DRIFT-ADMISSIBILITY VIA OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce structure-aware optimal transport diagnostics for assessing whether empirical time-series dynamics are compatible with a prescribed predictable-drift class. Our approach translates a pathwise property of stochastic processes into a finite-dimensional projection problem over empirical transition laws, encoding admissibility as row-wise moment constraints on local couplings and quantifying deviations via a Kullback–Leibler projection gap $G_{\lambda,k}$. This yields both a principled diagnostic and a scalable Bregman–Dykstra algorithm that retains the efficiency of Sinkhorn-type methods while incorporating structural constraints. Calibrated against a Brownian reference null, $G_{\lambda,k}$ separates process classes that are indistinguishable by second-order statistics alone—including fractional Brownian motion and BSS processes sharing identical Hurst indices—without asymptotic resampling or continuous-time recovery. This positions the method as a general tool for structure-aware model validation of sequential data.

1 INTRODUCTION

A central challenge in modelling time-series data is determining whether observed dynamics are compatible with a given structural hypothesis. In stochastic analysis, semimartingale structures are a key hypothesis, underpinning arbitrage-free modelling, stochastic integration, and continuous-time financial theory Jacod & Shiryaev (2003); Beiglböck et al. (2011); Rogers (1997). However, verifying such a structure from discrete observations is notoriously difficult: classical characterisations are pathwise or asymptotic and do not translate directly into finite-sample, data-driven diagnostics Jacod & Protter (2012); Barndorff-Nielsen & Shephard (2004); Aït-Sahalia & Jacod (2009); Bibinger et al. (2024). Optimal transport (OT) provides scalable tools for learning couplings, but standard formulations typically constrain only marginals and costs rather than conditional structure.

We bridge these perspectives with a *structure-aware OT projection* that encodes predictable-drift admissibility as row-wise moment constraints on empirical transition laws. Unlike standard Sinkhorn formulations, the constraints act on the *conditional* distribution of increments given the predictor state rather than on global marginals or aggregate costs. This gives a finite-dimensional, observation-driven diagnostic for a bounded predictable drift class inspired by special semimartingales. The resulting *projection gap* $G_{\lambda,k}$ measures the minimal KL deformation required to enforce the chosen constraints; it is a compatibility diagnostic, not a standalone statistical test (Section 6).

In this paper, we make four contributions:

(a) We formulate predictable-drift admissibility as a finite-dimensional OT projection with row-wise moment constraints on empirical transition laws.

(b) We develop an efficient entropic solver using Bregman–Dykstra updates, combining Sinkhorn scaling with exponential tilting, and validate it against `cvxpy`/MOSEK.

(c) We define the projection gap $G_{\lambda,k}$, calibrate it against a Brownian reference null, and establish its monotonicity under enlargement of the admissible class (Proposition 1).

(d) We show on simulated processes that $G_{\lambda,k}$ captures conditional memory structure beyond local second-order statistics—including separation of fBM and BSS processes with matched Hurst indices—and introduce a complementary spike proxy, $\mathrm{Var}_k(G_{\lambda,k})$, that distinguishes roughness-driven from jump-driven violations using the temporal variance of the gap sequence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1.1 Related Work

Entropy-regularised OT Cuturi (2013) can be extended beyond classical marginal constraints through iterative Bregman projections Benamou et al. (2015). Existing constrained OT formulations include unbalanced transport Chizat et al. (2018), capacity constraints Korman & McCann (2015), and moment- or feature-constraints Blondel et al. (2018); Buyl & De Bie (2022); Tang et al. (2024); these generally constrain global marginals or aggregate statistics. Flow-rate constraints bound aggregate mass through a region Dong et al. (2024): in contrast, our constraints act row-wise on the conditional first moment of the coupling, directly encoding local transition structure.

Semimartingales are fundamental in stochastic integration and financial modelling Jacod & Shiryaev (2003); Beiglböck et al. (2011); Rogers (1997), but discrete observations do not reveal their latent continuous-time decomposition. Existing diagnostics often rely on high-frequency asymptotics Jacod & Protter (2012); Barndorff-Nielsen & Shephard (2004); Aït-Sahalia & Jacod (2009); Bibinger et al. (2024). We instead use a finite-sample, distributional perspective, replacing the pathwise condition with a convex projection problem on empirical transition laws.

Throughout, semimartingale properties are understood relative to a fixed stochastic basis $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$; see Appendix A for background. We fix a single observed grid and filtration $\mathbb{G}$ and ask whether its one-step dynamics are compatible with a special semimartingale decomposition $X = X_0 + M + A$ when viewed through $\mathbb{G}$.

2 Drift-Admissibility Problem

We formulate a finite-dimensional diagnostic on the sampled increments $\Delta_i X$. A predictable drift increment is $\mathcal{G}_{i-1}$-measurable, so we prescribe lower and upper predictable box edges $d_{i,-}, d_{i,+} \in \mathcal{G}_{i-1}$ with $d_{i,-} \le d_{i,+}$.

Definition 1 (Predictable-box compatibility)**.** *Fix an observation grid π, the observed filtration $\mathbb{G}$, and predictable box edges $(d_{i,-}, d_{i,+})_{i=1}^N$. We say that the sampled increment system $(\Delta_i X)_{i=1}^N$ is* predictable-box compatible *with the box edges if, for every $i = 1, \dots, N$,*

$$d_{i,-} \le \mathbb{E}[\Delta_i X \mid \mathcal{G}_{i-1}] \le d_{i,+} \quad \textit{a.s.},$$

whenever the conditional expectation is well defined.

Remark 1 (Doob-decomposition equivalence)**.** *Predictable-box compatibility is equivalent to Doob decomposability of the sampled increments with predictable finite-variation component $A_i = \mathbb{E}[\Delta_i X \mid \mathcal{G}_{i-1}]$ lying in the prescribed box. It is deliberately weaker than asserting that the underlying continuous-time X is a special semimartingale, which cannot be certified from discrete data alone.*

A natural mesh-scaled specification sets $d_{i,\pm} = \Delta t\, \beta_{i,\pm}$ with $\beta_{i,\pm} \in \mathcal{G}_{i-1}$, consistent with Itô-type models with locally bounded drift rate. In finite samples, the conditional laws of $\Delta_i X$ are not known; the next section replaces them by a finite-dimensional coupling whose row-wise conditional means are constrained to lie in the drift box.

3 Constraining One-Step Drift with Optimal Transport

We encode the observable past through a predictor variable $Z_i \in \mathcal{G}_{i-1}$, $i = 1, \dots, N$. The predictor Z_i may be finite-dimensional and may contain, for example, the current observed state, lagged increments, realized variation estimates, or other $\mathcal{G}_{i-1}$-measurable summaries of the past. In particular, the framework is not restricted to the Markovian choice $Z_i = X_{t_{i-1}}$: a richer predictor encodes non-Markovian dependence, while a low-dimensional predictor yields a deliberately coarser diagnostic. We represent the predictable box edges as measurable functions of the predictor, $d_{i,-} = d_-(Z_i)$, $d_{i,+} = d_+(Z_i)$.

Let $Y_i := \Delta_i X$ denote the one-step increment. For computational purposes we discretize the predictor space and the increment space. Let $\mu = \sum_{k=1}^m \mu_k \delta_{z_k}$ and $\nu = \sum_{j=1}^\ell \nu_j \delta_{y_j}$ be discrete measures on predictor and increment space, respectively. A coupling $T = (T_{kj}) \in \mathbb{R}_+^{m \times \ell}$ between μ and ν

satisfies $T\mathbf{1} = \mu$, $T^\top \mathbf{1} = \nu$. The row indexed by k corresponds to the conditional distribution of the next increment given predictor value z_k. The predictable box-edge constraints therefore become

$$\mu_k \, d_-(z_k) \;\le\; \sum_{j=1}^{\ell} y_j \, T_{kj} \;\le\; \mu_k \, d_+(z_k), \quad k = 1, \dots, m.$$

Remark 2 (Discretisation trade-off)**.** *Discretisation makes the infinite-dimensional population problem tractable, exactly as in standard entropic OT Cuturi (2013). The main cost is that predictor bins pool states with different true conditional means, potentially diluting constraint violations. As argued in Section 4.1, this dilution is conservative—it suppresses $G_{\lambda,t}$, never inflates it.*

More generally, one may replace the raw increment y by a measurable increment operator $g^n(z, y)$. This allows for truncation, normalisation, or other predictor-dependent transformations of the one-step increment. The choice $g^n(z, y) = y_j - z_k$ yields the basic increment formulation used throughout this paper. This motivates the following feasible class of couplings:

$$\mathcal{C}_n(d_-, d_+; g^n) := \Big\{ T \in \mathbb{R}_+^{m \times \ell} : T\mathbf{1} = \mu,\ T^\top \mathbf{1} = \nu, \\ \mu_k d_-(z_k) \le \textstyle\sum_j g^n(z_k, y_j) T_{kj} \le \mu_k d_+(z_k),\ k = 1, \dots, m \Big\}.$$

