached nodewise level has p directions with N fitting observations. Maintain the empirical feature means, feature second moments, and feature–g_t cross moments. Adding a direction adds one row and column to the second moment matrix. Centering these moments gives $\widehat{\mathbf{G}}$ and $\widehat{\mathbf{b}}$ for every visited prefix.

When the empirical Gram matrices are positive definite, nested Cholesky updates give the total arithmetic and storage requirements, after feature evaluation,

$$\begin{aligned} &O(Np^2 + p^3) \quad \text{arithmetic operations,} \\ &O(p^2) \quad \text{additional sufficient-statistic storage.} \end{aligned} \tag{127}$$

To derive this bound, at basis dimension j the new sample feature is multiplied by the j features available at that level. Accumulating the new row and column costs $O(Nj)$, and summing over $j \le p$ gives $O(N \sum_j j) = O(Np^2)$. Feature means and cross moments cost $O(Np)$; subtracting products of means produces the centered covariance and cross-covariance.

A Cholesky extension from $j-1$ to j solves a triangular system and computes one additional diagonal entry, costing $O(j^2)$. Summation gives $O(\sum_{j=1}^{p} j^2) = O(p^3)$. Solving for the coefficient at each visited prefix has the same cubic total upper bound. The sufficient statistics and factorization occupy

$O(p^2)$ entries. Every smaller Gram matrix is a principal prefix of the largest one, so its products are available from these updates. If a node is incident to two unresolved edges, its fitted coefficient is computed once and used by both.

Proposition 4 gives a sample condition under which the Gram matrix is positive definite; a fixed positive ridge also supports Cholesky updates. Singular unregularized fits can use a pseudoinverse, with the cost determined by that factorization routine. Streaming sufficient statistics requires $O(p^2)$ working memory in addition to stored observations. Keeping all feature values instead adds $O(Np)$ storage.

Feature evaluation contributes a separate cost before the linear system is solved. For global, diagonal, and bounded-block affine directions, the features can be assembled in $O(d)$ work per sample, up to fixed block-size factors, once the spatial score is available. Evaluating the score is a target cost. Thus inexpensive flow application and covariance fitting have different scalings: a large diagonal class can have a cheap map but a much larger regression than a scalar global class. Early stopping saves the work of those unreached feature directions and linear systems.

H.2 Cost of sequential search and frozen production

Sequential selection pays for the candidates it visits, whereas production pays repeatedly for the map it selects. We first describe the shared acceptance computation and then express these two costs in terms of the selected levels. Each candidate is evaluated using the acceptance probability of its production kernel, so one log-ratio routine serves both stages. On stored validation pairs, we average $\exp\{\min(0, \log R_T)\}$. During production, we draw $U \sim \mathrm{Unif}(0,1)$ independently and accept when $\log U < \min(0, \log R_T)$. Averaging probabilities in validation gives the same mean acceptance as averaging accept/reject indicators, without the additional Bernoulli sampling noise.

Current-state log densities in the denominator can be cached. An accepted exchange replaces the states and cached destination-density values; after a rejection the old cache remains valid. For temperature annealing, one stored energy can be used at every inverse temperature. For likelihood tempering, a stored log-prior/log-likelihood pair serves the same role. The factorization of the path determines which target evaluations can be shared.

Once the selected level is frozen, a structured edge evaluates its map at every exchange opportunity, while a raw-selected edge evaluates no map. Let $\widehat{m}_k = \widehat{c}_k$ and $\mathcal{T} = \{k : \widehat{m}_k > 0\}$. Write $F^{\mathrm{screen}}_{k,m}$ for an upper budget for fitting, constructing, and validating level m on unresolved edge k. Counting a shared nodewise fit on both adjacent edges gives an upper bound; actual reuse lowers the work. Let $c_{k,m}$ bound the cost of one forward–inverse map pair and its log Jacobians, with $c_{k,0} = 0$ and $c_{k,m}$ nondecreasing in m.

To bound the levels that may be visited, we use the first candidate with enough acceptance margin to pass the held-out rule. Thus the threshold in the following definition is $a_\star + 2\omega$, rather than the $a_\star$ threshold used for exact recovery in equation (118):

$$m^\dagger_k := \min\{c \in \{0, \dots, M\} : A_{k,c} \geq a_\star + 2\omega\}, \tag{128}$$

with the empty minimum equal to infinity. Let F_0 be the total initial raw-validation cost and B_k the number of production opportunities on edge k.

