IDENTIFYING THE PREDICTABLE DRIFT OF A SEMIMARTINGALE FROM MARGINAL LAWS

Anonymous authors
Paper under double-blind review

ABSTRACT

A special semimartingale admits a unique decomposition $X = X_0 + M + A$ into a local martingale M and a predictable finite-variation part A. We consider the identification of A when X is observed only through its marginal laws at a succession of times, so that the pairing between observations, and with it the likelihood of the path, is lost. The estimand is then the projection of the sampled predictable compensator onto the observable feature filtration, namely the current state together with whatever randomness is shared across the population, so that at a fixed diffusion coefficient the marginal flow identifies the drift only up to a Markovian projection. We estimate it jointly with couplings between adjacent empirical marginals, constrained so that their conditional first moment agrees with a parametric drift, the parameter being identified by a rank condition upon the mean features. If the drift is an affine functional of an observed lag window, the joint problem is a convex quadratic programme whose solution is the pseudo-panel regression of econometrics. Our principal concern is the case, which we believe not to have been treated before, in which the drift is the output of a *hidden* linear dynamical system whose dynamics are themselves to be identified from the marginals. The joint problem is then a bilinear quadratically constrained programme, which we solve to certified global optimality by spatial branch and bound; with unpenalised state disturbances and a drift basis growing with the grid it is NP-hard already in latent dimension one, by reduction from ℓ^1 rank-one matrix approximation, whereas the complexity of the deterministic system at fixed latent dimension remains open. A block-coordinate decomposition offers a cheaper alternative. Its fixed point is available in closed form; it converges locally Q-linearly under either an exact global solution of the identification sub-problem or the acquiescence condition of Hardt, Ma and Recht; and it shows that the couplings inform the parameter through the mean displacements between marginals and through nothing else. For the estimator itself we obtain rates at a fixed mesh, separated into Monte-Carlo, estimation and grid contributions.

Special semimartingales form the natural structural class for continuous-time stochastic modelling within the classical Itô framework (Jacod & Protter, 2012; Jacod, 2012; Jacod & Shiryaev, 2003): the unique decomposition into a local-martingale and a predictable finite-variation component makes them the canonical domain for the stochastic calculus of mathematical finance (Beiglböck et al., 2011), where martingality is imposed under an equivalent pricing measure (Delbaen, 1992; Delbaen & Schachermayer, 1994) but is only a benchmark under the physical measure (Fama, 1970), for which semimartingality is the appropriate assumption (Rogers, 1997). How, then, is one to identify the predictable drift of an observed process?

When a full sample path is available the question is classical: the drift enters the likelihood through the Girsanov density (Liptser & Shiryaev, 2001a;b, Ch. 7, 17), or through the characteristics for general semimartingales (Jacod & Shiryaev, 2003, Ch. III–IV). Either way it is identified but not consistently estimable from a single path on a fixed horizon, the laws of W and $W + bt$ on $[0, T]$ being equivalent, so consistency needs a long time span, and under infill asymptotics only the volatility and the jump measure are identifiable (Aït-Sahalia & Jacod, 2014).

The question changes its character when the process is observed only through its *marginal laws* at successive times: population snapshots in which individuals are not tracked, cross-sectional panels, or samples whose pairing across times is unknown. The path is then unavailable, and its likelihood

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.

with it, leaving only the laws of X_{t_k} and $X_{t_{k+1}}$; yet many independent copies on a fixed horizon is precisely the regime in which a drift *can* be identified. The present paper develops an identification framework for that regime, resting on constrained optimal transport between adjacent empirical marginals: a coupling between the laws at t_k and t_{k+1} is the discrete analogue of the joint law of $(X_{t_k}, X_{t_{k+1}})$, and the requirement that its conditional first moment agree with a parametric drift is the empirical form of the Doob decomposition. The estimator recovers the drift from displacements between marginals, in the spirit of trajectory inference (Lavenant et al., 2024).

Settings in finance and insurance fit the framework. A (re)insurer observes at successive valuation dates the distribution of claims across a book, claims being unlinked across dates and sharing a persistent unobserved driver such as claims inflation, which is what the latent state represents. In options markets the risk-neutral law at each maturity follows from prices by the formula of Breeden & Litzenberger (1978) while the joint law does not; there the drift is fixed by the pricing measure and our constraint is that of martingale optimal transport (Beiglböck et al., 2013).

The principal contribution is the Non-Convex Setting of Section 2.2, in which the observed lag window gives way to a hidden linear state whose dynamics are to be identified from the marginals. We believe the problem not to have been posed before: trajectory inference fits a drift field non-parametrically, and the pseudo-panel literature of econometrics fits an autoregression upon cohort means. Section 4.2 writes the joint problem as a bilinear quadratically constrained programme, Section 4.3 says where its hardness resides, and Section 7 solves it to certified global optimality. Section 6 obtains in closed form the fixed point of a cheaper block-coordinate estimator, and that closed form explains how little the transport layer contributes to the parameter, what goes wrong when marginals are manufactured from a single path by windowing, and why the estimator is no rival to likelihood methods when paths are observed.

Throughout, $X = (X_t)_{t\in[0,T]}$ is a real-valued special semimartingale on $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, observed at n equally spaced times $t_k = k\Delta$, $k = 0, \dots, n-1$, $\Delta = T/(n-1)$. The samples $r_k := X_{t_k}$ inherit a discrete Doob decomposition upon the sub-filtration $\mathcal{G}_k := \mathcal{F}_{t_k}$, whose predictable compensator is the canonical discrete-time proxy for the continuous-time drift (Section 1). We identify it jointly with transport couplings $\{P_k\}$ between the empirical laws at adjacent times, as a single mathematical programme — convex in the first setting, a bilinear quadratically constrained programme in the second — and also by a block-coordinate scheme whose inner step is a drift-constrained Schrödinger bridge. Code reproducing every figure accompanies the submission.

Related work. Recovering dynamics from unpaired snapshots is trajectory inference (Hashimoto et al., 2016; Schiebinger et al., 2019; Chizat et al., 2022; Lavenant et al., 2024) and data-driven Schrödinger bridging (Pavon et al., 2021; Vargas et al., 2021; Bunne et al., 2023), which fit a bridge non-parametrically and read the drift off it; we impose instead a parametric predictable drift as a linear constraint upon the coupling (Section 6). The same observation scheme — repeated cross-sections with no linkage of individuals across dates — is the *pseudo-panel* setting of econometrics (Deaton, 1985; Moffitt, 1993), and Proposition 6.1 places our estimator within it (Appendix Q); neither literature admits a latent state with dynamics of its own. PO-MFL (Gu et al., 2025) does admit one, but imposes a prior upon its dynamics rather than identifying them. The M-step of the Non-Convex Setting is a classical output-error problem (Ho & Kálmán, 1966; Van Overschee & De Moor, 1996; Shumway & Stoffer, 1982; Ljung, 1999; Hazan et al., 2017; Simchowitz et al., 2018; Oymak & Ozay, 2019; Sarkar & Rakhlin, 2019; Tsiamis & Pappas, 2019); we draw on Hardt et al. (2018) for the acquiescence condition, which is what yields a *global* guarantee for a first-order method.

1 Setting

Definitions of stochastic basis, adapted, predictable and càdlàg processes, (local) martingales and (special) semimartingales are collected in Appendix A; we follow Protter (2005, Ch. I–III) and Jacod & Shiryaev (2003, Ch. I). The sampled process (r_k) admits the discrete Doob decomposition $r_k = r_0 + M_k^\Delta + A_k^\Delta$, with M^Δ a $\mathcal{G}$-martingale and A^Δ the unique $\mathcal{G}$-predictable process of finite variation,

$$A_k^\Delta - A_{k-1}^\Delta = \mathbb{E}\big[r_k - r_{k-1} \mid \mathcal{G}_{k-1}\big]. \tag{1}$$

This increment is the discrete target throughout, and it is consistent with the continuous-time object: if $A_t = \int_0^t a_s\,ds$ with a bounded and a.s. right-continuous at t, then $\Delta^{-1}(A^{\Delta}_k - A^{\Delta}_{k-1}) \to a_t$ in $L^1(\mathbb{P})$ along meshes containing t (Proposition C.1, Appendix C). From now on $\Delta = 1$. In discrete time, (1) exists for any integrable adapted process. We need an assumption on its relation to A:

Assumption 1.1 (Well-posedness). *X is an $\mathbb{F}$-special semimartingale with $\mathbb{E}\big[\sup_{t\le T}|X_t|\big] < \infty$ whose canonical decomposition $X = X_0 + M + A$ has M a martingale on $[0,T]$ and $A_t = \int_0^t a_s\,ds$ for a predictable a; the martingale property holds when a is bounded (Appendix B), and in particular it suffices that $X \in \mathcal{H}^1$ with A absolutely continuous.*

In discrete time the Doob decomposition of the integrable adapted sequence (r_k) is unique without further ado, a predictable martingale null at zero being zero. Assumption 1.1 relates it to the continuous-time drift: it furnishes a and makes M a true martingale, so that (1) is the $\mathcal{G}_{k-1}$-projection of $\int_{t_{k-1}}^{t_k} a_s\,ds$ (Proposition C.1), and it makes X special for any observer coarser than $\mathbb{F}$ (Appendix B). Without specialness the finite-variation part is not unique, and (8) would constrain an unidentified object while remaining perfectly solvable. The assumption thus secures for the *estimand* what (T1)–(T2) of Section 5 and Assumption P.1 secure for the *solver*.

The drift is defined once and for all by the $\mathbb{F}$-canonical decomposition, a being the density of the $\mathbb{F}$-predictable part A. Every other drift-like object occurring below — the sampled compensator (1), the model b_k, the drift implied by a coupling — is a *conditional expectation* of that one object upon a coarser σ-field, and not a rival definition; the constraints of Section 4 are therefore moment conditions on projections. Three σ-fields occur, nested,

$$\mathcal{G}_k := \mathcal{F}_{t_k} \;\supseteq\; \mathcal{H}_k \;\supseteq\; \mathcal{H}^N_k, \qquad \int_{t_k}^{t_{k+1}} a_s\,ds \xrightarrow{\mathbb{E}[\cdot|\mathcal{G}_k]} A^{\Delta}_{k+1} - A^{\Delta}_k \xrightarrow{\mathbb{E}[\cdot|\mathcal{H}_k]} b_k \xrightarrow{\Pi_N} \widehat{\cdot}, \quad (2)$$

with $\mathcal{H}_k$ the *feature* σ-field of Section 2.1 and $\mathcal{H}^N_k$ its image under the grid map. The first object is the integrated drift, not a_{t_k}; and the last arrow is the nearest-atom map Π_N, which alters the random variable rather than conditioning it, so the grid estimand is the compensator of the *rounded* process (Lemma R.3). The compensator conditions upon $\mathcal{G}_k$, which in the marginal-law regime is not observable; Section 2 selects the observable $\mathcal{H}_k$ which is to replace it. The first two arrows are L^1-contractions; the last is not. It depends upon N, whence the estimand is itself N-dependent and the grid must be counted part of the statistical model and not merely of the discretisation.

2 Two models for the predictable drift

The identification of A requires a parametrisation, and two are considered here: in the *Convex Setting* the projected drift is an affine functional of a finite window of *observed* past values; in the *Non-Convex Setting* it is the output of a linear dynamical system with an *unobserved* state. Upon this single choice turns the difference between a jointly convex identification problem and an NP-hard one.

2.1 Convex Setting: an autoregressive drift

Before a drift model is written down, it must be said upon what the model is permitted to depend. Fix for each k a σ-field

$$\mathcal{H}_k \;=\; \sigma\big(X_{t_k}\big) \vee \mathcal{S}_k \;\subseteq\; \mathcal{G}_k, \quad (3)$$

the join of the *individual* coordinate X_{t_k}, upon which the estimator conditions through the grid atom, with a *common* σ-field $\mathcal{S}_k$ carrying whatever is shared across the sampled population at time k. The requirement which governs all that follows is that *every feature entering b_k be $\mathcal{H}_k$-measurable*. A regressor lying outside $\mathcal{H}_k$ carries a coefficient attached to information upon which the estimator never conditions: well defined as part of a structural model, but not identified from this observation scheme, which is what matters here.

With $\mathcal{H}_k$ fixed, assume the sampled compensator admits a *Markovian-projection* representation tied to a finite past,

$$\mathbb{E}\big[X_{t_{k+1}} - X_{t_k} \mid \mathcal{H}_k\big] \;=\; b_k\big(X_{t_k};\theta_\star\big), \qquad b_k(x;\theta) := \alpha + \beta_1 x + \sum_{j=1}^{p-1} \beta_{j+1}\, s_{k-j}, \quad (4)$$

for some $\theta_\star \in \mathbb{R}^{p+1}$, in the spirit of Gyöngy (1986) and Brunick & Shreve (2013), the statistics $s_{k-1}, \ldots, s_{k-p+1}$ being $\mathcal{S}_k$-measurable. What these may be is settled by the observation scheme of Section 3 and is in no way free. (i) For a *single path*, $\mathcal{S}_k = \sigma(r_j : j \le k)$ and $s_j = r_j$: the classical autoregression, and the only case in which "the past of the process" is a regressor at all. (ii) For a *population with a common factor*, individuals are not tracked, so that no individual past exists; $\mathcal{S}_k$ is generated by the shared randomness, and s_j must be a statistic of it, such as the input u_j or a functional of the flow like $\bar{x}_j = \mathbb{E}_{\mu_j}[X]$. (iii) For *independent copies*, $\mathcal{S}_k$ is degenerate, so no *random* shared feature survives — though a known deterministic input is measurable with respect to a trivial σ-field and can still drive an identifiable transfer function, so it is a random common factor that is lost, not time dependence. Appendix E sets out why (ii) may not borrow the regressors of (i), a tempting error which quietly converts a well-posed estimand into a meaningless one.

We have here an autoregressive structural model whose parameter is to be identified from the observations. Its continuous-time reading is supplied by Proposition C.1: $b_k(x;\theta_\star) = \Delta\,\mathbb{E}[a_{t_k} \mid \mathcal{H}_k, X_{t_k} = x] + o(\Delta)$ as $\Delta \downarrow 0$. We work, however, at a *fixed* mesh, so that the target is the discrete compensator increment itself and Δ is absorbed into θ; the limit asserts that this target approaches the continuous-time drift as the mesh is refined, not that the two agree at the mesh employed.

2.2 Non-Convex Setting: a drift driven by a hidden linear state

The lag window in (4) is a memory of *finite* length, whereas a latent factor possessing dynamics of its own — a slowly mean-reverting expected return, a hidden regime variable, an unobserved liquidity state — gives rise to an impulse response of *infinite* length, and its truncation at lag p is a misspecification which no choice of θ will repair. In the Non-Convex Setting the observed lag window is accordingly replaced by a hidden state evolving as a linear system,

$$z_{k+1} = A\,z_k + B\,u_k, \qquad b_k(x;\Theta) = \varphi(x)^\top H\,z_k \;\xrightarrow[\varphi(x)=(1,x)^\top]{}\; b_k(x;\Theta) = \alpha + \beta x + c^\top z_k, \tag{5}$$

with $z_k \in \mathbb{R}^d$, u_k an observed exogenous input, $\varphi : \mathbb{R} \to \mathbb{R}^q$ a fixed basis and $\Theta = (A, B, H, z_0)$. In the affine specialisation on the right, which is used throughout the numerical work, the latent state shifts the intercept of an otherwise affine drift; it is the discrete analogue of $dX_t = (\alpha + \beta X_t + c^\top Z_t)dt + \sigma dW_t$ with Z unobserved.

Remark 2.1 (The Convex Setting is the memoryless case of the Non-Convex Setting)**.** *With $d = p-1$, $u_k = s_k$, $B = e_1$, A the nilpotent shift $Ae_j = e_{j+1}$ ($Ae_d = 0$) and $c = (\beta_2, \ldots, \beta_p)^\top$ one has $z_k = (s_{k-1}, \ldots, s_{k-p+1})^\top$, and (5) reduces to (4). The Convex Setting is thus the sub-model in which A is* known *and the state is an* observed *function of the data; it is the hiddenness of the state, and not its dimension, which destroys convexity.*

Remark 2.2 (Canonical coordinates do not identify the allocation of poles)**.** *Fixing the gauge removes the $GL(d)$ orbit but not every ambiguity of a mean-only estimator. With $d = 1$, known input and $z_0 = m_0 = 0$, write $z_{k+1} = rz_k + u_k$ and $m_{k+1} = sm_k + cz_k$ for the population mean, so $\beta = s - 1$ and $a = -r$. Eliminating z gives $m_{k+2} = (r+s)m_{k+1} - rs\,m_k + cu_k$, symmetric in r and s: interchanging them leaves every mean observation unchanged, for every input. So $(a, \beta, c) = (-0.2, -0.6, c)$ and $(-0.4, -0.8, c)$ are indistinguishable from the means, and both lie in the acquiescent region used below. This is not a similarity gauge, the allocation between observed and latent state having changed; and by Proposition 6.1 the estimator sees the means alone. Claims below about recovering (a, c, α, β) are to be read modulo this ambiguity, which only within-marginal information, or a prior upon the allocation, could break.*

Since Θ is identified only up to similarity, we fix the controllable canonical form of Hardt et al. (2018, Sec. 1.7): $A = \mathrm{comp}(a)$ is the companion matrix of $p_a(\zeta) = \zeta^d + a_1\zeta^{d-1} + \cdots + a_d$, $B = e_d$, and (5) is parametrised by $(a, c, \alpha, \beta) \in \mathbb{R}^{2d+2}$. Stability, $\rho(A) < \varrho$, is not a convex condition upon a, and we employ in its place the relaxation $\mathcal{B}_\varrho = \{a : p_a(\zeta)/\zeta^d \in \mathcal{W}\ \forall |\zeta| = \varrho\}$ of Hardt et al. (2018, Sec. 4), where

$$\mathcal{W} = \big\{\zeta \in \mathbb{C} :\ \mathrm{Re}\,\zeta \ge (1+\tau_0)|\mathrm{Im}\,\zeta|\big\} \cap \big\{\zeta :\ \tau_1 \le \mathrm{Re}\,\zeta \le \tau_2\big\} \tag{6}$$

is a truncated wedge; such a are called *ϱ-acquiescent*, and they satisfy $\rho(\mathrm{comp}(a)) < \varrho$. Since $p_a(\zeta)/\zeta^d = 1 + \sum_j a_j\zeta^{-j}$ is affine in a, sampling the circle at J points reduces $\mathcal{B}_\varrho$ to $4J$ linear inequalities (Appendix F).

3 CONSTRAINED TRANSPORT BETWEEN ADJACENT MARGINALS

Write μ_k, ν_k for the empirical laws of the sampled process at t_k and t_{k+1}, projected onto a common atomic grid $\{x_1, \dots, x_N\}$ by nearest-bin assignment and identified with probability vectors in Δ_{N-1} (Appendix D). A coupling $P_k \in \mathbb{R}_{\geq 0}^{N \times N}$ with marginals (μ_k, ν_k) is a one-step Markov kernel from the law of r_k to that of r_{k+1}, the discrete analogue of the joint law of $(X_{t_k}, X_{t_{k+1}})$. We identify the drift parameter jointly with a sequence $\{P_k\}$ of such couplings whose conditional first moment of the increment is the modelled compensator,

$$\sum_{j=1}^{N} P_k[i,j]\,(x_j - x_i) \;=\; b_k(x_i;\Theta)\,\mu_k[i], \qquad i:\ \mu_k[i] > 0, \tag{7}$$

the empirical version of (1) at $\Delta = 1$, and which minimise the total cost $\sum_k \langle C,\, P_k\rangle$ for the normalised quadratic $C_{ij} = (x_i - x_j)^2 / \max_{kl}(x_k - x_l)^2$. Writing $\mathrm{inc}_{ij} := x_j - x_i$ and $d_k := (P_k \odot \mathrm{inc})\mathbf{1}$ for the empirical compensator vector, (7) reads

$$d_k[i] \;=\; b_k(x_i;\Theta)\,\mu_k[i] \qquad \text{for every } i \text{ with } \mu_k[i] > 0. \tag{8}$$

Unpaired marginals determine no joint law, whereas the conditional moment in (7) is a functional of one, so a joint law must be posited before the constraint can be written down. At the population level the Fréchet class of couplings with the two marginal laws is the partially identified set for the one-step joint law, and the transportation polytope $\Pi(\mu_k, \nu_k)$ is its sample analogue; from it (8) cuts the elements consistent with the model, what remains is of order N^2, and it is the entropy which makes the selection, the cost being constant upon that set (Proposition T.1). The coupling is thus a modelling assumption, and infeasibility at finite M speaks of the empirical marginals, not of the population (Appendix G).

The pair index runs over $\mathcal{P} = \{p-1, \dots, n-2\}$, $n_{\text{pairs}} = |\mathcal{P}|$; x_k^-, x_k^+ are the extreme atoms charged by ν_k, so under any coupling the conditional mean increment out of x_i lies in $[x_k^- - x_i,\ x_k^+ - x_i]$.

The marginals may arise in three ways, upon which the randomness of the flow $(\mu_k)_k$, and hence the content of $\mathcal{S}_k$ in (3), depends. With *independent copies* the empirical marginals converge as $M \to \infty$ to the deterministic law of X_{t_k}, $\mathcal{S}_k$ is degenerate, and by Section 2.1(iii) there is no dynamic model left to fit. With a *common factor* the flow is random and measurable with respect to the filtration of the factor, which is precisely $\mathcal{S}_k$; this is the scheme we adopt, and the only one in which the Non-Convex Setting has content, the hidden state *being* the shared randomness. Under *rolling windows* the flow carries the randomness of a single path; we retain it as a diagnostic, but it carries an aggregation bias which does not vanish with the sample size (Section 6). It follows that M and n are not interchangeable: the M-step has $K = n_{\text{pairs}}$ residuals whatever M may be, and M enters only through the Monte-Carlo error of Theorem R.10, so $M = 600$ below is a noise level and not a sample size. Appendix D gives both schemes and the conventions for the grid and the cost.

4 THE TWO OPTIMISATION PROBLEMS

4.1 CONVEX SETTING: A JOINTLY CONVEX PROGRAMME

In the Convex Setting, the identification problem in $(\theta, \{P_k\})$ is a single penalised convex programme, the discrete predictable-compensator constraint (8) being relaxed by an ℓ^1 slack:

$$\begin{aligned} \min_{\theta, \{P_k\}, \{s_k^\pm\}} \quad & \sum_{k \in \mathcal{P}} \langle C,\, P_k\rangle + \lambda_{\text{drift}} \sum_{k \in \mathcal{P}} \mathbf{1}^\top (s_k^+ + s_k^-) + \lambda_{\text{reg}} \big(\alpha^2 + \|\beta\|^2\big) \\ \text{s.t.} \quad & (P_k \odot \mathrm{inc})\mathbf{1} - b_k(\mathrm{grid};\theta) \odot \mu_k = s_k^+ - s_k^-, \\ & P_k \geq 0,\ P_k\mathbf{1} = \mu_k,\ P_k^\top \mathbf{1} = \nu_k,\ s_k^+, s_k^- \geq 0 \qquad (k \in \mathcal{P}). \end{aligned} \tag{9}$$

The Tikhonov term is small ($\lambda_{\text{reg}} = 10^{-4}$ by default) and serves to keep the problem strictly convex in θ; letting $\lambda_{\text{reg}} \to 0$ recovers a pure linear programme, and letting $\lambda_{\text{drift}} \to \infty$ imposes (8) exactly whenever it is feasible, which is why the slack is there. The programme is *jointly* convex in $(\theta, \{P_k\})$, with $(p+1) + n_{\text{pairs}}(N^2 + 2N)$ variables, so that its global optimum is attainable by an interior-point solver; it is its $\mathcal{O}(nN^2)$ size which occasions the decomposition of Section 5.

4.2 Non-Convex Setting: a quadratically constrained programme

In the Non-Convex Setting, the drift entering (8) is no longer affine in the unknowns; it is the output $\varphi(x_i)^\top H z_k$ of the hidden state, and the state itself obeys $z_{k+1} = Az_k + Bu_k$. Both couplings are *bilinear*, and the complete problem becomes

$$\begin{aligned}\min_{\Theta,\{z_k\},\{w_k\},\{P_k\},\{s_k^\pm\}} \quad & \sum_{k\in\mathcal{P}} \langle C,\, P_k\rangle + \lambda_{\text{drift}} \sum_{k\in\mathcal{P}} \mathbf{1}^\top (s_k^+ + s_k^-) + \lambda_{\text{reg}} \|\Theta\|^2 + \lambda_w \sum_k \|w_k\|^2 \\ \text{s.t.} \quad & (P_k \odot \mathrm{inc})\mathbf{1} - \big(\Phi H z_k\big) \odot \mu_k = s_k^+ - s_k^-, \qquad z_{k+1} = Az_k + Bu_k + w_k, \\ & P_k \geq 0,\ P_k\mathbf{1} = \mu_k,\ P_k^\top \mathbf{1} = \nu_k,\ s_k^\pm \geq 0 \ \ (k \in \mathcal{P}), \quad a \in \mathcal{B}_\varrho,\ z_0 = 0.\end{aligned} \tag{10}$$

Here Φ stacks the basis rows $\varphi(x_i)^\top$, w_k is an optional state innovation with weight $\lambda_w \geq 0$ ($w \equiv 0$ giving the deterministic LDS), and $a \in \mathcal{B}_\varrho$ abbreviates the $4J$ inequalities of (6). The whole of the non-convexity resides in the $\mathcal{O}(K(d^2+d))$ bilinear terms Az_k and Hz_k across the horizon — in companion form $K(d+1)$ of them — the $\mathcal{O}(nN^2)$ transport variables, the marginals and the slacks being linear; and it is this structure which spatial branch and bound exploits. We solve (10) with Gurobi (Gurobi Optimization, LLC, 2024) under `NonConvex=2`, obtaining a solution which is globally optimal together with a certificate of optimality, up to numerical precision.

The innovation weight, and the initial condition. The weight λ_w is a matter of modelling and not a constant to be tuned; it places (10) within a one-parameter family whose two endpoints are the cases treated here. At $\lambda_w = \infty$, that is $w \equiv 0$, one has the deterministic system (5), whose complexity at fixed d remains open; at $\lambda_w = 0$, Theorem 4.1 asserts NP-hardness. The interior $\lambda_w \in (0, \infty)$ is the filtering model, and is better conditioned than either extreme: the objective is then strongly convex in z for given (a, c, α, β), so that a Kalman smoother eliminates the state exactly and leaves a non-convex problem in $2d+2$ variables with a C^1 value function, which supplies the uniqueness half of Assumption P.1. A prior $z_0 \sim N(m_0, P_0)$ — so that $\mathcal{F}_0$ is not trivial, as it cannot be when the factor possesses a distribution rather than a value — adds a Tikhonov term and makes the estimator a posterior mode, under which even a noiseless state yields a non-degenerate problem of learning (Appendix J). In the experiments $z_0 = 0$.

4.3 Hardness of the complete problem in the Non-Convex Setting

The gulf between the two settings is not an artefact of the formulation. Write $\mathrm{OPT}_{\mathrm{NC}}$ for the value of (10) at $\lambda_{\text{reg}} = \lambda_w = 0$.

Theorem 4.1 (NP-hardness)**.** *Deciding whether* $\mathrm{OPT}_{\mathrm{NC}} \leq V$ *for a given rational* V *is NP-hard, already for* $d = 1$ *and* $\lambda_w = 0$*, for the quadratic transport cost, for a drift basis* φ *whose dimension grows with the grid, and even when the supplied pairs are the consecutive marginals of a* single *chain upon a common grid of polynomial size — whereas on the same data problem (9) of the Convex Setting is a convex quadratic programme of the same size.*

The proof (Appendix S) proceeds by reduction from ℓ^1-norm rank-one matrix approximation of a *sign* matrix, shown NP-hard by Gillis & Vavasis (2018), in three steps which may be of some independent interest. The transport block is *pinned* by choosing marginals whose monotone matching is the unique optimal vertex, so that the plans remain optimal uniformly in Θ once λ_{drift} falls below an explicit threshold; what then remains is a weighted ℓ^1 fit of the prescribed displacements by $\big(b_k(x_i;\Theta)\big)_{k,i}$; and a latent state of dimension d confines that matrix to rank at most d. Since the data are the consecutive marginals of one process, the rows of the matrix cannot simply be read off as unrelated source–destination pairs: the K encoding pairs are interleaved with K *reset* pairs, upon whose source atoms every basis function vanishes, so that their contribution to the objective is a constant. Remark S.2 delimits the statement in three ways, of which one deserves emphasis here: the reduction needs a drift basis whose dimension grows with the grid, and for the fixed affine basis of (5) the programme at $\lambda_w = 0$ is a linear programme for every d. At $\lambda_w = \infty$ with d held fixed the complexity is open, and allowing d to grow renders the fit exact and the problem trivial.

The experiments of Section 7 use the fixed affine basis at $\lambda_w = \infty$ and are therefore *not* instances of the hard family. What branch and bound answers there is the bilinear non-convexity of $z_{k+1} = \mathrm{comp}(a) z_k + e_d u_k$ coupled to $c^\top z_k$, a genuine obstruction to convexity and the reason the M-step needs a global method, and not the hardness proved above.