3.1 Entropic Projection Formulation

Let $C \in \mathbb{R}^{m \times \ell}$ be a cost matrix. In all experiments reported in this paper, $C_{k,ij} \equiv (Y_j - Z_i)^2$, the matrix of pairwise squared distances between the one-dimensional predictor and increment grids, acting as a smoothness prior that favours cheap local transitions. Define the strictly positive Gibbs kernel $K_0 := \exp(-C/\varepsilon)$ for $\varepsilon > 0$. We use the generalised KL divergence $\mathrm{KL}(T \parallel K_0) := \sum_{k,j} \big(T_{kj} \log \frac{T_{kj}}{K_{0,kj}} - T_{kj} + K_{0,kj} \big)$. The hard-constrained entropic projection problem is

$$\min_{T \in \mathcal{C}_n(d_-, d_+; g^n)} \mathrm{KL}(T \parallel K_0). \tag{1}$$

Remark 3 (KL divergence as information-theoretic deformation)**.** *KL is chosen because its minimiser over a linear moment constraint is an exponential-family member (Lemma 3), enabling closed-form row projections that compose with Sinkhorn scaling (Lemma 1). By Sanov's theorem, the resulting gap is the exponential rate at which the observed law is improbable under the nearest admissible law—precisely why Hoeffding-type tests are KL-tests.*

We isolate the marginal constraints $\mathcal{C}_\mu := \{T \ge 0 : T\mathbf{1} = \mu\}$ and $\mathcal{C}_\nu := \{T \ge 0 : T^\top \mathbf{1} = \nu\}$, and the drift-constraint halfspaces

$$\mathcal{C}_+ := \Big\{ T \ge 0 : \textstyle\sum_j g^n(z_k, y_j) T_{kj} \le \mu_k d_+(z_k),\ k = 1, \dots, m \Big\},$$
$$\mathcal{C}_- := \Big\{ T \ge 0 : \textstyle\sum_j g^n(z_k, y_j) T_{kj} \ge \mu_k d_-(z_k),\ k = 1, \dots, m \Big\}.$$

The feasible set decomposes as $\mathcal{C}_n = \mathcal{C}_\mu \cap \mathcal{C}_\nu \cap \mathcal{C}_+ \cap \mathcal{C}_-$. This formulation makes explicit that predictable-drift admissibility is a feasibility problem over conditional couplings.

3.2 KL Projections with Predictable-Drift Constraints

The sets $\mathcal{C}_+$ and $\mathcal{C}_-$ are convex but not affine. Standard Sinkhorn iterations therefore no longer guarantee convergence; we consider the general projection problem

$$\min_{T \in \bigcap_{l=1}^{L} \mathcal{C}_l} \mathrm{KL}(T \parallel K_0), \tag{2}$$

where $\{\mathcal{C}_r\}_{r=1}^{L}$ denotes the collection of marginal and drift-constraint sets.

Lemma 1 (Dykstra's algorithm for KL projections Bauschke & Lewis (2000))**.** *Let $\{\mathcal{C}_l\}_{l=1}^{L}$ be closed convex sets with non-empty intersection, and suppose that the KL projection is finite. Then cyclic KL projections with Dykstra corrections converge to the unique minimiser of the projection problem.*

Since the row moment $\psi_k(\gamma_k)$ is strictly decreasing (see Lemma 3), at most one of $\mathcal{C}_+, \mathcal{C}_-$ will be active for a given row k. Corollary A2, shows that if the reference row already satisfies the prescribed bound, the KL projection onto that halfspace leaves the row unchanged. If the constraint is active, the scalar γ_k is updated by Newton's method (Lemma 3). Crucially, relaxing the marginal constraints via λ means that the Sinkhorn scalings (u, v) must be re-solved jointly with the tilt multipliers $\{\gamma_i\}$: the two sub-problems are coupled because the tilt changes the effective kernel seen by the Sinkhorn step, and the Sinkhorn step changes the row masses against which the tilt is computed. The complete alternating Bregman procedure is given in Appendix B.

4 Projection Gap and Diagnostic

Let $(X_t)_{t=0}^T$ be a real-valued process observed on a discrete grid $t_k = k\Delta t$, with associated natural filtration $\mathcal{F}_k \coloneqq \sigma(X_0, \dots, X_k)$. We interpret the observed path as a discretisation of an underlying continuous-time process, where $N = T/\Delta t$ governs the resolution of the empirical dynamics.

Empirical local transitions. To probe local one-step dynamics we construct empirical transition measures using a rolling window of size h. For each $k \geq h$, let

$$\hat{\mu}_k \coloneqq \frac{1}{h} \sum_{s=k-h+1}^{k} \delta_{Z_s}, \qquad \hat{\nu}_k \coloneqq \frac{1}{h} \sum_{s=k-h+1}^{k} \delta_{Y_s},$$

where $Z_s \in \mathcal{F}_{s-1}$ is the chosen predictor and $Y_s = X_s - X_{s-1}$ is the one-step increment. These empirical measures provide a finite-sample proxy for the joint law of predictor states and subsequent increments over the local window.

4.1 Marginal relaxation and the projection gap

The drift-admissibility constraint set $\mathcal{C}_k \coloneqq \mathcal{C}_- \cap \mathcal{C}_+$ may be *infeasible* for the hard-marginal problem (1) even when the underlying process is genuinely admissible. This occurs for two structurally distinct reasons:

(i) Discretisation-induced aggregation. At coarse predictor resolutions, each bin $\mathcal{B}_k$ pools states $z \in \mathcal{B}_k$ whose true conditional means $\mathbb{E}[Y \mid Z = z]$ need not be identical. The empirical row mean of the coupling for bin k is therefore the bin-averaged conditional mean

$$\hat{m}_k \;=\; \frac{1}{\hat{\mu}_k(\mathcal{B}_k)} \int_{\mathcal{B}_k} \mathbb{E}[Y \mid Z = z]\, \hat{\mu}_k(dz). \tag{3}$$

For a *memoryless* process, $\mathbb{E}[Y \mid Z = z] \approx 0$ for all z, so $\hat{m}_k \approx 0$ irrespective of bin width and the drift constraint remains inactive. For a process *with memory*, however, $\mathbb{E}[Y \mid Z = z]$ varies systematically with z: states within the same bin share a directional bias, and the bin average (3) inherits and preserves this bias rather than cancelling it. Coarse binning therefore acts as a *low-pass filter on the conditional mean structure*: it retains the memory signal while smoothing high-frequency noise. The only genuine loss occurs when a bin boundary splits a region where the conditional mean changes sign, in which case the averaging is conservative—it suppresses $G_{\lambda,k}$ rather than inflating it, so the diagnostic never falsely detects memory that is not present.

(ii) Structural incompatibility. The process genuinely violates the prescribed admissible class: the conditional mean of the increment distribution lies outside the drift box $[d_-^n, d_+^n]$ for a positive fraction of predictor states, independently of the discretisation resolution.

Remark 4 (Discretisation as a memory filter)**.** *Cause (i) is not a weakness of the diagnostic but a structural feature: it means that coarser predictor bins* amplify *the separation between process families with different memory structures. For BSS processes, the power kernel introduces long-range dependence that creates stronger directional persistence within each predictor bin than fBM at the same equivalent Hurst index. Coarse binning therefore makes the BSS–fBM separation* more *visible, not less—a property confirmed empirically in Section 5.1.*

Replacing the hard marginal constraints with soft penalties handles both cases uniformly: the solver is always well-posed, and the resulting diagnostic reflects the structural distance to admissibility. As such, we adopt the following structure for our diagnostic.

Relaxed-marginal objective. Given a cost matrix C_k on the local predictor–increment grid, define the Gibbs kernel $K_{0,k} \coloneqq \exp(-C_k/\varepsilon)$. In all experiments we use the quadratic cost $C_{k,ij} \coloneqq (Y_j - Z_i)^2$, a smoothness prior that favours cheap local transitions. The *marginal-relaxed objective* is

$$\mathcal{J}_{\varepsilon,\lambda}(T;\, \hat{\mu}_k,\, \hat{\nu}_k) \;\coloneqq\; \varepsilon \mathrm{KL}(T \parallel K_{0,k}) \;+\; \lambda\big[\mathrm{KL}(T\mathbf{1} \parallel \hat{\mu}_k) + \mathrm{KL}(T^\top \mathbf{1} \parallel \hat{\nu}_k)\big]\,, \tag{4}$$

where $\lambda > 0$ controls how strictly the empirical marginals are enforced. The *reference coupling* and *drift-constrained projection* are then

$$T_k^{\mathrm{ref}} \;\coloneqq\; \operatorname*{arg\,min}_{T \geq 0}\; \mathcal{J}_{\varepsilon,\lambda}(T;\, \hat{\mu}_k,\, \hat{\nu}_k), \qquad T_k^{\mathrm{proj}} \;\coloneqq\; \operatorname*{arg\,min}_{T \in \mathcal{C}_k}\; \mathcal{J}_{\varepsilon,\lambda}(T;\, \hat{\mu}_k,\, \hat{\nu}_k), \tag{5}$$

where $\mathcal{C}_k$ is the local predictable-box constraint set (Definition 1). By Lemma 2, T_k^{ref} has the scaling form $\mathrm{diag}(u)\, K_{0,k}\, \mathrm{diag}(v)$.

The complete rolling-window procedure, including the row-wise tilt step and warm-starting scheme, is given in Algorithm 1.