Proposition 8 (Adaptive screening and direct deployment cost)**.** *If $m^\dagger_k < \infty$ on all edges, then on $\mathsf{V}(\omega)$,*

$$C^{\text{CAT-PT}}_{\text{screen}} \leq F_0 + \sum_k \sum_{m=1}^{m^\dagger_k} F^{\mathrm{screen}}_{k,m}, \qquad C^{\text{CAT-PT}}_{\text{maps}} \leq \sum_{k \in \mathcal{T}} B_k c_{k,\widehat{m}_k} \leq \sum_k B_k c_{k,m^\dagger_k}. \tag{129}$$

Here B_k is the number of production opportunities on edge k; an empty inner sum is zero. The opportunity counts and cost bounds are evaluated conditional on the frozen choices.

Proof. Proposition 3(2) gives $\widehat{m}_k \leq m^\dagger_k$: the margin-sufficient candidate must be accepted by the selection rule when reached. Thus an edge screens only levels up to $m^\dagger_k$, proving the first inequality after adding raw validation. On each of the B_k production opportunities, a selected structured map is

evaluated once forward and once inversely, regardless of whether the proposal is accepted. An edge with choice zero has no map evaluation. This gives the first production sum. Finally, $c_{k,\widehat{m}_k} \le c_{k,m_k^\dagger}$ and $c_{k,0} = 0$ give the second bound. □

The bound ties work to the complexity needed on each edge, rather than the largest class in the library. Screening stops once a sufficient candidate passes. Production repeatedly uses only the chosen map; raw-selected edges use none. Every transported attempt evaluates its map even when the proposal is rejected, so the production saving comes from which map is selected, not from the rejection frequency. These costs can now be compared with fitting and deploying a fixed expressive map on every edge.

The finite-library rule can also include a fitted expressive map after the structured hierarchy. Its fitting routine is fixed independently of validation, and its direct exchange is evaluated by the same log-ratio routine. Including it in J extends the uniform validation and stopping bound of Proposition 3. The invariance argument in Appendix C.2 applies to its invertible map, while the Stein residual bounds describe candidates generated by the structured velocity classes. Such a candidate provides an additional level for edges whose changes are poorly represented by the inexpensive hierarchy.

H.3 Cost comparison with fixed expressive transports

The comparison asks how much transport work each method needs to meet the same acceptance target on the same ladder. Suppose the baseline fits a prescribed expressive map on each edge, as in a fixed neural-transport APT implementation (Zhang et al., 2026), and uses direct transported Metropolis–Hastings with the same production opportunity counts B_k. Let F_k^{exp} include its fitting, construction, and validation, and let c_k^{exp} be the cost of a forward–inverse pair and its log Jacobians. Evaluate the two fitted libraries under a simultaneous validation-accuracy event covering both methods, including the baseline candidates in the finite-library bound. Passing the same acceptance target then gives a common DEO/ELE round-trip lower bound. The comparison below concerns the work used to obtain and deploy these choices, not equality of their realized round-trip rates.

Proposition 9 (Transport-cost comparison at a common acceptance target)**.** *Under the conditions of Proposition 8, suppose the expressive baseline also passes the held-out rule at $a_\star$. On the simultaneous validation-accuracy event for both methods, they have the same guaranteed DEO/ELE round-trip lower bound in equation* (8)*, while their screening-plus-map costs obey*

$$C^{\mathrm{acc}}_{\text{CAT-PT}} \le F_0 + \sum_k \sum_{m=1}^{m_k^\dagger} F^{\mathrm{screen}}_{k,m} + \sum_k B_k c_{k,m_k^\dagger}, \tag{130}$$

$$C^{\mathrm{acc}}_{\mathrm{exp}} = \sum_k F_k^{\mathrm{exp}} + \sum_k B_k c_k^{\mathrm{exp}}. \tag{131}$$

Along a problem sequence, if

$$\frac{F_0 + \sum_{k,m\le m_k^\dagger} F^{\mathrm{screen}}_{k,m}}{\sum_k F_k^{\mathrm{exp}}} \longrightarrow 0, \qquad \frac{\sum_k B_k c_{k,m_k^\dagger}}{\sum_k B_k c_k^{\mathrm{exp}}} \longrightarrow 0, \tag{132}$$

with positive denominators, then $C^{\mathrm{acc}}_{\text{CAT-PT}} = o(C^{\mathrm{acc}}_{\mathrm{exp}})$.