5 A BLOCK-COORDINATE DECOMPOSITION

Both programmes carry $\mathcal{O}(nN^2)$ transport variables, and (10) carries them into every node of the branch and bound. The decomposition set out here keeps them out of the parameter problem. It is not a solver for (9) or (10): it imposes (8) exactly rather than through a slack and regularises the transport entropically, and is therefore a distinct estimator, whose fixed point Section 6 identifies. Its E-step is indifferent to which setting is in force, so that the non-convexity is confined to the M-step.

Adding the entropy $\epsilon\mathsf{H}(P_k) = -\epsilon\sum_{ij} P_k[i,j]\log P_k[i,j]$ to each pairwise cost and treating (8) as a hard constraint gives the discrete Schrödinger bridge between (μ_k, ν_k) with prescribed first conditional moment (Föllmer, 1988; Léonard, 2014), with Lagrangian

$$\mathcal{L}_k(P_k, \Gamma_k; \Theta) = \langle C,\, P_k\rangle - \epsilon\mathsf{H}(P_k) + \sum_{i:\,\mu_k[i]>0} \Gamma_k[i]\Big(\sum_j P_k[i,j](x_j - x_i) - b_k(x_i;\Theta)\mu_k[i]\Big), \quad (11)$$

the marginal constraints being imposed by Sinkhorn scalings. The **E-step** is the saddle point of $\mathcal{L}_k$ in (P_k, Γ_k) at fixed Θ; the **M-step** revises Θ from the compensators d_k. Two facts govern the E-step. With the marginals and the row means fixed the quadratic cost is constant, so that the plan is the maximum-entropy element of the feasible set for every $\epsilon > 0$ (Proposition T.1); and the tower property forces $\sum_i \mu_k[i] b_k(x_i;\Theta) = \mathbb{E}_{\nu_k}[X] - \mathbb{E}_{\mu_k}[X]$ upon any coupling with those marginals, so the E-step projects not upon b_k but upon $\tilde{b}_k = b_k + \delta_k$,

$$\delta_k = \big(\mathbb{E}_{\nu_k}[X] - \mathbb{E}_{\mu_k}[X]\big) - \sum_i \mu_k[i]\, b_k(x_i;\Theta), \quad (12)$$

clipped into the reachable range $[x_k^- - x_i,\, x_k^+ - x_i]$. Two regularity hypotheses recur below: (T1) the marginals are strictly positive on their support, and (T2) the clipped target lies strictly inside the reachable range in every active row, which the clipping enforces; neither guarantees that the constraint sets intersect, which is a convex-order condition (Remark G.1). The projection interleaves log-domain Sinkhorn rescalings with an exact row-wise Newton step upon Γ (Appendix I).

In the Convex Setting the M-step is the closed-form weighted least squares

$$\theta^{(\ell+1)} = \arg\min_\theta \sum_{k\in\mathcal{P},\, i:\mu_k[i]>0} \mu_k[i]\Big(\frac{d_k[i]}{\mu_k[i]} - b_k(x_i;\theta)\Big)^2 + \lambda_\theta\|\theta\|^2 = (X^\top W X + \lambda_\theta I)^{-1} X^\top W y, \quad (13)$$

a regression of the empirical predictable density $y[k,i] = d_k[i]/\mu_k[i]$ upon the features $\phi_{ki} = (1, x_i, s_{k-1}, \ldots, s_{k-p+1})$ with weights $W = \mathrm{diag}(\mu_k[i])$. In the Non-Convex Setting the same criterion is no longer a linear regression, z_k being unobserved and generated by the unknowns:

$$\min_{a,c,\alpha,\beta} \sum_{k\in\mathcal{P}} \sum_{i:\mu_k[i]>0} \mu_k[i]\Big(y_k[i] - \alpha - \beta x_i - c^\top z_k\Big)^2 \text{ s.t. } z_{k+1} = \mathrm{comp}(a) z_k + e_d u_k,\ a \in \mathcal{B}_\varrho. \quad (14)$$

The sum over atoms separates (Appendix K),

$$(14) \;=\; \min_{a,c,\alpha,\beta} \sum_{k\in\mathcal{P}} \big(\alpha + \beta\bar{x}_k + c^\top z_k - \bar{y}_k\big)^2 \;+\; \big(Q_2\beta^2 + Q_1\beta + Q_0\big), \quad (15)$$

with $\bar{x}_k = \sum_i \mu_k[i] x_i$, $\bar{y}_k = \mathbf{1}^\top d_k$ and Q_2, Q_1, Q_0 second moments of the grid and of y, so that the sub-problem has $2d + 2$ parameters and $K = n_{\text{pairs}}$ residuals whatever N may be. Introducing $s_k = c^\top z_k$ confines every bilinear term to the $K(d+1)$ equalities $z_{k+1} = \mathrm{comp}(a) z_k + e_d u_k$, $s_k = c^\top z_k$. We solve (15) **(a) globally**, by spatial branch and bound; **(b) by projected gradient**, Algorithm 1 of Hardt et al. (2018); or **(c) from a Ho–Kalman realisation** (Ho & Kálmán, 1966) of the Markov parameters of $u \mapsto \bar{y}$ (Appendix K).

6 FIXED POINTS AND CONVERGENCE OF THE DECOMPOSITION

An exact E-step with no clipping active gives $\mathbf{1}^\top d_k = \gamma_k := \mathbb{E}_{\nu_k}[X] - \mathbb{E}_{\mu_k}[X]$, the *mean displacement* between adjacent marginals, and gives it for *every* plan with the prescribed marginals. The coupling therefore influences the M-step only in so far as it fails to be feasible, and the E-step enters through $(\bar{x}_k, \gamma_k)$ and nothing besides.

Proposition 6.1 (Fixed points). *Write $\bar{\phi}_k = \sum_i \mu_k[i]\phi_{ki}$ and $Q = \sum_k \mathrm{Var}_{\mu_k}(X)$. With $\lambda_\theta = 0$, $\bar{\theta}$ is a fixed point in the Convex Setting if and only if it is the ordinary least squares regression (30) of the mean displacements upon the mean features; and $\bar{\Theta}$ is a fixed point of a single-valued globally solved M-step in the Non-Convex Setting if and only if it is the selected global minimiser of $J(\Theta) + Q(\beta - \bar{\beta})^2$ subject to the constraints of*

$$J(\Theta) = \sum_{k \in \mathcal{P}} \big(\alpha + \beta \bar{x}_k + c^\top z_k - \gamma_k\big)^2, \quad z_{k+1} = \mathrm{comp}(a) z_k + e_d u_k, \;\; a \in \mathcal{B}_\varrho. \tag{16}$$

In neither case do the couplings $\{P_k\}$ appear.

In the population scheme $\gamma_k = \bar{\phi}_k^\top \theta_\star$ exactly, so the estimator of the Convex Setting is Fisher consistent when the mean-feature Gram matrix is non-singular, whereas under windowing it carries an aggregation bias which does not vanish with n (Corollary Q.1). The regression (30) is the pseudo-panel estimator of Deaton (1985) and Moffitt (1993) with the population as one cohort (Appendix Q); of KN moment conditions it draws upon K, the E-step having annihilated the within-marginal variation, so the shape of μ_k governs the rate but never the limit. It is the second half of the proposition which makes the Non-Convex Setting tractable by decomposition: the sub-problem is *precisely* the identification problem of Hardt et al. (2018) for the scalar output γ_k, so that Ho–Kalman, $\mathcal{B}_\varrho$ and branch and bound apply without modification.

As to convergence, at an interior feasible point satisfying (T1)–(T2) the E-step is a cyclic projection upon three sets affine in P_k and converges locally linearly at a rate $c_E < 1$ (Theorem L.1). In the Convex Setting the outer loop is locally Q-linear with a contraction $\kappa_0 < 1$ given in closed form, to the fixed point of the inexact map, which lies within $\mathcal{O}(c_E^{L_{\mathrm{inner}}})$ of $\bar{\theta}$ for L_{inner} inner sweeps (Theorem O.2). With a hidden state the M-step is no longer a fixed linear operator and is invariant under $GL(d)$, so everything is read in the canonical coordinates of Appendix F. Solved to global optimality it is again locally Q-linear (Theorem P.3) under Assumption P.1, which requires a *separation* between the global minimiser and every other local one, failing which the iteration map is discontinuous; for an acquiescent system solved by projected gradient the rate is preserved at $\mathcal{O}(\varepsilon^{-2})$ gradients (Theorem P.2), though the weak quasi-convexity of Hardt et al. (2018) holds for their idealised risk and is not claimed for the criterion actually minimised. Remark 2.2 shows that a latent pole and the observed-state pole may be interchanged without altering any mean, so that Assumption P.1 can fail for the generating model of Section 7 and parameter errors there are read modulo that allocation.

Under hypotheses (R1)–(R6) of Appendix R, which define the pseudo-true $\theta_\star(\Delta)$, the residual variance ς^2 and the truncated mass τ_R,

$$\|\hat{\theta} - \theta_\star(\Delta)\| = \mathcal{O}_p\big(\Delta^{-1} M^{-1/2}\big) + \mathcal{O}_p\big(\varsigma K^{-1/2}\big) + \mathcal{O}\big(\Delta^{-1}(h^{\,r} + \tau_R)\big), \tag{17}$$

the three terms being Monte-Carlo, estimation and grid error, with $r = 1$ in general and $r = s + 1$ when the conditional marginal densities are of class C^s. The factors Δ^{-1} belong to the velocity normalisation of the joint limit and are absent at the fixed mesh employed here; the bound upon $\|\hat{\Theta} - \Theta_\star\|$, the mesh term $\mathcal{O}(\Delta)$ arising only when a continuous-time coefficient is insisted upon, and the asymptotic normality are in Appendix R.

7 Numerical experiments

We generate data by the Euler–Maruyama recursion of an Itô diffusion whose drift carries a hidden factor,

$$X_{k+1} = X_k + \big(\alpha + \beta X_k + c^\top Z_k\big)\Delta + \sigma\sqrt{\Delta}\,\xi_k, \qquad Z_{k+1} = \mathrm{comp}(a) Z_k + e_d u_k, \qquad \xi_k \sim N(0,1), \tag{18}$$

the recursion being adopted on purpose, since the exact transition of the diffusion would have increment slope $e^{\beta\Delta} - 1 = -0.330$ rather than -0.4, and it is the coefficients of the *discrete* model that we seek. We take u observed, $d = 2$, $\sigma = 0.35$, $\Delta = 1$, $(\alpha, \beta, c) = (0, -0.4, (0.4, 0.4))$, and the common-factor scheme: $M = 600$ copies driven by one factor path, observed at $n = 25$ times without pairing, giving $K = 24$ pairs upon a grid of $N = 14$ atoms, with $\epsilon = 0.05$ in the E-step. Here $a = (-0.6, 0.08)$ is ϱ-acquiescent and $a = (-1.2, 0.35)$ is not. The figures and the remaining tables are in Appendix Y, each naming the script which produces it.

Table 1: Forecasting the drift from the observed input: each method is fitted upon the first K_{tr} pairs and predicts b_k for the next 12 steps from u and the current snapshot mean; $\|b_k - b_k^{\mathrm{true}}\|_{L^2(\mu_k)}$ averaged over those steps, median over 5 seeds with the range, at $N = 14$, $M = 600$ (`code/compare_regimes.py`).

	$K_{\mathrm{tr}} = 24$		$K_{\mathrm{tr}} = 96$	
method	median	range	median	range
ours, Non-Convex Setting (identified (a, c, α, β))	0.043	[0.008, 0.051]	0.012	[0.002, 0.013]
FIR of u, four lags, with $(1, \bar{x}_k)$	0.050	[0.045, 0.074]	0.042	[0.042, 0.047]
PO-MFL-style AR(1) prior on the residual, run forward	0.351	[0.275, 0.740]	0.489	[0.295, 0.526]
Waddington-OT surrogate, last field carried forward	0.589	[0.452, 0.893]	0.813	[0.496, 1.016]
Waddington-OT surrogate *given the next snapshot*	0.073	[0.069, 0.083]	0.080	[0.076, 0.088]

On instances small enough for (10) to be closed ($N = 10$, $K = 11$, 1361 variables) spatial branch and bound returns a certified global optimum in under a second, and the parameter so obtained lies within 0.05–0.15 of the truth (Table 7). Upon the non-acquiescent system the restriction to $\mathcal{B}_\varrho$ costs two orders of magnitude in the objective, and projected gradient fails to solve even the restricted problem on two instances in five, a good many random restarts coming to rest above the certified value (Figure 4). The complete programme and the decomposition agree in their distance from the truth to within 10%, and Figure 6 compares their running times as N grows.

What the identification of the dynamics is worth is best seen out of sample (Table 1). Having been fitted upon the first K_{tr} pairs, our estimator predicts the drift over the following twelve steps from the observed input and the current snapshot mean alone. A per-step method has no forecast to offer beyond repeating its last field, and PO-MFL itself cannot enter, since it needs the snapshots it would be asked to predict; what can enter is its prior, an AR(1) fitted to the observable residual and run forward, which with a finite impulse response of u (Remark 2.1) is the competitor that deserves to be taken seriously. Our error falls by a factor of nearly four as K_{tr} grows from 24 to 96 while none of the competitors moves; at 24 the M-step settles, on some seeds, upon the interchanged allocation of poles of Remark 2.2, which the longer record resolves, and at 96 our error is less than a sixth of what the per-step method attains when it is handed the very snapshot whose drift it is asked to predict. Table 5 of Appendix U shows the same pooling of information across the horizon in sample, against M and K; Appendix Y records the decomposition upon this instance.

In sample, the hidden state earns its place against the finite-memory alternatives, namely lags of $\bar{x}$ and a finite impulse response of the input u, the latter being by Remark 2.1 the proper observed-feature competitor: the latent model reaches a given residual with fewer parameters, and the more slowly the factor mixes the greater the saving (Figure 3(b,c)). Appendix U sets the estimator beside surrogates of Waddington-OT (Schiebinger et al., 2019), of the path-space mean-field Langevin estimator (Chizat et al., 2022) and of PO-MFL (Gu et al., 2025), each receiving the same snapshots upon the same grid and each scored against the known drift, with running times (Table 4). There we improve upon the two that carry no latent state by a factor of three at $N = 100$, whereas PO-MFL is granted a prior upon the latent dynamics (!) where we must identify them, and against it the comparison is close, 0.026 against 0.020 with overlapping seed ranges; a misspecification of that prior's innovation scale by a factor of two removes its advantage.

8 Conclusion

The recovery of the predictable drift from marginal laws is a convex quadratic programme, and at bottom a pseudo-panel regression, when the drift is an affine functional of an observed lag window. When it is the output of a hidden linear system whose dynamics are themselves to be identified, a problem which we believe to be new, it is a bilinear programme which spatial branch and bound closes at the sizes considered here, and which is NP-hard in latent dimension one once the state disturbances are unpenalised and the drift basis grows with the grid. The decomposition is a cheap alternative, its local rate in the latter case resting upon a separation assumption without convex counterpart; the transport layer is idle for the estimate (Proposition 6.1) yet returns the better kernels (Appendix W).

REPRODUCIBILITY STATEMENT

Every numerical claim in this paper is produced by the code accompanying the submission, and each claim names the script which produces it. The generating model is (18) with the parameters of Section 7; the grid, cost and marginal conventions are in Appendix D; the E-step, including its sweep budget and the residual we monitor, is in Appendix I; and the three solvers for the M-step are in Appendix K. The fine-grid study of Appendix T, run upon a window fixed in advance of the sample, the E-step feasibility diagnostics of that appendix, including the convex-order check of Remark G.1, the comparison against trajectory inference of Appendix U together with its misspecification sweep, the forecasting and pooling experiments (Tables 1 and 5, `code/compare_regimes.py`), the kernel-fidelity study of Appendix W and the check upon the asymptotic variance in Appendix R each state the script, the seeds and the sweep budget used. Assumptions are collected as Assumptions 1.1 and P.1, hypotheses (T1)–(T2) and (R1)–(R6); every theorem is proved in the appendices, and Remark S.2 delimits the hardness claim. Solving the non-convex programmes to certified global optimality requires a Gurobi licence; everything else runs on the standard scientific Python stack.

AI DISCLOSURE

In this work, we used generative AI tools to aid writing, research ideation and discovery, drafting proofs and generating synthetic datasets. We have reviewed all AI-assisted work. AI-assisted code for the block decomposition has been validated against solutions obtained with Gurobi Optimizer. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Yacine Aït-Sahalia and Jean Jacod. *High-Frequency Financial Econometrics*. Princeton University Press, 2014. doi: 10.23943/princeton/9780691161433.001.0001.

Donald W. K. Andrews. Non-strong mixing autoregressive processes. *Journal of Applied Probability*, 21(4):930–934, 1984.

Heinz H. Bauschke and Jonathan M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996. doi: 10.1137/S0036144593251710.

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.

Mathias Beiglböck, Pierre Henry-Labordère, and Friedrich Penkner. Model-independent bounds for option prices: a mass transport approach. *Finance and Stochastics*, 17(3):477–501, 2013.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. doi: 10.1561/2200000016.

Douglas T. Breeden and Robert H. Litzenberger. Prices of state-contingent claims implicit in option prices. *The Journal of Business*, 51(4):621–651, 1978.

Gerard Brunick and Steven E. Shreve. Mimicking an Itô process by a solution of a stochastic differential equation. *Annals of Applied Probability*, 23(4):1584–1628, 2013. doi: 10.1214/12-AAP881.

Charlotte Bunne, Ya-Ping Hsieh, Marco Cuturi, and Andreas Krause. The Schrödinger bridge between Gaussian measures has a closed form. In *AISTATS*, volume 206 of *PMLR*, pp. 5802–5833, 2023.

Caihua Chen, Bingsheng He, Yinyu Ye, and Xiaoming Yuan. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, 155(1–2):57–79, 2016. doi: 10.1007/s10107-014-0826-5.

Lénaïc Chizat, Stephen Zhang, Matthieu Heitz, and Geoffrey Schiebinger. Trajectory inference via mean-field Langevin in path space. In *NeurIPS*, volume 35, pp. 16731–16742, 2022. doi: 10.52202/068431-1217.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.

Angus Deaton. Panel data from time series of cross-sections. *Journal of Econometrics*, 30(1–2): 109–126, 1985. doi: 10.1016/0304-4076(85)90134-4.

Freddy Delbaen. Representing martingale measures when asset prices are continuous and bounded. *Mathematical Finance*, 2(2):107–130, 1992. doi: 10.1111/j.1467-9965.1992.tb00041.x.

Freddy Delbaen and Walter Schachermayer. A general version of the fundamental theorem of asset pricing. *Mathematische Annalen*, 300(1):463–520, 1994. doi: 10.1007/BF01450498.

Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970. doi: 10.2307/2325486.

Anthony V. Fiacco and Garth P. McCormick. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. Wiley, New York, 1968.

Hans Föllmer. Random fields and diffusion processes. In P.-L. Hennequin (ed.), *École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87*, volume 1362 of *Lecture Notes in Mathematics*, pp. 101–203. Springer, Berlin, 1988. doi: 10.1007/BFb0086180.

Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1): 17–40, 1976. doi: 10.1016/0898-1221(76)90003-1.

Nicolas Gillis and Stephen A. Vavasis. On the complexity of robust PCA and ℓ_1-norm low-rank matrix approximation. *Mathematics of Operations Research*, 43(4):1072–1084, 2018. doi: 10.1287/moor.2017.0895.

Anming Gu, Edward Chien, and Kristjan Greenewald. Partially observed trajectory inference using optimal transport and a dynamics prior. In *ICLR*, 2025.

Gurobi Optimization, LLC. *Gurobi Optimizer Reference Manual*, 2024. URL `https://www.gurobi.com`.

István Gyöngy. Mimicking the one-dimensional marginal distributions of processes having an Itô differential. *Probability Theory and Related Fields*, 71(4):501–516, 1986. doi: 10.1007/BF00699039.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018. Preprint arXiv:1609.05191.

Tatsunori Hashimoto, David Gifford, and Tommi Jaakkola. Learning population-level diffusions with generative RNNs. In *ICML*, volume 48 of *PMLR*, pp. 2417–2426, 2016.

Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. In *NeurIPS*, volume 30, pp. 6702–6712, 2017.

B. L. Ho and R. E. Kálmán. Effective construction of linear state-variable models from input/output functions. *at – Automatisierungstechnik*, 14(1–12):545–548, 1966. doi: 10.1524/auto.1966.14.112.545.

Jean Jacod. Statistics and high frequency data. In Mathieu Kessler, Alexander Lindner, and Michael Sørensen (eds.), *Statistical Methods for Stochastic Differential Equations*, volume 124 of *Monographs on Statistics and Applied Probability*, chapter 3, pp. 191–310. Chapman and Hall/CRC, Boca Raton, FL, 2012. doi: 10.1201/b12126-8.

Jean Jacod and Philip Protter. *Discretization of Processes*, volume 67 of *Stochastic Modelling and Applied Probability*. Springer, Heidelberg, 2012. doi: 10.1007/978-3-642-24127-7.

Jean Jacod and Albert N. Shiryaev. *Limit Theorems for Stochastic Processes*, volume 288 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 2 edition, 2003. doi: 10.1007/978-3-662-05265-5.

Hugo Lavenant, Stephen Zhang, Young-Heon Kim, and Geoffrey Schiebinger. Toward a mathematical theory of trajectory inference. *The Annals of Applied Probability*, 34(1A):428–500, 2024. doi: 10.1214/23-AAP1969.

Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34(4):1533–1574, 2014. doi: 10.3934/dcds.2014.34.1533.

Adrian S. Lewis, D. Russell Luke, and Jérôme Malick. Local linear convergence for alternating and averaged nonconvex projections. *Foundations of Computational Mathematics*, 9(4):485–513, 2009. doi: 10.1007/s10208-008-9036-y.

Robert S. Liptser and Albert N. Shiryaev. *Statistics of Random Processes I: General Theory*, volume 5 of *Stochastic Modelling and Applied Probability*. Springer, 2nd edition, 2001a. doi: 10.1007/978-3-662-13043-8.

Robert S. Liptser and Albert N. Shiryaev. *Statistics of Random Processes II: Applications*, volume 6 of *Stochastic Modelling and Applied Probability*. Springer, 2nd edition, 2001b. doi: 10.1007/978-3-662-10028-8.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall, 2nd edition, 1999.

Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, Cambridge, 2nd edition, 2009. doi: 10.1017/CBO9780511626630.

Robert Moffitt. Identification and estimation of dynamic models with a time series of repeated cross-sections. *Journal of Econometrics*, 59(1–2):99–123, 1993. doi: 10.1016/0304-4076(93)90041-3.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of LTI systems from a single trajectory. In *American Control Conference*, pp. 5655–5661, 2019. doi: 10.23919/ACC.2019.8814438.

Michele Pavon, Giulio Trigila, and Esteban G. Tabak. The data-driven Schrödinger bridge. *Communications on Pure and Applied Mathematics*, 74(7):1545–1573, 2021. doi: 10.1002/cpa.21975.

Philip E. Protter. *Stochastic Integration and Differential Equations*, volume 21 of *Stochastic Modelling and Applied Probability*. Springer, 2nd edition, 2005. doi: 10.1007/978-3-662-10061-5.

Stephen M. Robinson. Strongly regular generalized equations. *Mathematics of Operations Research*, 5(1):43–62, 1980. doi: 10.1287/moor.5.1.43.

L. C. G. Rogers. Arbitrage with fractional Brownian motion. *Mathematical Finance*, 7(1):95–105, 1997. doi: 10.1111/1467-9965.00025.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *ICML*, volume 97 of *PMLR*, pp. 5610–5618, 2019.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.e22, 2019. doi: 10.1016/j.cell.2019.01.006.

Robert H. Shumway and David S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4):253–264, 1982. doi: 10.1111/j.1467-9892.1982.tb00349.x.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning without mixing: towards a sharp analysis of linear system identification. In *COLT*, volume 75 of *PMLR*, pp. 439–473, 2018.

Christophe Stricker. Quasimartingales, martingales locales, semimartingales et filtration naturelle. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 39(1):55–63, 1977.

Anastasios Tsiamis and George J. Pappas. Finite sample analysis of stochastic system identification. In *IEEE Conference on Decision and Control*, pp. 3648–3654, 2019. doi: 10.1109/CDC40024.2019.9029499.

Peter Van Overschee and Bart De Moor. *Subspace Identification for Linear Systems*. Kluwer Academic Publishers, Boston, MA, 1996. doi: 10.1007/978-1-4613-0465-4.

Francisco Vargas, Pierre Thodoroff, Austen Lamacraft, and Neil Lawrence. Solving Schrödinger bridges via maximum likelihood. *Entropy*, 23(9):1134, 2021. doi: 10.3390/e23091134.

Marno Verbeek. Pseudo-panels and repeated cross-sections. In László Mátyás and Patrick Sevestre (eds.), *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, pp. 369–383. Springer, Berlin, 3rd edition, 2008. doi: 10.1007/978-3-540-75892-1_11.

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 centred local-martingale innovations and a predictable finite-variation component.

B PROJECTIONS, AND WELL-POSEDNESS OF THE ESTIMAND

The hierarchy (2). Each arrow is a conditional expectation, so the three compose by the tower property and each is a contraction on L^1, and on L^2 when the relevant variable is square-integrable. The first coarsens continuous time to the mesh: A^Δ is the discrete predictable projection of the increments of $A^{\mathbb{G}}$, and Proposition C.1 is what connects it back to a_t as $\Delta \downarrow 0$. The second replaces $\mathcal{G}_k$, which is not observable in the marginal-law regime, by the feature σ-field $\mathcal{H}_k$ of Appendix E. The third applies the grid map Π_N. Because composition of projections is again a projection, there is a single estimand throughout and no risk of two constraints referring to incompatible objects — a risk that is real if one imposes conditions indexed by σ-fields that are not nested, since the feasible set can then be empty for reasons having nothing to do with the data.

The N-dependence deserves emphasis. The estimator targets $\mathbb{E}[a_{t_k} \mid \mathcal{H}_k^N]$, not $\mathbb{E}[a_{t_k} \mid \mathcal{H}_k]$, so refining the grid changes the estimand and not only the numerical error. Reported quantities are therefore comparable across methods at fixed N and across N only up to that bias.

Why Assumption 1.1 is the right hypothesis. Two things are needed of the estimand, and the assumption supplies both. First, the object (1) must exist and be unique: in discrete time this asks only that (r_k) be integrable, since the predictable part of a discrete Doob decomposition is $\mathcal{G}_{k-1}$-measurable by construction and therefore unique. Second, it must be the projection of the continuous-time drift and not of something else. Writing $r_k - r_{k-1} = (M_{t_k} - M_{t_{k-1}}) + (A_{t_k} - A_{t_{k-1}})$, the first bracket has zero conditional mean given $\mathcal{G}_{k-1} = \mathcal{F}_{t_{k-1}}$ precisely when M is a true

martingale on the sampling grid, which a local martingale need not be; and the second is integrable when A has integrable variation. Both hold for $X \in \mathcal{H}^1$ (Protter, 2005, Ch. IV), whose canonical decomposition has $M \in \mathcal{H}^1$ and $\mathbb{E}\int_0^T |dA_s| < \infty$. Without specialness the decomposition is not unique, the finite-variation part being defined only up to a local martingale of finite variation, and no estimand is singled out. The assumption is stated on X and its $\mathbb{F}$-decomposition, not on the sampling scheme, and therefore survives any further coarsening of the observation times. Stricker's theorem (Stricker, 1977), which passes semimartingality to a genuine continuous-time subfiltration to which X is adapted, is not what is needed here: the sampled σ-fields $\mathcal{G}_k = \mathcal{F}_{t_k}$ shrink nothing, and the discrete decomposition is unique for any integrable adapted sequence.

C Discrete-time observation scheme and Doob decomposition

We assume the càdlàg special semimartingale $X = (X_t)_{t\in[0,T]}$ is observed at n equally spaced times

$$0 = t_0 < t_1 < \cdots < t_{n-1} = T, \qquad t_k = k\Delta, \qquad \Delta := T/(n-1).$$

Set $r_k := X_{t_k}$ for $k = 0, \ldots, n-1$ and let $\mathcal{G}_k := \mathcal{F}_{t_k}$ denote the discretely sampled filtration. The discrete-time process $(r_k)_{k=0}^{n-1}$ is integrable whenever X is of class (D) on $[0,T]$, in particular whenever $X \in \mathcal{H}^1$ (Protter, 2005, Ch. IV), and on the discrete time index it admits a *Doob decomposition*

$$r_k \;=\; r_0 + M_k^\Delta + A_k^\Delta, \qquad A_k^\Delta - A_{k-1}^\Delta \;=\; \mathbb{E}\big[r_k - r_{k-1} \mid \mathcal{G}_{k-1}\big], \qquad k = 1, \ldots, n-1, \quad (19)$$

where M^Δ is a $\mathcal{G}$-martingale with $M_0^\Delta = 0$ and A^Δ is the unique $\mathcal{G}$-predictable process of finite variation with $A_0^\Delta = 0$. Equation (19) restates (1); it is the discrete-time analogue of the continuous-time decomposition $X_t = X_0 + M_t + A_t$ and reduces to it as the mesh refines (Proposition C.1). From now on $\Delta = 1$, any rescaling being absorbed into the drift parameters.