Lemma 2 (Relaxed Sinkhorn as KL projection with soft marginals Chizat et al. (2018); Langbridge et al. (2026))**.** *Let $K_0 \in \mathbb{R}_+^{m \times \ell}$ and μ, ν be the observed marginals. For $\lambda > 0$, the relaxed objective* (4) *for $T \in \mathbb{R}_+$ admits a unique minimiser of the form $T^* = \mathrm{diag}(u) K_0 \mathrm{diag}(v)$, for some scaling vectors $(u, v) \in \mathbb{R}_+^m \times \mathbb{R}_+^\ell$.*

Lemma 3 (Row-wise KL projection and exponential tilt)**.** *Fix (u, v) and define $\tilde{T}_{kj} \coloneqq u_k K_{kj} v_j$. Consider the row-wise KL projection*

$$\min_{T_{k\cdot} \geq 0} \sum_{j=1}^{\ell} T_{kj} \log \frac{T_{kj}}{\tilde{T}_{kj}} \quad \textit{s.t.} \quad \sum_{j=1}^{\ell} g^n(z_k, y_j) T_{kj} = c_k.$$

Assume $g^n(z_k, y_j)$ is finite and non-constant on the support of $\tilde{T}_{k\cdot}$. Then the following hold.

- *(i)* ***Exponential tilt.*** *The unique solution has the form $T_{kj}(\gamma_k) \propto \tilde{T}_{kj} \exp(-\gamma_k g^n(z_k, y_j)/\varepsilon)$, where γ_k is the unique scalar satisfying $\psi_k(\gamma_k) = c_k$ and $\psi_k(\gamma) \coloneqq \sum_j g^n(z_k, y_j) T_{kj}(\gamma)$.*
- *(ii)* ***Strong monotonicity.*** *The map ψ_k is C^∞, strictly decreasing, and satisfies $\psi_k'(\gamma) = -\varepsilon^{-1} m_k \mathrm{Var}_{\pi_k(\gamma)}(g^n(z_k, Y))$.*
- *(iii)* ***Stability and root-finding.*** *The inverse map $\gamma_k = \psi_k^{-1}(c_k)$ is Lipschitz: $|\gamma_k(c) - \gamma_k(c')| \leq \varepsilon (m_k \sigma_k^2)^{-1} |c - c'|$. The scalar equation $\psi_k(\gamma_k) = c_k$ is solved by bisection (global) or Newton's method (local, quadratic convergence; Appendix C.2).*

Theorem 1 (Convergence of relaxed predictable-drift projection)**.** *Assume $\varepsilon > 0$, $\lambda > 0$, and that $\mathcal{C}_+$ and $\mathcal{C}_-$ are closed convex sets with nonempty intersection. Then the relaxed objective* (4) *admits a unique minimiser for T_k^{ref} and T_k^{proj}. Moreover, the iterative scheme combining relaxed Sinkhorn updates with Bregman projections onto $\mathcal{C}_+$ and $\mathcal{C}_-$ converges to this minimiser.*

If additionally $\mathrm{Var}_{\pi_k(\gamma)}(g^n(z_k, Y)) \geq \sigma_k^2 > 0$ uniformly, then the algorithm converges linearly.

Proof. The proof of Theorem 1 is deferred to Appendix C.2. □

Definition 2 (Projection gap)**.** *The* projection gap *at time t is*

$$G_{\lambda,k} := \mathit{KL}(T_k^{\mathrm{proj}} \parallel T_k^{\mathrm{ref}}). \tag{6}$$

*By strict convexity and attainment, $G_{\lambda,k} = 0$ if and only if $T_k^{\mathrm{ref}} \in \mathcal{C}_k$, i.e. the unconstrained reference coupling already satisfies the prescribed drift constraints over the window (**Mode I**). The non-trivial cases are:*

Mode II *(Tilt channel, $G_{\lambda,k} > 0$, feasible). The set $\mathcal{C}_\mu \cap \mathcal{C}_\nu \cap \mathcal{C}_k$ is non-empty but $T_k^{\mathrm{ref}} \notin \mathcal{C}_k$. The projection T_k^{proj} is obtained by exponential tilting of the reference rows (Lemma A1): each active row is reweighted to shift its conditional mean to the nearest admissible boundary, preserving the row support. This mode is characteristic of processes whose increment distribution is* feasibly *admissible but whose entropic optimum lies outside the drift box.*

Mode III *(Mass reallocation, $G_{\lambda,k} > 0$, infeasible). The hard-marginal constraint set $\mathcal{C}_\mu \cap \mathcal{C}_\nu \cap \mathcal{C}_k$ is empty for some rows: no coupling with the prescribed marginals can satisfy the drift box for those predictor states. Under the soft-marginal objective* (4)*, the solver responds by reducing the mass assigned to infeasible predictor states, effectively down-weighting windows where the empirical increment distribution is structurally incompatible with the admissible class. This mode is characteristic of rough processes ($H < 0.5$) at fine mesh resolutions, where the increment variance dominates the drift envelope.*

By Sanov's theorem, $G_{\lambda,k}$ is the exponential rate at which the observed law is improbable under the nearest admissible law. Larger values of $G_{\lambda,k}$ therefore represent stronger evidence of structural incompatibility.

The quadratic lower bound underlying the gap scaling is given in Appendix D; the monotonicity property needed for the envelope sensitivity analysis follows directly from the nested feasible sets below.

Proposition 1 (Monotonicity of the projection gap)**.** *Fix k and $\varepsilon, \lambda > 0$, and let $\{\mathcal{C}_k^{(r)}\}_{r\geq 1}$ be a nested family of local predictable-box constraint sets satisfying $\mathcal{C}_k^{(r+1)} \subseteq \mathcal{C}_k^{(r)}$.*

For each r, define $T_k^{\mathrm{proj}(r)} := \arg\min_{T\in\mathcal{C}_k^{(r)}} \mathcal{J}_\lambda(T; \hat{\mu}_k, \hat{\nu}_k)$, and let $G_{\lambda,k}^{(r)} := \mathit{KL}(T_k^{\mathrm{proj}(r)} \parallel T_k^{\mathrm{ref}})$.

Then $\mathcal{C}_k^{(r+1)} \subseteq \mathcal{C}_k^{(r)}$ implies the constrained infimum of $\mathcal{J}_{\varepsilon,\lambda}$ over $\mathcal{C}_k^{(r+1)}$ is no smaller than over $\mathcal{C}_k^{(r)}$, giving $G_{\lambda,k}^{(r+1)} \geq G_{\lambda,k}^{(r)}$ directly from the definition of $G_{\lambda,k}$ as a difference of infima.

4.2 Calibration against a Brownian null

To calibrate the drift envelope in a resolution-invariant manner, we use the reflection principle for standard Brownian motion. Let $M_W := \sup_{0\leq s\leq W} B_s$ denote the running supremum of a standard Brownian motion over a window of duration $h\Delta t$. Then

$$\mathbb{P}(M_W \geq y) \;=\; 2\,\mathbb{P}(B_W \geq y) \;=\; 2\left(1 - \Phi\left(\frac{y}{\sqrt{h\Delta t}}\right)\right), \tag{7}$$

where Φ denotes the standard normal CDF. Setting $y = c\sqrt{h\Delta t}$ gives a drift envelope

$$d_\pm(x_i) \;\equiv\; \pm c\sqrt{h\cdot\Delta t}, \tag{8}$$

where the critical coefficient $c = \Phi^{-1}(1-\alpha/2)$ is invariant to both window length and sampling rate for a targeted Brownian violation probability α. Throughout this paper we adopt the canonical calibration $c = 0.50$, corresponding to $\alpha = 2(1-\Phi(0.5)) \approx 0.617$: approximately 62% of Brownian windows activate the constraint boundary, ensuring separation between process classes while remaining a deliberately loose threshold.

5 Experiments

We evaluate the diagnostic on synthetic stochastic processes spanning a range of local regularity, memory structure, and jump behaviour[1]. The primary comparison uses fractional Brownian motion (fBM) and Brownian semistationary (BSS) processes with a power convolution kernel, both parametrised by an equivalent Hurst index $H \in [0.10, 0.90]$.

Table 1 summarises the processes, and unless otherwise stated, all experiments use $N = 400$ observations ($\Delta t = 0.0025$), window size $h = 30$ ($h\Delta t = 0.075$), entropic regularisation $\varepsilon = 20/N$, marginal penalty $\lambda = 20$, envelope scale $c = 0.50$ ($\alpha \approx 62\%$ Brownian violation probability), and 10 independent MC trials (error bars: 95% CI).

[1]Code to reproduce all experiments is available at `https://anonymous.4open.science/r/drift-admissibility-F36B/`

Table 1: Synthetic processes used in the primary experiments. All processes are simulated over $t \in [0, 1]$ with $\Delta t = 1/N$.