Proof. Proposition 8 bounds CAT-PT's two cost components. The baseline pays F_k^{exp} once and c_k^{exp} on each of the B_k opportunities, yielding equation (131). Both methods pass the rule at the same threshold, so Proposition 1 supplies the same lower bound on their round-trip rates. To prove the final statement, write S and P for the two nonnegative baseline components. For every $\epsilon > 0$, equation (132) eventually bounds CAT-PT's screening term by ϵS and its map term by ϵP. Their sum is at most $\epsilon(S + P)$, proving the claimed little-o relation. □

The two ratios in equation (132) concern separate sources of savings. The first requires the entire search through the sufficient structured levels to be cheap relative to fitting the expressive maps. The second requires the selected maps to be cheap over their production opportunities. Explicit affine flows determine the structured map costs; the baseline architecture, optimizer, and training budget determine its costs. An accelerator with its own adaptive architecture can be compared by inserting

its search and deployment costs in the same expressions. Thus inexpensive sufficient classes give a transport-side saving under the stated cost ratios. A comparison of total runtime also needs the operations that these ratios do not count.

We separate those operations in the total-cost calculation:

$$C_{\text{total}} = C_{\text{pilot+calibration}} + C_{\text{screen}} + C_{\text{local}} + C_{\text{maps}} + C_{\text{exchange targets}} + C_{\text{other}}. \tag{133}$$

Each evaluation is charged to the stage where it occurs. Target calls made while fitting and validating belong to C_{screen}; the same calls are not counted again as production work. Local MCMC contributes C_{local}, and computing densities at transformed production states contributes $C_{\text{exchange targets}}$.

For an energy path $\widetilde{\pi}_t(\mathbf{x}) = e^{-\beta(t)U(\mathbf{x})}$, current-state energies can be reused at all temperatures. Raw exchange then needs no new energy call, while direct transported exchange needs the energies of its two transformed states. A generic path requires the two destination log densities once current-state densities are cached; whether they can be computed from shared target quantities depends on the path factorization. This is why map arithmetic, target-call counts, and wall-clock time are useful separate measurements. Their relative contribution depends on the target and implementation.

A simple amortization calculation illustrates the role of production length. Suppose the two methods attain equal stationary communication and have initialization costs S_C, S_E and average production costs per sweep p_C, p_E. After B sweeps, CAT-PT costs less exactly when

$$S_C - S_E + B(p_C - p_E) < 0. \tag{134}$$

If both differences are nonpositive and one is strict, it costs less at every positive horizon. If $S_C < S_E$ but $p_C > p_E$, its initial saving lasts until $B < (S_E - S_C)/(p_C - p_E)$. When the measured round-trip rates differ, each method's resource use is instead divided by its own number of completed journeys. Reporting pre-production work, production throughput, and total throughput therefore distinguishes savings in fitting from savings during sampling. The quantities in equation (133) also show when cheaper transports are a substantial part of the overall computational budget.

I Additional interpretations of the Stein fitting objective

The squared Stein residual has two further interpretations. For a short annealing step, it determines the leading quadratic term of the symmetric KL mismatch between the transported source and destination distribution. For a family with several annealing parameters, minimizing the same residual defines a matrix that measures how much change remains in each parameter direction. We derive these two connections and explain what they imply for transport fitting and possible path design.

I.1 Local symmetric-KL error after transport

Does minimizing the Stein residual also reduce a familiar measure of transport error? For a short step, the answer is yes for the leading term of symmetric KL divergence. We compare the transported source with the destination by undoing the map on the latter. The resulting law is $Q^{\mathbf{v}}_{t,h} = (T^{\mathbf{v}}_{t\to t+h})^{-1}_{\#}\pi_{t+h}$, with log-density ratio $D^{\mathbf{v}}_{t,h}$ relative to π_t, as in equation (63). To control the KL expansion, suppose in addition to Assumption 5 that

$$D^{\mathbf{v}}_{t,h} = h\rho^{\mathbf{v}}_t + O_{L^3(\pi_t)}(h^2), \qquad \mathbb{E}_{\pi_t}[|D^{\mathbf{v}}_{t,h}|^3 e^{|D^{\mathbf{v}}_{t,h}|}] = O(h^3). \tag{135}$$

The first condition controls the log ratio beyond its linear term; the second controls the exponential remainder when converting it back into a density ratio. Together they allow the expansion to be integrated in the KL calculation.

Define $\mathrm{SKL}(P,Q) = \frac{1}{2}\{\mathrm{KL}(P\|Q) + \mathrm{KL}(Q\|P)\}$. Under these conditions,

$$\mathrm{SKL}\big(\pi_{t+h}, (T^{\mathbf{v}}_{t\to t+h})_{\#}\pi_t\big) = \frac{h^2}{2}\mathbb{E}_{\pi_t}[(s_t + \mathcal{S}_t\mathbf{v}_t)^2] + O(h^3). \tag{136}$$

The coefficient of $h^2/2$ is the population Stein regression objective. Within a velocity class, the population fit therefore minimizes the leading local symmetric-KL mismatch. The conclusion

concerns the short-step expansion in equation (136); a finite-gap rejection guarantee is supplied separately by Theorem 1. The rejection expansion in equation (66) instead involves the mean absolute difference of the same residual. These two local quantities describe different aspects of one remaining density change: its second moment controls local KL sensitivity, while its absolute differences determine first-order exchange rejection.