Proposition C.1 (Mesh-refinement consistency)**.** *Suppose A is absolutely continuous in time with $\mathbb{F}$-predictable density a, that is $A_t = \int_0^t a_s\, ds$ with a bounded (or uniformly integrable), and let $t \in [0,T)$ be a time at which a is a.s. right-continuous. Along meshes containing t as a node t_{k-1},*

$$\Delta^{-1}\big(A_k^\Delta - A_{k-1}^\Delta\big) \;=\; \mathbb{E}\Big[\Delta^{-1}\int_t^{t+\Delta} a_s\, ds \;\Big|\; \mathcal{F}_t\Big] \;\longrightarrow\; a_t \quad \text{in } L^1(\mathbb{P}) \text{ as } \Delta \downarrow 0.$$

If a admits a càdlàg version the statement holds at every t; in general it holds for Lebesgue-a.e. t.

Proof in Appendix N.

D Discretisation, marginals and observation schemes

The three regimes in detail. Write $X^{(m)}$, $m = 1, \ldots, M$, for the copies. Under independent copies the $X^{(m)}$ are i.i.d., $\mu_k \to \mathrm{Law}(X_{t_k})$ weakly a.s., and the limit is deterministic: conditioning on $\mathcal{S}_k$ is conditioning on a trivial σ-field, and the estimator recovers $\mathbb{E}[a_{t_k} \mid X_{t_k}]$, the Gyöngy projection, with no room for lagged structure. Under a common factor Z the copies are only *conditionally* i.i.d. given $\sigma(Z)$; $\mu_k \to \mathrm{Law}(X_{t_k} \mid Z)$, a random measure, and $\mathcal{S}_k \subseteq \sigma(Z_j : j \le k)$. Under rolling windows there is one trajectory and μ_k is a window average along it, so $\mathcal{S}_k = \sigma(r_j : j \le k)$ and the flow is a functional of the path.

The asymptotics separate accordingly. In the common-factor scheme the data handed to the M-step are the K pairs $(\bar{x}_k, \gamma_k)$, each measured with error $O_p(M^{-1/2})$. Increasing M shrinks that error; it does not add residuals. Since (16) has $2d+2$ unknowns and K residuals, consistency for (a, c) is a statement about $n \to \infty$ along a single realisation of the factor, and therefore requires an ergodicity hypothesis on Z — more precisely, upon the pair formed by Z and the population conditional mean of X, which is what the mean features are functions of. Proposition R.9 verifies it for the generating model (18) of the experiments, initialisation included; what remains a hypothesis there is persistent excitation, (R5), which is checked numerically and not proved.

Symbols with two meanings. M denotes the local-martingale part of X in Section 1 and Appendices A–B and the number of copies everywhere else; A denotes the predictable part of X in those places and the state matrix of (5) elsewhere; $\mathsf{H}(P)$ is the entropy of a coupling, H the output matrix of (5) and $\mathcal{H}_k$ the feature σ-field; the sign matrix of Appendix S is written M.

Let $(r_k)_{k=0}^{n-1} = (X_{t_k})_{k=0}^{n-1}$ be the discrete samples of Section C. Denote the rolling empirical laws of the sampled process over a window of length w by

$$\mu_k \;=\; \frac{1}{w}\sum_{j=k-w+1}^{k} \delta_{r_j}, \qquad \nu_k \;=\; \frac{1}{w}\sum_{j=k-w+2}^{k+1} \delta_{r_j},$$

where, for a probability measure η on $\mathbb{R}$, $\Pi_N\eta := \sum_{i=1}^{N} \eta(B_i)\,\delta_{x_i}$ denotes its projection onto the common atomic grid $\{x_1, \dots, x_N\} \subset \mathbb{R}$ by nearest-bin assignment and B_i denotes the cell of points closer to x_i than to any other atom. We identify $\Pi_N\eta$ with the probability vector $(\eta(B_i))_{i\le N} \in \Delta_{N-1}$ and write μ_k, ν_k for the projected laws throughout.

Empirical adjacent marginals. For each valid index $k \ge p-1$, the projected rolling laws $\mu_k = \Pi_N\big(w^{-1}\sum_{j=k-w+1}^{k}\delta_{r_j}\big)$ and $\nu_k = \Pi_N\big(w^{-1}\sum_{j=k-w+2}^{k+1}\delta_{r_j}\big)$ are empirical surrogates for the laws of X_{t_k} and $X_{t_{k+1}}$, respectively, under the assumption of approximate local stationarity over the window. We write $x_k^- := \min\{x_j : \nu_k[j] > 0\}$ and $x_k^+ := \max\{x_j : \nu_k[j] > 0\}$ for the extreme atoms charged by ν_k: under any coupling with second marginal ν_k, the conditional mean increment from atom x_i lies in $[x_k^- - x_i,\, x_k^+ - x_i]$, which reduces to $[x_1 - x_i,\, x_N - x_i]$ when ν_k charges every atom.

Observation scheme. The estimators take as input the empirical marginals $(\mu_k, \nu_k)_{k\in\mathcal{P}}$ and, through the lagged features in b_k, summary statistics of the past; two schemes produce such marginals. In the *population* scheme μ_k is the empirical law of X_{t_k} across M copies of the process whose pairing across times is unknown or ignored. This is the setting of trajectory inference and the one in which the estimators are consistent (Section 6). In the *rolling-window* scheme the marginals are formed from consecutive blocks of a single path under an assumption of local stationarity. We retain it for the simulation study, where a single path is the natural object, but Section 6 shows that it introduces a window-aggregation bias that does not vanish with the sample size, so it should be read as a diagnostic device rather than as the intended use of the method.

State grid. A shared grid of N equally spaced atoms is chosen on $[\min_k r_k - 10^{-8}, \max_k r_k + 10^{-8}]$. The transport cost is the normalised quadratic cost $C_{ij} = (x_i - x_j)^2 / \max_{kl}(x_k - x_l)^2$, $C \in \mathbb{R}^{N\times N}$; with the time unit $\Delta = 1$ this is the natural empirical quadratic-variation cost for the increment $r_{k+1} - r_k$.

Index set of consecutive pairs. The set of valid source indices is

$$\mathcal{P} \;=\; \{p-1, p, \dots, n-2\}, \qquad n_{\text{pairs}} \;=\; |\mathcal{P}|.$$

Each $k \in \mathcal{P}$ corresponds to one consecutive pair (t_k, t_{k+1}) of observation times, and the coupling P_k is the discrete analogue of the law of $(X_{t_k}, X_{t_{k+1}})$.

Increment matrix. The signed-increment matrix $\mathrm{inc} \in \mathbb{R}^{N\times N}$ is defined by $\mathrm{inc}_{ij} = x_j - x_i$. The empirical predictable-compensator increment vector at pair k is the row-sum

$$d_k \;=\; (\,P_k \odot \mathrm{inc}\,)\,\mathbf{1} \;\in\; \mathbb{R}^N, \tag{20}$$

i.e. $d_k[i]$ is the μ_k-weighted conditional expectation of the increment given $r_k = x_i$ implied by the coupling P_k.

Centred increment and discrete martingale condition. For a target drift $b \in \mathbb{R}^N$ define $h_b \in \mathbb{R}^{N\times N}$ by $h_b[i,j] = (x_j - x_i) - b[i]$. The predictable-compensator constraint (8) is then the row-wise discrete martingale condition

$$\sum_j P_k[i,j]\, h_{b_k}[i,j] \;=\; 0 \qquad \text{for every } i \text{ with } \mu_k[i] > 0, \tag{21}$$

i.e. the centred increment $r_{k+1} - r_k - b_k(r_k;\theta)$ has zero conditional expectation under P_k on every observed atom – exactly the empirical realisation of the discrete Doob decomposition (19) on the grid.

E THE FEATURE σ-FIELD, IN DETAIL

This appendix expands (3). The point is narrow but it decides whether the parameter $\theta_\star$ of (4) exists, so it is worth setting out slowly.

What the estimator conditions on. The constraint (8) is imposed row by row: for each active atom x_i it equates the conditional mean increment out of x_i under P_k with $b_k(x_i;\theta)$. The conditioning is therefore on the event $\{\Pi_N X_{t_k} = x_i\}$ together with whatever the index k carries. Nothing else is available: the coupling has marginals (μ_k, ν_k) and no memory of which sample moved where at earlier times. So the σ-field the estimator actually conditions on is $\mathcal{H}_k^N = \sigma(\Pi_N X_{t_k}) \vee \mathcal{S}_k$, and the model may only use features measurable with respect to it.

Why the single-path regressors cannot be borrowed. In the single-path scheme the samples at time k are $r_{k-w+1}, \dots, r_k$, successive values of one trajectory, and the lag r_{k-j} is a number known at time k: it is $\mathcal{S}_k$-measurable, and (4) is the classical autoregression. In the population scheme the samples at time k are $X_{t_k}^{(1)}, \dots, X_{t_k}^{(M)}$, one per copy, and the pairing across times is unavailable by hypothesis. Then:

- the individual past $X_{t_{k-j}}^{(m)}$ is not $\mathcal{H}_k$-measurable — indeed the estimator cannot even name the index m, since μ_k is an unordered empirical measure — so a coefficient on it is undefined;
- writing "r_{k-j}" is then not shorthand for anything: there is no distinguished trajectory whose past could be substituted.

What survives is the part of the history that is *common* to the population: the exogenous input u_j if one is observed, the latent factor if the model posits one, or a functional of the marginal flow such as $\bar{x}_j = \mathbb{E}_{\mu_j}[X]$, which is $\mathcal{S}_k$-measurable because it is a statistic of μ_j and the whole flow is shared. Substituting $s_j = \bar{x}_j$ turns the lag model into a dependence on the recent *history of the population*, which is a different — and, in this observation scheme, the only available — statement.

Consequences. Three follow. First, the design matrix X of (13) must be assembled from the individual coordinate x_i and from $\mathcal{S}_k$-measurable columns; the reference implementation does exactly this, using $\bar{x}_j$ or u_j according to the scheme. Second, Remark 2.1 is exact rather than approximate: the Convex Setting is the Non-Convex Setting with A a known nilpotent shift driven by s_k, so the honest finite-memory baseline against a latent factor driven by u is a finite-impulse-response filter *of* u, which is what Section 7 reports. Third, when $\mathcal{S}_k$ is degenerate — independent copies, no common factor — the lag coefficients $\beta_2, \dots, \beta_p$ multiply deterministic functions of time and the model has no dynamic content; that regime is discussed in Section 3.

Remark E.1 (Markovian projection)**.** *For $p = 1$ and $\mathcal{H}_k = \sigma(X_{t_k})$, (4) is the projection of Gyöngy (1986), which yields a Markovian Itô process with the same one-dimensional marginal laws as X. For $p > 1$ it is the mimicking theorem of Brunick & Shreve (2013) for drifts depending on path functionals, here the lag window. Both are statements that a coarser model reproduces the marginal flow; read in the direction we need them, they say that the marginal flow does* not *determine more than the projection, which is the identification limit of Section 8.*

F GAUGE, CANONICAL FORM AND THE RELAXATION OF THE SPECTRAL RADIUS

Identifiability and gauge. For any invertible $T \in \mathbb{R}^{d\times d}$ the parameters $(TAT^{-1}, TB, HT^{-1}, Tz_0)$ generate exactly the same drift sequence, so Θ is identified only up to similarity. The identified object is the transfer function $G(\zeta) = H(\zeta I - A)^{-1}B$ together with the free response $HA^k z_0$, equivalently the Markov parameters HA^kB. Following Hardt et al. (2018, Sec. 1.7) we remove the gauge by fixing the controllable canonical form: $A = \mathrm{comp}(a)$ is

the companion matrix of $p_a(\zeta) = \zeta^d + a_1\zeta^{d-1} + \cdots + a_d$ and $B = e_d$, so that Θ is parametrised by $(a, c, \alpha, \beta) \in \mathbb{R}^{2d+2}$ in (5). All statements about convergence *of the parameters* below are understood in these coordinates; statements about the drift itself are gauge free.

Stability. A predictable drift that neither explodes nor is driven by an explosive latent factor requires $\rho(A) < 1$. The set $\{a : \rho(\mathrm{comp}(a)) < \varrho\}$ is not convex. We use instead the convex relaxation $\mathcal{B}_\varrho$ of Hardt et al. (2018, Sec. 4): with

$$\mathcal{W} \;=\; \big\{\zeta \in \mathbb{C} :\; \mathrm{Re}\,\zeta \;\geq\; (1+\tau_0)\,|\mathrm{Im}\,\zeta|\big\} \;\cap\; \big\{\zeta :\; \tau_1 \leq \mathrm{Re}\,\zeta \leq \tau_2\big\}, \tag{22}$$

a truncated wedge in the complex plane, one sets $\mathcal{B}_\varrho = \{a \in \mathbb{R}^d :\; p_a(\zeta)/\zeta^d \in \mathcal{W} \;\; \forall |\zeta| = \varrho\}$ and calls the system *ϱ-acquiescent* when $a \in \mathcal{B}_\varrho$. Since $p_a(\zeta)/\zeta^d = 1 + \sum_{j \leq d} a_j \zeta^{-j}$ is affine in a, $\mathcal{B}_\varrho$ is an intersection of half-planes indexed by the circle, hence convex, and $a \in \mathcal{B}_\varrho$ implies $\rho(\mathrm{comp}(a)) < \varrho$ (Hardt et al., 2018, Lem. 4.2). Discretising the circle at J points turns $\mathcal{B}_\varrho$ into $4J$ linear inequalities in a, which we add verbatim to the mathematical programmes of Section 4 and use as the projection set of the gradient method of Section 5.

G Feasibility of the drift constraint

The equality-constrained version of (8) can be infeasible on a finite grid even when the continuous-time problem is feasible, which is why the ℓ^1 slack is there. Two mechanisms operate. The first is a boundary effect of the atomisation: under any coupling with marginal μ_k the leftmost atom x_1 can only send mass to the right and the rightmost atom x_N only to the left, so a monotone affine target $b(x) = \alpha + \beta x$ with $\beta > 0$ makes the constraint at $i = N$ infeasible whatever θ is. The second is the support of ν_k: an active atom can only send mass to $\mathrm{supp}\,\nu_k$, so the constraint at i is infeasible whenever $b(x_i) \notin [x_k^- - x_i,\, x_k^+ - x_i]$. The global-balance shift of (12) addresses the first but not the second; Remark G.1 characterises exact feasibility.

Remark G.1 (Exact feasibility of the drift constraint)**.** *Fix k and an active atom x_i. Under any coupling with second marginal ν_k the conditional mean increment out of x_i lies in $[x_k^- - x_i,\; x_k^+ - x_i]$, so the equality (8) is feasible at i only if $b_k(x_i;\theta)$ belongs to that interval; and feasibility at every active atom jointly requires in addition the tower identity $\sum_i \mu_k[i] b_k(x_i;\theta) = \mathbb{E}_{\nu_k}[X] - \mathbb{E}_{\mu_k}[X]$ holds. The two conditions together are necessary and not sufficient. Write $m_i := x_i + b_k(x_i;\theta)$ for the proposed conditional means and $\eta_k := \sum_i \mu_k[i]\,\delta_{m_i}$ for their law. A coupling with marginals (μ_k, ν_k) and the prescribed row means exists if and only if $\eta_k \preceq_{\mathrm{cx}} \nu_k$ in the convex order — $\sum_i \mu_k[i] f(m_i) \leq \sum_j \nu_k[j] f(x_j)$ for every convex f — by Strassen's theorem, a coupling of η_k and ν_k with $\mathbb{E}[Y \mid M] = M$ being exactly a martingale coupling; upon the line, given equal means, it suffices to check $\sum_i \mu_k[i](m_i - t)_+ \leq \sum_j \nu_k[j](x_j - t)_+$ at the finitely many knots t of the two supports. The interval condition is the case $f = (\cdot - x_k^+)_+$ and its mirror image; the tower identity is the case $f = \pm x$; what the two miss is the capacity of the destination marginal. Upon $(-1, 0, 1)$ with $\mu = (0.45, 0.10, 0.45)$, $\nu = (0.05, 0.90, 0.05)$ and $m = (-0.9, 0, 0.9)$ every target is interior and the means balance, yet $\sum_i \mu_i m_i^2 = 0.729 > 0.1 = \sum_j \nu_j x_j^2$, which conditional Jensen forbids. Both necessary conditions are used below: the interval is enforced by the clipping step of the E-step (which is exactly assumption (T2) of Lemma M.5), the identity by the global-balance shift (12); neither, nor both together, guarantees $\mathcal{A} \cap \mathcal{B} \cap \mathcal{D} \neq \emptyset$, and the local convergence statements of Appendix L are conditional upon that intersection being non-empty (Appendix V). The criterion is checked numerically upon every E-step of the reported runs in Appendix T.*

The ℓ^1 slack of (9) restores feasibility uniformly in θ, at the price of a piecewise-linear penalty in the residual $d_k[i] - b_k(x_i;\theta)\mu_k[i]$.

H THE ALGORITHM IN FULL

Algorithm 1 Block-coordinate identification of the predictable drift (both settings)

Input: marginals $(\mu_k, \nu_k)_{k\in\mathcal{P}}$; inputs (u_k) (Non-Convex Setting); cost C; increments inc; order p resp. d; ϵ, λ, tolerances.
Initialise Θ; $P_k \leftarrow$ entropic coupling of (μ_k, ν_k); $\Gamma_k \leftarrow 0$.
for $\ell = 1, \dots, L$ **do**
  for each pair $k \in \mathcal{P}$ **do** ▷ E-step: identical in both settings
    $\tilde{b}_k \leftarrow \mathrm{clip}\big(b_k(\mathrm{grid};\Theta) + \delta_k\big)$ by (12); $(P_k, \Gamma_k) \leftarrow \mathrm{proj}_{\mathcal{D}}\,\mathrm{proj}_{\mathcal{B}}\,\mathrm{proj}_{\mathcal{A}}$ warm-started at Γ_k; $d_k \leftarrow (P_k \odot \mathrm{inc})\mathbf{1}$
  end for
  M-step: Convex Setting — $\theta \leftarrow$ weighted least squares (13) (convex, closed form); Non-Convex Setting — $\Theta \leftarrow$ global solution of the QCQP (15), or projected gradient on $\mathcal{B}_\varrho$.
  Test convergence.
end for
return Θ, $\{P_k\}$.

I IMPLEMENTATION OF THE E-STEP

REFORMULATION AS A DISCRETE MARTINGALE CONDITION.

With centred increment $h_{ij}^{(k)} = (x_j - x_i) - b_k(x_i;\Theta)$, the predictable-compensator constraint becomes the row-wise discrete martingale condition $\sum_j P_k[i,j]h_{ij}^{(k)} = 0$: the empirical statement that $r_{k+1} - r_k - b_k(r_k;\Theta)$ is, conditionally on the discrete past, a martingale increment under P_k. The Kullback–Leibler projection of the entropic kernel onto this affine slice is parameterised by one Lagrange multiplier $\Gamma_k[i]$ per active row,

$$P_k[i,j] \;=\; u[i]\,v[j]\,\exp\Big(-\tfrac{1}{\epsilon}\big(C_{ij} + \Gamma_k[i]\,h_{ij}^{(k)}\big)\Big), \tag{23}$$

where u, v are the Sinkhorn scalings enforcing the marginal constraints. Equation (23) is the closed form of the saddle point of $\mathcal{L}_k$ and coincides with the Radon–Nikodym density of the entropic Schrödinger bridge with prescribed first conditional moment of the increment.

GLOBAL-BALANCE CORRECTION.

The candidate drift need not satisfy the compatibility identity $\sum_i \mu_k[i]b_k(x_i;\Theta) = \mathbb{E}_{\nu_k}[X] - \mathbb{E}_{\mu_k}[X]$ that the tower property forces on any coupling with marginals (μ_k, ν_k); without the shift δ_k of (12) the drift manifold $\mathcal{D}$ may miss the marginal slice $\mathcal{A} \cap \mathcal{B}$ altogether and the alternating projection oscillates. The clipping into $[x_k^- - x_i,\, x_k^+ - x_i]$ is what enforces assumption (T2) of Lemma M.5, and in the reference implementation it is applied with a small relative margin so that Γ stays finite.

SWEEP BUDGET AND THE RESIDUAL WE MONITOR.

Any plan with the prescribed marginals satisfies $\mathbf{1}^\top d_k = \mathbb{E}_{\nu_k}[X] - \mathbb{E}_{\mu_k}[X] = \gamma_k$ identically, whatever drift target it is projected against; the deviation $\mathrm{res}_k := |\mathbf{1}^\top d_k - \gamma_k|$ is therefore a pure measure of how far the Sinkhorn scalings are from enforcing $P_k\mathbf{1} = \mu_k$, $P_k^\top\mathbf{1} = \nu_k$, and it is exactly the quantity the M-step consumes (Proposition 6.1). It is also the diagnostic that matters, because a drift target outside the achievable set makes $\mathcal{A} \cap \mathcal{B} \cap \mathcal{D}$ empty and the alternating projection then cannot enforce the marginals either.

The cyclic projection is run for at most 2000 sweeps with an early exit at a KKT residual of 10^{-10}. Well-conditioned instances exit long before the cap — at $N = 14$, $\epsilon = 0.05$ the residual reaches 10^{-10} in 43 sweeps — but the budget matters in the stiff regime $\epsilon \approx \epsilon^\star$ on a fine grid, where the entropic kernel is as narrow as the transition kernel it models. There $\max_k \mathrm{res}_k/\|\gamma\|_\infty$ falls as 42.5%, 12.0%, 3.3%, 1.4%, 0.4% at 60, 200, 500, 1000, 2000 sweeps ($N = 100$, $\epsilon = \epsilon^\star$), so 2000 is the first budget that brings it below the 1% we require. An ϵ-scaling continuation, warm-starting Γ from a larger ϵ, would reach the same accuracy more cheaply.

LOG-DOMAIN SINKHORN AND NEWTON UPDATE OF Γ.

Inside the projection (`equality_drift_projection` in `code/core.py`), each outer iteration interleaves

- *Sinkhorn-style row/column rescalings* (in log-space, with soft-marginal exponent λ):
$$\log u \leftarrow (1-\eta)\log u + \eta\big(\log\mu - \log\textstyle\sum_j e^{\log K_{ij}+\log v_j}\big), \qquad \eta = \tfrac{\lambda}{1+\lambda},$$
$$\log v \leftarrow (1-\eta)\log v + \eta\big(\log\nu - \log\textstyle\sum_i e^{\log K_{ij}+\log u_i}\big),$$
the retained term $(1-\eta)\log u$ being what makes this an under-relaxation of the exact row projection rather than a soft-marginal update: dropping it would leave a fixed point with $u_i(Kv)_i = \mu_i^{\eta}(Kv)_i^{1-\eta} \neq \mu_i$, an unbalanced-transport iteration with a different feasible set. The reference implementation retains it. where $\log K_{ij} = -C_{ij}/\epsilon - \Gamma[i]\,h_{ij}/\epsilon$.
- *Row-wise Newton update of* Γ that solves $\sum_j h_{ij}\,P[i,j] = 0$ in one scalar unknown $\Delta\Gamma_i$ per row by Newton's method on the smooth, monotone, 1-D residual:
$$\Delta\Gamma_i \;=\; \epsilon\,\frac{\sum_j h_{ij}\,e^{\,L_{ij}}}{\sum_j h_{ij}^2\,e^{\,L_{ij}}}, \qquad L_{ij} = \log u_i + \log K_{ij}^{(0)} - \tfrac{\Gamma_i h_{ij}}{\epsilon} + \log v_j,$$
(`code/core.py`). The denominator is the weighted second moment of h, strictly positive whenever the row support is non-degenerate; Newton therefore converges quadratically in ~ 5 steps.

J THE INNOVATION WEIGHT λ_w

For $\lambda_w \in (0,\infty)$ the reduced objective (15) acquires the term $\lambda_w \sum_k \|z_{k+1} - \mathrm{comp}(a) z_k - e_d u_k\|^2$ and becomes, for fixed (a,c,α,β), a strictly convex quadratic in $z = (z_0,\dots,z_K)$ with block-tridiagonal Hessian. Its minimiser is the Rauch–Tung–Striebel smoother of the linear-Gaussian model with state matrix $\mathrm{comp}(a)$, observation $\gamma_k - \alpha - \beta\bar{x}_k = c^\top z_k$, process covariance $\lambda_w^{-1} I$ and observation variance 1; it costs $\mathcal{O}(Kd^3)$ and is exact. Eliminating z gives the prediction-error objective

$$J(a,c,\alpha,\beta) \;=\; \big(Q_2\beta^2 + Q_1\beta + Q_0\big) \;+\; \min_z \Big\{ \textstyle\sum_k (\alpha + \beta\bar{x}_k + c^\top z_k - \gamma_k)^2 \\ + \lambda_w \textstyle\sum_k \|z_{k+1} - \mathrm{comp}(a) z_k - e_d u_k\|^2 \Big\},$$

which by Danskin's theorem is continuously differentiable in (a,c,α,β) with gradient evaluated at the smoothed state, and whose inner minimiser is unique — the half of Assumption P.1 that must otherwise be assumed. Three regimes are then worth separating in an empirical study: $\lambda_w \to \infty$ recovers the deterministic recursion as a hard constraint; $\lambda_w \to 0$ makes z free and recovers the rank-constrained fit of Theorem 4.1; and intermediate λ_w trades the two, with the bilinearity in (a,z) surviving throughout, so global solution still requires spatial branch and bound. Whether the interior is easier than the $\lambda_w = 0$ endpoint — the hardness proof needs the disturbances free — is open, and is the natural next complexity question.

K THE M-STEP OF THE NON-CONVEX SETTING: REDUCTION, AND SOLVERS

Reduction to per-pair statistics. Write $y_k[i] = d_k[i]/\mu_k[i]$, $\bar{x}_k = \sum_i \mu_k[i] x_i$, $\bar{y}_k = \sum_i \mu_k[i] y_k[i] = \mathbf{1}^\top d_k$ and $s_k = c^\top z_k$. Expanding the square in (14) and using $\sum_i \mu_k[i] = 1$,

$$\sum_i \mu_k[i]\big(y_k[i] - \alpha - \beta x_i - s_k\big)^2 \\ = \big(\alpha + \beta\bar{x}_k + s_k - \bar{y}_k\big)^2 + \underbrace{\textstyle\sum_i \mu_k[i]\big(y_k[i] - \beta x_i\big)^2 - \big(\bar{y}_k - \beta\bar{x}_k\big)^2}_{\text{free of } \alpha, s_k},$$

and summing over k gives (15) with

$$Q_2 = \sum_k \Big(\sum_i \mu_k[i] x_i^2 - \bar{x}_k^2\Big), \qquad Q_1 = -2\sum_k \Big(\sum_i \mu_k[i] y_k[i] x_i - \bar{y}_k\bar{x}_k\Big),$$

$$Q_0 = \sum_k \Big(\sum_i \mu_k[i] y_k[i]^2 - \bar{y}_k^2 \Big).$$

The three constants are the within-marginal variances and covariance of the grid and of the empirical predictable density; they depend on the couplings but not on Θ, and only Q_2, Q_1 affect the minimiser, through β alone. The resulting programme has $2d + 2 + (K+1)d + K$ variables and $K(d+1)$ bilinear equalities, independently of N.

(a) Global branch and bound. The programme handed to Gurobi is

$$\min_{a,c,\alpha,\beta,z,s} \sum_k w_k \big(\alpha + \beta \bar{x}_k + s_k - \bar{y}_k\big)^2 + Q_2 \beta^2 + Q_1 \beta \text{ s.t. } \begin{cases} z_{k+1,j} = z_{k,j+1}, & j < d, \\ z_{k+1,d} = -\sum_j a_j z_{k,d-j} + u_k, \\ s_k = \sum_j c_j z_{k,j}, \qquad z_0 = 0, \\ Ga \le h \ \ (4J \text{ rows}), \end{cases}$$

with box bounds $|a_j|, |c_j|, |\alpha|, |\beta| \le 5$ and $|z_{k,j}| \le 25$ (the bounds matter: spatial branch and bound needs finite ranges to build McCormick envelopes for the products $a_j z_{k,d-j}$ and $c_j z_{k,j}$). Only the two families of bilinear equalities are non-convex, $K(d+1)$ of them; we set `NonConvex=2` and `MIPGap=` 10^{-8}. For $d = 2$, $K = 24$ this is 80 variables and 74 constraints and closes in one to two seconds.

(b) Projected gradient. Gradients of the loss with respect to (a, c, α, β) are obtained by back-propagation through the state recursion: the adjoint satisfies $g_{z_k} = 2\varpi_k \varrho_k c + A^\top g_{z_{k+1}}$, with ϖ_k the weight of pair k (unity throughout, since $\bar{y}_k = \mathbf{1}^\top d_k$ already carries the mass) and ϱ_k the scalar residual — neither is the state innovation w_k of (10) nor the sample r_k of Section 1 — and $\partial J / \partial a$ is the reversed last row of $\sum_k g_{z_{k+1}} z_k^\top$ with a sign change. We use a decaying step $\eta_t = \eta_0/(1 + t/500)$ and, after each step, the projection onto $\mathcal{B}_\varrho$ described next. The iterate with the smallest full loss is returned, so a diverging run (which happens without the projection on non-acquiescent systems) reports its best point rather than an overflow.