Process	**Evolution**	**Parameters**
Brownian motion	$dX_t = \sigma\, dW_t$	$\sigma = 1.0$
Fractional BM (fBM)	$X_t = B_t^H$	$H \in (0, 1)$
BSS (power kernel)	$X_t = \kappa \int_0^t g(t-s)\, dW_s$ $g(x) = x^\alpha (1+x)^{-\gamma-\alpha}$	$\alpha \in (-0.5, 0.5)$, $\gamma = \kappa = 1.0$
Compound Poisson	$\Delta X_t = \sum_{s \le t} J_s,\ J_s \sim \mathcal{N}(0, \sigma_J^2)$	$\lambda_J \in \{1, 3, 5, 7, 9\},\ \sigma_J = 2$
CP Diffusion	$dX_t = \sigma\, dW_t + \sum_{s \le t} J_s$	$\lambda_J \in \{1, 3, 5, 7, 9\},\ \sigma_J = 2,\ \sigma = 1$

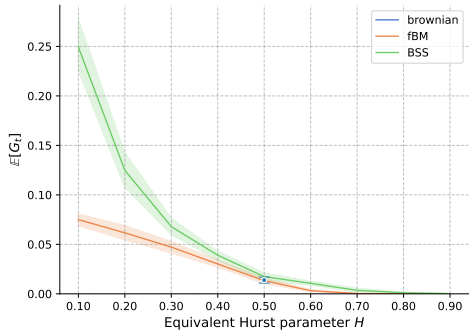


Figure 1: Mean projection gap $\mathbb{E}[G_{\lambda,t}]$ against equivalent Hurst parameter H for Brownian motion, fBM, and BSS (power kernel). Shaded ribbons: pointwise 95% CI over 10 MC trials.

5.1 Mean projection gap across the Hurst spectrum

Figure 1 reports the expected projection gap $\mathbb{E}[G_{\lambda,t}]$ against the equivalent Hurst index for fBM and BSS processes, together with the Brownian reference. Three regimes emerge cleanly, each corresponding to a distinct mode of the projection gap characterised in Section 4.1.

Figure 1 shows three regimes. For $H < 0.5$, the increment variance exceeds the Brownian scaling and the gap increases as H decreases, through the tilt and mass-reallocation channels. At $H = 0.5$, Brownian motion gives the expected small baseline gap induced by the 62% calibration. For $H > 0.5$, the envelope becomes non-binding and the gap collapses towards zero.

Most importantly, BSS has a larger gap than fBM at matched H in the rough regime. Since the two processes share the same local scaling exponent, this separation cannot be attributed to local regularity alone; it is consistent with sensitivity to the kernel-dependent conditional memory structure. The BSS curve also remains above Brownian at $H = 0.5$, illustrating that matched Hurst scaling does not imply drift admissibility.

5.2 Sensitivity and grid refinement

Envelope sensitivity. Figure 2 reports $\log \mathbb{E}[G_{\lambda,t}]$ against the Brownian violation probability α for rough fBM ($H = 0.2$), Brownian motion, and smooth fBM ($H = 0.8$). The profiles are non-increasing in the envelope width, as predicted by Proposition 1. Brownian motion is inactive for wide envelopes, enters Mode II as the envelope tightens, while rough fBM activates the constraint earlier and smooth fBM later.

Grid refinement. Figure 3 reports the mean gap against resolution $n \in [50, 750]$ for the same three processes, with $c = 0.5$ and $h = 30$ fixed. The rough process shows increasing gap with resolution as more windows enter Mode III; Brownian motion stabilises to its calibrated baseline, while the smooth process approaches Mode I.

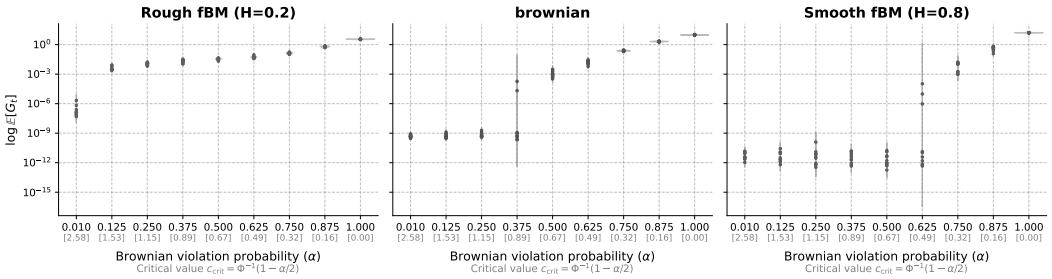


Figure 2: Natural logarithm of the mean gap against Brownian violation probability α for rough fBM ($H = 0.2$), Brownian motion, and smooth fBM ($H = 0.8$).

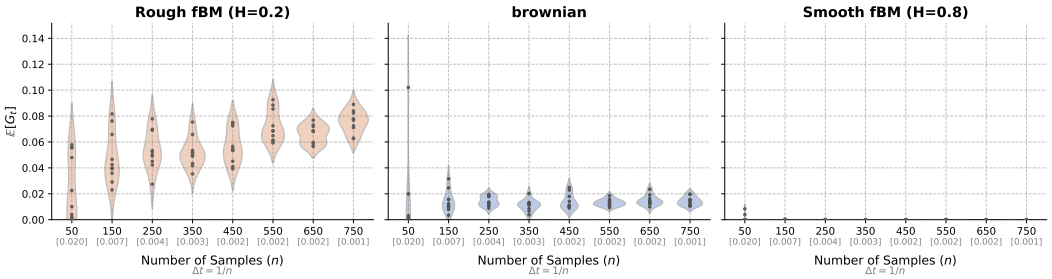


Figure 3: Mean projection gap against grid resolution n for rough fBM ($H = 0.2$), Brownian motion, and smooth fBM ($H = 0.8$).

5.3 Mean gap and variance as a two-dimensional diagnostic

We augment the fBM and BSS sweep of Section 5.1 with compound Poisson (CP) processes and CP diffusions, parametrised by jump rate $\lambda_J \in \{1, 3, 5, 7, 9\}$ and jump size $\sigma_J = 2$, using identical calibration. At $\lambda_J = 9$ and $h = 30$, the expected number of jumps per window is $\lambda_J h \Delta t = 0.675$; each jump contributes $\mathbb{E}[|J|] \approx 1.60 \approx 11.7\times$ the envelope half-width, producing a large but isolated gap spike whenever it occurs.

Figure 4 reports $\mathbb{E}[G_{\lambda,k}]$ (left) and $\mathrm{Var}_k(G_{\lambda,k})$ (right) for all process families. Two findings emerge. The mean gap (left) orders processes by roughness as in Section 5.1, but jump processes at high λ_J also achieve large means—through window-averaged jump contributions rather than persistent increment variance—with substantially wider confidence intervals reflecting trial-to-trial variability in jump occurrence. The within-trial variance (right) cleanly separates the two failure modes: rough processes exhibit *sustained elevation* (large gap in nearly every window, moderate variance relative to mean), while jump processes exhibit *isolated spikes* (near-zero baseline punctuated by rare large events, variance far exceeding the mean). At $\lambda_J = 9$, the Compound Poisson variance is approximately $40\times$ that of Brownian motion; rough fBM ($H = 0.2$) is approximately $7\times$—a clear separation despite comparable mean gaps. CP Diffusion interpolates between the two regimes: its diffusion component adds a roughness floor that reduces spike-to-baseline contrast relative to pure CP.

The joint $(\mathbb{E}[G_{\lambda,k}],\ \mathrm{Var}_k(G_{\lambda,k}))$ plane therefore provides a two-dimensional diagnostic that no single second-order statistic can replicate. Representative gap time-series for both regimes are shown in Appendix E.

5.4 Solver validation

We validate the Bregman–Dykstra implementation against an equivalent CVXPY/MOSEK formulation over 200 independent transport problems, using identical costs, regularisation, marginal penalties, and constraints. Table 2 reports complementary objective, plan, constraint, and drift discrepancies.

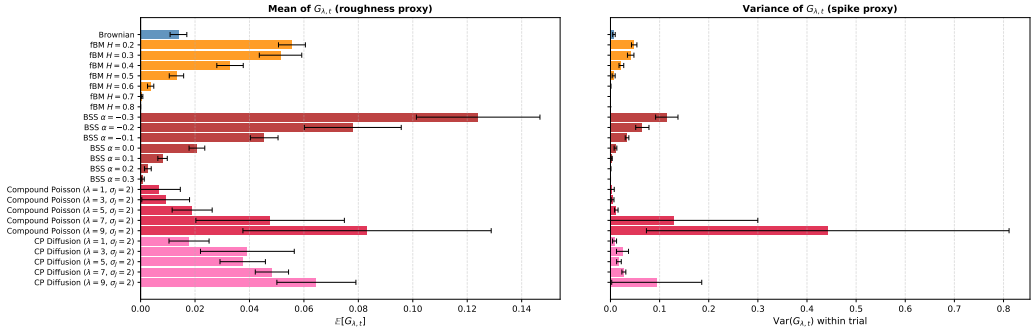


Figure 4: **Left:** Mean projection gap $\mathbb{E}[G_{\lambda,k}]$ and **Right:** within-trial variance $\mathrm{Var}_k(G_{\lambda,k})$ for Brownian motion, fBM ($H \in [0.2, 0.8]$), BSS ($\alpha \in [-0.3, 0.3]$), Compound Poisson ($\lambda_J \in \{1, 3, 5, 7, 9\}$, $\sigma_J = 2$), and CP Diffusion ($\lambda_J \in \{1, 3, 5, 7, 9\}$, $\sigma_J = 2$, $\sigma = 1$). Error bars: 95% CI over 10 MC trials. Calibration: $c = 0.50$ ($\alpha \approx 62\%$), $N = 400$, $h = 30$, $\varepsilon = 20/N$, $\lambda = 20$.