To derive equation (136), with $D = \log(dQ/dP)$, we have $dQ = e^D dP$ and $\mathbb{E}_P e^D = 1$. The two directions of KL are therefore

$$\mathrm{KL}(P\|Q) = -\mathbb{E}_P D,$$
$$\mathrm{KL}(Q\|P) = \mathbb{E}_Q D = \mathbb{E}_P[e^D D].$$

Their average gives

$$\mathrm{SKL}(P,Q) = \tfrac{1}{2}\mathbb{E}_P[(e^D - 1)D]. \tag{137}$$

This form is convenient because the leading term of $(e^D - 1)D$ is D^2. The scalar bound $|(e^u - 1)u - u^2| \le C|u|^3 e^{|u|}$ and the second part of equation (135) yield

$$\mathbb{E}_P[(e^D - 1)D] = \mathbb{E}_P D^2 + O(h^3).$$

For $P = \pi_t$, write $D = h\rho_t^{\mathbf{v}} + R_h$ with $\|R_h\|_{L^3(\pi_t)} = O(h^2)$. Its second moment expands as

$$\mathbb{E}_P D^2 = h^2 \mathbb{E}_P[(\rho_t^{\mathbf{v}})^2] + 2h\mathbb{E}_P[\rho_t^{\mathbf{v}} R_h] + \mathbb{E}_P R_h^2.$$

On a probability space, $\|R_h\|_2 \le \|R_h\|_3 = O(h^2)$. Cauchy–Schwarz bounds the cross term by $2h\|\rho_t^{\mathbf{v}}\|_2\|R_h\|_2 = O(h^3)$, and $\mathbb{E}_P R_h^2 = O(h^4)$. Hence

$$\mathbb{E}_P[(e^D - 1)D] = h^2 \mathbb{E}_{\pi_t}[(\rho_t^{\mathbf{v}})^2] + O(h^3).$$

Finally, KL is unchanged when the same invertible map is applied to both measures: the two Jacobian factors cancel in the density ratio. Applying $T^{\mathbf{v}}_{t\to t+h}$ to π_t and $Q^{\mathbf{v}}_{t,h}$ therefore gives the endpoint expression in equation (136).

I.2 Stein fitting with multiple annealing parameters

With several annealing parameters, such as multiple likelihood weights, some directions may be easier to transport than others. We extend Stein fitting to quantify this difference. At each point in the parameter family, a matrix records the minimum residual variance for every possible direction of change. This provides a transport-adjusted measure of path variation while keeping the same velocity class.

Let $\boldsymbol{\lambda} \in \mathbb{R}^q$ index the density family and define $\mathbf{s}_{\boldsymbol{\lambda}}(\mathbf{x}) = \nabla_{\boldsymbol{\lambda}} \log \pi_{\boldsymbol{\lambda}}(\mathbf{x})$. For a path direction $\mathbf{u} = \dot{\boldsymbol{\lambda}}$, its scalar score is $\mathbf{u}^\top \mathbf{s}_{\boldsymbol{\lambda}}$. Let $\boldsymbol{\psi}_m$ collect the Stein features and write

$$\mathbf{I}_{\boldsymbol{\lambda}} = \mathbb{E}[\mathbf{s}_{\boldsymbol{\lambda}}\mathbf{s}_{\boldsymbol{\lambda}}^\top], \qquad \mathbf{G}_m = \mathbb{E}[\boldsymbol{\psi}_m\boldsymbol{\psi}_m^\top], \qquad \mathbf{C}_m = \mathbb{E}[\boldsymbol{\psi}_m\mathbf{s}_{\boldsymbol{\lambda}}^\top]. \tag{138}$$

The first matrix is the covariance of the parameter scores: its quadratic form gives the variance of a directional score before transport. The second is the feature covariance, and the third records the correlations between Stein features and parameter scores. The projection formula subtracts the score variation explained by those features.