Projection onto $\mathcal{B}_\varrho$. Sampling $|\zeta| = \varrho$ at J points, $p_a(\zeta)/\zeta^d = 1 + \sum_j a_j \varrho^{-j} e^{-\mathrm{i}j\omega}$ gives $\mathrm{Re} = 1 + \sum_j a_j \varrho^{-j} \cos(j\omega)$ and $\mathrm{Im} = -\sum_j a_j \varrho^{-j} \sin(j\omega)$, both affine in a, so $\mathcal{W}$ of (6) becomes the four families of half-spaces $\mathrm{Re} \mp (1+\tau_0)\mathrm{Im} \ge 0$, $\mathrm{Re} \ge \tau_1$, $\mathrm{Re} \le \tau_2$. We use $\varrho = 0.98$, $\tau_0 = 0.15$, $\tau_1 = 0.2$, $\tau_2 = 3$ and $J = 32$. The projection is Dykstra's algorithm on those half-spaces, preceded by a feasibility test that makes the common case free; inside the gradient loop we use the vectorised simultaneous (Cimmino) variant, which returns a point of $\mathcal{B}_\varrho$ close to the argument at a fraction of the cost.

(c) Spectral initialisation. The first L Markov parameters of $u \mapsto \bar{y}$ are estimated by the least-squares finite-impulse-response fit $\min_m \|\Phi_u m - \bar{y}\|^2$ with Φ_u the lower-triangular Toeplitz matrix of u; the Hankel matrix $H_{ij} = m_{i+j}$ is formed, truncated to rank d by its singular value decomposition, and split as $H_{\mathcal{H}} = \mathcal{O}\mathcal{R}$ with $\mathcal{O} = U_d \Sigma_d^{1/2}$, $\mathcal{R} = \Sigma_d^{1/2} V_d^\top$; then $A = \mathcal{O}^\dagger H^\uparrow \mathcal{R}^\dagger$, $B = \mathcal{R}_{:,1}$, $H = \mathcal{O}_{1,:}$, where $H^\uparrow$ is the shifted Hankel matrix. The characteristic polynomial of A gives a, which is then projected onto $\mathcal{B}_\varrho$ before being used as a warm start.

L THE INHERITED RATES, IN FULL

THE E-STEP, IN BOTH SETTINGS

Nothing here depends on the drift model: the E-step sees Θ only through the vector $b_k(\mathrm{grid}; \Theta) \in \mathbb{R}^N$. In the Kullback–Leibler geometry it is a cyclic Bregman projection of one coupling onto the two marginal sets and the drift set $\mathcal{D}$, which are affine in the probability coordinates P — not in $L = \log P$, where a row constraint reads $\sum_j e^{L_{ij}} = \mu_i$ — hence smooth embedded submanifolds of the positive cone (Lemma M.1), and pairwise transversal at an interior intersection point provided (T1) the marginals are strictly positive on their support and (T2) $\tilde{b}[i] \in (x_k^- - x_i,\, x_k^+ - x_i)$ in every active row, which the clipping enforces.

Theorem L.1 (Local linear convergence of the E-step)**.** *Let $\bar{P} \in \mathcal{S} := \mathcal{A} \cap \mathcal{B} \cap \mathcal{D}$ be an interior solution at which* (T1)–(T2) *hold. There are $\mathcal{N} \ni \bar{P}$ and $c_E \in (0,1)$, determined by the W-orthogonal projections onto the tangent spaces of the three sets at $\bar{P}$, where $W = \operatorname{diag}(1/\bar{P}_{ij})$ is the Hessian of the entropy, and for every $c' \in (c_E, 1)$ a constant L, with $\operatorname{dist}(P^{(\ell)}, \mathcal{S}) \le L\,(c')^{\ell} \operatorname{dist}(P^{(0)}, \mathcal{S})$ for $P^{(0)} \in \mathcal{N}$.*

Conditions (T1)–(T2) ensure the local regularity used in the proof but do not themselves imply $\mathcal{S} \neq \emptyset$; joint feasibility is characterised in Remark G.1. Appendix M gives the geometry, and Appendix N the argument, which is carried out in the Kullback–Leibler geometry at $\bar{P}$ and not by transfer from the Euclidean theorem of Lewis et al. (2009, Thm. 5.2). The theorem concerns the exact cyclic projection, which is what the implementation performs, without acceleration.

THE OUTER LOOP

Let F be one E-step followed by one M-step, $\bar{\Theta}$ a fixed point. In the Convex Setting the M-step is the closed-form weighted least squares (13), so F is C^1 wherever the E-step is — which (T1)–(T2) guarantee — and Ostrowski's theorem gives local Q-linear convergence with $\kappa \le \kappa_0 + \mathcal{O}(c_E^{L_{\text{inner}}}) + \mathcal{O}(\epsilon)$ towards the fixed point $\bar{\theta}_L$ of the inexact map, which lies within the same distance of $\bar{\theta}$, where κ_0 is the contraction of the exact block map, computed in closed form by Proposition O.3, and the second term is the inexactness of L_{inner} inner sweeps (Theorem O.2, Appendix O). Sweeping harder drives that term down but leaves κ_0, which therefore dominates in practice.

M ALTERNATING PROJECTION: THE GEOMETRY BEHIND THE E-STEP RATE

M.1 GEOMETRIC SETTING

Throughout this section θ is fixed, the pair index k is suppressed, and we work in the relative interior of the coupling simplex $\Delta_{N\times N} = \{P \in \mathbb{R}^{N\times N}_{\ge 0} : \mathbf{1}^\top P \mathbf{1} = 1\}$. Define

$$\mathcal{A} = \big\{P \in \Delta_{N\times N} \,:\, P\mathbf{1} = \mu\big\}, \tag{24}$$

$$\mathcal{B} = \big\{P \in \Delta_{N\times N} \,:\, P^\top \mathbf{1} = \nu\big\}, \tag{25}$$

$$\mathcal{D} = \Big\{P \in \Delta_{N\times N} \,:\, \textstyle\sum_j P[i,j]\, h_{ij} = 0 \;\forall i \text{ with } \mu_i > 0\Big\}, \tag{26}$$

where $h_{ij} = (x_j - x_i) - \tilde{b}[i]$ is the centred increment with the clipped target drift $\tilde{b}$.

The three constraint sets are affine slices *in probability coordinates*: each of $P\mathbf{1} = \mu$, $P^\top\mathbf{1} = \nu$ and $\sum_j P[i,j]h_{ij} = 0$ is linear in P. They are not affine in $\log P$, a row constraint reading $\sum_j e^{L_{ij}} = \mu_i$ there, and they should not be described as e-flat submanifolds in L-space. What the exponential family $P = \exp(L)/Z$ with $L = \log K^{(0)} + \alpha\mathbf{1}^\top + \mathbf{1}\beta^\top - \Gamma h$ does describe is the set of *dual* potentials traced out by the projections; the Sinkhorn and Newton updates move within that family, and each is the exact KL projection onto one of the primal affine slices.

Lemma M.1 (Smoothness of the constraint manifolds)**.** *Each of $\mathcal{A}$, $\mathcal{B}$, $\mathcal{D}$ is a smooth embedded submanifold of $\Delta_{N\times N} \cap \operatorname{int}\mathbb{R}^{N\times N}_{>0}$ of codimension $N-1$, $N-1$ and N_+ respectively, where $N_+ = |\{i : \mu_i > 0\}|$; the codimensions are those of the three sets taken separately, and $\mathcal{D}$ is here taken upon all N_+ active rows, as the implementation does. Their tangent spaces at any interior point are the linear subspaces of $\mathbb{R}^{N\times N}$ defined by the corresponding row-, column-, or row-drift constraints, respectively.*

Proof in Appendix N.

M.2 SINKHORN–DYKSTRA AS ALTERNATING PROJECTION

In KL geometry, one Sinkhorn u-update is the exact KL projection of the current coupling onto $\mathcal{A}$; one v-update is the exact projection onto $\mathcal{B}$; and one Newton Γ-update is the exact KL projection onto $\mathcal{D}$. The row-marginalisation of (23) in the dual coordinates is monotone in Γ_i, and the Newton

step $\Delta\Gamma_i = \epsilon\, f_i / f_i''$ is the Bregman projection step. The combined inner iteration is therefore the cyclic alternating Bregman projection

$$P^{(\ell+1)} \;=\; \mathrm{proj}_{\mathcal{D}}\big(\mathrm{proj}_{\mathcal{B}}\big(\mathrm{proj}_{\mathcal{A}}(P^{(\ell)})\big)\big), \tag{27}$$

where $\mathrm{proj}_{\mathcal{S}}$ denotes KL projection onto $\mathcal{S}$. The fixed points are $\mathcal{A} \cap \mathcal{B} \cap \mathcal{D}$, i.e. entropic transport plans with prescribed marginals and prescribed predictable-compensator increment – the drift-constrained Schrödinger bridges of Section 5.

Remark M.2. *Strictly speaking the soft-marginal exponent $\lambda/(1+\lambda) < 1$ gives an* averaged *(under-relaxed) projection $\frac{\lambda}{1+\lambda}\,\mathrm{proj}_{\mathcal{S}} + \frac{1}{1+\lambda} I$. Since averaged projections inherit the rate of Theorem L.1 (with a softer constant: the derivative of the under-relaxed step is $\frac{\lambda}{1+\lambda}\Pi_T^W + \frac{1}{1+\lambda} I$, which still has norm one only upon T), we treat λ as the exact-projection limit $\lambda \to \infty$ in what follows and absorb the extra factor into the contraction constant.*

M.3 Super-regularity and transversality

Definition M.3 ((Lewis et al., 2009, Def. 4.3))**.** *A closed set $S \subset \mathbb{R}^n$ is* super-regular *at $\bar{x} \in S$ if for every $\eta > 0$ there exists $\delta > 0$ such that*

$$\langle y,\, z - x \rangle \;\le\; \eta\, \|y\|\, \|z - x\|$$

for all $x, z \in S \cap B(\bar{x}, \delta)$ and all $y \in N_S^p(x)$, where N_S^p denotes the proximal normal cone. Equivalently, the angle between proximal normals at x and chords $(z - x)$ is uniformly close to $\pi/2$ locally.

Smooth submanifolds are super-regular at every point: this is immediate from the definition because tangent vectors are normal to normal vectors. By Lemma M.1, each of $\mathcal{A}$, $\mathcal{B}$, $\mathcal{D}$ is super-regular at any $\bar{P} \in \mathcal{A} \cap \mathcal{B} \cap \mathcal{D}$.

Definition M.4 (Transversality)**.** *Two closed sets $A, B \subset \mathbb{R}^n$ are* strongly regular *(or* transversal*) at $\bar{x} \in A \cap B$ if $N_A(\bar{x}) \cap (-N_B(\bar{x})) = \{0\}$. The* constant of regularity *is*

$$c_{AB} \;=\; \max\big\{\langle u,\, -v\rangle :\, u \in N_A(\bar{x}), v \in N_B(\bar{x}), \|u\| = \|v\| = 1\big\} \;<\; 1.$$

For smooth manifolds, $N_S(\bar{x})$ is the orthogonal complement of $T_S(\bar{x})$; transversality reduces to $T_A(\bar{x}) + T_B(\bar{x}) = \mathbb{R}^n$.

Lemma M.5 (Transversality of the OT-drift system)**.** *Let $\bar{P} \in \mathcal{A} \cap \mathcal{B} \cap \mathcal{D}$ be an interior coupling (i.e. $\bar{P}_{ij} > 0$ for all i, j with $\mu_i, \nu_j > 0$). Suppose further that*

- (T1) *all marginals are strictly positive, $\mu_i > 0$ and $\nu_j > 0$;*
- (T2) *the centred-increment matrix h has, for every active row i, both positive and negative entries (i.e. $\tilde{b}[i] \in (x_k^- - x_i,\, x_k^+ - x_i)$, which is guaranteed by the clipping step in the E-step).*

Here, for the purpose of this lemma only, $\mathcal{D}$ is taken on any $N_+ - 1$ of the active rows, the remaining one being implied on $\mathcal{A} \cap \mathcal{B} \cap \mathcal{D}$ by the tower property once the balance identity holds; the codimension in Lemma M.1 is then $N_+ - 1$. Remark N.1 explains why the implementation may, and should, keep all N_+ rows. Then the pairwise tangent sums $T_{\mathcal{A}}(\bar{P}) + T_{\mathcal{B}}(\bar{P})$, $T_{\mathcal{A}}(\bar{P}) + T_{\mathcal{D}}(\bar{P})$ and $T_{\mathcal{B}}(\bar{P}) + T_{\mathcal{D}}(\bar{P})$ all equal the tangent space of $\Delta_{N \times N}$ at $\bar{P}$. Consequently the three manifolds are pairwise transversal at $\bar{P}$ and the joint intersection regularity constant

$$c_{\mathcal{ABD}} \;=\; \max_{\substack{u \in N_{\mathcal{AB}}(\bar{P}) \\ v \in N_{\mathcal{D}}(\bar{P}) \\ \|u\| = \|v\| = 1}} \langle u,\, -v\rangle \;<\; 1,$$

where $N_{\mathcal{AB}} = N_{\mathcal{A}} + N_{\mathcal{B}}$ is the normal cone to $\mathcal{A} \cap \mathcal{B}$ (which is itself a smooth manifold of codimension $2(N-1)$).

Proof in Appendix N.

M.4 Lewis–Luke–Malick theorem and application, for reference

The theorem below is recorded because it remains the tool for the inequality-bounded variant of the projection discussed in Appendix V, whose feasible set is not affine. The proof of Theorem L.1 given in Appendix N does not use it: the three sets being affine in P, the local rate follows from the geometry of W-orthogonal projections directly, and the transversality constant $c_{\mathcal{ABD}}$ of Lemma M.5 is not claimed to be the Kullback–Leibler contraction factor.

Theorem M.6 (Lewis, Luke, Malick (Lewis et al., 2009, Thm. 5.2))**.** *Let A, B be closed subsets of $\mathbb{R}^n$, super-regular at $\bar{x} \in A \cap B$, and strongly regular at $\bar{x}$ with regularity constant $c < 1$. Let $c' \in (c, 1)$. Then there exist $\delta > 0$ and $L > 0$ such that for any starting point $x_0 \in B(\bar{x}, \delta)$ the alternating-projection iteration $x_{\ell+1} \in \mathrm{proj}_A(\mathrm{proj}_B(x_\ell))$ is well-defined and*

$$\mathrm{dist}(x_\ell, A \cap B) \;\le\; L\,(c')^\ell\ \mathrm{dist}(x_0, A \cap B), \qquad \ell = 0, 1, 2, \dots$$

The same conclusion holds for the cyclic projection $x_{\ell+1} \in \mathrm{proj}_{S_1} \circ \cdots \circ \mathrm{proj}_{S_m}(x_\ell)$ over a finite collection of closed sets that are pairwise super-regular and collectively strongly regular at $\bar{x} \in \bigcap_j S_j$.

N Proofs: mesh refinement and the transport step

Proof of Proposition C.1. The first equality is immediate from (1) and the integral form of A, once M is seen to be a true martingale on $[0, T]$: with a bounded, $\sup_{t \le T} |M_t| \le \sup_{t \le T} |X_t| + |X_0| + T\|a\|_\infty$, which is integrable by Assumption 1.1, so the local martingale M is dominated by an integrable variable and is uniformly integrable, whence its conditional increments vanish. Right-continuity of a at t gives $\Delta^{-1} \int_t^{t+\Delta} a_s\, ds \to a_t$ a.s., dominated convergence upgrades this to L^1, and conditional expectation is an L^1-contraction. For the last claim apply Lebesgue's differentiation theorem pathwise.

Proof of Lemma M.1. Each set is the level set of an affine map on the open positive cone: row sums, column sums, and row-drifts. The Jacobians of these maps are the corresponding selection operators, which are surjective onto their images for $\mu, \nu > 0$ on the active rows. The constant-rank theorem gives the manifold structure and tangent space.

Proof of Lemma M.5. The three-set constant lives in normal space. At an interior $\bar{P}$ we may write $N_{\mathcal{A}} = \{a\mathbf{1}^\top : a \in \mathbb{R}^N\}$, $N_{\mathcal{B}} = \{\mathbf{1}c^\top : c \in \mathbb{R}^N\}$ and

$$N_{\mathcal{D}} = \{\mathrm{diag}(\gamma)h :\ \gamma \text{ supported on the } N_+ - 1 \text{ rows defining } \mathcal{D}\},$$

as well as $N_{\mathcal{AB}} = N_{\mathcal{A}} + N_{\mathcal{B}}$. Suppose that $a\mathbf{1}^\top + \mathbf{1}c^\top = \mathrm{diag}(\gamma)h$ with γ so supported, and let i_0 denote the omitted active row, so that $\gamma_{i_0} = 0$. Row i_0 reads $a_{i_0}\mathbf{1}^\top + c^\top = 0$, whence $c = -a_{i_0}\mathbf{1}$ and $a\mathbf{1}^\top + \mathbf{1}c^\top = (a - a_{i_0}\mathbf{1})\mathbf{1}^\top$. Row i then reads $(a_i - a_{i_0})\mathbf{1}^\top = \gamma_i h_{i\cdot}$; by (T2) the row $h_{i\cdot}$ takes both signs and is therefore not constant, so $\gamma_i = 0$ and $a_i = a_{i_0}$. Upon the remaining rows $\gamma_i = 0$ holds by construction and $a_i = a_{i_0}$ follows at once. Both sides vanish, so $N_{\mathcal{AB}}(\bar{P}) \cap \big(-N_{\mathcal{D}}(\bar{P})\big) = \{0\}$, and since the unit spheres are compact, $c_{\mathcal{ABD}} < 1$. The pairwise statements follow from the same computation with one of a, c absent, (T1) supplying $T_{\mathcal{A}} + T_{\mathcal{B}} = \mathbb{R}^{N \times N}$ modulo the mass constraint.

Remark N.1. *The restriction of $\mathcal{D}$ to $N_+ - 1$ rows is essential to the constant $c_{\mathcal{ABD}}$, and to nothing else. Were $\mathcal{D}$ taken upon all N_+ active rows, the choice $\gamma = \mathbf{1}$ would give $\mathrm{diag}(\gamma)h = h = \mathbf{1}x^\top - (x + \tilde{b})\mathbf{1}^\top$, a row-constant plus a column-constant matrix and hence an element of $N_{\mathcal{A}} + N_{\mathcal{B}}$; the intersection above would contain $h \neq 0$, and $c_{\mathcal{ABD}}$ would equal 1, so that the Lewis–Luke–Malick route to Theorem L.1 would be closed. The redundancy is the tower property once again: given both marginals, $\sum_i \sum_j P[i, j](x_j - x_i) = \mathbb{E}_\nu[x] - \mathbb{E}_\mu[x]$ holds identically, so that when the balance identity holds a one row constraint is implied by the others. The proof of Theorem L.1 given below is indifferent to the choice: the intersection $\mathcal{S}$ is the same for both versions of $\mathcal{D}$, and so is its tangent space, since upon $T_{\mathcal{A}} \cap T_{\mathcal{B}}$ the N_+ tangent row-drift constraints sum to $\sum_j x_j \sum_i \dot{P}_{ij} - \sum_i (x_i + \tilde{b}_i) \sum_j \dot{P}_{ij} = 0$; the contraction argument therefore applies verbatim with all N_+ rows, and the implementation, which projects upon every active row, is correct as it stands. The two versions differ only when the clipping has disturbed the balance identity, and then they are different problems: with $N_+ - 1$ rows the iteration converges to a plan which violates the omitted row by*

exactly the balance defect, whereas with all rows the intersection is empty and the iteration stalls at a best-effort compromise. This is a reason to keep every row and to test feasibility (Remark G.1) rather than to omit a constraint and assume it.

Proof of Theorem L.1. A caution before the argument. Lewis et al. (2009) treat *Euclidean* alternating projections, and local equivalence of the Kullback–Leibler divergence to a squared Euclidean distance does not by itself transfer their theorem to KL projections: equivalent divergences need not have the same projections. We therefore argue directly in the W-geometry at the positive feasible point $\bar{P}$, with $W := \operatorname{diag}(1/\bar{P}_{ij})$ the Hessian of the negative entropy $P \mapsto \sum_{ij} P_{ij} \log P_{ij}$, and write $\langle u, v\rangle_W := \sum_{ij} u_{ij} v_{ij} / \bar{P}_{ij}$, $\|\cdot\|_W$ for the corresponding norm and Π_T^W for the W-orthogonal projection onto a subspace T. The argument is in three steps and uses only that $\mathcal{A}, \mathcal{B}, \mathcal{D}$ are affine in P (Lemma M.1); it does not use Lemma M.5, nor Theorem M.6.

Step 1: the derivative of one KL projection. Let $\mathcal{S}$ be one of the three sets, $\mathcal{S} = \{P : \Lambda P = \lambda\}$ with Λ linear, and $T_{\mathcal{S}} = \ker \Lambda$ its tangent space. For Q in the positive cone the KL projection $\Pi_{\mathcal{S}}(Q) := \arg\min_{P \in \mathcal{S}} \mathrm{KL}(P \,\|\, Q)$ exists, is unique by strict convexity, and by the KKT conditions is the point of $\mathcal{S}$ with $\log P - \log Q \in \operatorname{range} \Lambda^\top = T_{\mathcal{S}}^\perp$, the Euclidean orthogonal complement; the implicit function theorem makes $Q \mapsto \Pi_{\mathcal{S}}(Q)$ smooth near $\bar{P}$, where $\Pi_{\mathcal{S}}(\bar{P}) = \bar{P}$. Differentiating $\log P - \log Q \in T_{\mathcal{S}}^\perp$ at $Q = \bar{P}$ in a direction $\dot{Q}$ gives $W(\dot{P} - \dot{Q}) \in T_{\mathcal{S}}^\perp$, that is $\langle \dot{P} - \dot{Q}, v\rangle_W = 0$ for every $v \in T_{\mathcal{S}}$, while $\dot{P} \in T_{\mathcal{S}}$ since P stays in $\mathcal{S}$. Hence

$$D\Pi_{\mathcal{S}}(\bar{P}) = \Pi_{T_{\mathcal{S}}}^W.$$

Step 2: one cycle contracts off the common tangent space. Let $\Phi := \Pi_{\mathcal{D}} \circ \Pi_{\mathcal{B}} \circ \Pi_{\mathcal{A}}$ be the cycle (27); Φ fixes every point of $\mathcal{S} = \mathcal{A} \cap \mathcal{B} \cap \mathcal{D}$ and is smooth near $\bar{P}$, and by the chain rule

$$J := D\Phi(\bar{P}) = \Pi_{T_{\mathcal{D}}}^W \, \Pi_{T_{\mathcal{B}}}^W \, \Pi_{T_{\mathcal{A}}}^W.$$

Set $T := T_{\mathcal{A}} \cap T_{\mathcal{B}} \cap T_{\mathcal{D}}$, the tangent space of $\mathcal{S}$ at $\bar{P}$. Then $J = I$ upon T, and J maps $T^{\perp_W}$ into itself, each factor being self-adjoint for $\langle\cdot,\cdot\rangle_W$ and fixing T. If $v \in T^{\perp_W}$ has $\|Jv\|_W = \|v\|_W$, then each of the three projections preserved the W-norm of its argument, which for an orthogonal projection means that it acted as the identity: $v \in T_{\mathcal{A}}$, then $v \in T_{\mathcal{B}}$, then $v \in T_{\mathcal{D}}$, so $v \in T \cap T^{\perp_W} = \{0\}$. By compactness of the unit sphere of $T^{\perp_W}$,

$$c_E := \max\big\{\|Jv\|_W : \; v \in T^{\perp_W}, \; \|v\|_W = 1\big\} < 1.$$

No transversality is used: a product of orthogonal projections onto finitely many subspaces of a finite-dimensional space contracts strictly off their intersection, whatever the angles between them. Transversality would make c_E computable from those angles; without it the constant is what the display says and no more.

Step 3: from the derivative to the iterates. $\mathcal{S}$ is an affine set, hence a smooth manifold with the same tangent space T at every point. For P near $\bar{P}$ let $s \in \mathcal{S}$ be its W-nearest point and $v := P - s \in T^{\perp_W}$, so $\operatorname{dist}_W(P, \mathcal{S}) = \|v\|_W$. Since $\Phi(s) = s$ and Φ is C^1, $\Phi(P) = s + J_s v + o(\|v\|_W)$ with $J_s := D\Phi(s)$, and $s \mapsto J_s$ is continuous with $J_{\bar{P}} = J$; hence $\operatorname{dist}_W(\Phi(P), \mathcal{S}) \le \|J_s v\|_W + o(\|v\|_W) \le c' \operatorname{dist}_W(P, \mathcal{S})$ for any $c' \in (c_E, 1)$, upon a neighbourhood $\mathcal{N}$ of $\bar{P}$ small enough that $\|J_s|_{T^{\perp_W}}\|_W$ and the remainder together stay below c'. Iterating gives $\operatorname{dist}_W(P^{(\ell)}, \mathcal{S}) \le (c')^\ell \operatorname{dist}_W(P^{(0)}, \mathcal{S})$ as long as the iterates remain in $\mathcal{N}$, which they do once $P^{(0)}$ is close enough to $\bar{P}$, the distance to $\mathcal{S}$ decreasing and the iterates moving by at most a constant times that distance. The Euclidean distance and the square root of the Kullback–Leibler divergence are each equivalent to $\|\cdot\|_W$ upon $\mathcal{N}$, the cone being open, which supplies the constant L of the statement. □

Three remarks upon the scope. First, the argument is for the exact cyclic projection: finite Newton solves, the under-relaxed marginals of Appendix I and any acceleration are perturbations of Φ and are not covered, although the under-relaxed step, whose derivative is $\frac{\lambda}{1+\lambda}\Pi_T^W + \frac{1}{1+\lambda} I$, contracts off T by the same reasoning. Second, since the three sets are affine in P, cyclic I-projections converge globally to the I-projection of the starting kernel onto $\mathcal{S}$ whenever $\mathcal{S}$ contains a strictly positive point (Csiszár, 1975); the local statement above is what supplies a rate. Third, the argument gives nothing when $\mathcal{S} = \emptyset$, which is why the feasibility question of Remark G.1 precedes it.

O THE OUTER LOOP IN THE CONVEX SETTING

Let F denote one E-step followed by one M-step and let $\bar{\theta}$ be a fixed point. In the Convex Setting the M-step is the closed-form weighted least squares (13), so F is C^1 wherever the E-step is, which is guaranteed by (T1)–(T2) through implicit differentiation of (23) (Lemma O.4).

Assumption O.1. *$B + \lambda_\theta I \succ 0$, where $B = \sum_k \bar{\phi}_k \bar{\phi}_k^\top$ is the mean-feature Gram matrix; at $\lambda_\theta > 0$ this holds unconditionally, and at $\lambda_\theta = 0$ it is the Gram condition of Proposition 6.1.*

Theorem O.2 (Convex Setting)**.** *Under Assumption O.1 and* (T1)–(T2) *there are a neighbourhood $\mathcal{N}_\theta \ni \bar{\theta}$ and $\kappa \in (0,1)$ with $\|\theta^{(\ell)} - \bar{\theta}_L\| \le \kappa^\ell \|\theta^{(0)} - \bar{\theta}_L\|$, and*

$$\kappa \;\le\; \kappa_0 \;+\; \mathcal{O}\big(c_E^{L_{\mathrm{inner}}}\big) \;+\; \mathcal{O}(\epsilon), \qquad \kappa_0 := \rho\big(F'_\infty(\bar{\theta})\big),$$

where F_∞ is the block map formed with an exact *E-step, L_{inner} is the number of inner sweeps per outer iteration, and c_E is the rate of Theorem L.1. Here, $\bar{\theta}_L$ is the fixed point of the map actually run, and $\|\bar{\theta}_L - \bar{\theta}\| = \mathcal{O}(c_E^{L_{\mathrm{inner}}}) + \mathcal{O}(\epsilon)$.*

The two roles must be kept apart. The quantity c_E is the rate at which the cyclic projection approaches its own limit *at a fixed parameter*; it governs how inexact the inner solve is, and enters additively. The contraction of the outer iteration is κ_0, the spectral radius of the derivative of the exact block map, and it is a matter of parametric sensitivity, not of algorithmic speed. Sweeping harder drives the second term down but leaves κ_0 where it is — and κ_0 is bounded away from zero, as the next proposition shows.

Proposition O.3 (The nominal contraction)**.** *Write $B = \sum_k \bar{\phi}_k \bar{\phi}_k^\top$, $V = \sum_k \mathrm{Cov}_{\mu_k}(\phi_k)$, $q = \sum_k \bar{\phi}_k \gamma_k$ and $H_\theta = B + V + \lambda_\theta I$. In the Convex Setting with an exact, unclipped E-step the block map is*

$$F_\infty(\theta) = H_\theta^{-1}\big(V\theta + q\big), \qquad F'_\infty(\theta) = H_\theta^{-1} V,$$

its fixed points solve $(B + \lambda_\theta I)\theta = q$, and if $B + \lambda_\theta I \succ 0$ the unique fixed point is the ridge *regression $\bar{\theta} = (B + \lambda_\theta I)^{-1} q$ upon the mean features, approached as a contraction in the H_θ-norm with factor $\kappa_0 = \lambda_{\max}(H_\theta^{-1/2} V H_\theta^{-1/2}) < 1$.*

Proof. Substituting the exact unclipped response $y_{ki} = \phi_{ki}^\top \theta + \delta_k$ with $\delta_k = \gamma_k - \bar{\phi}_k^\top \theta$ into (13) gives $\sum_{ki} \mu_k[i] \phi_{ki} y_{ki} = (B+V)\theta + q - B\theta = V\theta + q$, whence the displayed map. Subtracting the fixed-point equation gives $F_\infty(\theta) - \bar{\theta} = H_\theta^{-1} V(\theta - \bar{\theta})$, an operator represented in the H_θ-norm by the symmetric positive semi-definite matrix $H_\theta^{-1/2} V H_\theta^{-1/2}$, whose eigenvalues lie strictly below one precisely when $H_\theta - V = B + \lambda_\theta I \succ 0$. □

Two points deserve emphasis. The derivative is $H_\theta^{-1} V$ and not $H_\theta^{-1}(V + \lambda_\theta I)$: the term $-\lambda_\theta H_\theta^{-1}\theta$ in passing from (13) to the update is harmless at $\lambda_\theta = 0$ and material otherwise, and the limit is then the ridge and not the ordinary least-squares regression. And the relevant positivity is that of $B + \lambda_\theta I$, not of the full weighted Gram matrix H_θ: at $\lambda_\theta = 0$ one may have $H_\theta \succ 0$ while B is singular, in which case a whole line of fixed points exists and no contraction to a specified one is possible. With $\lambda_\theta > 0$ the map contracts even when B is singular.