Table 2: Quantitative comparison of the proposed Bregman–Dykstra projection algorithm against CVXPY (MOSEK) over 200 independent transport problems.

Metric	**Mean**	**Std**	**Min**	**Max**
Relative objective gap $\frac{\lvert\mathcal{J}_{\varepsilon,\lambda}(T)-\mathcal{J}_{\varepsilon,\lambda}(T^{\mathrm{MOSEK}})\rvert}{\lvert\mathcal{J}_{\varepsilon,\lambda}(T^{\mathrm{MOSEK}})\rvert}$	0.0203	0.0112	0.0016	0.0656
Max plan discrepancy $\max_{ij}\left\lvert \frac{T_{ij}}{\lvert T\rvert_1} - \frac{T^{\mathrm{MOSEK}}_{ij}}{\lvert T^{\mathrm{MOSEK}}\rvert_1}\right\rvert$	0.0022	0.0016	0.0003	0.0106
Max constraint violation	2.17×10^{-5}	2.76×10^{-4}	0	0.0035
Mean absolute drift error $\frac{1}{n}\sum_i \lvert d_i - d_i^{\mathrm{MOSEK}}\rvert$	0.0033	0.0020	0.0003	0.0099

The methods occasionally select different primal plans, but the resulting conditional drift functions are nearly indistinguishable and satisfy the constraints to high accuracy. Runtime scaling is reported in Appendix F.

6 Discussion and Limitations

Our experiments show that the projection gap captures differences between stochastic regimes not explained by local regularity: in particular, the BSS–fBM separation at matched Hurst indices indicates sensitivity to kernel-dependent conditional memory. The within-trial variance $\mathrm{Var}_k(G_{\lambda,k})$ further separates roughness-driven from jump-driven violations (Section 5.3): sustained elevation versus isolated spikes reflects a qualitative difference in the temporal structure of constraint violations that the mean gap alone cannot resolve. The diagnostic is nevertheless model-relative: $G_{\lambda,k}$ depends on the data, predictor, admissible class, discretisation, window, and transport geometry, so scores from different specifications are not directly comparable. Its computational cost is manageable for the settings considered here but may require parallelisation for long series. Finally, $G_{\lambda,k}$ is a compatibility diagnostic rather than a calibrated hypothesis test. A formal test requires joint control of empirical discretisation, rolling-window approximation, entropic regularisation, and envelope scaling; moment-inequality nulls also require least-favourable calibration Andrews & Shi (2013); Chetverikov (2018). Learning the admissible class and extending to richer multivariate constraints are natural directions for future work.

7 Conclusion

We presented a structure-aware OT diagnostic for assessing whether empirical time-series dynamics are compatible with a prescribed predictable-drift class. By casting admissibility as row-wise constraints on empirical transition laws, we obtain a finite-dimensional projection problem with an efficient Bregman–Dykstra solver. The resulting projection gap, calibrated against a Brownian reference null, detects conditional structure beyond second-order statistics and separates process classes with matched local regularity.

Reproducibility Statement

The code used to produce the results in the paper and appendices is provided at `https://anonymous.4open.science/r/drift-admissibility-F36B/`. The repository is organised around the same computational pipeline as the paper: the transport solver implements the entropic OT objective and Bregman–Dykstra projections described in Section 3; the row-wise exponential tilt implements Lemma 3; and the rolling-window procedure computes the projection-gap sequence $\{G_{\lambda,k}\}$ defined in Section 4. The main experimental figures are generated from the corresponding process simulators and diagnostic routines: Figure 1 uses the fBM/BSS comparison of Section 5.1, while the sensitivity and grid-refinement experiments moved to the appendix correspond to the envelope and resolution analyses described in Section 5.2. The solver validation in Section 5.4 compares the same constrained transport problem against the `cvxpy`/MOSEK formulation. The `README_reproducibility.md` file provides the mapping from these paper components to the relevant scripts and commands. Proofs for the key lemmas and theorems are provided in the appendix.

AI Disclosure

LLM-based code agents were used to implement standard numerical routines (Sinkhorn updates, Newton root-finding); all generated code was reviewed and validated against the cvxpy/MOSEK baseline in Section 5.4. LLM assistance was also used to check draft text for clarity. All scientific content and experimental design are solely the work of the authors.

References

Yacine Aït-Sahalia and Jean Jacod. Estimating the degree of activity of jumps in high frequency data. *The Annals of Statistics*, 37(5A):2202–2244, 2009.

Donald WK Andrews and Xiaoxia Shi. Inference based on conditional moment inequalities. *Econometrica*, 81(2):609–666, 2013.

Ole E Barndorff-Nielsen and Neil Shephard. Power and bipower variation with stochastic volatility and jumps. *Journal of financial econometrics*, 2(1):1–37, 2004.

Heinz H Bauschke and Adrian S Lewis. Dykstras algorithm with Bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

Mathias Beiglböck, Walter Schachermayer, and Bezirgen Veliyev. A direct proof of the Bichteler–Dellacherie theorem and connections to arbitrage. *The Annals of Probability*, 39(6):2424–2440, 2011. doi: 10.1214/10-AOP602.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Markus Bibinger, Nikolaus Hautsch, and Alexander Ristig. Jump detection in high-frequency order prices. *arXiv preprint arXiv:2403.00819*, 2024.

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *International conference on artificial intelligence and statistics*, pp. 880–889. PMLR, 2018.

Maarten Buyl and Tijl De Bie. Optimal transport of classifiers to fairness. *Advances in Neural Information Processing Systems*, 35:33728–33740, 2022.

Denis Chetverikov. Adaptive tests of conditional moment inequalities. *Econometric Theory*, 34(1): 186–227, 2018.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of computation*, 87(314):2563–2609, 2018.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Anqi Dong, Arthur Stephanovitch, and Tryphon T. Georgiou. Mongekantorovich optimal transport through constrictions and flow-rate constraints. *Automatica*, 160:111448, 2024. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2023.111448. URL https://www.sciencedirect.com/science/article/pii/S0005109823006155.

Jean Jacod and Philip Protter. *Discretization of Processes*, volume 67 of *Stochastic Modelling and Applied Probability*. Springer, Heidelberg, 2012.

Jean Jacod and Albert N. Shiryaev. *Limit Theorems for Stochastic Processes*. Springer, Berlin, 2 edition, 2003.

J. Korman and R. McCann. Optimal transportation with capacity constraints. *Transactions of the American Mathematical Society*, 367(3):1501–1521, 2015.

Abigail Langbridge, Martin Corless, and Robert Shorten. Sinkhorn-knopp balancing with generalised martingale-type constraints. *arXiv preprint arXiv:2606.26739*, 2026.

L. C. G. Rogers. Arbitrage with fractional brownian motion. *Mathematical Finance*, 7(1):95–105, 1997.

X. Tang, H. Rahmanian, M. Shavlovsky, K.K. Thekumparampil, T. Xiao, and L. Ying. A sinkhorn-type algorithm for constrained optimal transport. *arXiv preprint arXiv:2403.05054*, 2024.

A Stochastic Analysis Background

A *stochastic basis* is a quadruple $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and where $\mathbb{F} := (\mathcal{F}_t)_{t\in[0,T]}$ denotes a filtration, that is, an increasing family of sub-σ-fields of $\mathcal{F}$:

$$\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}, \qquad 0 \le s \le t \le T.$$

Throughout the paper we take the filtration to be right-continuous, meaning that

$$\mathcal{F}_t = \bigcap_{u>t} \mathcal{F}_u, \qquad 0 \le t < T.$$

The σ-field $\mathcal{F}_t$ represents the information available up to time t. Accordingly, martingale, local-martingale, predictability, and semimartingale properties are always understood relative to the chosen stochastic basis.

A process $X = (X_t)_{t\in[0,T]}$ on this basis is said to be $\mathbb{F}$-*adapted* if X_t is $\mathcal{F}_t$-measurable for every $t \in [0,T]$ (often written $X_t \in \mathcal{F}_t$). A process is $\mathbb{F}$-*predictable* if it is measurable with respect to the predictable sigma-field $\mathcal{P}$ on $\Omega \times [0,T]$, generated by the left-continuous adapted processes; in particular, left-continuous adapted processes are predictable. A *càdlàg*, or RCLL, process is a process whose sample paths are right-continuous and admit left limits. A process X is an $\mathbb{F}$-*martingale* if it is an adapted integrable process such that

$$\mathbb{E}[X_s \mid \mathcal{F}_t] = X_t, \qquad 0 \le t \le s \le T.$$

Equivalently, conditional on the information available at time t, the future value of the process has expectation equal to its present value. A *local martingale* on $[0,T]$ is an adapted process M for which there exists a sequence of stopping times $(\tau_k)_k$ with $\tau_k \uparrow T$ a.s. such that M^{τ_k} is a martingale on $[0,T]$ for every k. A real-valued, càdlàg $\mathbb{F}$-adapted process X is a *semimartingale*, with respect to $\mathbb{F}$, if it admits a decomposition

$$X_t = X_0 + M_t + A_t, \qquad t \in [0,T],$$

where M is an $\mathbb{F}$-local martingale and A is an $\mathbb{F}$-adapted process of finite variation. In this paper, we focus on the *special semimartingale* case, in which such a decomposition can be chosen with A predictable. See Jacod & Shiryaev (2003) for background and for the refinement of this structure in terms of predictable characteristics for Itô semimartingales. For the purposes of the present paper, the essential point is the special semimartingale separation between conditionally centered local-martingale innovations and a predictable finite-variation component.