Proposition 10 (Minimum Stein residual for a multiparameter path)**.** *For every path direction* $\mathbf{u}$*,*

$$\inf_{\boldsymbol{\theta}} \mathbb{E}[(\mathbf{u}^\top \mathbf{s}_{\boldsymbol{\lambda}} + \boldsymbol{\psi}_m^\top \boldsymbol{\theta})^2] = \mathbf{u}^\top \mathbf{M}_m(\boldsymbol{\lambda})\mathbf{u}, \qquad \mathbf{M}_m := \mathbf{I}_{\boldsymbol{\lambda}} - \mathbf{C}_m^\top \mathbf{G}_m^\dagger \mathbf{C}_m. \tag{139}$$

Moreover,

$$\mathbf{0} \preceq \mathbf{M}_m(\boldsymbol{\lambda}) \preceq \mathbf{I}_{\boldsymbol{\lambda}}, \tag{140}$$

and nested velocity classes decrease $\mathbf{M}_m$ *in positive-semidefinite order.*

Proof. Apply Proposition 2 to the scalar score $s_{\mathbf{u}} = \mathbf{u}^\top \mathbf{s}_{\boldsymbol{\lambda}}$. Its variance is $\mathbf{u}^\top \mathbf{I}_{\boldsymbol{\lambda}} \mathbf{u}$, and its cross-covariance with $\boldsymbol{\psi}_m$ is $\mathbf{C}_m \mathbf{u}$. The projection formula therefore gives

$$\begin{aligned}\inf_{\boldsymbol{\theta}} \mathbb{E}[(s_{\mathbf{u}} + \boldsymbol{\psi}_m^\top \boldsymbol{\theta})^2] &= \mathbf{u}^\top \mathbf{I}_{\boldsymbol{\lambda}} \mathbf{u} - (\mathbf{C}_m \mathbf{u})^\top \mathbf{G}_m^\dagger (\mathbf{C}_m \mathbf{u}) \\ &= \mathbf{u}^\top (\mathbf{I}_{\boldsymbol{\lambda}} - \mathbf{C}_m^\top \mathbf{G}_m^\dagger \mathbf{C}_m)\mathbf{u},\end{aligned}$$

which proves equation (139). The left-hand side is nonnegative for every $\mathbf{u}$, so $\mathbf{M}_m \succeq 0$. Choosing the zero velocity is feasible and gives residual variance $\mathbf{u}^\top \mathbf{I}_{\boldsymbol{\lambda}} \mathbf{u}$, so the infimum is no larger; hence $\mathbf{M}_m \preceq \mathbf{I}_{\boldsymbol{\lambda}}$. Finally, nesting enlarges the feasible velocity set, so the quadratic form decreases for every $\mathbf{u}$, which is exactly positive-semidefinite ordering. □

The optimal coefficient for direction $\mathbf{u}$ is $-\mathbf{G}_m^\dagger \mathbf{C}_m \mathbf{u}$. The projection result in Appendix B ensures that $\mathbf{C}_m \mathbf{u}$ belongs to the range of $\mathbf{G}_m$, so this expression also covers redundant Stein features. The subtracted term $\mathbf{C}_m^\top \mathbf{G}_m^\dagger \mathbf{C}_m$ measures the part of the parameter-score variation removed by structured transport.

In a direction $\mathbf{u}$ with $\mathbf{u}^\top \mathbf{M}_m \mathbf{u} = 0$, the available velocity explains the entire directional score in L^2. In a direction with $\mathbf{C}_m \mathbf{u} = 0$, the Stein features are orthogonal to that score, and the residual magnitude remains $\mathbf{u}^\top \mathbf{I}_{\boldsymbol{\lambda}} \mathbf{u}$. Thus the positive-semidefinite matrix $\mathbf{M}_m$ distinguishes directions that transport can simplify from directions that still require overlap between neighboring distributions.

Along a differentiable path $\boldsymbol{\lambda}(t)$, the corresponding residual length is

$$\int_0^1 \sqrt{\dot{\boldsymbol{\lambda}}(t)^\top \mathbf{M}_m(\boldsymbol{\lambda}(t)) \dot{\boldsymbol{\lambda}}(t)}\, dt. \tag{141}$$

The path direction determines the score to be matched, and the velocity class determines which part can be removed. Enlarging the class decreases the quadratic form in every direction. The resulting matrix can be viewed as a projected thermodynamic metric. It suggests choosing paths that favor parameter directions whose density changes are well represented by the available motions. The matrix thus provides an objective for comparing paths according to the variation that a specified transport class leaves unexplained.

J Additional synthetic experiments

This section gives the implementation and evaluation details for the synthetic experiments, together with additional results that are not shown in the main text. We first describe the common sampling protocol, the implementations of all compared methods, and the evaluation metrics. We then give the complete construction of each benchmark, followed by the shifted-quartic experiment and the additional transported baselines on the 40-mode Gaussian mixture.