This is worth stating because it explains what is seen in Section 7: the outer iterate falls by a large factor at the first step and then settles, the settling being κ_0 and not a failure of the inner solver. For the affine $\phi = (1, x)^\top$ at $\lambda_\theta = 0$ the expression collapses to a quantity one may read off the data,

$$\kappa_0 \;=\; \frac{\bar{\sigma}^2}{\bar{\sigma}^2 + s_{\bar{x}}^2},$$

the mean within-marginal variance over that plus the variance across k of the marginal means: the outer map contracts quickly when the means move a great deal relative to the spread of each snapshot, and slowly when they do not. At the data of Section 7, upon the grid of $N = 14$ atoms in use there, the two quantities are $\bar{\sigma}^2 = 0.206$ and $s_{\bar{x}}^2 = 1.40$, giving $\kappa_0 = 0.128$; the closed form agrees with $\lambda_{\max}(H_\theta^{-1/2} V H_\theta^{-1/2})$ computed from B and V directly. Upon a grid of 100 atoms the within-marginal variance loses its binning inflation of $h^2/12 \approx 0.026$ and the figures are 0.181, 1.40 and 0.114, the values one also obtains from the copies themselves without any grid. The exact

block map thus removes some 87% of the error at every outer step: quick, but not the near-constancy which a single-step fall by a factor of several hundred would require, which needs $\kappa_0 \approx 10^{-3}$. What is seen in Section 7 is of another origin, recorded in Appendix Q.

It also shows the two constants are governed by different quantities — κ_0 by the ratio of within-marginal covariance to total second moment, c_E by the normal-cone angle of the three constraint sets — so that no amount of sweeping makes the outer iteration superlinear.

The M-step of the Convex Setting in full

For fixed $\{P_k\}$, the empirical predictable-compensator vector is $d_k[i] = (P_k \odot \mathrm{inc})\mathbf{1}_i$, cf. (20). The M-step solves the weighted least-squares problem

$$\theta^{(\ell+1)} \;=\; \arg\min_{\theta} \sum_{k\in\mathcal{P},\, i:\mu_k[i]>0} \mu_k[i] \Big(\frac{d_k[i]}{\mu_k[i]} - b_k(x_i;\theta)\Big)^2 \;+\; \lambda_\theta \,\|\theta\|^2, \tag{28}$$

with feature row $\phi_{ki} = [1, x_i, s_{k-1}, \ldots, s_{k-p+1}]$ and weight $\mu_k[i]$. This is the closed-form weighted least-squares update $\theta^{(\ell+1)} = (X^\top W X + \lambda_\theta I)^{-1} X^\top W y$, where X stacks the feature rows, $W = \mathrm{diag}(\mu_k[i])$ and $y[k,i] = d_k[i]/\mu_k[i]$ is the empirical predictable density. The optimisation step here outputs the Markovian-projection coefficients of (4) regressed against the empirical compensator densities and is the empirical analogue of an L^2 regression of the predictable density onto the lag features. An optional exponential-moving-average step $\theta \leftarrow (1-m)\theta_{\text{raw}} + m\theta_{\text{prev}}$ damps oscillations induced by Sinkhorn numerical noise.

Lemma O.4 (Differentiability of the inner-projection map)**.** *Under (T1)–(T2) of Lemma M.5, the map $\theta \mapsto \{P_k(\theta)\}$ that returns the exact Schrödinger-bridge projection at parameter θ is continuously differentiable in a neighbourhood of $\bar\theta$, with Jacobian computable by implicit differentiation of the KKT system*

$$P_k[i,j] = u_i v_j \exp\Big(-\tfrac{1}{\epsilon}\big(C_{ij} + \Gamma_i\, h_{ij}^{(k)}(\theta)\big)\Big), \quad h_{ij}^{(k)}(\theta) = (x_j - x_i) - b_k(x_i;\theta).$$

Proof in Appendix N.

Corollary O.5 (Rate of the composite iteration)**.** *The combined inner-outer scheme contracts with factor*

$$\kappa_{\textit{total}} \;=\; \kappa_0 \;+\; \mathcal{O}\big(c_E^{L_{\textit{inner}}}\big)$$

up to an additive inexactness of the same order, where κ_0 is the nominal contraction of Proposition O.3 and $L_{\textit{inner}}$ is the number of inner sweeps per outer iteration. At $L_{\textit{inner}} = 15$ and $c_E \approx 0.7$ one has $c_E^{15} \approx 4.7 \cdot 10^{-3}$, so the rate is dominated by κ_0; this is consistent with the outer loop settling within 5–15 iterations on the instances of Section 7, which run the E-step to a residual of 10^{-10} (Appendix I) and so make the inexactness negligible. The second term can be made negligible, whereas κ_0 cannot: the composite rate is bounded below by the nominal one however hard the inner loop is driven.

Proof of Lemma O.4. $h_{ij}^{(k)}$ is affine in θ, hence smooth. The implicit KKT system (Sinkhorn marginal equations plus predictable-compensator equations) defines (u, v, Γ) smoothly as long as the Jacobian of the constraints w.r.t. (u, v, Γ) is invertible at $\bar P$, which is the non-degeneracy condition that $T_{\mathcal{A}} + T_{\mathcal{B}} + T_{\mathcal{D}}$ span the relevant tangent space – precisely Lemma M.5.

Proof of Theorem O.2. By Lemma O.4 the inner-projection map is C^1 near $\bar\theta$; denote its Jacobian $J_E \in \mathbb{R}^{N^2 n_{\text{pairs}} \times (p+1)}$. The drift-identification step is the WLS solver (13), whose closed form is $\theta \mapsto H_\theta^{-1}\,(X^\top W y(\theta))$, a C^1 map of the empirical compensator densities $\{d_k\} = \{(P_k \odot \mathrm{inc})\mathbf{1}\}$, and hence of θ through J_E.

Apply the chain rule to the *exact* block map: $F'_\infty(\bar\theta) = J_M J_E$, where J_M is the M-step Jacobian and J_E is the sensitivity of the *limit* of the inner projection with respect to θ. Note that J_E is rectangular, so it has no spectral radius; it is bounded in operator norm by Lemma O.4, and it is not c_E that bounds it. The rate c_E of Theorem L.1 is the speed at which the cyclic projection approaches that limit at a fixed θ, which is a different matter altogether: the algorithmic rate of an inner solver does

not bound the parametric derivative of the solution it converges to. Ostrowski's theorem applied to F_∞ gives a local contraction with any constant exceeding $\kappa_0 = \rho(F'_\infty(\bar\theta))$, which Proposition O.3 computes in closed form.

For the iteration actually run, write F_L for the map obtained from L_{inner} sweeps. By Theorem L.1 the inner iterate is within $\mathcal{O}(c_E^{L_{\text{inner}}})$ of its limit, uniformly upon a neighbourhood, so $\|F_L - F_\infty\|_\infty = \mathcal{O}(c_E^{L_{\text{inner}}})$ there. A standard perturbation of a contraction then gives, for $\Theta^{(\ell+1)} = F_L(\Theta^{(\ell)})$,

$$\|\theta^{(\ell+1)} - \bar\theta\| \;\le\; \big(\kappa_0 + \mathcal{O}(c_E^{L_{\text{inner}}})\big)\|\theta^{(\ell)} - \bar\theta\| \;+\; \mathcal{O}(c_E^{L_{\text{inner}}}),$$

so that F_L is itself a local contraction of modulus $\kappa_0 + \mathcal{O}(c_E^{L_{\text{inner}}})$ with a unique fixed point $\bar\theta_L$ in the neighbourhood, at which the displayed inequality forces $(1-\kappa_0)\|\bar\theta_L - \bar\theta\| = \mathcal{O}(c_E^{L_{\text{inner}}})$. Convergence is geometric to $\bar\theta_L$ and not to $\bar\theta$. The further $\mathcal{O}(\epsilon)$ is the displacement of the computed E-step limit from the exact one where the constraints are not met exactly — by Proposition T.1 an exactly feasible plan does not depend upon ϵ — and it likewise moves $\bar\theta_L$ and not the rate. The inexactness is therefore additive and can be made negligible by sweeping; the contraction κ_0 cannot.

P PROOFS: CONVERGENCE WITH A HIDDEN STATE

Assumption P.1 (Regularity of the M-step in the Non-Convex Setting)**.** *At the M-step data generated by the fixed point $\bar\Theta$, the programme (15) has $\bar\Theta$ as its unique global minimiser, LICQ and strict complementary slackness hold for the active constraints of $\mathcal{B}_\varrho$ there, the second-order sufficient condition holds, and the minimiser is* separated*, in the sense that every other local minimiser has value at least $\bar J + \Delta_{\text{sep}}$ with $\Delta_{\text{sep}} > 0$.*

The two parts of the argument are these. Robinson's strong regularity together with the classical sensitivity theorem (Robinson, 1980; Fiacco & McCormick, 1968) render the localised solution map single-valued and C^1; and the separation Δ_{sep} places the *global* minimiser upon that branch in a neighbourhood of the fixed-point data, so that Ostrowski's theorem applies. In the absence of separation the global minimiser may pass from one branch to another and the iteration map is discontinuous. This is a genuine feature of globally solved non-convex sub-problems, and not an artefact of the argument.

Theorem P.2 (Acquiescent systems and inexact M-steps)**.** *Suppose in addition that $\bar a \in \mathcal{B}_\varrho$ and that the M-step is solved by projected gradient (Hardt et al., 2018, Alg. 1). Then* (i) *the* idealised *risk of Hardt et al. (2018, Lem. 3.3) — the infinite-horizon transfer-function risk under white-noise excitation and without static regressors — is τ-weakly-quasi-convex upon $\mathcal{B}_\varrho$, so upon* that *objective the only stationary point is the global minimiser and projected gradient attains it at the* sub*linear rate $\mathcal{O}(1/T)$. The criterion (15) actually solved is a finite-sample one with an arbitrary input and with (α, β) fitted alongside (a, c); we do* not *claim the property for it, and spurious stationary points are observed (Appendix Y). And* (ii) *the outer iteration keeps its Q-linear rate once the inner accuracy obeys $\eta_\ell \le (\kappa - \kappa_0)\|\Theta^{(\ell)} - \bar\Theta\|$, at a total cost of $\mathcal{O}(\varepsilon^{-2})$ gradients rather than $\mathcal{O}(\log(1/\varepsilon))$ (Appendix P).*

The gap between the idealised risk and the criterion actually minimised is what the branch-and-bound solver exists to close, and part (ii) should be read as conditional upon an outer contraction holding and upon inner accuracy being measured in the parameter metric. The linear rate is thus preserved, though the guarantee is of a different character, and acquiescence is restrictive, failing at $d = 2$ for the poles $\{0.7, 0.5\}$.

Theorem P.3 (Exact global M-step)**.** *Under* (T1)–(T2) *and Assumption P.1, if the M-step is solved to global optimality then there are $\mathcal{N} \ni \bar\Theta$ and $\kappa \in (0,1)$ with $\|\Theta^{(\ell)} - \bar\Theta\| \le \kappa^\ell \|\Theta^{(0)} - \bar\Theta\|$, for the exact iteration, and κ admits the same decomposition $\kappa \le \kappa_0 + \mathcal{O}(c_E^{L_{\text{inner}}}) + \mathcal{O}(\epsilon)$ as in Theorem O.2, the iteration with L_{inner} inner sweeps converging to the fixed point $\bar\Theta_L$ of the inexact map at that distance from $\bar\Theta$, with κ_0 now the spectral radius of the exact block map built upon the* non-linear *least-squares solution map.*

Proof of Theorem P.3. Write $E : \Theta \mapsto \{P_k(\Theta)\}$ for the exact E-step and $\Psi : (\bar x, \bar y) \mapsto \arg\min(15)$ for the M-step, so that the block map is $F = \Psi \circ \Lambda \circ E$ with $\Lambda\{P_k\} =$

$(\bar{x}_k, \mathbf{1}^\top d_k, \sum_i d_k[i]x_i)_k$ the (linear) statistics map — the third component being what Q_1 of (15) depends upon — of Section 5. Lemma O.4 gives $E \in C^1$ near $\bar{\Theta}$ under (T1)–(T2), and Λ is linear, so it suffices to show that Ψ is single-valued and C^1 near the fixed-point data $(\bar{x}, \bar{y})$, and that the solver returns its value.

The programme (15) is a smooth non-linear programme in (a, c, α, β, z) with the linear inequality constraints of $\mathcal{B}_\varrho$ and the smooth equality constraints $z_{k+1} = \mathrm{comp}(a)z_k + e_d u_k$. Assumption P.1 states LICQ and the second-order sufficient condition at $\bar{\Theta}$. By Robinson's theorem (Robinson, 1980) the associated generalised equation is strongly regular, hence the KKT system can be solved for the primal-dual pair as a Lipschitz function of the data, and by the classical sensitivity theorem (Fiacco & McCormick, 1968, Thm. 2.3.2), or by the implicit function theorem applied to the KKT system on the active set, this localised solution map Ψ_{loc} is continuously differentiable on a neighbourhood $\mathcal{U}$ of $(\bar{x}, \bar{y})$.

It remains to know that the *global* minimiser coincides with Ψ_{loc} on a possibly smaller neighbourhood. Let $J(\Theta; \xi)$ denote the objective at data $\xi = (\bar{x}, \bar{y})$. J is jointly continuous and, on the compact set to which the box constraints confine Θ, uniformly so; hence the value function $\xi \mapsto \min_\Theta J$ and the value restricted to a neighbourhood of any other local minimiser are continuous. By the separation hypothesis, at $\xi = \bar{\xi}$ every local minimiser other than $\bar{\Theta}$ has value at least $\bar{J} + \Delta_{\mathrm{sep}}$; by uniform continuity there is $\varepsilon > 0$ such that for $\|\xi - \bar{\xi}\| < \varepsilon$ all such values remain above $\bar{J} + \Delta_{\mathrm{sep}}/2$ while $J(\Psi_{\mathrm{loc}}(\xi); \xi) < \bar{J} + \Delta_{\mathrm{sep}}/4$. Hence the global minimiser at ξ is $\Psi_{\mathrm{loc}}(\xi)$, and a solver certifying global optimality returns it.

Therefore F_∞ is C^1 near $\bar{\Theta}$ with $F'_\infty(\bar{\Theta}) = \Psi'_{\mathrm{loc}}\, \Lambda\, E'$, in which E' is the sensitivity of the *limit* of the inner projection and is bounded in operator norm, not by c_E. Writing $\kappa_0 = \rho(F'_\infty(\bar{\Theta})) < 1$ under the stated smallness, Ostrowski's theorem gives local Q-linear convergence of the exact iteration with any rate exceeding κ_0; the iteration actually run differs from it by $\mathcal{O}(c_E^{L_{\mathrm{inner}}})$, which enters additively exactly as in the proof of Theorem O.2, so that the convergence is to the fixed point of the inexact map, at distance $\mathcal{O}(c_E^{L_{\mathrm{inner}}}) + \mathcal{O}(\epsilon)$ from $\bar{\Theta}$. □

Proof of Theorem P.2. (i) is Hardt et al. (2018, Lem. 3.3) applied to (16): on $\mathcal{B}_\varrho$ the idealised risk of an ϱ-acquiescent system is τ-weakly quasi-convex, i.e. $\langle \nabla J(\Theta),\, \Theta - \bar{\Theta} \rangle \geq \tau\big(J(\Theta) - J(\bar{\Theta})\big)$, and for such functions projected gradient descent with a fixed step on an ℓ-smooth objective satisfies $\min_{t \leq T} J(\Theta_t) - J(\bar{\Theta}) = \mathcal{O}(1/(\tau^2 T))$ (Hardt et al., 2018, App. A); in particular every stationary point in $\mathcal{B}_\varrho$ is a global minimiser, which is the property we use.

(ii) Let $\Theta^{(\ell+1)} = F(\Theta^{(\ell)}) + e_\ell$ with $\|e_\ell\| \leq \eta_\ell$ the inexactness of the inner solve, and let F be a local contraction with factor κ_0 as in Theorem P.3. Then

$$\|\Theta^{(\ell+1)} - \bar{\Theta}\| \leq \|F(\Theta^{(\ell)}) - F(\bar{\Theta})\| + \eta_\ell \leq \kappa_0 \|\Theta^{(\ell)} - \bar{\Theta}\| + \eta_\ell,$$

so $\eta_\ell \leq (\kappa - \kappa_0)\|\Theta^{(\ell)} - \bar{\Theta}\|$ gives $\|\Theta^{(\ell+1)} - \bar{\Theta}\| \leq \kappa \|\Theta^{(\ell)} - \bar{\Theta}\|$ and the claim follows by induction.

(iii) Assumption P.1 implies quadratic growth of J around $\bar{\Theta}$, so an inner accuracy η_ℓ in the parameter metric corresponds to an objective accuracy of order η_ℓ^2; by (i) this requires $T_\ell = \mathcal{O}(\tau^{-2}\eta_\ell^{-2})$ gradient steps. With $\eta_\ell \asymp \kappa^\ell$ this is $T_\ell = \mathcal{O}(\kappa^{-2\ell})$, and $\sum_{\ell \leq L} T_\ell$ is dominated by its last term, which is $\mathcal{O}(\varepsilon^{-2})$ for $\kappa^L = \varepsilon$. Under local strong convexity the inner projected gradient iteration is itself locally Q-linear with some factor $q < 1$, so $T_\ell = \mathcal{O}(\log(1/\eta_\ell)/\log(1/q)) = \mathcal{O}(\ell)$ and $\sum_{\ell \leq L} T_\ell = \mathcal{O}(L^2) = \mathcal{O}(\log^2(1/\varepsilon))$. □

Q Proofs: the fixed point of the block iteration

Consistent, but not efficient. The programme imposes one moment condition per active atom per pair, KN in all (336 at the $N = 14$, $K = 24$ of Appendix K), and one might expect the estimator to use them. It does not, and the mechanism is not that the M-step ignores the within-marginal variation but that the E-step *annihilates* it: projecting onto $\mathcal{D}$ leaves

$$y_k[i] - b_k(x_i; \Theta) = \delta_k \qquad \text{for every active } i, \tag{29}$$

a residual constant in i. Hence $\mathrm{Cov}_{\mu_k}(y_k, X) = \beta\, \mathrm{Var}_{\mu_k}(X)$ exactly, the terms $Q_2\beta^2 + Q_1\beta$ of (15) merely reproduce the current β, and the shape of μ_k enters only through the preconditioner H_θ,

affecting the rate but never the limit. In GMM terms the KN conditions carry the information of K scalars. The estimator is therefore consistent, $\gamma_k = \bar{\phi}_k^\top \theta_\star$ making the surviving condition correctly specified, with the asymptotic variance of a K-observation regression however large M and N may be; and the loss falls upon β, which enters $\bar{\phi}_k$ only through $\bar{x}_k$.

Proof of Proposition 6.1, Convex Setting. The fixed-point equation of the Convex Setting is, in full,

$$\bar{\theta} = \Big(\sum_k \bar{\phi}_k \bar{\phi}_k^\top \Big)^{-1} \sum_k \bar{\phi}_k \gamma_k, \tag{30}$$

whenever the Gram matrix is invertible.

Substituting $d_k[i]/\mu_k[i] = \phi_{ki}^\top \theta + \delta_k$ into (13) gives $\theta' = \theta + H_\theta^{-1} \sum_{k,i} \mu_k[i] \phi_{ki} \delta_k = \theta + H_\theta^{-1} \sum_k \bar{\phi}_k \delta_k$, and by (12), $\delta_k = \gamma_k - \bar{\phi}_k^\top \theta$. Hence $\theta' = \theta$ if and only if $\sum_k \bar{\phi}_k (\gamma_k - \bar{\phi}_k^\top \theta) = 0$.

Corollary Q.1 (Consistency and the windowing bias)**.** *In the population scheme, μ_k, ν_k are the laws of $X_{t_k}, X_{t_{k+1}}$, so $\gamma_k = \mathbb{E}[b_k(X_{t_k}; \theta_\star)] = \bar{\phi}_k^\top \theta_\star$ exactly and (30) returns $\theta_\star$: the estimator is Fisher consistent. In the rolling-window scheme, $\gamma_k = (r_{k+1} - r_{k-w+1})/w$ is a telescoping window difference and $\bar{\phi}_k$ is a window average of features, so (30) is a regression of time-aggregated increments on time-aggregated features. The resulting aggregation bias is $\mathcal{O}(1)$ in w and does not vanish as $n \to \infty$; the rolling-window scheme should be read as a diagnostic device, not as the intended use of the method.*

Relation to pseudo-panel estimation. The fixed point (30) will be familiar to econometricians. When a population is observed through repeated cross-sections and individuals cannot be followed, Deaton (1985) proposed grouping the individuals into cohorts, replacing each individual by the cohort mean at each date, and treating the sequence of cohort means as a panel — a *pseudo-panel*. Moffitt (1993) extended the device to dynamic models, in which the individual lagged dependent variable is unavailable, and showed that what is identified from a time series of cross-sections is the dynamics of the cohort means together with whatever aggregate regressors are common to the cohort; Verbeek (2008) surveys the literature. Our setting is the single-cohort case of theirs, and three of our statements are pseudo-panel statements under another name.

First, the closed form (30) regresses the mean displacement $\gamma_k = \mathbb{E}_{\nu_k}[X] - \mathbb{E}_{\mu_k}[X]$ upon the mean feature $\bar{\phi}_k$: it is Deaton's between-cohort regression with one cohort and K dates. The transport layer and the KN moment conditions add nothing to it, exactly as the within-cohort variation adds nothing to Deaton's estimator once the cohort means are formed. Second, the feature σ-field of Section 2.1 — that an individual lag is not a regressor and only a common statistic s_j is — is Moffitt's identification argument: a model with an individual lagged dependent variable is not identified from repeated cross-sections, and what may be fitted is its aggregate counterpart, which is why the lag of $\bar{x}_j$ or the input u_j appears in (4) and not r_{k-j}. Third, the Monte-Carlo term $\mathcal{O}_p(M^{-1/2})$ of Theorem R.10, with M read as a noise level and not as a sample size, is the cell-size errors-in-variables problem which occupies the second half of Deaton (1985): the cohort means are estimated from finitely many individuals, their sampling error enters the regressors, and consistency is a statement about the number of dates, here K, and not about the number of individuals. Deaton's measurement-error correction, applied to $\bar{\phi}_k$, would be the analogue here of an adjustment for finite M; we have not pursued it, the term being dominated by the estimation term in our experiments.

What the present paper adds to that literature is the identification of the *estimand* as a projection of the predictable compensator of a semimartingale (Section 1), the latent-state parametrisation and its complexity (Sections 2.2–4.3), and the couplings themselves, which the pseudo-panel literature does not construct and which Appendix W shows to be informative as transition kernels even though they are inert for the parameter.

The proximal term, and the implementation. The proximal term of Proposition 6.1 is not decoration: the exact E-step returns $y_k[i] = \gamma_k + \beta_{\text{old}}(x_i - \bar{x}_k)$, so the M-step of (15) minimises J with a proximal penalty upon β alone, and every fixed point is a stationary point of J but not conversely. The reference implementation, however, solves the reduced problem without that penalty: `bcd.fit_bcd_lds` passes only $(\bar{x}_k, \bar{y}_k)$ to the M-step solver, and the terms $Q_2\beta^2 + Q_1\beta + Q_0$ are evaluated afterwards, to report the objective. By the proposition the fixed points of the two M-steps

coincide, so no estimate in this paper depends upon the choice; what differs is the path to them. Without the penalty the M-step depends upon the E-step through $(\bar{x}_k, \gamma_k)$ alone, so that after one exact E-step it is already at its fixed point and the outer iteration has nothing left to do but absorb the inexactness of the E-step. This is the fall by a factor of several hundred at the first outer step in Section 7. It is a property of the reduced M-step, not of the map F_∞ of Proposition O.3, whose contraction $\kappa_0 = 0.128$ governs the weighted least squares of the Convex Setting, which does retain the within-marginal terms; with the penalty in place the slope β would contract at a comparable rate and the first step would be less dramatic.

Proof of Proposition 6.1, Non-Convex Setting. Immediate from $\bar{y}_k = \mathbf{1}^\top d_k = \gamma_k$ and (15); the β-only terms $Q_2\beta^2 + Q_1\beta + Q_0$ are the within-marginal variance of y and do not affect stationarity in (a, c, α).

R RATES IN THE JOINT LIMIT

Proposition 6.1 identifies the estimand; it does not say at what rate the estimator approaches it. We record here the missing statement. The four sources of error are separated deliberately, since they are governed by four different quantities and are balanced by the experimenter in four different ways: the number of copies M, the number of adjacent pairs $K = n_{\text{pairs}}$, the grid spacing h and the mesh Δ.

HYPOTHESES

Throughout this section Π_N denotes the nearest-atom projection of Appendix D onto a uniform grid of spacing h, and $\mathcal{Z} := \sigma(Z_k, u_k : k \geq 0)$ the σ-field generated by the common factor. We write $\mathbb{E}_{\mathcal{Z}}$ for $\mathbb{E}[\cdot \mid \mathcal{Z}]$.

(R1) *Observation scheme.* Conditionally on $\mathcal{Z}$ the copies $X^{(1)}, \dots, X^{(M)}$ are independent and identically distributed, each satisfying $dX_t^{(m)} = a_t^{(m)}\, dt + \sigma\, dW_t^{(m)}$ with $a_t^{(m)} = a(X_t^{(m)}, Z_{\lfloor t/\Delta \rfloor})$ and with independent driving Brownian motions, themselves independent of $\mathcal{Z}$.

(R2) *Moments, binning and truncation.* $\sup_{t \leq T} \mathbb{E}[X_t^4] < \infty$ and $\sup_{t \leq T} \mathbb{E}|a_t| < \infty$. The grid is the uniform partition of $[-R, R]$ of spacing h, with $N = \Theta(R/h)$ atoms and with mass falling outside $[-R, R]$ assigned to the nearer endpoint. We keep the two approximations separate: the *binning* error, of order h, and the *truncation* error $\tau_R := \sup_k \big(\mathbb{E}_{\mathcal{Z}}[(|X_{t_k}| - R)_+] + h\, \mathbb{P}_{\mathcal{Z}}(|X_{t_k}| > R)\big)$, which is not zero for a law of unbounded support — as the Gaussian transitions of (18) have — however far R be taken. For an integer $r \geq 1$ we assume, when $r \geq 2$, that the conditional densities f_k of X_{t_k} given $\mathcal{Z}$ possess $r-1$ derivatives, integrable over the line, with $\sup_k \|\partial^{r-1} f_k\|_{L^1} < \infty$ almost surely; for $r = 1$ nothing is assumed. Two conventions deserve notice. First, $N = \Theta(h^{-1})$ only for a *fixed* window; an expanding window has $N = \Theta(R/h)$, and the cost of driving τ_R down is paid in N. Second, a grid fitted to the observed sample would be random, and its truncation error would require a tail analysis of its own which we do not give. The fine-grid study of Appendix T, which is the experiment offered in illustration of Theorem R.10, accordingly uses a window $[-R, R]$ fixed before the data are drawn, so that its grid is deterministic and the theorem applies to it as stated; the remaining experiments, which make no appeal to that theorem, fit the grid to the observed sample.

(R3) *Regularity of the drift within a mesh interval.* There is $L_a < \infty$ with $\mathbb{E}\big|\mathbb{E}[a_s - a_{t_k} \mid \mathcal{F}_{t_k}]\big| \leq L_a(s - t_k)$ for every k and every $s \in [t_k, t_{k+1}]$. The restriction to a single interval is deliberate: the factor of (18) is piecewise constant and jumps at the mesh points, so a bound global in $t \leq s$ would fail across a boundary, and the one-step argument of Lemma R.4 needs no more than the local form.