B Algorithm Pseudocode

Algorithm 1 gives the complete rolling-window procedure for computing the projection gap sequence $\{G_{\lambda,k}\}_{k=h}^{N-1}$ from a single observed path. The three structural cases correspond directly to the three-mode characterisation of Definition 2: rows satisfying $\psi_k(0) \in [d_-, d_+]$ contribute $G_{\lambda,k} = 0$ (Mode I); rows requiring a tilt ($\gamma_i^\star \neq 0$) but with a non-empty feasible set contribute via the exponential-tilt channel (Mode II); and rows for which no coupling with the prescribed marginals can satisfy the drift box are handled by the soft-marginal objective, which reduces their mass contribution (Mode III).

Algorithm 1 Projection Gap Diagnostic

Require: Observed path $X_{t_0}, \dots, X_{t_N}$; window size h; envelope scale c; entropic regularisation $\varepsilon > 0$; marginal penalty $\lambda > 0$; convergence tolerance $\delta > 0$

Ensure: Projection gap sequence $\{G_{\lambda,k}\}_{k=h}^{N-1}$

Set drift envelope half-width based on Brownian calibration: ▷ Section 4.2

$$d_{\pm}(x_i) \;\equiv\; \pm c\sqrt{h \cdot \Delta t}$$

for $k = h$ **to** $N-1$ **do**

  Build rolling-window empirical measures over $[k-h+1,\, k]$: ▷ Section 4.1

$$\hat{\mu}_k \;:=\; \frac{1}{h}\sum_{s=k-h+1}^{k} \delta_{Z_s}, \qquad \hat{\nu}_k \;:=\; \frac{1}{h}\sum_{s=k-h+1}^{k} \delta_{Y_s}$$

  where $Z_s \in \mathcal{F}_{s-1}$ is the predictor and $Y_s = X_s - X_{s-1}$ is the one-step increment.

  Discretise predictor grid $\{z_i\}_{i=1}^{m}$ and increment grid $\{y_j\}_{j=1}^{\ell}$ from the window data.

  Set quadratic cost matrix: $C_{k,ij} \;\equiv\; (Y_j - Z_i)^2$

  Set Gibbs kernel: $K_{0,k} \;:=\; \exp(-C_k/\varepsilon)$

  Compute T_k^{ref} by relaxed Sinkhorn iterations until convergence: ▷ Lemma 2

$$T_k^{\mathrm{ref}} \;:=\; \operatorname*{arg\,min}_{T \geq 0}\; \varepsilon\,\mathrm{KL}(T \parallel K_{0,k}) + \lambda\Big[\mathrm{KL}(T\mathbf{1} \parallel \hat{\mu}_k) + \mathrm{KL}(T^{\top}\mathbf{1} \parallel \hat{\nu}_k)\Big]$$

  Retain the converged Sinkhorn scalings $(u^{\star}, v^{\star}) \in \mathbb{R}_+^m \times \mathbb{R}_+^{\ell}$.

  Initialise Sinkhorn scalings: $u \leftarrow u^{\star},\; v \leftarrow v^{\star}$ ▷ warm-start from T_k^{ref}

  Initialise tilt multipliers: $\gamma_i \leftarrow 0$ for all $i = 1, \dots, m$

  repeat ▷ outer Bregman loop

    Form the tilt-adjusted kernel:

$$\widetilde{K}_{ij} \;:=\; K_{0,k,ij}\exp\big(-\gamma_i\, g^n(z_i, y_j)/\varepsilon\big), \quad i = 1, \dots, m,\; j = 1, \dots, \ell$$

    Soft-marginal Sinkhorn updates on $\widetilde{K}$: ▷ Lemma 2

$$u \;\leftarrow\; \left(\frac{\hat{\mu}_k}{\widetilde{K}\, v}\right)^{\lambda/(\lambda+\varepsilon)}, \qquad v \;\leftarrow\; \left(\frac{\hat{\nu}_k}{\widetilde{K}^{\top} u}\right)^{\lambda/(\lambda+\varepsilon)}$$

    Set current coupling iterate: $T_k^{\mathrm{proj}} \leftarrow \mathrm{diag}(u)\,\widetilde{K}\,\mathrm{diag}(v)$

    for each predictor bin $i = 1, \dots, m$ **do**

      Compute current row moment:

$$\psi_k(\gamma_i) \;=\; \sum_{j=1}^{\ell} g^n(z_i, y_j)\,[T_k^{\mathrm{proj}}]_{ij} \quad \text{where} \quad g^n(z, y) = y_j - z_i$$

      if $\psi_k(\gamma_i) \notin \big[d_-(z_i),\; d_+(z_i)\big]$ **then** ▷ Constraint active: Corollary A2

        Set boundary target: $d_i \;=\; \begin{cases} d_-(z_i) & \text{if } \psi_k(0) < d_-(z_i) \\ d_+(z_i) & \text{if } \psi_k(0) > d_+(z_i) \end{cases}$

        Update γ_i by one Newton step on $\psi_k(\gamma) = d_i$: ▷ Lemma 3,(ii,iv)

$$\gamma_i \;\leftarrow\; \gamma_i \;-\; \frac{\psi_k(\gamma_i) - c_i}{\psi_k'(\gamma_i)} \quad \text{where} \quad \psi_k'(\gamma_i) = -\varepsilon^{-1} m_i \mathrm{Var}_{\pi_k(\gamma_i)}\big(g^n(z_i, Y)\big) < 0$$

      else ▷ Row already admissible, Corollary A2

        $\gamma_i \leftarrow 0$

      end if

    end for

  until $\max_i |\psi_k(\gamma_i) - d_i^{\mathrm{active}}| \leq \delta$ *and* Sinkhorn marginal residuals $\leq \delta$ ▷ joint convergence; guaranteed by Theorem 1

  Compute projection gap $G_{\lambda,k} \;:=\; \mathrm{KL}(T_k^{\mathrm{proj}} \parallel T_k^{\mathrm{ref}})$ ▷ Definition 2

end for

return $\{G_{\lambda,k}\}_{k=h}^{N-1}$

Complexity. Each outer iteration (line 3) requires one relaxed Sinkhorn solve and m one-dimensional root-finding steps. The Sinkhorn solve converges in $\mathcal{O}(\varepsilon^{-1}\log\varepsilon^{-1})$ iterations, each costing $\mathcal{O}(m\ell)$ operations; the root-finding steps each cost $\mathcal{O}(\ell\log(1/\delta))$ for bisection tolerance δ. The overall complexity of Algorithm 1 is therefore $\mathcal{O}\big((N-h)\cdot h^3\varepsilon^{-1}\log\varepsilon^{-1}\big)$, where the h^3 factor reflects the $m=\ell=h$ discretisation used throughout the experiments. The outer loop is embarrassingly parallel: windows share no information and can be evaluated concurrently, providing significant acceleration when $N \gg h$. To avoid redundant computation, the constrained solver T_k^{proj} is warm-started from the Kantorovich potentials (u, v) recovered during the T_k^{ref} solve.

C Supplementary Lemmas and Proofs

C.1 KL Projection for Conditional Bounds

Lemma A1 (Row-wise exponential tilt and scalar dual variable)**.** *Fix a row index k, and write $g_{kj} := g_n(z_k, y_j), j = 1,\dots,\ell$. Let $\widetilde{T}_{k\cdot} \in \mathbb{R}_+^\ell$ be a strictly positive reference row with total mass $m_k := \sum_{j=1}^{\ell} \widetilde{T}_{kj} > 0$, and assume that g_{kj} is finite and nonconstant on the support of $\widetilde{T}_{k\cdot}$. For $\gamma \in \mathbb{R}$, define the tilted row*

$$T_{kj}(\gamma) := m_k \frac{\widetilde{T}_{kj}\exp(-\gamma g_{kj}/\varepsilon)}{\sum_{r=1}^{\ell}\widetilde{T}_{kr}\exp(-\gamma g_{kr}/\varepsilon)}, \qquad j = 1,\dots,\ell.$$

Then $\sum_j T_{kj}(\gamma) = m_k$, and the row moment $\psi_k(\gamma) := \sum_{j=1}^{\ell} g_{kj}T_{kj}(\gamma)$ is continuous and strictly decreasing. Moreover, $\lim_{\gamma\to+\infty}\psi_k(\gamma) = m_k\min_j g_{kj}$ and $\lim_{\gamma\to-\infty}\psi_k(\gamma) = m_k\max_j g_{kj}$. Consequently, for every $c_k \in (m_k\min_j g_{kj},\, m_k\max_j g_{kj})$ there exists a unique scalar $\gamma_k \in \mathbb{R}$ such that $\psi_k(\gamma_k) = c_k$. The row $T_{k\cdot}(\gamma_k)$ is the unique solution of

$$\min_{q\in\mathbb{R}_+^\ell}\sum_{j=1}^{\ell}\left(q_j\log\frac{q_j}{\widetilde{T}_{kj}} - q_j + \widetilde{T}_{kj}\right)$$

subject to $\sum_{j=1}^{\ell} q_j = m_k$, and $\sum_{j=1}^{\ell} g_{kj}q_j = c_k$.