Common sampling and calibration protocol. Within each target, all methods use the same annealing path, temperature ladder, local MCMC kernel, production length, and production initialization. We use deterministic even–odd (DEO) communication throughout. Before transport fitting, Raw DEO-PT is used to calibrate the common temperature ladder and the local HMC step size using rejection-rate feedback. These quantities are then frozen and shared by Raw PT, CAT-PT, NF-APT, and the corresponding ablations. Consequently, differences in production performance are due to the exchange kernels rather than independently tuned temperature ladders or local samplers.

An *offline fit* denotes one complete repetition of this pre-production procedure. Each fit independently performs calibration and transport construction before running a collection of independent production populations. Whenever methods are compared within a fit, we use the same production seeds when available.

CAT-PT implementation. The hierarchy is fixed before fitting. Level 0 is Raw exchange. Level 1 combines an arbitrary d-dimensional translation with a single global isotropic scaling direction, giving $d+1$ parameters. Level 2 replaces the global scale by independent coordinate-wise scaling, giving $2d$ parameters. Level 3 additionally includes all off-diagonal linear directions within prespecified blocks. For blocks of sizes m_b, its parameter count is

$$2d + \sum_b m_b(m_b - 1).$$

With two-dimensional blocks, Levels 1–3 therefore contain $d+1$, $2d$, and $3d$ parameters, respectively. The affine classes used in the experiments have closed-form forward maps, inverses, and log Jacobians, so production does not require a numerical ODE solve.

Table 2: Baseline implementations used in the synthetic experiments. All non-oracle methods share the same target-specific temperature ladder and local MCMC kernel.

Method	Implementation
Raw DEO-PT	Deterministic even–odd PT with the standard untransported swap proposal. The Raw run is also used to calibrate the common ladder and HMC step size before these quantities are frozen.
CAT-PT	Sequential structured hierarchy described above. Candidates are fitted by Stein regression on the fitting block, evaluated by held-out exchange acceptance, and the first certified candidate is frozen on each edge.
NF-APT	Independent implementation of the deterministic normalizing-flow variant of accelerated PT (Zhang et al., 2026). One 20-layer RealNVP is trained separately for every neighboring temperature edge. A frozen checkpoint is selected using held-out symmetric KL. Production uses the resulting deterministic invertible map in the same transported-exchange Metropolis correction used by CAT-PT.
Diag-Affine	Fixed affine ablation obtained by matching coordinate-wise empirical moments between adjacent temperatures. It uses a diagonal linear part and is followed by the exact transported-exchange Metropolis correction.
Full-Affine	Fixed affine ablation obtained by matching the empirical mean and full covariance between adjacent temperatures. This is an internal structured baseline rather than a reproduction of an external algorithm.
Max-Structured	CAT-PT ablation that always deploys the most expressive structured candidate available in the library, rather than stopping at the first sufficient level.
CMCD-APT	Independent adaptation of the controlled-diffusion APT baseline (Zhang et al., 2026). We use a two-step stochastic bridge with a shared neural network that learns the bridge drift, temperature schedule, and diffusion scale. Production uses the corresponding path-space Metropolis–Hastings correction.
Oracle-score Diff-APT	Diffusion APT adaptation (Zhang et al., 2026) used only on the 40-mode GMM. It uses the analytic variance-preserving diffusion score of the mixture and a two-step stochastic path, and therefore represents an oracle-assisted diffusion baseline.

Baseline implementations. Table 2 summarizes the compared exchange mechanisms.

The Diag-Affine, Full-Affine, Max-Structured, and Oracle methods are included only as controlled synthetic ablations. NF-APT is the primary expressive transport baseline in the main text. CMCD-APT and Oracle-score Diff-APT are reported in the additional GMM comparison below.

Evaluation metrics. We evaluate communication under a fixed MCMC budget using completed round trips. For offline fit f and production population p, let $R_{f,p}$ denote the number of completed endpoint-to-endpoint-and-back round trips by all replica identities. We report

$$R = \frac{1}{F}\sum_{f=1}^{F}\frac{1}{P_f}\sum_{p=1}^{P_f} R_{f,p},$$

the mean number of completed round trips per production population. Since all methods within a benchmark use the same number of production iterations, R directly compares temperature-space communication at equal MCMC budget.

We additionally report round trips per second, R/s, obtained by dividing the total number of round trips within each fit by its production wall-clock time. Sampling efficiency is measured by ESS/s for prespecified scalar observables of the cold chain. ESS is computed separately for each independent production population, summed across populations, and divided by production time. When a single summary over several observables is needed, we report the geometric mean of their ESS/s values. ESS has the usual interpretation as the equivalent number of independent samples with the same Monte Carlo precision (Vehtari et al., 2021).