(R4) *Two targets, and the gap between them.* With $\gamma_k^\infty := \mathbb{E}_{\mathcal{Z}}[X_{t_{k+1}}] - \mathbb{E}_{\mathcal{Z}}[X_{t_k}]$ and $\bar{\phi}_k^\infty$ the corresponding exact mean features, let

$$\theta_\star(\Delta) := \arg\min_\theta \mathbb{E}\big[(\Delta^{-1}\gamma_k^\infty - \bar{\phi}_k^{\infty\top}\theta)^2\big], \qquad \zeta_k := \Delta^{-1}\gamma_k^\infty - \bar{\phi}_k^{\infty\top}\theta_\star(\Delta),$$

$$\varsigma^2 := \mathbb{E}[\zeta_k^2].$$

This is the *discrete-time* pseudo-true parameter at mesh Δ, and it is what the estimator approaches; the mesh is already inside it, and inside ζ_k. Whenever a continuous-time coefficient θ_0 is the object of interest one must in addition assume

$$\big\|\theta_\star(\Delta) - \theta_0\big\| \;\le\; L_\theta\,\Delta \tag{31}$$

for Δ below some Δ_0. For (18), whose data are generated by the Euler recursion itself, the discrete model holds exactly, $\varsigma = 0$, $\theta_\star(\Delta) = \theta_0$ and (31) is trivial. For data generated by the *exact* transition of the corresponding Ornstein–Uhlenbeck equation the picture is otherwise: the finite-interval conditional mean is not Δ times the instantaneous drift but $(e^{\beta\Delta} - 1)x + q_k F_\beta(\Delta)$ with $F_b(\Delta) = (e^{b\Delta} - 1)/b$, so that ς need not vanish at fixed Δ and the slope of $\theta_\star(\Delta)$ is $\Delta^{-1}(e^{\beta\Delta} - 1) = \beta + \frac{1}{2}\beta^2\Delta + \mathcal{O}(\Delta^2)$; (31) then holds, with L_θ read off from that expansion, but it is a hypothesis about the generator and not a consequence of the estimator. It is for this reason that the assertion "exact specification of (4) makes $\varsigma = 0$" must be read of the *discrete* model at the mesh in use.

(R5) *Persistent excitation.* $\mathbb{E}\|\bar\phi_k^\infty\|^4 < \infty$ and $\lambda_{\min}\big(G\big) \ge \lambda_0 > 0$, where $G := \mathbb{E}[\bar\phi_k^\infty \bar\phi_k^{\infty\top}]$; and the sequence $(\bar\phi_k^\infty)_k$ is stationary and ergodic, so that $\widehat G_K := K^{-1}\sum_k \bar\phi_k^\infty \bar\phi_k^{\infty\top} \to G$. When the mesh is allowed to vary, λ_0 is required to hold *uniformly in* Δ — a demand which Remark R.11 shows is not automatic.

(R6) *The score.* The array $\eta_k := \bar\phi_k^\infty \zeta_k$ is strictly stationary, centred, and strongly mixing with $\sum_j \alpha_\eta(j)^{\delta/(2+\delta)} < \infty$ and $\mathbb{E}\|\eta_0\|^{2+\delta} < \infty$ for some $\delta > 0$. This is stated of the score and not of the factor. Assuming it of (Z_k, u_k) alone would not do: $\bar\phi_k^\infty$ is the conditional mean of a driven diffusion, hence a functional of the whole past of the factor together with the initial condition, and a function of the entire past of a mixing process need not inherit its rate. The initialisation $z_0 = 0$ of the experiments is likewise not a stationary start; Proposition R.9(ii) supplies the transient analysis, the effect of the initialisation decaying geometrically and altering neither the law of large numbers nor the central limit theorem, so that no burn-in is required.

The geometry of (R5). The second requirement of (R5) is the *persistent excitation* of classical system identification, and it is worth saying what it asks geometrically. For a unit vector v one has $v^\top G v = \mathbb{E}\big[(v^\top \bar\phi_k^\infty)^2\big]$, the mean square of the projection of the mean-feature vector upon the direction v; and $\lambda_{\min}(G)$ is the least such quantity. The hypothesis therefore says that the cloud of feature vectors $\{\bar\phi_k^\infty\}$ has width in *every* direction — that it is not contained, even asymptotically, in any slab about a hyperplane through the origin. Were it so contained, with normal v, then $v^\top \bar\phi_k^\infty \equiv 0$ and the parameter combination $v^\top\theta$ would alter no prediction: it would be unidentified, and no amount of data would recover it.

The condition is not a technicality attached to consistency alone. In the decomposition (34) the whole error is premultiplied by G_K^{-1}, so $\lambda_{\min}(G)$ is the constant standing in front of all four terms of (32); halving it doubles the Monte-Carlo, estimation, grid and mesh contributions alike. Three ways of losing it are worth recording. Under *independent copies* the flow of marginals is deterministic and $\mathcal{S}_k$ is degenerate. This does not by itself make G singular, a deterministic curve of means being free to vary with k and a known input free to drive it; what it does is to make the lag coordinates a deterministic sequence, so that any exact recursion that sequence satisfies renders them linearly dependent, and excitation becomes fortuitous rather than structural. This is Section 2.1(iii): it is a random common factor that is lost, not time dependence. If the marginal flow is *stationary*, so that $\bar x_k$ does not vary with k, the intercept and the $\bar x$ coordinate are collinear and β is unidentified however rich the spatial spread of each μ_k may be; identifying a state-dependent drift requires temporal variation in the cross-sectional mean, not merely variation within a marginal. And if the factor is left to *decay without renewal* — the input switched off after a non-zero initial excitation, so that $Z_k = \mathrm{comp}(a)^k z_0 \to 0$ geometrically — its coordinate contributes a vanishing amount and G is singular in the direction of c. It is excitation, and not hiddenness alone, that the Non-Convex Setting requires.

The stability region already in use supplies the factor with *some* geometric rate of mixing, though — and this is the point of the counterexample below — not with the rate ϱ itself.

Lemma R.1 (Acquiescence implies geometric mixing)**.** *Let $a \in \mathcal{B}_\varrho$ with $\varrho < 1$ and let (u_k) be independent and identically distributed with $\mathbb{E}|u_0|^4 < \infty$ and with a component absolutely continuous with respect to Lebesgue measure — without which a stable autoregression need not be strongly mixing at all (Andrews, 1984). Then the stationary solution of $Z_{k+1} = \mathrm{comp}(a)Z_k + e_d u_k$ is geometrically ergodic and strongly mixing: there are $C < \infty$ and $\varrho_0 \in (0,1)$,* depending upon the law of u_0 as well as upon *a, with $\alpha(j) \le C\varrho_0^{\,j}$.*

Proof. By Appendix F, $a \in \mathcal{B}_\varrho$ entails $\rho(\mathrm{comp}(a)) < \varrho < 1$, so Z is a stable vector autoregression driven by independent innovations with an absolutely continuous component; such a chain is geometrically ergodic, and geometric ergodicity implies absolute regularity, hence strong mixing, at a geometric rate. □

Remark R.2 (The rate is not the spectral radius)**.** *One might expect $\alpha(j) \le C\varrho^{\,j}$, with the wedge parameter itself as the mixing rate. That is not so: the contraction of the conditional mean is one thing, the decay of dependence quite another, and the latter is governed by the innovation law as much as by a. Take the scalar recursion $Z_{k+1} = 0.1\, Z_k + u_k$ in stationarity, with*

$$u_k \;\sim\; 0.9\,\delta_0 + 0.1\,N(0,1)$$

independent. The innovations have every moment and an absolutely continuous component, and $a = -0.1$ lies in $\mathcal{B}_{0.5}$. The stationary law of Z_0 is continuous and symmetric — condition upon the first non-zero Gaussian innovation in the infinite-past series. Upon the event that the next j innovations all vanish, of probability $0.9^{\,j}$, one has $Z_j = 0.1^{\,j} Z_0$, so that $\operatorname{sgn} Z_j = \operatorname{sgn} Z_0$; upon the complementary event the intervening noise is symmetric and independent of Z_0, contributing a non-negative amount. Hence $\mathbb{E}[\operatorname{sgn} Z_0 \operatorname{sgn} Z_j] \ge 0.9^{\,j}$, and taking the events $\{Z_0 > 0\}$ and $\{Z_j > 0\}$,

$$\alpha(j) \;\ge\; \tfrac{1}{4}\,0.9^{\,j}.$$

No finite C makes $\alpha(j) \le C(0.5)^{\,j}$. The chain is still geometrically mixing, at a slower rate; what fails is the passage from control of the spectral radius to the same *advertised rate, and* (R6) *is accordingly stated of the score directly.*

THE FOUR ERRORS, SEPARATELY

Lemma R.3 (Binning and truncation)**.** *Under* (R2)*, for every k,*

$$\big|\mathbb{E}_{\mathcal{Z}}[\Pi_N X_{t_k}] - \mathbb{E}_{\mathcal{Z}}[X_{t_k}]\big| \;\le\; C_r\, h^{\,r} \;+\; \mathbb{E}_{\mathcal{Z}}\big[(|X_{t_k}| - R)_+\big] \qquad \textit{almost surely,}$$

with $C_1 = \frac{1}{2}$ and, for $r \ge 2$, $C_r = 2(2\pi)^{-r}\zeta(r)\sup_k \|\partial^{r-1} f_k\|_{L^1}$; the last term is the boundary correction due to the clip and is absent when no mass lies outside $[-R, R]$.

Proof. Write $\Pi_N = \Pi_h^{\mathrm{per}} \circ \kappa_R$, where κ_R is the clip to $[-R, R]$ and Π_h^{per} is nearest-atom rounding upon the whole line at spacing h; the two agree upon $[-R, R]$ and the clip is what the grid endpoints do. Then $|\mathbb{E}_{\mathcal{Z}}[\kappa_R(X) - X]| \le \mathbb{E}_{\mathcal{Z}}[(|X| - R)_+]$, which is the second term, and it remains to bound $g_h(x) := \Pi_h^{\mathrm{per}} x - x$, the h-periodic sawtooth of amplitude $h/2$ and mean zero over a period, in expectation under the law of $\kappa_R(X_{t_k})$. For $r = 1$ the bound is $|g_h| \le h/2$. For $r \ge 2$ we bound instead under the law of X_{t_k} itself, whose density f_k is by (R2) integrable upon the line with $r-1$ integrable derivatives — this is why the periodic extension, and not a function supported upon $[-R, R]$, is the right object to expand: $g_h(x) = \sum_{m \ne 0} c_m e^{2\pi i m x/h}$ with $|c_m| \le h/(2\pi|m|)$, so that $\mathbb{E}_{\mathcal{Z}}[g_h(X_{t_k})] = \sum_{m \ne 0} c_m\, \overline{\hat{f}_k(m/h)}$ and $|\hat{f}_k(\xi)| \le \|\partial^{r-1} f_k\|_{L^1}(2\pi|\xi|)^{-(r-1)}$, whence

$$\big|\mathbb{E}_{\mathcal{Z}}[g_h(X_{t_k})]\big| \;\le\; \frac{h^{\,r}\,\|\partial^{r-1} f_k\|_{L^1}}{(2\pi)^r} \sum_{m \ne 0} |m|^{-r} \;=\; \frac{2\zeta(r)}{(2\pi)^r}\|\partial^{r-1} f_k\|_{L^1}\, h^{\,r};$$

it remains to pass from the law of X_{t_k} to that of $\kappa_R(X_{t_k})$. The two agree on $\{|X| \le R\}$, and on its complement both $g_h(X)$ and $g_h(\kappa_R X)$ lie in $[-h/2, h/2]$, so $|\mathbb{E}_{\mathcal{Z}}[g_h(X) - g_h(\kappa_R X)]| \le h\,\mathbb{P}_{\mathcal{Z}}(|X| > R)$, which is the third term. A Lipschitz argument would not do here, g_h having jumps of size h at the cell boundaries: a point just beyond R can change the rounding error by h while its tail excess is arbitrarily small. Two cautions. The order $h^{\,s+1}$ for a C^s density requires the derivative to be integrable upon the line, not merely smooth locally; and when that L^1 bound is random after conditioning upon the factor, the unconditional statement carries the corresponding probabilistic qualification. □

Lemma R.4 (Mesh error). *Under* (R2) *and* (R3), $\big\|\mathbb{E}[X_{t_{k+1}} - X_{t_k} \mid \mathcal{G}_k] - \Delta\, a_{t_k}\big\|_{L^1} \le \frac{1}{2} L_a \Delta^2$ *for every* k.

Proof. The local-martingale part has vanishing conditional mean, so $\mathbb{E}[X_{t_{k+1}} - X_{t_k} \mid \mathcal{G}_k] = \int_{t_k}^{t_{k+1}} \mathbb{E}[a_s \mid \mathcal{G}_k]\, ds$; subtract Δa_{t_k} and apply (R3) under the integral. □

Remark R.5 (Why the exponent is one and not one-half). *Had one bounded* $\mathbb{E}|a_s - a_{t_k}|$ *rather than* $\mathbb{E}|\mathbb{E}[a_s - a_{t_k} \mid \mathcal{G}_k]|$, *an affine drift evaluated along a diffusion would have yielded only* $\mathcal{O}(\sqrt{s - t_k})$, *and the mesh error would have been* $\mathcal{O}(\sqrt{\Delta})$. *Conditioning removes the diffusive fluctuation, and it is for this reason that* (R3) *is stated in conditional form. For the model (18) the hypothesis holds with* $L_a = |\beta| \sup_t \mathbb{E}|a_t|$, *since* Z *is constant over a mesh interval and* $\mathbb{E}[X_s - X_t \mid \mathcal{F}_t] = \int_t^s \mathbb{E}[a_v \mid \mathcal{F}_t]\, dv$.

Lemma R.6 (Monte-Carlo error). *Under* (R1) *and* (R2), $\big(K^{-1} \sum_k (\bar{x}_k^M - \mathbb{E}_{\mathcal{Z}} X_{t_k})^2\big)^{1/2} = \mathcal{O}_p(M^{-1/2})$, *and the same holds for the mean features.*

Proof. Conditionally on $\mathcal{Z}$ the summands within a given k are independent with common variance $\mathrm{Var}_{\mathcal{Z}}(X_{t_k})$, so the conditional second moment of $\bar{x}_k^M - \mathbb{E}_{\mathcal{Z}} X_{t_k}$ is $\mathrm{Var}_{\mathcal{Z}}(X_{t_k})/M$; average over k and use (R2). □

Lemma R.7 (Estimation error). *Under* (R4) *and* (R6), $\big\|K^{-1} \sum_k \eta_k\big\| = \mathcal{O}_p(\varsigma K^{-1/2})$ *and* $K^{-1/2} \sum_k \eta_k \Rightarrow \mathcal{N}(0, \Omega)$ *with* $\Omega = \sum_{j \in \mathbb{Z}} \mathrm{Cov}(\eta_0, \eta_j)$, *the series converging absolutely.*

Proof. Hypothesis (R6) is precisely what Ibragimov's central limit theorem for strongly mixing sequences requires, and the summability of $\alpha_\eta(j)^{\delta/(2+\delta)}$ together with the $2 + \delta$ moment gives the absolute convergence of the covariance series by Davydov's inequality; η_k is centred by the definition of $\theta_\star(\Delta)$ in (R4). □

Remark R.8 (What (R6) does and does not follow from). *At a fixed* $\Delta > 0$, (R2) *and conditional Jensen do give a fourth moment of* $\Delta^{-1}\gamma_k^\infty$, *hence of* ζ_k *once* $\theta_\star(\Delta)$ *is finite, and Cauchy–Schwarz then gives the* second *moment of the score,* $\mathbb{E}\|\eta_k\|^2 \le (\mathbb{E}\|\bar{\phi}_k^\infty\|^4)^{1/2} (\mathbb{E}|\zeta_k|^4)^{1/2}$. *That is not the* $2 + \delta$ *moment a strong-mixing central limit theorem needs, nor is it uniform as* Δ *varies, and it says nothing at all of the mixing of* η. *The honest course is to assume* (R6) *of the score, or to verify it for an explicitly augmented state under a stated input and noise model; we do the former, and, for the generating model of the experiments, the latter as well: there the relevant score is a quadratic function of the finite-dimensional state* (m_k, Z_k) *formed by the population conditional mean and the factor, and Proposition R.9 verifies* (R6) *directly, together with the ergodicity of the mean features which* (R5) *asks for.*

Proposition R.9 (Ergodicity of the generating model). *Let the data be generated by (18) at a fixed mesh* Δ, *with* $|1 + \beta\Delta| < 1$, $\rho(\mathrm{comp}(a)) < 1$, *and* (u_k) *independent and identically distributed, independent of the Brownian increments, with* $\mathbb{E}|u_0|^{4+\delta} < \infty$ *for some* $\delta > 0$ *and with a law possessing a component absolutely continuous with respect to Lebesgue measure. Write* $m_k := \mathbb{E}_{\mathcal{Z}}[X_{t_k}]$ *for the population conditional mean,* $\tilde{m}_k := m_k + \alpha/\beta$ *for its centring, and* $S_k := (\tilde{m}_k, Z_k) \in \mathbb{R}^{1+d}$. *Then*

$$S_{k+1} = F S_k + G u_k, \qquad F = \begin{pmatrix} 1 + \beta\Delta & \Delta c^\top \\ 0 & \mathrm{comp}(a) \end{pmatrix}, \qquad G = \begin{pmatrix} 0 \\ e_d \end{pmatrix},$$

$$\rho(F) = \max\{|1 + \beta\Delta|, \rho(\mathrm{comp}(a))\} < 1.$$

Suppose the pair (F, G) *controllable; since* $(\mathrm{comp}(a), e_d)$ *is controllable, this holds if and only if* $c^\top \big((1 + \beta\Delta) I - \mathrm{comp}(a)\big)^{-1} e_d \neq 0$, *that is, unless the transfer function from the input to* $c^\top Z$ *has a zero at the pole* $1 + \beta\Delta$ *of the mean recursion.*

(i) S *possesses a unique stationary law* π, *with* $\int |s|^{4+\delta}\, d\pi < \infty$, *and the stationary chain is geometrically ergodic and strongly mixing:* $\alpha_S(j) \le C \varrho_0^j$ *for some* $C < \infty$ *and* $\varrho_0 \in (0, 1)$ *depending upon* a, β, Δ, c *and the law of* u_0.

(ii) *Let S° be the stationary chain driven by the same innovations and S the chain from any S_0 with $\mathbb{E}|S_0|^{4+\delta} < \infty$, in particular from the $z_0 = 0$, $m_0 = \mathbb{E}[X_0]$ of the experiments. Then $S_k - S_k^\circ = F^k(S_0 - S_0^\circ)$, and for every $g : \mathbb{R}^{1+d} \to \mathbb{R}^q$ with $|g(s) - g(s')| \le L(1 + |s| + |s'|)\,|s - s'|$ — which covers every polynomial of degree at most two —*

$$\frac{1}{K}\sum_{k\le K}\big\{g(S_k) - g(S_k^\circ)\big\} = \mathcal{O}_p(K^{-1}), \qquad \frac{1}{\sqrt{K}}\sum_{k\le K}\big\{g(S_k) - g(S_k^\circ)\big\} = o_p(1).$$

(iii) *Consequently the mean features $\bar\phi_k^\infty$, which are affine in S_k — $(1, m_k)$ in the Convex Setting, $(1, m_k, z_k(a))$ with $z_k(a) = Z_k$ at the true a in the Non-Convex Setting, and $(1, m_k, \dots, m_{k-p+1})$ under a lag window — form a stationary ergodic sequence under π, and $\widehat{G}_K \to G$ in probability from the initialisation of the experiments; and any score $\eta_k = g(S_k, \dots, S_{k-p})$ polynomial of degree at most two in finitely many lags and centred under π satisfies* (R6)*, with $\alpha_\eta(j) \le C\varrho_0^{j-p}$ and $\mathbb{E}\|\eta_0\|^{2+\delta/2} < \infty$.*

Proof. The recursion for m_k follows by conditioning (18) upon $\mathcal{Z}$: $m_{k+1} = m_k + \Delta(\alpha + \beta m_k + c^\top Z_k)$, whence the block form of F after centring, and $\rho(F)$ is read off the block triangle. (i) S is a stable linear state-space model driven by independent innovations whose law has an absolutely continuous component; controllability of (F, G) makes it a ψ-irreducible aperiodic T-chain (Meyn & Tweedie, 2009, Ch. 4 and 6–7), and with a norm $\|\cdot\|_*$ in which $\|F\|_* < 1$, which exists since $\rho(F) < 1$, the function $V(s) = 1 + \|s\|_*^{4+\delta}$ satisfies the geometric drift condition $\mathbb{E}[V(S_{k+1}) \mid S_k = s] \le \lambda V(s) + b$ with $\lambda < 1$, by the moment bound upon u_0. Geometric ergodicity follows (Meyn & Tweedie, 2009, Thm. 15.0.1), and geometric ergodicity implies β-mixing, hence strong mixing, at a geometric rate; the moment of π is the stationary form of the drift inequality. (ii) Subtracting the two recursions leaves $F^k(S_0 - S_0^\circ)$, with $\|F^k\|_* \le \|F\|_*^k =: r^k$. For g as stated, $|g(S_k) - g(S_k^\circ)| \le L(1 + |S_k| + |S_k^\circ|)\,Cr^k|S_0 - S_0^\circ|$, and $\sum_k r^k(1 + |S_k| + |S_k^\circ|)|S_0 - S_0^\circ|$ has finite expectation, so the sum $\sum_k\{g(S_k) - g(S_k^\circ)\}$ converges absolutely almost surely and is $\mathcal{O}_p(1)$; dividing by K and by $\sqrt{K}$ gives the two claims. (iii) A measurable function of $(S_k, \dots, S_{k-p})$ is strongly mixing with $\alpha_\eta(j) \le \alpha_S(j - p)$ for $j > p$; the moment follows from $\mathbb{E}\|\eta_0\|^{2+\delta/2} \le C\,\mathbb{E}(1 + |S_0|)^{4+\delta}$ under π; the ergodic theorem and Ibragimov's central limit theorem apply to the stationary chain, and (ii) transfers both to the chain started at the initialisation of the experiments. □

At the parameters of Section 7, $1 + \beta\Delta = 0.6$, $\mathrm{comp}(a)$ has eigenvalues 0.2 and 0.4, the input is Gaussian, and $c^\top(0.6I - \mathrm{comp}(a))^{-1}e_d \neq 0$, so the hypotheses hold; the two indistinguishable parameter points of Remark 2.2 are exactly where the controllability condition is at risk, a pole of the mean recursion being cancelled against a zero of the factor's transfer function.

THE THEOREM

Theorem R.10 (Joint rate)**.** *Assume* (T1)–(T2) *and* (R1)–(R6)*, and let $M, K \to \infty$ and $h, R^{-1} \to 0$, with λ_0 of* (R5) *and the mixing and moment constants of* (R6) *holding uniformly along the sequence. Then the fixed point $\hat\theta$ of the block iteration in the Convex Setting satisfies, at a fixed mesh Δ,*

$$\|\hat\theta - \theta_\star(\Delta)\| \;=\; \underbrace{\mathcal{O}_p\big(\Delta^{-1}M^{-1/2}\big)}_{\textit{Monte Carlo}} \;+\; \underbrace{\mathcal{O}_p\big(\varsigma K^{-1/2}\big)}_{\textit{estimation}} \;+\; \underbrace{\mathcal{O}\big(\Delta^{-1}(h^{\,r} + \tau_R)\big)}_{\textit{grid}}, \tag{32}$$

and if in addition $\Delta \to 0$ along the sequence with (31) in force and with λ_0, the mixing and the moment constants held uniformly in Δ, then

$$\|\hat\theta - \theta_0\| \;=\; \mathcal{O}_p\big(\Delta^{-1}M^{-1/2}\big) \;+\; \mathcal{O}_p\big(\varsigma K^{-1/2}\big) \;+\; \mathcal{O}\big(\Delta^{-1}(h^{\,r} + \tau_R)\big) \;+\; \underbrace{\mathcal{O}(\Delta)}_{\textit{mesh}}, \tag{33}$$

provided $\Delta^{-1}M^{-1/2} \to 0$ and $\Delta^{-1}(h^{\,r} + \tau_R) \to 0$. The fourth term is the gap (31) between the two targets and appears only in (33); it is not *a further discretisation residual added to the first three, the mesh being already inside $\theta_\star(\Delta)$ and ζ_k. The factors Δ^{-1} are those of numerical differentiation: γ_k is itself of order Δ, and it is $\Delta^{-1}\gamma_k$ which estimates a velocity, so an error in γ_k is magnified by Δ^{-1} in the estimate. The estimand of this appendix is accordingly the* velocity *parameter, and not the*

discrete θ of Section 2.1, into which Δ was absorbed; the two differ by a factor Δ. At the fixed mesh $\Delta = 1$ used throughout Sections 1–7 the estimand reverts to the discrete one, the Δ^{-1} factors go with it, (32) is the operative statement, and the mesh term is absent altogether, the discrete recursion (18) approximating no continuous-time drift. In the Non-Convex Setting let $\Theta^\star := \arg\min_{\mathcal{B}_\varrho \times \mathcal{C}} J_\infty$ be the identified set *of the population objective J_∞ of (16), which by Remark 2.2 need not be a singleton. If Assumption P.1 holds* for *J_∞ at every point of $\Theta^\star$ — so that $\Theta^\star$ is finite, its points being isolated minimisers within a compact box — if J_∞ exceeds its minimum by $\Delta_{\mathrm{sep}} > 0$ outside a neighbourhood of $\Theta^\star$, and if the parameter is confined to a compact box $\mathcal{B}_\varrho \times \mathcal{C}$ upon which the sensitivities ∂_Θ are uniformly bounded, then the same bound holds for $\operatorname{dist}(\hat{\Theta}, \Theta^\star)$, in canonical coordinates, with constants depending also upon the smallest reduced-Hessian eigenvalue over $\Theta^\star$ and upon Δ_{sep}. When an identifying convention — within-marginal information, or a prior upon the allocation of poles — makes $\Theta^\star$ the singleton $\{\Theta_\star\}$, this is ordinary consistency for $\Theta_\star$ at the same rate. See Remark R.12 for what that proviso costs.*

Remark R.11 (Uniform conditioning under infill is not automatic)**.** *Statement (33) asks that $\lambda_{\min}(G) \geq \lambda_0$ hold uniformly as $\Delta \downarrow 0$. One might think that the factor of (18) supplies this, its recursion being specified in the index k rather than as the discretisation of a continuous-time process. It does not, for the design is built not from the factor but from the* conditional mean*, which does inherit the mesh. Let Z_k be independent, centred, of unit variance, and let the diffusion have drift $\beta X + cZ_k$ with $\beta < 0$; the conditional mean then obeys*

$$m_{k+1} = e^{\beta\Delta} m_k + cF_\beta(\Delta) Z_k, \qquad F_\beta(\Delta) = \frac{e^{\beta\Delta} - 1}{\beta}, \qquad \operatorname{Var}(m_k) = \frac{c^2 F_\beta(\Delta)^2}{1 - e^{2\beta\Delta}} \;\sim\; \frac{c^2\Delta}{-2\beta}.$$

For the features $(1, m_k)$ the smallest population Gram eigenvalue is therefore of order Δ, not bounded away from zero; and the mean process acquires persistence $e^{\beta\Delta} \to 1$ in observation-index time although the factor is independent across indices. One may of course posit a triangular array with uniform design and score controls, but it must be verified for the scaling at hand. Where the physical dynamics are held fixed, the effective horizon and the slowness of the mean process cannot be waved away, and it is (32), at fixed Δ, upon which we rely.

Proof. Convex Setting. We work from an exact algebraic identity rather than from any convergence theorem for the transport layer. Write $\hat{y}_k := \Delta^{-1}\gamma_k$ for the *measured velocity* and $\hat{\phi}_k$ for the measured mean feature — the empirical quantities actually computed, from M copies upon a grid of spacing h over $[-R, R]$. Proposition 6.1 is stated for the regression of γ_k, whose coefficient is Δ times the velocity coefficient; we work throughout in the velocity parametrisation, and record that this rescaling applies to θ and not to the dynamic coefficients a of the Non-Convex Setting, which do not scale with Δ, so that Θ is not a uniform scalar rescaling of the discrete parameter. In that parametrisation the fixed point is

$$\hat{\theta} = (\widehat{G}_K + \lambda_K I)^{-1} \frac{1}{K} \sum_k \hat{\phi}_k \hat{y}_k, \qquad \widehat{G}_K = \frac{1}{K} \sum_k \hat{\phi}_k \hat{\phi}_k^\top,$$

with $\lambda_K = \lambda_\theta / K \geq 0$ the normalised ridge, $\lambda_K = 0$ recovering (30). For *any* fixed target θ,

$$\hat{\theta} - \theta \;=\; (\widehat{G}_K + \lambda_K I)^{-1} \Big\{ \frac{1}{K} \sum_k \hat{\phi}_k \big(\hat{y}_k - \hat{\phi}_k^\top \theta\big) - \lambda_K \theta \Big\}, \tag{34}$$

an identity, with no hypothesis whatever. Take first $\theta = \theta_\star(\Delta)$ and decompose the residual as

$$\hat{y}_k - \hat{\phi}_k^\top \theta_\star(\Delta) \;=\; \zeta_k \;+\; e_k^{\mathrm{mc}} \;+\; e_k^{\mathrm{gr}},$$

where e_k^{mc} collects the replacement of $\mathbb{E}_{\mathcal{Z}}$ by the average over the M copies and e_k^{gr} the replacement of the exact marginals by their binned and clipped projections. There is no fourth term: the mesh is inside ζ_k by (R4). By the ergodic theorem, which the stationarity and ergodicity of the features in (R5) licenses, and (R5), $\widehat{G}_K \to G$ in probability and $\lambda_{\min}(G) \geq \lambda_0$, so $\|(\widehat{G}_K + \lambda_K I)^{-1}\| \leq 2/\lambda_0$ with probability tending to one; the perturbation of $\hat{\phi}_k$ itself is of order $M^{-1/2} + h^r + \tau_R$ by Lemmas R.3 and R.6, and enters (34) only through terms already present or of higher order. The three sums are then bounded in turn: the first by Lemma R.7; the second by Cauchy–Schwarz with Lemma R.6, since $(K^{-1}\sum_k \|\hat{\phi}_k\|^2)^{1/2} = \mathcal{O}_p(1)$; the third by Lemma R.3, the grid error of γ_k

being the difference of two such terms. Both are errors in γ_k and are multiplied by Δ^{-1} upon passing to $\Delta^{-1}\gamma_k$, which is the source of the Δ^{-1} factors; and the ridge contributes $\lambda_K\|\theta_\star\|$, absent at $\lambda_K = 0$. This is (32). For (33) add the triangle inequality with (31), which supplies the single $\mathcal{O}(\Delta)$ term. Lemma R.4 is what makes (31) plausible for a given generator; it is an L^1 estimate upon the increment, and we note that it does not by itself control the product of the mesh residual with an unbounded feature vector — for that one needs either bounded features, a weighted estimate, or the L^2 bound which (R2) supplies at fixed Δ.