Proof of Lemma A1. The normalisation in the definition of $T_{k\cdot}(\gamma)$ gives $\sum_j T_{kj}(\gamma) = m_k$. Define the probability weights

$$\pi_{kj}(\gamma) := \frac{\widetilde{T}_{kj}\exp(-\gamma g_{kj}/\varepsilon)}{\sum_{r=1}^{\ell}\widetilde{T}_{kr}\exp(-\gamma g_{kr}/\varepsilon)}.$$

Then

$$\psi_k(\gamma) = m_k\sum_{j=1}^{\ell} g_{kj}\pi_{kj}(\gamma).$$

Since the support is finite and all g_{kj} are finite, ψ_k is smooth. Differentiating gives

$$\psi_k'(\gamma) = -\frac{m_k}{\varepsilon}\operatorname{Var}_{\pi_k(\gamma)}(g_{kY}).$$

By the nonconstancy assumption, the variance is strictly positive for all finite γ, so ψ_k is strictly decreasing and continuous. As $\gamma\to+\infty$, the probability vector $\pi_k(\gamma)$ concentrates on the minimizers of g_{kj}, and as $\gamma\to-\infty$, it concentrates on the maximizers. This proves the two endpoint limits. Existence and uniqueness of γ_k for any interior target c_k then follow from the intermediate value theorem and strict monotonicity.

It remains to identify the minimizer. Introduce Lagrange multipliers α and η for the mass and moment constraints. The first-order condition for an interior minimizer is

$$\log\frac{q_j}{\widetilde{T}_{kj}} + \alpha + \eta g_{kj} = 0,$$

so

$$q_j = \widetilde{T}_{kj}\exp(-\alpha)\exp(-\eta g_{kj}).$$

Writing $\eta = \gamma/\varepsilon$ and choosing α to enforce $\sum_j q_j = m_k$ gives exactly the tilted form above. Strict convexity of the KL objective on the positive orthant gives uniqueness. □

Corollary A1 (Compatibility of row-wise bounds)**.** *For a row with total mass $m_k \geq 0$, the attainable values of $\sum_{j=1}^{\ell} g_n(z_k, y_j) T_{kj}$ are exactly $[m_k \min_j g_n(z_k, y_j), m_k \max_j g_n(z_k, y_j)]$. Therefore a target interval $[a_k, b_k]$ is feasible for row k if and only if $[a_k, b_k] \cap [m_k \min_j g_n(z_k, y_j), m_k \max_j g_n(z_k, y_j)] \neq \varnothing$. In particular, under the hard marginal constraint $T\mathbf{1} = \mu$, one has $m_k = \mu_k$, $a_k = \mu_k d_-^n(z_k)$, and $b_k = \mu_k d_+^n(z_k)$. Where $m_k = 0$, the attainable set reduces to $\{0\}$ and the same feasibility condition applies.*

Proof of Corollary A1. Fix a row k and write $g_{kj} := g^n(z_k, y_j)$, $g_k^{\min} := \min_{1 \leq j \leq \ell} g_{kj}$, $g_k^{\max} := \max_{1 \leq j \leq \ell} g_{kj}$. Let $T_{k\cdot} \in \mathbb{R}_+^{\ell}$ be any nonnegative row with total mass $\sum_{j=1}^{\ell} T_{kj} = m_k$. Then

$$g_k^{\min} T_{kj} \leq g_{kj} T_{kj} \leq g_k^{\max} T_{kj} \qquad \text{for every } j.$$

Summing over j gives

$$m_k g_k^{\min} \leq \sum_{j=1}^{\ell} g_{kj} T_{kj} \leq m_k g_k^{\max}.$$

Thus every feasible row moment belongs to

$$[m_k g_k^{\min},\, m_k g_k^{\max}].$$

Conversely, let

$$c \in [m_k g_k^{\min},\, m_k g_k^{\max}].$$

If $g_k^{\min} = g_k^{\max}$, then all g_{kj} are equal and the only attainable moment is $m_k g_k^{\min} = m_k g_k^{\max}$, so the claim is immediate. Otherwise, choose indices j_- and j_+ such that

$$g_{kj_-} = g_k^{\min}, \qquad g_{kj_+} = g_k^{\max}.$$

There exists $\theta \in [0, 1]$ such that

$$c = m_k\big(\theta g_k^{\min} + (1 - \theta) g_k^{\max}\big).$$

Define a row $T_{k\cdot}$ by

$$T_{kj_-} := m_k \theta, \qquad T_{kj_+} := m_k(1 - \theta),$$

with all other entries equal to zero. Then

$$\sum_{j=1}^{\ell} T_{kj} = m_k$$

and

$$\sum_{j=1}^{\ell} g_{kj} T_{kj} = m_k\big(\theta g_k^{\min} + (1 - \theta) g_k^{\max}\big) = c.$$

Hence the set of attainable row moments is exactly

$$[m_k g_k^{\min},\, m_k g_k^{\max}].$$

It follows that a target interval $[a_k, b_k]$ is feasible for row k if and only if

$$[a_k, b_k] \cap [m_k g_k^{\min},\, m_k g_k^{\max}] \neq \varnothing.$$

Indeed, feasibility requires the row moment to lie both in the prescribed target interval and in the attainable moment interval, and the converse follows from the attainability argument above.

In the hard-marginal case $T\mathbf{1} = \mu$, the row mass is $m_k = \mu_k$. Taking

$$a_k = \mu_k d_-^n(z_k), \qquad b_k = \mu_k d_+^n(z_k),$$

gives the stated specialisation. □

Corollary A2 (Complementary slackness for row-wise drift bounds)**.** *Fix a row k with mass $m_k > 0$, and define $\psi_k(\gamma)$ as in Lemma A1. Then $\psi_k : \mathbb{R} \to \mathbb{R}$ is continuous and strictly decreasing. Let $([a_k, b_k])$ be a feasible target interval as in Corollary A1. Then the constrained projection onto this interval is characterised as follows:*

- *If $\psi_k(0) < a_k$, then the lower bound is enforced by a unique $\gamma_k^\star < 0$ satisfying $\psi_k(\gamma_k^\star) = a_k$.*
- *If $a_k \le \psi_k(0) \le b_k$, then the untilted row already satisfies the admissibility interval, and no row tilt is required: $\gamma_k^\star = 0$.*
- *If $\psi_k(0) > b_k$, then the upper bound is enforced by a unique $\gamma_k^\star > 0$ satisfying $\psi_k(\gamma_k^\star) = b_k$.*

Therefore, the projection onto $\mathcal{C}_k$ reduces to a KL projection onto the intersection of two affine constraints, which is given in closed form by a single exponential tilting step (Lemma A1).

C.2 PROOFS

Proof of Lemma 2. The objective is strictly convex in T and coercive due to the KL penalties on both T and its marginals, hence admits a unique minimiser.

Introducing dual variables for the marginal penalties, the first-order optimality conditions yield

$$\log \frac{T_{kj}}{K_{0,kj}} + \lambda \log \frac{(T\mathbf{1})_k}{\mu_k} + \lambda \log \frac{(T^\top \mathbf{1})_j}{\nu_j} = 0.$$

Rearranging gives

$$T_{kj} = K_{0,kj}\, u_k\, v_j,$$

where

$$u_k = \exp\left(-\lambda \log \frac{(T\mathbf{1})_k}{\mu_k}\right), \quad v_j = \exp\left(-\lambda \log \frac{(T^\top \mathbf{1})_j}{\nu_j}\right).$$

Thus the solution has the stated scaling form. □

Proof of Theorem 1. The objective of the relaxed problem is a sum of KL divergences, hence strictly convex in T. The marginal penalties relax only the marginal constraints. Consequently, the relaxed problem admits a unique minimiser whenever $C_+ \cap C_-$ is nonempty, even if the hard marginal-constrained feasible set $C_\mu \cap C_\nu \cap C_+ \cap C_-$ is empty.

If the row-wise predictable-drift constraints are temporarily omitted, the relaxed problem reduces to the soft-marginal entropic OT subproblem

$$\min_{T \ge 0} \mathrm{KL}(T \| K_0) + \lambda \left\{ \mathrm{KL}(T\mathbf{1} \| \mu) + \mathrm{KL}(T^\top \mathbf{1} \| \nu) \right\}.$$

Its minimiser has the scaling form

$$T = \mathrm{diag}(u) K_0 \,\mathrm{diag}(v),$$

as in unbalanced Sinkhorn scaling.

The row-wise predictable-drift constraints are then imposed through Bregman projections onto C_+ or C_- due to complementary slackness (Corollary A2). When a row constraint is active, Lemma A1 shows that the corresponding projection is a one-dimensional exponential tilt of that row.