For additional diagnostics, we report minimum edge acceptance in the ambient-dimension experiment, where it directly measures the communication bottleneck along the ladder.

Unless stated otherwise, R/s and ESS/s use production time only.

Table 3: Main synthetic benchmark settings. Populations are independent production populations within each offline fit.

Target	d	Replicas	HMC steps	Fits	Pop./fit
Shifted Gaussian	10	10	5	5	128
Structured Many-Well	32	10	5	5	128
40-mode GMM	10	10	5	5	128
Shifted Quartic	12	6	5	3	128

Main benchmark settings. Table 3 collects the basic run sizes. The target constructions and observables are specified below.

Shifted Gaussian. The Gaussian path is

$$\pi_t = \mathcal{N}\left((-1+2t)\mathbf{1},\, 0.1^2 I_{10}\right), \qquad t \in [0,1].$$

Only the mean changes with t, so the exact finite transport between two annealing levels is a translation. This motion is contained in the first structured CAT-PT class and provides a controlled case in which a low-complexity transport is exactly sufficient. Each offline fit runs 128 independent populations, each with 2,000 settling iterations followed by 100,000 production iterations. The three cold-chain observables are the mean over all coordinates, x_0, and the mean square $d^{-1}\sum_i x_i^2$; all three target expectations are available analytically.

Structured Many-Well. The target is $d = 32$. We partition the canonical coordinates into 16 consecutive two-dimensional blocks and designate the first four blocks as the residual region,

$$\mathcal{J} = \{0, 2, 4, 6\}.$$

Writing

$$v = \exp(-\beta A)x,$$

the annealed target is

$$\log \widetilde{\pi}_\beta(x) = -\frac{1}{2}\|v\|^2 + \beta \sum_{j\in\mathcal{J}} \left(-v_j^4 + 6.5 v_j^2 + 0.5 v_j\right) - \beta\, \mathrm{tr}(A) + C_\beta.$$

The implementation permits general two-dimensional generator blocks, but the reported experiment uses the global structure

$$A = 0.7 I_{32}.$$

Hence the known cross-temperature linear change is a global isotropic rescaling acting on all 32 coordinates.

The nonlinear residual is restricted to the first four two-dimensional blocks, giving a supplied residual region of dimension $r_{\mathrm{str}} = 8$. The double-well potential acts on the first coordinate of each such block, while its paired coordinate remains Gaussian. The four double-well coordinates therefore generate $2^4 = 16$ sign modes at the target distribution. We use ten replicas and five-step HMC with frozen step size 0.35. Each of five independent offline fits uses 128 matched production populations for every method, with 2,000 settling iterations followed by 100,000 production iterations. All methods within a fit share the temperature ladder, local HMC parameters, initialization rule, production seeds, and deterministic even–odd communication schedule. We evaluate the mean of the four active coordinates, their second moment, and the fraction of active coordinates that are positive. The second-moment ESS saturates the trajectory-length cap for all methods, so the Many-Well GM-ESS/s summary uses the first and third statistics only.

40-mode Gaussian mixture. We follow the 40-mode Gaussian-mixture setup used in APT (Midgley et al., 2022; Zhang et al., 2026). The 10-dimensional target is $\pi_1(x) = \frac{1}{40}\sum_{j=1}^{40} \mathcal{N}\left(x; \mu_j,\, 0.025^2 I_{10}\right)$, where the 40 benchmark component locations occupy the first two coordinates and the remaining eight entries of every μ_j are zero. The reference is $\pi_0 = \mathcal{N}(0, I_{10})$, and intermediate targets use the geometric bridge $\widetilde{\pi}_t(x) = \pi_0(x)^{1-t}\pi_1(x)^t$. We use ten replicas and five-step HMC. Each of five offline fits contains 128 independent populations with 2,000 settling iterations followed by 100,000 production iterations. The three cold-chain observables are x_0, x_1, and the mean square. For mode diagnostics, each target sample is assigned to its nearest known mixture component; we then report the number of visited components.