Non-Convex Setting. Write J_K for the objective of (16) formed from the computed $(\bar{x}_k, \gamma_k)$ and J_∞ for its population counterpart formed from $(\bar{x}_k^\infty, \gamma_k^\infty)$. Confine Θ to the compact box $\mathcal{B}_\varrho \times \mathcal{C}$ of the statement, upon which $z_k(a)$ and its derivatives are bounded uniformly, the relevant series being geometric because $\rho(\mathrm{comp}(a)) < \varrho < 1$. Assume the second-order sufficient condition and the separation *of* J_∞ at each point of the identified set $\Theta^\star$ and the separation of $\Theta^\star$, as in the statement: these supply a quadratic growth $J_\infty(\Theta) - \min J_\infty \geq \lambda_1 \operatorname{dist}(\Theta, \Theta^\star)^2$ upon a neighbourhood of $\Theta^\star$, and the separation confines every global minimiser of J_K to that neighbourhood once $\sup_{\mathcal{B}_\varrho \times \mathcal{C}} |J_K - J_\infty| < \Delta_{\mathrm{sep}}/2$. Uniform convergence over the box, and not merely a pointwise bound, is what is needed here, and it follows from the pointwise bounds of the first part together with the equicontinuity which the uniform bound upon the sensitivities supplies upon a compact box. The standard argument for M-estimators, applied at the point $\Theta^j \in \Theta^\star$ nearest to $\hat{\Theta}$, then gives $\operatorname{dist}(\hat{\Theta}, \Theta^\star) \leq \lambda_1^{-1} \max_j \|\nabla J_K(\Theta^j) - \nabla J_\infty(\Theta^j)\| + o_p(\cdot)$, the gradient difference being a sum of the same terms multiplied by the bounded sensitivities $\partial_\Theta(\alpha + \beta\bar{x}_k + c^\top z_k)$; where a constraint of $\mathcal{B}_\varrho$ is active at $\Theta_\star$ the unconstrained inverse-Hessian expansion must be replaced by the corresponding constrained local expansion under LICQ and strict complementarity. □

Remark R.12 (The latent extension is conditional, and upon what)**.** *Assumption P.1 is stated of the M-step loss (15), whereas the argument above needs quadratic growth and separation of the population mean loss J_∞ built upon (16). These are not the same function: Proposition 6.1 exhibits the former as the latter plus a slope-proximal quadratic, so a Hessian bound for one does not transfer to the other. Nor is the transfer merely technical — the stationary-versus-fixed distinction of Remark R.11 and the pole ambiguity of Remark 2.2 both intervene before any statistical argument can be brought to bear, the second of them showing that the identified set $\Theta_\star$ need not be a singleton in canonical coordinates. We therefore state the latent half of Theorem R.10 as conditional upon quadratic growth and separation of J_∞ about $\Theta^\star$, as a statement about* $\operatorname{dist}(\hat{\Theta}, \Theta^\star)$*, and do not claim to have derived it from Assumption P.1 as stated of the M-step loss. The logical structure is then: ergodicity and the effect of the initialisation are proved for the generating model (Proposition R.9); persistent excitation is assumed and checked numerically; quadratic growth and separation of J_∞ are assumed; and uniqueness holds only modulo the equivalence of Remark 2.2. For the Euler generator of the experiments $\varsigma = 0$ (Remark R.14), so the $K^{-1/2}$ term is not the principal stochastic error there: finite M and the grid generate the estimation noise, while a longer K supplies excitation and identification horizon rather than averaging. Two further cautions. Stability of* $\mathrm{comp}(a)$ *controls the response to* bounded *inputs and furnishes moment bounds under stochastic assumptions; it does not make the states and their sensitivities uniformly bounded pathwise for an unbounded Gaussian input sequence, which is why the box $\mathcal{C}$ appears in the statement. And the boxes used in the proof belong in the statement, where we have now put them.*

Corollary R.13 (Balancing, and the K-observation variance)**.** *Suppose the premises of (33) — the four bounds, the mesh-gap hypothesis (31) and uniform conditioning along an expanding-horizon sequence — are in force, and take τ_R negligible beside h^r. Balancing $\Delta^{-1}M^{-1/2}$ against Δ gives $\Delta \asymp M^{-1/4}$; requiring $\Delta^{-1}h^r \lesssim \Delta$ then gives $h^r \asymp M^{-1/2}$, that is $h \asymp M^{-1/(2r)}$, and $K \gtrsim M^{1/2}$ makes the estimation term no larger. The four contributions of (32) are then balanced at $\mathcal{O}_p(M^{-1/4})$. Since $K = n_{pairs} \asymp T/\Delta$, the requirement $K \gtrsim M^{1/2}$ is compatible with $\Delta \asymp M^{-1/4}$ only if the horizon T grows; the balancing is therefore a statement about an expanding horizon, and not about the fixed $[0, T]$ of Section 1. It is a balancing of the stated* upper bounds *under those premises, and no evidence of minimax optimality; nor is it a prescription until the premises are established for the design at hand, which Remark R.11 shows is a real question and not a formality. If instead the grid, mesh and Monte-Carlo errors are each $o(K^{-1/2})$, then*

$$\sqrt{K}\,\big(\hat{\theta} - \theta_\star\big) \implies \mathcal{N}\big(0,\ G^{-1}\Omega G^{-1}\big),$$

which is the sandwich form for a regression on K observations. This is the precise sense of the assertion in Section 6 that the estimator attains the asymptotic variance of a K-observation regression however large M and N may be: neither M nor N appears in the limit.

The sandwich, checked against replications. Corollary R.13 is an asymptotic statement, and it is worth asking whether it describes the dispersion actually observed. In the Convex Setting the estimator is the closed form (30) and needs no transport, so replications are cheap; the model there omits the latent factor, so $\varsigma > 0$ and the limit is non-degenerate, which is the case the corollary describes. We drew 150 independent realisations at $K = 999$, $M = 2000$, computed the empirical covariance of $\sqrt{K}(\hat{\theta} - \theta_\star)$, and compared it against $G^{-1}\Omega G^{-1}$ with Ω estimated by Bartlett-weighted autocovariances at bandwidth L:

	empirical	$L = 10$	$L = 25$	$L = 50$	$L = 100$
intercept	0.184	0.560	0.342	0.250	0.205
slope	0.083	0.141	0.103	0.090	0.086

At $L = 100$ the predicted variances exceed the observed by 11% and 4% respectively, which we regard as agreement, the empirical variances themselves carrying a standard error of some 8% upon 150 replications. The instructive part is the bandwidth. The usual automatic choice $4(K/100)^{2/9}$ gives $L \approx 6$ here and overstates the asymptotic variance three-fold, because the score $\phi_k\zeta_k$ carries the omitted factor together with the mean-reverting state and its autocovariances remain appreciable, and partly negative, well beyond a dozen lags. A practitioner forming confidence intervals from Corollary R.13 should choose the bandwidth generously; the error is in the conservative direction, but by a factor of three rather than a few per cent. The script is `code/sandwich.py`.

Remark R.14 (What the theorem does not cover)**.** *Three restrictions deserve mention. Under exact specification of the* discrete *model at the mesh in use — which is the case for the Euler generator of (18), though not for the exact transition of the corresponding diffusion — one has $\varsigma = 0$, the estimation term vanishes and the fit is exact in the population limit; this is why the residuals observed in Section 7 are so small, and it is a property of the simulation rather than of the method. Since $\varsigma = 0$ there, the finite-seed spread reported in Appendix Y cannot be attributed to the estimation term: with M fixed at 600 the Monte-Carlo, conditioning, clipping and solver contributions are not extinguished by refining the grid, and the experiment does not identify which of them dominates. The hypothesis $\lambda_{\min}(G) \geq \lambda_0$ is a condition of persistent excitation on the common factor, and fails precisely in the regime of Section 2.1(iii), where $\mathcal{S}_k$ is degenerate. Finally, the Monte-Carlo term does not improve with K in (32), because the same M copies are used at every time; when the conditional fluctuations decorrelate over the mesh, the bound sharpens to $\mathcal{O}_p((MK)^{-1/2})$ for that component, but we have not assumed enough to claim it.*

S Proof of the NP-hardness theorem

We reduce from ℓ^1-norm rank-one matrix approximation, which is NP-hard by Gillis & Vavasis (2018, Thm. 2) already for sign matrices: given $\mathsf{M} \in \{\pm 1\}^{K \times n}$ with $n \geq 2$ and a rational V, it is NP-hard to decide whether

$$\min_{v \in \mathbb{R}^n,\, z \in \mathbb{R}^K} \|\mathsf{M} - z v^\top\|_1 \;\leq\; V. \tag{35}$$

The restriction to $\{\pm 1\}$ entries matters twice over: it keeps the grid of the constructed instance of *polynomial* cardinality — a grid of $\mathcal{O}(nB)$ atoms would be only pseudo-polynomial in a binary-encoded entry bound B, since polynomial bit size of the endpoints does not make the enumeration of the intervening atoms polynomial — and it makes every matched displacement of unit magnitude, which is what allows the marginals to be arranged into a single chain.

That last point is the substance of the construction. The pairs (μ_k, ν_k) supplied to (10) are not an arbitrary list of source–destination pairs: they are the *consecutive* marginals of one observed process, so that $\nu_k = \mu_{k+1}$. A reduction which chose ν_k freely from row k of M while holding μ_k fixed would not produce an instance of the problem we have posed. We therefore interleave the encoding pairs with *reset* pairs, whose contribution to the objective is a constant.

The instance. Given such an M, put

$$x_i := 4(i-1)+1 \quad (1 \le i \le n), \qquad \mathcal{X} := \{0, 1, \dots, 4n-2\}, \qquad D := 4n-2, \qquad S := D^2,$$

and define the probability measures

$$\mathsf{A} := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}, \qquad \mathsf{B}_r := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i + \mathsf{M}[r,i]} \quad (1 \le r \le K).$$

All of these atoms lie in $\mathcal{X}$, whose cardinality $4n-1$ is polynomial in the size of the instance. The atoms of A are congruent to 1 modulo 4 and those of every B_r are even, so the two supports are disjoint; and since $\mathsf{M}[r,i] \in \{\pm 1\}$, the atoms of B_r are $4(i-1)$ or $4(i-1)+2$ and hence strictly increasing in i, with consecutive gaps at least 2. The gaps of A are exactly 4.

Take as the observed chain of marginals the sequence of $2K+1$ measures

$$\mathsf{B}_1,\ \mathsf{A},\ \mathsf{B}_1,\ \mathsf{A},\ \mathsf{B}_2,\ \mathsf{A},\ \mathsf{B}_3,\ \dots,\ \mathsf{A},\ \mathsf{B}_K,$$

whose $2K$ consecutive pairs are K *encoding* pairs $\mathsf{A} \to \mathsf{B}_r$, $r = 1, \dots, K$, and K *reset* pairs $\mathsf{B}_r \to \mathsf{A}$. Take φ to be the indicator basis $\varphi_i(x) = \mathbf{1}\{x = x_i\}$ of the n atoms of A, so that $q = n$; take $d = 1$, $H = v \in \mathbb{R}^n$, $a = 0$ — an admissible acquiescent coefficient — and $u \equiv 0$, $z_0 = 0$, $\lambda_{\text{reg}} = \lambda_w = 0$, and the quadratic cost $C_{ij} = (x_i - x_j)^2/S$.

Because $\lambda_w = 0$ the recursion $z_{k+1} = a z_k + w_k$ places no restriction whatever upon $z_1, z_2, \dots$; we write z_r for the state at the rth encoding time. The first pair of the chain is a reset pair, so the forced $z_0 = 0$ constrains none of the encoding factors. On a reset pair the source atoms are those of B_r, where every φ_i vanishes, so the model drift is identically zero there whatever Θ may be; on the encoding pair $\mathsf{A} \to \mathsf{B}_r$ the model drift at x_i is $z_r v_i$.

Step 1: the transport block is pinned, uniformly in Θ. For every pair both marginals are uniform upon n sorted atoms, so the vertices of $\Pi(\mu_k, \nu_k)$ are the scaled permutation matrices $\frac{1}{n}\Pi_\pi$. For a strictly convex cost upon the line the monotone matching $P_k^\star = \frac{1}{n}\Pi_{\text{id}}$ is optimal, and undoing a single inversion $i < j$ changes the cost by

$$\tfrac{2}{nS}\,(x_j - x_i)\big(T(x_j) - T(x_i)\big) \;\ge\; \sigma, \qquad \sigma := \frac{16}{nS},$$

since on either kind of pair one of the two gaps is at least 4 and the other at least 2. Hence $P_k^\star$ is the unique optimal vertex and every other vertex is worse by at least σ. Writing a general $P = \sum_v \theta_v v$ as a convex combination of vertices,

$$\langle C,\, P\rangle - \langle C,\, P_k^\star\rangle \;\ge\; \sigma \sum_{v \ne P_k^\star} \theta_v \;\ge\; \tfrac{\sigma}{2}\,\|P - P_k^\star\|_1,$$

because $\|P - P_k^\star\|_1 \le 2\sum_{v \ne P_k^\star} \theta_v$. The slack term of (10) at pair k is $\pi_k(P) = \big\|(P \odot \text{inc})\mathbf{1} - g \odot \mu_k\big\|_1$ for the drift vector g, and $|\pi_k(P) - \pi_k(P')| \le D\|P - P'\|_1$. Therefore, choosing

$$\lambda_{\text{drift}} \;:=\; \frac{\sigma}{4D} \;=\; \frac{4}{nSD},$$

a number of polynomial bit size, we get for every $P \ne P_k^\star$ and every drift vector g

$$\langle C,\, P\rangle + \lambda_{\text{drift}}\pi_k(P) - \langle C,\, P_k^\star\rangle - \lambda_{\text{drift}}\pi_k(P_k^\star) \;\ge\; \Big(\tfrac{\sigma}{2} - \lambda_{\text{drift}} D\Big)\|P - P_k^\star\|_1 \;>\; 0.$$

So the inner minimisation over P_k returns the monotone matching whatever Θ may be.

Step 2: what is left. Under the monotone matchings every matched displacement has magnitude one, so the total transport cost is $2K/S$, a constant. Upon a reset pair the realised row displacements are $-\mathsf{M}[r,i]$ and the model drift is zero, so the slack term is $\frac{1}{n}\sum_i |\mathsf{M}[r,i]| = 1$, again a constant. Upon the encoding pair $\mathsf{A} \to \mathsf{B}_r$ the realised row displacement at x_i is $\mathsf{M}[r,i]$ and the model drift is $z_r v_i$, so the slack term is $\frac{1}{n}\sum_i |\mathsf{M}[r,i] - z_r v_i|$. Adding up,

$$\text{OPT}_{\text{NC}} = \frac{2K}{S} + \lambda_{\text{drift}} K + \frac{\lambda_{\text{drift}}}{n} \min_{z \in \mathbb{R}^K,\, v \in \mathbb{R}^n} \big\|\mathsf{M} - z v^\top\big\|_1, \tag{36}$$

the minimisation being unconstrained because, with $d = 1$ and the indicator basis, $G(\Theta)[r,i] = z_r v_i$ ranges over exactly the matrices of rank at most one, and because $\lambda_w = 0$ leaves the z_r free.

Step 3: the threshold. Setting $V' := 2K/S + \lambda_{\text{drift}}K + \lambda_{\text{drift}}V/n$, a rational of polynomial bit size, we have $\mathrm{OPT}_{\mathrm{NC}} \le V'$ if and only if (35) holds. The grid has $4n-1$ atoms, the chain has $2K+1$ marginals, and every rational datum has polynomial encoding size, so the construction is polynomial. □

Remark S.1 (Numerical check)**.** *The pinning of Step 1 and the identity (36) were verified upon small instances by linear programming. For $(n, K) \in \{(3,3),(4,3),(5,2)\}$ and adversarial drift vectors drawn uniformly from $[-8,8]^n$, the optimal coupling of every pair was the monotone matching in every one of* 240 *trials per instance; and at random (z, v) the value of the constructed programme agreed with the right-hand side of (36) to 10^{-17}. The proof, and not the check, establishes the general claim.*

Remark S.2 (Scope: the basis must be rich)**.** *Three restrictions should be stated plainly, and the first most narrows the theorem.*

The basis dimension grows with the instance. *The reduction takes φ to be the indicator basis of the N_0 active atoms, so that $q = N_0$ and $b_k(x_i;\Theta) = (Hz_k)_i$ ranges over all rank-one matrices; that richness is what is exploited. For a* fixed *basis the picture is quite different. With the affine $\varphi(x) = (1,x)^\top$ of (5) one has $b_k(x_i;\Theta) = \alpha + \beta x_i + c^\top z_k$, and at $\lambda_w = 0$ the vector z_k is unconstrained, so that $s_k := c^\top z_k$ is merely a free scalar attached to pair k. The programme then reduces, for every d, to*

$$\min_{\alpha,\beta,s} \sum_k \sum_i \mu_k[i]\,\big|y_k[i] - \alpha - \beta x_i - s_k\big|,$$

a linear programme, solvable in polynomial time. The hardness is therefore a statement about the general model with a basis whose dimension grows, and not about the affine specialisation used in Section 7. The tractability just described belongs to the intercept-shift *model $\alpha + \beta x_i + c^\top z_k$, in which the latent state enters additively; it should not be transferred to the general bilinear parametrisation $\varphi(x)^\top H z_k$ with $\varphi = (1,x)^\top$, in which the state modulates the slope as well as the intercept, and whose complexity at fixed d we have not settled. What renders the latter tractable at $\lambda_w = 0$ is precisely what renders it uninteresting there — a free scalar per pair fits the mean displacements exactly — and it is the constraint tying the s_k to a low-order recursion, $\lambda_w > 0$ or $\lambda_w = \infty$, which restores the difficulty.*

The innovations are unconstrained. *Step 2 is where they are used: for a* deterministic $d = 1$ *system $z_k = \lambda^k z_0$, the row factor is geometric rather than free, and we do not know the complexity of that sub-case. Neither do we claim hardness for $\lambda_w > 0$. One may bound the perturbation of the objective by $\lambda_w K(1 + \|A\|)^2 Z^2$ upon a box $\|z_k\|_\infty \le Z$, but a threshold reduction is not preserved by a small perturbation without a promise gap and an a-priori bound upon the optimal states, neither of which the construction supplies; we leave the statement at $\lambda_w = 0$. For fixed d and $\lambda_w = \infty$ the complexity of (10) is likewise open.*

The hidden state is small. *Letting d grow makes the model so expressive that the fit is exact and the problem trivial. Hardness is thus a statement about a* small *hidden state, with unconstrained innovations and a rich output basis.*

Remark S.3. *Every quantity above has polynomial bit size: $|\mathcal{X}| = 4n-1$, $S = (4n-2)^2$, and σ^{-1}, D and $\lambda_{\mathit{drift}}^{-1}$ are polynomial in n. The reduction uses λ_{drift} small, that is, the regime in which the transport term dominates. We make no claim about the complementary hard-constrained regime $\lambda_{\mathit{drift}} \to \infty$: the construction specifies every entry of M, so exact agreement with a rank-one matrix is a rank test rather than a completion problem, and in a pure feasibility problem the optimality of one transport vertex does not by itself force that vertex.*

T THE EXACT E-STEP: INVARIANCE IN ϵ, AND FINE GRIDS

INVARIANCE OF THE EXACT E-STEP

Proposition T.1 (The exact E-step does not see ϵ)**.** *Fix μ_k, ν_k and feasible row destination means $m_i = x_i + \tilde{b}_k[i]$. The transport cost and the mass-weighted conditional variance then take the same value at* every *coupling with those marginals and row means, given in closed form by (37)–(38) of Appendix T; and for every $\epsilon > 0$ the minimiser of $\langle C, P\rangle - \epsilon \mathsf{H}(P)$ over that set is the same maximum-entropy coupling.*

So neither the quadratic cost nor ϵ selects the exact E-step coupling; an observed ϵ-dependence is evidence of residual infeasibility.

For feasible row means m_i, every coupling with the prescribed marginals satisfies

$$\sum_{ij} P[i,j](x_j - x_i)^2 = \sum_j \nu_k[j]x_j^2 + \sum_i \mu_k[i]x_i^2 - 2\sum_i \mu_k[i]x_i m_i, \tag{37}$$

$$\sum_i \mu_k[i]\,\mathrm{Var}_P(Y \mid X = x_i) = \sum_j \nu_k[j]x_j^2 - \sum_i \mu_k[i]m_i^2. \tag{38}$$

Proof of Proposition T.1. Expand $(x_j - x_i)^2$ and use the row sums, the column sums and $\sum_j P[i,j]x_j = \mu_k[i]m_i$; (38) follows by subtracting the squared row means from the second destination moment. The cost being constant, the programme reduces to maximising H over a non-empty convex set, and strict concavity gives uniqueness. Normalising C by the grid diameter, or replacing the entropy by $\mathrm{KL}(P \,\|\, \mu_k\nu_k^\top)$, changes only constants. □

Two consequences must be stated plainly. The quadratic transport cost plays no part in selecting the exact E-step coupling, and neither does ϵ: what the E-step returns is the maximum-entropy element of the feasible set. The entropic parameter therefore cannot be tuned to make P_k a particular bridge, and the average conditional variance of P_k is determined by the marginals and the drift before any transport problem is solved. Where an ϵ-dependence is observed numerically, it is evidence that the computed plan does not satisfy the constraints exactly — through clipping, through the soft marginal update of Appendix I, or through an unconverged sweep — and not evidence about the bridge.

The scope is the exact E-step. The slack-penalised programmes (9) and (10) do not fix the row means exactly and are not covered, nor is the transport-pinning mechanism of Appendix S; and a finite-tolerance or soft-marginal computation may depend upon ϵ strongly. Identities (37) and (38) are used in the reference implementation as unit tests upon the E-step.

THE BEHAVIOUR AT FINE GRIDS

The experiments reported above use grids of a few tens of atoms. Since the estimand is N-dependent by (2), and since the grid error of Theorem R.10 is the term which refinement is supposed to reduce, it is worth asking what happens when the grid is made very much finer. We therefore repeated the study at $N \in \{100, 200, 300, 400\}$, upon five seeds and at two regularisations — the $\epsilon = 0.05$ used throughout, and the Föllmer value $\epsilon^\star = 2\sigma^2\Delta/S$ — holding everything else at the values of Section 7, so that $M = 600$, $K = 24$. This study, and only this one, is offered in illustration of Theorem R.10, and its grid is therefore the uniform partition of a window $[-R, R]$ with $R = 6$, fixed before the data are drawn from the stationary law of (18) and not from the sample: three standard deviations of the factor-driven conditional mean $\mathbb{E}[X \mid Z]$, whose stationary standard deviation is 1.61, plus three of the conditional spread $X \mid Z$, whose standard deviation is $\sigma\sqrt{\Delta}/\sqrt{1-(1+\beta\Delta)^2} = 0.44$, which gives 6.13, rounded to the integer. Mass beyond R is assigned to the nearer endpoint, as (R2) prescribes, so that $h = 2R/(N-1)$ falls from $0.346\,\sigma\sqrt{\Delta}$ at $N = 100$ to $0.086\,\sigma\sqrt{\Delta}$ at $N = 400$, and the truncation mass τ_R of (R2) is a known quantity: given the factor path, X_{t_k} is Gaussian with mean m_k and variance v_k, both computable from (u_k), and $\mathbb{E}_{\mathcal{Z}}[(|X_{t_k}| - R)_+]$ and $\mathbb{P}_{\mathcal{Z}}(|X_{t_k}| > R)$ are the corresponding Gaussian tail integrals (`code/window_R.py`). Upon four of the five seeds τ_R is below 10^{-9}; upon the fifth (seed 7), whose factor path drives the conditional mean to -4.72 at the last pair, it is $4.3 \cdot 10^{-4}$ at $N = 100$ and $2.7 \cdot 10^{-4}$ at $N = 400$, of which $2.2 \cdot 10^{-4}$ is the tail excess and the remainder the boundary term $h\,\mathbb{P}_{\mathcal{Z}}(|X| > R)$ with $\mathbb{P}_{\mathcal{Z}}(|X| > R) = 1.7 \cdot 10^{-3}$. The grid term of (32) is therefore h^r and not τ_R throughout this study, the window having been chosen so. All forty runs completed; none was abandoned.

Refinement of the grid changes nothing. Every column of Table 2 is flat. The comparison across seeds is uninformative here, the seed-to-seed spread being far the larger; the sensitive test is the

Table 2: Medians over five seeds. rel is the relative discrepancy between the block iteration and the closed form (30); err is $\|\hat{\Theta} - \Theta_\star\|$; icv is the implied conditional variance of the coupling in units of $\sigma^2\Delta$; the clip is reported both as the fraction of pairs at which some atom is clipped and as the μ_k-mass actually displaced.

	N	$h/\sigma\sqrt{\Delta}$	clip (pairs)	clip (mass)	rel	icv	err
$\epsilon = 0.05$	100	0.346	0.29	$6.9 \cdot 10^{-4}$	$3.3 \cdot 10^{-4}$	0.493	0.0634
	200	0.172	0.33	$6.3 \cdot 10^{-4}$	$3.3 \cdot 10^{-4}$	0.491	0.0612
	300	0.115	0.33	$6.3 \cdot 10^{-4}$	$3.2 \cdot 10^{-4}$	0.490	0.0597
	400	0.086	0.33	$7.6 \cdot 10^{-4}$	$3.1 \cdot 10^{-4}$	0.490	0.0597
$\epsilon = \epsilon^\star$	100	0.346	0.29	$6.9 \cdot 10^{-4}$	$1.4 \cdot 10^{-4}$	0.491	0.0625
	200	0.172	0.33	$6.3 \cdot 10^{-4}$	$1.6 \cdot 10^{-4}$	0.489	0.0611
	300	0.115	0.33	$6.3 \cdot 10^{-4}$	$1.6 \cdot 10^{-4}$	0.488	0.0607
	400	0.086	0.33	$7.6 \cdot 10^{-4}$	$1.5 \cdot 10^{-4}$	0.488	0.0621

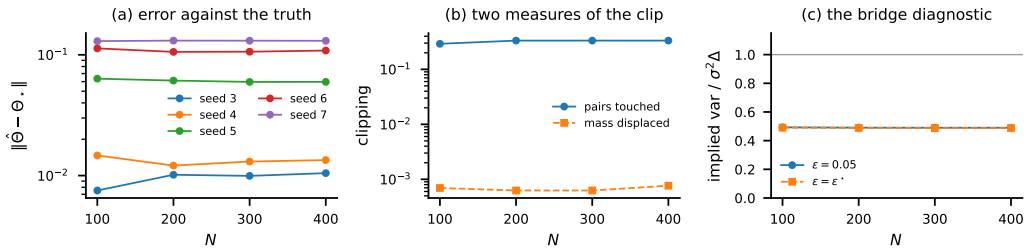

Figure 1: The fine-grid study. (a) the error against the truth, one line per seed: flat in N, and spread over a decade between seeds. (b) the two ways of measuring the clip. (c) the implied conditional variance at the two regularisations, with $\sigma^2\Delta$ marked.

paired one, in which each seed is followed across N. At $\epsilon = 0.05$ the error $\|\hat{\Theta} - \Theta_\star\|$ is

$$\begin{array}{lllll} \text{seed } 3: & 0.0075 & 0.0102 & 0.0099 & 0.0105 \\ \text{seed } 4: & 0.0147 & 0.0121 & 0.0131 & 0.0134 \\ \text{seed } 5: & 0.0634 & 0.0612 & 0.0597 & 0.0597 \\ \text{seed } 6: & 0.1130 & 0.1057 & 0.1061 & 0.1085 \\ \text{seed } 7: & 0.1298 & 0.1311 & 0.1309 & 0.1307 \end{array} \qquad (N = 100, 200, 300, 400),$$

so that quadrupling the number of atoms moves the estimate by between 0.0013 and 0.0073, whereas changing the seed at fixed N moves it from 0.0075 to 0.1311, a factor of seventeen. This is Theorem R.10 with the grid term extinguished and, since the Euler generator makes $\varsigma = 0$ (Remark R.14), the estimation term absent as well. What remains at fixed $M = 600$ and $K = 24$ is the Monte-Carlo term, whose constant $\lambda_{\min}(\widehat{G}_K)^{-1}$ depends upon the realisation of the factor path, together with the clipping; the seed spread is that of the design, and the grid term has long since fallen beneath it. There is accordingly a point — here at about $N = 20$, by Figure 5(a) — beyond which refinement of the grid is wasted effort, and the sentence of Section 7 should be read with that in mind.