Convergence of the algorithm. Each update in the proposed algorithm corresponds to a KL (Bregman) projection onto one of the constraint sets.

Since these sets are convex and the KL divergence is a Bregman divergence, and due to the complementary slackness of the drift constraints, the alternating projection scheme is a series of Bregman projections. Standard results on Bregman projections (e.g., Bauschke & Lewis (2000); Benamou et al. (2015)) ensure convergence to the unique minimiser T^*.

If, in addition, the variance lower bound

$$\mathrm{Var}_{\pi_k(\gamma)}(g_n(z_k, Y)) \ge \sigma_k^2 > 0$$

holds uniformly, then the algorithm converges linearly:

$$\mathrm{KL}(T^{(t)} \parallel T^*) \le \rho^t \,\mathrm{KL}(T^{(0)} \parallel T^*)$$

for some $\rho \in (0, 1)$ depending on $(\varepsilon, \lambda, \{\sigma_k\})$. □

Proof of Lemma A2. Fix a row k and write $g_j := g_n(z_k, y_j)$. By Lemma A1, the tilted plan takes the form

$$T_{kj}(\gamma) = m_k \frac{\tilde{T}_{kj} \exp(-\gamma g_{kj}/\varepsilon)}{Z(\gamma)}, \quad Z(\gamma) = \sum_r \tilde{T}_{kr} \exp(-\gamma g_{kr}/\varepsilon)\,.$$

Then

$$\log \frac{T_{kj}(\gamma)}{\tilde{T}_{kj}} = -\frac{\gamma}{\varepsilon} g_{kj} + \log m_k - \log Z(\gamma),$$

and hence

$$\mathrm{KL}(T(\gamma) \parallel \tilde{T}) = -\frac{\gamma}{\varepsilon}\psi_k(\gamma) + m_k \left[\log m_k - \log Z(\gamma)\right],$$

where

$$\psi_k(\gamma) := \sum_j g_j T_{kj}(\gamma).$$

Differentiating with respect to γ yields

$$\frac{d}{d\gamma}\mathrm{KL} = -\frac{1}{\varepsilon}\psi_k(\gamma) - \frac{\gamma}{\varepsilon}\psi_k'(\gamma) - \frac{Z'(\gamma)}{Z(\gamma)}.$$

A direct computation gives

$$\frac{Z'(\gamma)}{Z(\gamma)} = -\frac{1}{m_k \varepsilon}\psi_k(\gamma),$$

so the first and last terms cancel, leaving

$$\frac{d}{d\gamma}\mathrm{KL} = -\frac{\gamma}{\varepsilon}\psi_k'(\gamma).$$

By Lemma 3, we have

$$\psi_k'(\gamma) = -\frac{m_k}{\varepsilon}\,\mathrm{Var}_{\pi_k(\gamma)}(g_n(z_k, Y)),$$

and therefore

$$\frac{d}{d\gamma}\mathrm{KL}(T(\gamma) \parallel \tilde{T}) = \frac{m_k \gamma}{\varepsilon^2}\,\mathrm{Var}_{\pi_k(\gamma)}(g_n(z_k, Y)).$$

Since $T_k(0) = \widetilde{T}_k$, we have $\mathrm{KL}(T_k(0) \parallel \tilde{T}_k) = 0$. Integrating from 0 to γ then yields

$$\mathrm{KL}(T_k(\gamma) \parallel \tilde{T}_k) = \frac{m_k}{\varepsilon^2}\int_0^\gamma \mathrm{Var}_{\pi_k(s)}(g_n(z_k, Y))\; s\, ds.$$

Applying the uniform lower bound $\mathrm{Var}_{\pi_k(\gamma)}(g_n(z_k, Y)) \geq \sigma_k^2$ from Lemma 3 yields

$$\mathrm{KL}(T_k(\gamma) \parallel \tilde{T}_k) \geq \frac{m_k}{\varepsilon^2}\int_0^\gamma s\, ds = \frac{m_k\, \sigma_k^2}{2\varepsilon^2}\gamma^2.$$

□

D Additional Theoretical Result

Lemma A2 (Quadratic lower bound on row projection gap)**.** *Fix a row k with row mass $m_k \geq 0$, and let $T_k(\gamma)$ be the exponential tilt of a reference row $\widetilde{T}_k$ as in Lemma 3. Then*

$$\frac{d}{d\gamma}KL(T_k(\gamma) \parallel \widetilde{T}_k) = \frac{m_k \gamma}{\varepsilon^2}\mathrm{Var}_{\pi_k(\gamma)}(g^n(z_k, Y)).$$

Consequently, if $\mathrm{Var}_{\pi_k(\gamma)}(g^n(z_k, Y)) \geq \sigma_k^2 > 0$ for $0 \leq \gamma \leq \bar{\gamma}$, then $G_{\lambda,k} \geq m_k \sigma_k^2 \gamma^2/(2\varepsilon^2)$.

E Gap time-series: roughness and spike regimes

Figure A1 shows representative gap sequences $\{G_{\lambda,k}\}_{k=h}^{N-1}$ for selected processes from Section 5.3, illustrating the qualitative difference between the two failure modes identified in the main text.

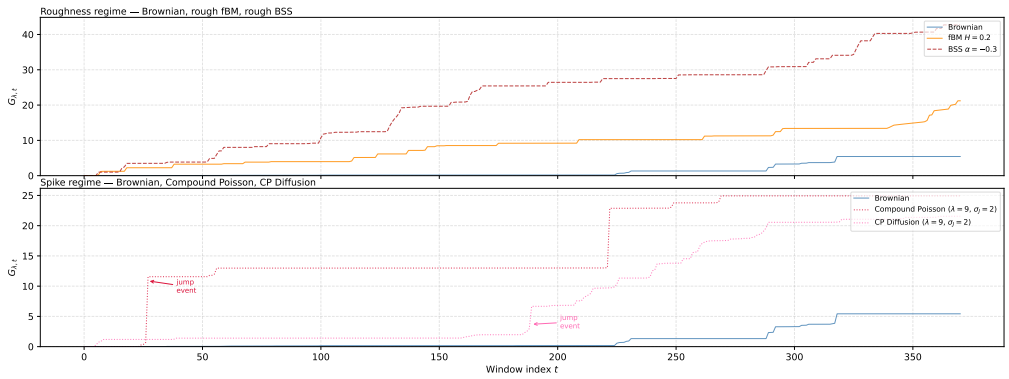
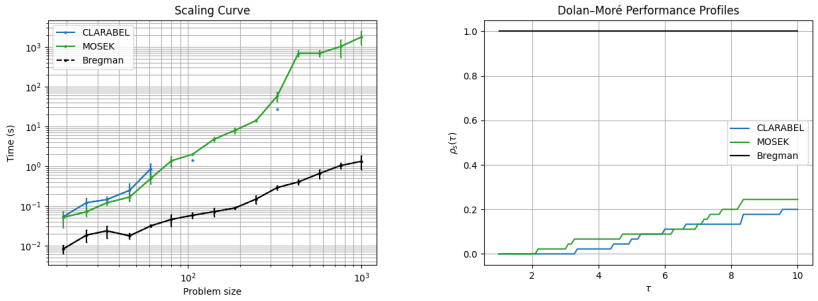


Figure A1: **Top:** Gap sequence $G_{\lambda,k}$ for Brownian motion, rough fBM ($H = 0.2$), and rough BSS ($\alpha = -0.3$). Rough processes exhibit sustained elevation: the gap is large in nearly every window, growing monotonically as the path accumulates constraint violations. **Bottom:** Gap sequence for Brownian motion, Compound Poisson ($\lambda_J = 9$, $\sigma_J = 2$), and CP Diffusion ($\lambda_J = 9$, $\sigma_J = 2$). Jump processes exhibit isolated spikes: the gap is near zero between jump events and jumps abruptly to a large value when a window contains one or more jumps (annotated). The two panels share the same x-axis (window index t) but differ in y-axis scale, reflecting the different magnitudes of the two regimes. Same hyperparameters as Figure 4.

F Computational complexity of the Bregman-Dykstra scheme

We compare the proposed Bregman–Dykstra implementation with a generic convex-optimisation baseline implemented in `cvxpy`, both solving (5) for the drift-constrained T^{proj}. For each problem size, both methods solve the same finite-dimensional projection problem, with the same cost matrix, entropic regularisation parameter, marginal penalty, and constraints.

Figure A2 shows the runtimes, in seconds, on an Apple M2 Pro (16 GB) machine running MacOS Tahoe 26.4.1. No parallelisation or warm-starting is used in either setting. The specialized algorithm uses the structure of the entropic projection problem: marginal updates reduce to Sinkhorn-type scalings, while predictable-drift constraints reduce to row-wise one-dimensional dual updates. This structure is not exploited by a generic solver. The resulting scaling behaviour supports the use of the Bregman–Dykstra formulation as a practical implementation of the diagnostic.



(a) Scaling curves.

(b) Dolan-Moré performance profiles.

Figure A2: Computational comparison between the proposed Bregman–Dykstra projection algorithm and a generic `cvxpy` solver. Panel (a) reports scaling curves as the discretisation resolution increases. Panel (b) reports Dolan–Moré performance profiles, showing the fraction of instances solved within a given runtime factor of the fastest method.