Ambient-dimension scaling. The experiment in Figure 2 isolates ambient dimension from the dimension of the supplied structured correction. For each $d \in \{16, 32, 64, 128, 256\}$, we consider the zero-mean Gaussian path

$$\pi_t = \mathcal{N}(0, \Sigma_t), \qquad \Sigma_t = \operatorname{diag}\left(e^{2ta_1}, \ldots, e^{2ta_d}\right), \qquad t \in [0, 1].$$

All coordinates undergo the same global scale change with coefficient $a_j = 0.75$. In addition, a fixed set S of $r_{\mathrm{str}} = 8$ coordinates has extra scale variation,

$$a_j = 0.75 + \delta_\ell, \qquad j = S_\ell,$$

where

$$\delta = (1.8,\ 1.6,\ 1.4,\ 1.2,\ 1.0,\ 0.9,\ 0.8,\ 0.7).$$

Thus increasing d adds coordinates that follow only the common global scale change, while the number of coordinates requiring the additional structured correction remains fixed at eight.

The structured coordinate set S is sampled separately for each random seed and is supplied to CAT-PT as problem structure. CAT-PT is not given the transport coefficients: both the common scale motion and the correction on S are estimated from samples by Stein regression. We use five random seeds at each ambient dimension, giving 25 independent repetitions in total.

All dimensions use ten replicas equally spaced along the same annealing parameter interval. CAT-PT compares the raw candidate with a rank-8 structured candidate. Kernel selection uses the held-out criterion with target acceptance $a_\star = 0.8$ and validation margin $\omega = 0.015$. At every temperature, the fitting, validation, and final test sets contain $4{,}096$, $8{,}192$, and $16{,}384$ independent samples, respectively, with no sample reuse across the three stages.

Communication is evaluated using 128 independent populations and 10,000 deterministic even–odd exchange sweeps. Raw PT and CAT-PT use the same temperature ladder and sweep budget. To isolate the exchange mechanism from finite local-mixing effects, states at each annealing level are independently refreshed from the corresponding stationary Gaussian distribution between exchange sweeps. The experiment therefore measures temperature-space communication under the ELE setting used in the theoretical analysis, rather than the wall-clock performance of a full local-MCMC implementation.

Shifted quartic product. We additionally consider a 12-dimensional shifted quartic product family. Let

$$q(z) \propto \exp(-z^4/4), \qquad \mu_t = -1 + 2t, \qquad s = 0.6.$$

The annealing path is the product location family $\pi_t(x) = \prod_{i=1}^{12} \frac{1}{s} q\left(\frac{x_i - \mu_t}{s}\right)$. The marginal shape is quartic rather than Gaussian, but the exact cross-temperature motion is again a translation. The experiment therefore tests whether the behavior of the structured translation candidate depends on Gaussianity of the target rather than on the structure of its motion.

We use 6 replicas and five-step HMC. Each of three independent offline fits run 128 populations with 2,000 settling iterations and 50,000 production iterations. The observables are the coordinate mean, x_0, and the mean square; their target values are obtained from the analytic quartic moments.

CAT-PT reaches the schedule-limited round-trip count and substantially improves wall-clock efficiency over NF-APT, showing that the translation structure remains useful beyond the Gaussian setting.

CAT-PT and Max-Structured attain the same schedule-limited communication, but CAT-PT stops at the first structured level rather than requiring the largest candidate. The quartic experiment therefore provides an additional structured case in which the inexpensive CAT-PT candidate is sufficient despite the target having a non-Gaussian marginal shape.

Additional transported baselines on the 40-mode mixture. The main text uses Raw PT and NF-APT as the primary baselines. To broaden the comparison without duplicating Table 1, we additionally evaluate CMCD-APT and Oracle-score Diff-APT on the 40-mode GMM. Table 5 reports only these additional methods; the primary baselines are reported in Table 1. All five methods visit all 40 target components during production, so the differences in R and ESS/s are not explained by differences in observed mode coverage.

Table 4: Shifted-quartic results. R is the mean round-trip count per production population. GM-ESS/s is the geometric mean over the three prespecified cold-chain observables.

Method	R	R/s	GM-ESS/s
Raw PT	36.5	32.6	1,804.1
Diag-Affine	17,145.9	13,395.6	9,533.9
Full-Affine	13,662.8	10,841.8	7,894.5
NF-APT	17,724.7	1,924.0	1,403.0
Max-Structured	24,995.0	15,714.2	16,237.5
CAT-PT	**24,995.0**	**15,975.3**	**16,507.1**

Table 5: Additional transported baselines on the 40-mode GMM.

Method	R	R/s	GM-ESS/s
CMCD-APT	2,686.1	180.1	146.2
Oracle-score Diff-APT	3,179.9	2,364.5	2,525.9

Oracle-score Diff-APT is a particularly strong reference because it uses the analytic diffusion score of the target. Nevertheless, the CAT-PT values in Table 1 remain higher in both estimated round-trip throughput and geometric-mean ESS/s.