The clip is not what a count of pairs reports. The fraction of pairs at which some active atom is clipped does not fall with N: it is 0.29 to 0.33 throughout. This is an artefact of the measure. The indicator fires when *any* active atom is clipped, and as N grows with the window held fixed there are ever more atoms in the tails carrying almost no mass, each of which may be clipped without consequence. The μ_k-mass actually displaced is $7 \cdot 10^{-4}$, equally flat, and it is that quantity which bounds the influence of the clip upon the estimate. Counts of pairs elsewhere in this paper should be read as reporting the support of the clip and not its size.

The bridge diagnostic does not discriminate here. The implied conditional variance is 0.490–$0.493\,\sigma^2\Delta$ at $\epsilon = 0.05$ and 0.488–$0.491\,\sigma^2\Delta$ at $\epsilon^\star$, agreeing to within 0.003 at every N, which

is Proposition T.1 and not a property of fine grids, although the two values of ϵ differ by an order of magnitude. It is not a fine-grid phenomenon in which the marginals dominate the entropic blur. By (38) the mass-weighted conditional variance is a function of the marginals and the row means alone, upon every grid, so the agreement is exact wherever the constraints are met and the apparent ϵ-dependence at $N = 14$ (Appendix Y) measures the residual infeasibility there. What the grid does govern is how nearly the row means can be realised at all; the value 0.49 is low here because the drift target used in this study is the *misspecified* closed form of (30), which carries no latent factor. Evaluated at the true drift the same quantity is $1.15\,\sigma^2\Delta$ at $N = 14$ and $1.02\,\sigma^2\Delta$ at $N = 100$: the constrained coupling reproduces the conditional spread of the generating model without any tuning of ϵ, which is the statement that survives.

The projection plateaus — but in which norm? At these grids the quantity we had been reporting, namely the largest row-conditional drift error $\max_i |d_k[i]/\mu_k[i] - \tilde{b}_k[i]|$ over active rows, settles at $5 \cdot 10^{-2}$ to $5 \cdot 10^{-1}$ and does not improve with further sweeping. That is a maximum over rows, unweighted by mass, and it is a poor summary. So too is the scalar check $|\mathbf{1}^\top d_k - \gamma_k|$ used elsewhere: it tests a single moment and is no certificate of feasibility, since a plan with large compensating row errors passes it. We therefore report the constraints separately. At 250 sweeps, over the 24 pairs of a run upon the sample-fitted grid of the other experiments (`code/estep_diag.py` and `code/cx_diag.py` both use it), the medians and worst cases are

	$\|P\mathbf{1} - \mu\|_1$	$\|P^\top\mathbf{1} - \nu\|_1$	$\|(P \odot \mathrm{inc})\mathbf{1} - \mu \odot \tilde{b}\|_1$	$\max_i \|\mathbb{E}_i h\|$	$\sum_i \mu_i \|\mathbb{E}_i h\|$	$\|\mathbf{1}^\top d - \gamma\|$
$N = 100$, median	$3 \cdot 10^{-16}$	$2 \cdot 10^{-5}$	$1 \cdot 10^{-16}$	$1 \cdot 10^{-16}$	$2 \cdot 10^{-17}$	$1 \cdot 10^{-16}$
worst	$4 \cdot 10^{-16}$	$1.4 \cdot 10^{-3}$	$4.9 \cdot 10^{-4}$	$1.2 \cdot 10^{-1}$	$4.9 \cdot 10^{-4}$	$7.6 \cdot 10^{-4}$
$N = 200$, median	$4 \cdot 10^{-16}$	$1 \cdot 10^{-6}$	$2 \cdot 10^{-16}$	$1 \cdot 10^{-16}$	$2 \cdot 10^{-17}$	$2 \cdot 10^{-16}$
worst	$5 \cdot 10^{-16}$	$1.3 \cdot 10^{-3}$	$4.8 \cdot 10^{-4}$	$1.1 \cdot 10^{-1}$	$4.8 \cdot 10^{-4}$	$7.2 \cdot 10^{-4}$
$N = 400$, median	$6 \cdot 10^{-16}$	$2 \cdot 10^{-7}$	$2 \cdot 10^{-16}$	$1 \cdot 10^{-16}$	$2 \cdot 10^{-17}$	$2 \cdot 10^{-16}$
worst	$6 \cdot 10^{-16}$	$1.4 \cdot 10^{-3}$	$5.2 \cdot 10^{-4}$	$1.1 \cdot 10^{-1}$	$5.2 \cdot 10^{-4}$	$7.9 \cdot 10^{-4}$

The row marginal is satisfied to machine precision by construction, the column marginal and the drift constraint to 10^{-3} and $5 \cdot 10^{-4}$ upon the worst pair of the run; and the 10^{-1} plateau is confined to the unweighted maximum, the same quantity weighted by mass being $5 \cdot 10^{-4}$. The plateau is an artefact of tail rows, but not a harmless one: the convex-order check below shows that on those pairs the constraint set is empty by a margin of the same order, so that the residual is the infeasibility and not the solver. Two further seeds at $N = 200$ give unweighted maxima $4 \cdot 10^{-2}$ and $6 \cdot 10^{-2}$ against mass-weighted $1.4 \cdot 10^{-4}$ and $1.6 \cdot 10^{-4}$, so the pattern is not particular to the seed. On fine grids the two should always be distinguished. The script is `code/estep_diag.py`.

The convex-order check. Remark G.1 makes joint feasibility of an E-step a convex-order statement, $\eta_k \preceq_{\mathrm{cx}} \nu_k$ with $\eta_k = \sum_i \mu_k[i]\delta_{m_i}$ and $m_i = x_i + \tilde{b}_k[i]$ the shifted and clipped targets, and gives it a finite test: equality of means, and

$$V_k := \max_t \Big\{ \sum_i \mu_k[i](m_i - t)_+ - \sum_j \nu_k[j](x_j - t)_+ \Big\} \;\le\; 0,$$

the maximum being attained at a knot of the two supports. We computed V_k and the mean defect $D_k := \sum_i \mu_k[i] m_i - \mathbb{E}_{\nu_k}[X]$ for every E-step of the runs reported in this paper (`code/cx_diag.py`), the targets being formed exactly as the E-step forms them, by the shift (12) followed by the clip. At the configuration of Section 7 — seed 3, $M = 600$, $K = 24$ — and at the fixed point of the iteration, $\max_k V_k$ is below 10^{-15} at $N = 100$: every E-step of that run is exactly feasible. At $N = 14$ it is $2.4 \cdot 10^{-4}$, positive upon 2 of the 24 pairs, and upon both it coincides with $|D_k|$: the violation is entirely a defect of the mean, and it has a single cause, a sample atom of mass $1/600$ at the edge of the support of μ_k whose target lies outside its reachable interval and is clipped by 0.07–0.14, which undoes the balance that (12) had just restored. At the true parameter the same figures are $2.0 \cdot 10^{-4}$ at $N = 14$ and below 10^{-15} at $N = 100$. Upon the fine grids of this appendix, with the misspecified target used here, $\max_k V_k$ is 3.2, 2.6 and $2.5 \cdot 10^{-4}$ at $N = 100$, 200, 400, positive upon 7–8 of the 24 pairs, again equal to the mean defect except upon one pair where a shape violation of $4 \cdot 10^{-5}$ remains; at the true parameter all three are below 10^{-15}. The plateau of the table above is therefore accounted for: a mean defect of 2–3 $\cdot 10^{-4}$ makes $\mathcal{A} \cap \mathcal{B} \cap \mathcal{D}$

empty, the cyclic projection settles at a best-effort plan whose drift residual, 4.9–$5.2 \cdot 10^{-4}$ when weighted by mass, is of the order of that defect, and no sweep budget can remove it. The defect enters the estimate only through γ_k, at 10^{-4} against displacements of order 10^{-1}, which is why the estimates of this appendix are unaffected; but it is an infeasibility of the constraint set and should be named as such. Iterating the shift and the clip to a common fixed point — redistributing the mass the clip has moved over the rows it did not touch — removes the defect to machine precision and leaves $\max_k V_k \le 4 \cdot 10^{-5}$ upon the fine grids and below 10^{-15} at $N = 14$; the runs reported here do not use it.

It does not follow that the budget matters. A study at $N = 200$ gives, for $100, 250, 500, 1000$ sweeps, a relative discrepancy of $3.467, 3.447, 3.412, 3.347$ in units of 10^{-4}, a deviation $\|\bar{y} - \gamma\|_\infty$ of $3.068, 3.042, 3.010, 2.952$ in units of 10^{-4}, and an implied variance of $0.4797, 0.4796, 0.4796, 0.4796$, at a cost rising from 7.4 to 52 seconds; every reported quantity is stable to better than one per cent from a hundred sweeps onwards. The runs above therefore use 250 sweeps, and the experiments of Section 7 are unaffected, retaining the 2000 of Appendix I. Since three sweep budgets appear in this paper, they are collected here once and for all:

experiment	E-step sweeps	where set
Section 7 and Appendices T, U ($N = 14, 100$)	2000	default of `bcd.fit_bcd_lds`
this appendix, large grids ($N = 100$–400)	250	`code/big_n.py`
L_{inner} of Corollary O.5	15	a rate bound, not a run

The third is not the configuration of any experiment but the number of inner sweeps per *outer* iteration at which Corollary O.5 evaluates the composite rate; the two former are the budgets of the E-step itself. Cost grows as $N^{1.6}$ to $N^{1.9}$, from 25 seconds a run at $N = 100$ to 317 at $N = 400$, as the $\mathcal{O}(N^2)$ sweep of Appendix I would suggest.

U Comparison with trajectory inference

The methods closest to ours in input are those of trajectory inference, which also take unpaired snapshots and return something one may call a drift. They do not, however, return the same object, and the differences are instructive enough to be worth setting out. We compare estimands and hypotheses first, and numbers afterwards, with the caveats that the numbers require.

What each procedure estimates. Write $\mathcal{H}_k = \sigma(X_{t_k}) \vee \mathcal{S}_k$ for the feature σ-field of (3).

- *Waddington-OT* (Schiebinger et al., 2019) couples consecutive snapshots by entropically regularised transport with growth, and reads a velocity from the resulting displacement. The drift it reports at a cell is a conditional mean displacement per unit time.
- *Mean-field Langevin in path space* (Chizat et al., 2022), with the identifiability theory of Lavenant et al. (2024), estimates a path measure by minimising an entropy against a Wiener reference subject to the observed marginals. Its drift is that of the minimising measure, obtained non-parametrically, and is the Markovian projection $\mathbb{E}[a_t \mid X_t]$ of Gyöngy (1986).
- *PO-MFL* (Gu et al., 2025) extends this to a partially observed state, regularising the path measure towards a dynamics prior on a latent process.
- The present paper estimates $\mathbb{E}[a_{t_k} \mid \mathcal{H}_k]$ within a parametric family, by constraining the first conditional moment of the coupling.

Why a comparison on mean displacements is vacuous. It is tempting to compare methods by the displacement they attribute to a marginal. That comparison cannot separate them. As observed before Proposition 6.1, $\mathbf{1}^\top d_k = \gamma_k$ for *every* coupling with marginals (μ_k, ν_k), whatever drift it was projected against and whatever cost, growth term or entropic penalty produced it. The mean displacement is a functional of the two marginals alone. Any two procedures which report it will therefore agree exactly, and by Proposition 6.1 our own estimator sees the couplings through γ_k and nothing else. This is borne out to machine precision below: the three coupling-based procedures reproduce γ_k to $4 \cdot 10^{-16}$, $4 \cdot 10^{-16}$ and $6 \cdot 10^{-11}$ respectively. What distinguishes the procedures

is not the displacement but what is regressed upon it, and what is done with the variation *within* a marginal — which, by (29), our E-step annihilates.

The feature σ-field is what separates the estimands. The Gyöngy projection $\mathbb{E}[a_t \mid X_t]$ estimated by the path-space methods is our $\mathbb{E}[a_{t_k} \mid \mathcal{H}_k]$ in the case $\mathcal{S}_k$ degenerate, that is case (iii) of Section 2.1: independent copies, a deterministic curve of measures, and a mean-field drift. This is precisely the regime in which the identifiability theory of Lavenant et al. (2024) operates, and in which, by Section 3, there is no dynamic model left for us to fit. When the copies share a factor, $\mathcal{S}_k$ is not degenerate, our estimand is strictly finer, and the two procedures are estimating different things — neither wrongly. It is worth being explicit that the finer estimand is bought by a parametric hypothesis, and that a non-parametric method cannot be faulted for failing to recover an object its observation model does not identify.

The latent-state case. PO-MFL (Gu et al., 2025) is the nearest relative of our Non-Convex Setting, both admitting a hidden state which the marginals do not observe. The difference is the direction of the assumption: a *prior* upon the latent dynamics is imposed there, whereas here those dynamics are the object to be identified. In the language of (10) a dynamics prior is a penalty upon the innovations, and so corresponds to $\lambda_w \in (0, \infty)$, the filtering case of Section 4.2; our hardness result concerns $\lambda_w = 0$, where no such prior is available, and Theorem 4.1 may be read as saying what the prior is buying.

The comparison, and its design. Every procedure receives the same input, namely the snapshots $\{\mu_k\}$ upon a common grid, and returns a drift field $b_k(\cdot)$ at each step. Since the generating model (18) is known, so is the truth, $b_k(x) = \alpha + \beta x + c^\top Z_k$ exactly, and each method may be scored by $\|b_k - b_k^{\mathrm{true}}\|_{L^2(\mu_k)}$ averaged over k.

What we have implemented are *surrogates* for the three published procedures, sharing their transport layer but not their statistical formulation, and the entries of Table 3 should be read as such. Precisely:

- *Waddington-OT surrogate*: *balanced* entropic transport between consecutive snapshots, with the drift read off as the conditional mean displacement, at $\epsilon = 0.05$ and at $\epsilon^\star$. The procedure of Schiebinger et al. (2019) carries a growth term and is unbalanced; we have not implemented growth, there being none in (18) to estimate. One consequence should be flagged: the mean-displacement identity $\mathbf{1}^\top d_k = \gamma_k$ holds for any plan with the *prescribed* marginals, and does not transfer unchanged to an unbalanced formulation whose marginal constraints are different.
- *Mean-field Langevin surrogate*: the multi-marginal Schrödinger bridge upon the whole time chain against a discrete Markov reference, every snapshot imposed *exactly*, solved by iterative proportional fitting in the log domain to a residual below 10^{-11}. The estimator of Chizat et al. (2022) does something different: it trades a data-fitting term against path-space relative entropy, and its algorithm evolves point clouds in a grid-free formulation. Imposing every empirical marginal exactly removes the marginal-denoising problem which their statistical formulation is built to address, and our surrogate therefore tests the transport layer alone.
- *PO-MFL surrogate*: an innovation penalty upon the latent recursion of the small linear system of (10). Gu et al. (2025) instead carry a *known* driving vector field together with an unknown potential and an observation map, and their latent is attached to each individual whereas ours is a factor common to the population. We make no claim that their estimator is equivalent to (10) at any particular λ_w; the correspondence drawn above is one of modelling posture, not of objective.

This is a comparison of surrogates upon a common grid with a common scoring rule, which is the most that can be said without reimplementing three papers; the numbers below should not be quoted as the performance of the published estimators.

The two baselines are one baseline. The path-space and pairwise estimators do not merely agree closely; they agree to $2 \cdot 10^{-13}$ at $N = 14$ and $2 \cdot 10^{-12}$ at $N = 100$, upon every seed. That is a theorem and not a coincidence.

Proposition U.1 (The path-space and pairwise bridges have the same drift)**.** *Let R be a Markov reference measure upon $x_0, \ldots, x_n$ with transition kernels K_k, and let π minimise $\mathrm{KL}(\cdot \,\|\, R)$ subject to $\pi_k = \mu_k$ for every k. Then for each k the joint law of (x_k, x_{k+1}) under π is the solution of*

Table 3: Drift error $\|b_k - b_k^{\mathrm{true}}\|_{L^2(\mu_k)}$, averaged over k; median over five seeds, with the range beneath. The last column is the largest deviation of the implied mean displacement from γ_k. The three comparators are the surrogates described above, not the published estimators.

	$N = 14$		$N = 100$		mean-disp.
method	median	range	median	range	gap
ours, Non-Convex Setting	0.039	0.006–0.048	0.020	0.003–0.045	$3 \cdot 10^{-2}$
PO-MFL surrogate, data prior	0.024	0.022–0.033	0.026	0.022–0.033	$1 \cdot 10^{-8}$
PO-MFL surrogate, oracle prior	0.027	0.017–0.047	0.023	0.022–0.032	$1 \cdot 10^{-8}$
mean-field Langevin surrogate (path space)	0.074	0.069–0.077	0.058	0.056–0.063	$6 \cdot 10^{-11}$
Waddington-OT surrogate, $\epsilon = \epsilon^\star$	0.074	0.069–0.077	0.058	0.056–0.063	$4 \cdot 10^{-16}$
Waddington-OT surrogate, $\epsilon = 0.05$	0.193	0.167–0.210	0.190	0.167–0.202	$4 \cdot 10^{-16}$
ours, Convex Setting	0.559	0.490–0.722	0.557	0.482–0.722	1.3

the two-marginal Schrödinger problem between μ_k and μ_{k+1} with reference $R_{k,k+1}(dx, dy) = R_k(dx)K_k(x, dy)$. In particular the two estimators return the same drift, provided the pairwise method uses that same reference kernel, which for the quadratic cost means $\epsilon = \epsilon^\star$.

Proof. The constraints being upon the one-dimensional marginals, π has density $\prod_k f_k(x_k)$ with respect to R, so π is Markov and its $(k, k+1)$-marginal is $A_k(x)f_k(x)\,K_k(x, y)\,f_{k+1}(y)B_{k+1}(y)$, where A_k and B_{k+1} are the leftward and rightward messages. This is of the form $u(x)K_k(x, y)v(y)$ and has marginals μ_k, μ_{k+1} by construction; the two-marginal Schrödinger problem has exactly one solution of that form with those marginals, the Sinkhorn scalings being unique up to a reciprocal constant. □

Coupling the entire path therefore buys nothing at the level of the drift, and what separates the two surrogates in Table 3 is the choice of ϵ alone: at $\epsilon = 0.05$ the error is three times that at $\epsilon^\star$.

The scope of that statement wants care. Proposition U.1 is a statement about a finite positive Markov reference and *exactly* prescribed marginals; under those hypotheses the entropy chain rule gives

$$\mathrm{KL}(Q \,\|\, R) \;\geq\; \mathrm{KL}(\mu_0 \,\|\, r_0) + \sum_k \mathrm{KL}\big(Q_{k,k+1} \,\|\, \mu_k \otimes K_k\big),$$

with equality for the Markov law glued from the adjacent conditionals, so that each pair may be optimised alone. What the near-machine-precision agreement in Table 3 validates is therefore this restricted identity between two hard-constraint implementations. It does *not* show that fitting noisy marginals jointly buys nothing in the statistically regularised path-space procedures of Chizat et al. (2022) and Lavenant et al. (2024), where the marginals are not imposed exactly and the joint problem does not decouple. A reader who takes Proposition U.1 for a general redundancy of those methods will have taken more from it than it contains.

Against the two procedures without a latent state the estimator of the Non-Convex Setting is the more accurate by a factor of two to ten. Against PO-MFL, which has one, the comparison is close and its direction depends upon the grid: at $N = 14$ PO-MFL is the better, 0.024 against 0.039, and at $N = 100$ ours is, 0.020 against 0.026. The crossing is instructive. Refinement of the grid improves our estimate, whose M-step is parametric and gains from better resolved marginals, and leaves PO-MFL's unchanged, its latent being itself discretised. Estimating the latent prior from observables rather than granting it the truth changes little — 0.024 against 0.027 at $N = 14$ — so the comparison does not rest upon that concession. Three qualifications should be entered, and the last is the most important. First, the data are generated by the model being fitted, so the comparison is of a correctly specified parametric procedure against non-parametric ones; the latter are not thereby shown to be inferior at what they are for. Secondly, PO-MFL is designed for a latent carried by each individual, whereas the factor here is common to the population, and its benchmarks accordingly lie outside the scheme of Section 3; we have run it upon our data and not ours upon theirs. Thirdly — and this is visible in the table — the advantage is not attributable to the transport layer at all. The Convex Setting uses the same machinery and is the *worst* entry, at 0.56, because it has no latent factor and no input with which to follow $c^\top Z_k$; whereas the baselines, which estimate a fresh drift field at every step, track it to 0.06. What the table measures is the value of a correct parametric hypothesis and

of an observed exogenous input, exactly as Proposition 6.1 would lead one to expect, the couplings having been shown there to be inert. A procedure which had our model class but no drift constraint would, by that proposition, do as well.

Two consequences follow for the reading of the table, and we would rather state them than let a reader find them. The Convex Setting entry is not a fair test of convex against non-convex identification, because it is denied the very input the others exploit; the appropriate observed-feature competitor is a finite impulse response driven by the same u, with its order chosen upon held-out data, and that is the comparison of Figure 3(b,c), which is the better starting point. And the essential baseline throughout is the *mean-only* estimator — ordinary or ridge least squares for the convex model, and direct fitting of (16) for the latent one — since by Proposition 6.1 it is exactly what our estimator reduces to once the couplings are seen to be inert. It is reported in Appendix Y, where the block decomposition and the direct fit are compared upon twenty seeds; it tests the contribution claimed for transport more directly than any of the entries above, and a reader chiefly interested in that question should go there first.

Table 4: Drift error $\|b_k - b_k^{\mathrm{true}}\|_{L^2(\mu_k)}$ averaged over k at $N = 100$, and wall-clock seconds per fit: medians over five seeds, with ranges. The comparators are the surrogates of Appendix U, not the published estimators; PO-MFL is granted a prior upon the latent dynamics, which we must identify. Our Non-Convex time includes the 2000-sweep E-step; the Convex Setting is the closed form (30).

	drift error		time (s)	
method	median	range	median	range
ours, Non-Convex Setting	0.020	[0.003, 0.045]	31.5	[27.7, 43.2]
PO-MFL surrogate, data prior	0.026	[0.022, 0.033]	5.9	[5.1, 7.5]
PO-MFL surrogate, oracle prior	0.023	[0.022, 0.032]	6.3	[5.9, 7.8]
mean-field Langevin surrogate (path space)	0.058	[0.056, 0.063]	0.5	[0.5, 0.6]
Waddington-OT surrogate, $\epsilon = \epsilon^\star$	0.058	[0.056, 0.063]	1.7	[1.7, 2.7]
Waddington-OT surrogate, $\epsilon = 0.05$	0.190	[0.167, 0.202]	1.7	[1.6, 1.8]
ours, Convex Setting	0.557	[0.482, 0.722]	< 0.01	—

Pooling across the horizon. Table 5 varies the number of copies M and the number of pairs K at $N = 14$. The two surrogates estimate a fresh field at every step, so their error is that of one displacement field, of order $M^{-1/2}$, and does not fall with K: the PO-MFL surrogate stands between 0.024 and 0.028 for $K = 24$, 48 and 96 at $M = 600$. The parametric fit pools the K residuals into $2d + 2$ parameters and falls from 0.039 to 0.011 over the same range, and at $M = 50$ it is about twice as accurate as the PO-MFL surrogate at every K. The comparison is less regular at $K = 24$ and at $M = 100$, because on some seeds the M-step lands upon the interchanged allocation of poles of Remark 2.2, which leaves every mean intact but alters β and with it the drift field; the parameter error upon those seeds is 0.12 against 0.01–0.03 upon the others, and the cells are a numerical instance of that ambiguity rather than of estimation noise. Two variants which we tried, and which the tables do not report, may be recorded here. With an unobserved disturbance upon the latent state (standard deviation 0.2 per step) the per-step surrogates tie with or beat the parametric fit in sample, since the realised disturbance is visible in the next snapshot and only a method which commits to no recursion can follow it; and with the slow factor of Figure 3 (poles 0.9, 0.8) neither 24 nor 96 training pairs identify the near-unit poles well enough at $N = 14$ for the forecast of Table 1 to separate the methods. The scripts are `code/compare_regimes.py` and `code/collect_regimes.py`.

Everywhere above PO-MFL is granted a prior which is close to correct, fitted either to the true latent or to the observable residuals of γ upon $(1, \bar{x})$ — that is, to the residual of the very regression which Proposition 6.1 says our own fixed point performs; and since our own estimator is granted no prior at all but must identify (a, c, α, β), it is fair to ask what that concession is worth. We therefore misspecify the prior deliberately, moving the mean-reversion ϕ and then the innovation scale τ away from their fitted values, holding the latent grid and the input coefficient fixed so that only the assumed dynamics change. Medians over five seeds:

Table 5: Drift error $\|b_k - b_k^{\mathrm{true}}\|_{L^2(\mu_k)}$, averaged over k, against the number of copies M and the number of pairs K, at $N = 14$; medians over 5 seeds with the range in brackets. The surrogates are those of Table 3; the PO-MFL surrogate uses the prior estimated from observables. A per-step method sees each pair once and sits at the Monte-Carlo floor of one displacement field whatever K is; the parametric fit pools the K residuals. The script is `code/compare_regimes.py`.

M	K	Waddington-OT, $\epsilon = \epsilon^\star$	PO-MFL surrogate, data prior	ours, Non-Convex Setting
50	24	0.097 [0.082, 0.101]	0.070 [0.053, 0.076]	0.038 [0.029, 0.046]
50	48	0.099 [0.090, 0.111]	0.071 [0.068, 0.084]	0.033 [0.027, 0.058]
50	96	0.103 [0.097, 0.112]	0.072 [0.067, 0.081]	0.038 [0.015, 0.062]
100	24	0.087 [0.079, 0.091]	0.052 [0.047, 0.055]	0.039 [0.024, 0.053]
100	48	0.090 [0.079, 0.092]	0.053 [0.046, 0.058]	0.043 [0.023, 0.048]
100	96	0.093 [0.089, 0.098]	0.052 [0.050, 0.060]	0.041 [0.004, 0.047]
200	24	0.079 [0.073, 0.085]	0.037 [0.035, 0.049]	0.024 [0.010, 0.056]
200	48	0.082 [0.074, 0.088]	0.043 [0.035, 0.044]	0.019 [0.009, 0.040]
200	96	0.087 [0.086, 0.091]	0.043 [0.037, 0.048]	0.015 [0.006, 0.046]
600	24	0.074 [0.069, 0.077]	0.024 [0.022, 0.033]	0.039 [0.006, 0.048]
600	48	0.076 [0.071, 0.082]	0.028 [0.022, 0.031]	0.011 [0.005, 0.044]
600	96	0.082 [0.077, 0.083]	0.028 [0.021, 0.036]	0.011 [0.001, 0.045]

ϕ	0.00	0.30	0.50	fitted ≈ 0.65	0.85	0.95
$N = 14$	0.044	0.029	0.029	0.024	0.027	0.029
$N = 100$	0.053	0.035	0.027	0.026	0.024	0.022

$\tau\times$	$\frac{1}{4}$	$\frac{1}{2}$	1	2	4
$N = 14$	0.068	0.056	0.024	0.092	0.193
$N = 100$	0.054	0.043	0.026	0.099	0.191

The dependence is strongly asymmetric between the two parameters, and it is the scale and not the persistence which matters. Misspecifying ϕ over its whole admissible range costs about a factor of two, and not even monotonically: at $N = 100$ the error falls as ϕ is raised past its fitted value, so a prior which is too persistent is no worse than the right one, and only the memoryless choice $\phi = 0$ is clearly harmful. Misspecifying τ is another matter. Understating it by a factor of four doubles to trebles the error; *over*stating it by a factor of four multiplies the error by eight, to 0.19, which is worse than Waddington-OT at $\epsilon^\star$ and worse than every other entry of Table 3 save our own Convex Setting, which has no latent factor at all. The advantage PO-MFL holds at $N = 14$, and the parity it holds at $N = 100$, both disappear once the assumed innovation scale is wrong by a factor of two.

This is the price of the trade. A dynamics prior buys a great deal when it is approximately right — enough to beat us upon coarse grids — and it is the quantity hardest to know in advance, namely how much the latent state moves per step, that it is most sensitive to. Our own estimator makes no such assumption and cannot be misspecified in that way — the dynamics are estimated rather than assumed, though the order d and the lag length p remain choices; what it pays instead is the requirement of persistent excitation, hypothesis (R5) of Appendix R, without which it degrades in its own fashion. Neither procedure is free, and the two are charged in different coin. The script is `code/pomfl_misspec.py`.

The dynamics prior of PO-MFL occupies the middle ground of Section 4.2 at $\lambda_w \in (0, \infty)$, and Theorem 4.1, which concerns $\lambda_w = 0$, may be read as saying what that prior buys. We have no data set in which a common factor is present and known, so the comparison remains one upon simulated data.

V Discussion of the convergence proofs

A few comments on the proof are in order.

WHICH GEOMETRY THE E-STEP RATE LIVES IN.

The constraint sets $\mathcal{A}, \mathcal{B}, \mathcal{D}$ are affine in the probability coordinates P and, at an interior feasible point, smooth embedded submanifolds of the positive cone; they are not affine in $L = \log P$, where a marginal constraint reads $\sum_j e^{L_{ij}} = \mu_i$. For convex alternating projections in Euclidean geometry Bauschke & Borwein (1996) yield a linear rate, and for cyclic I-projections onto affine families Csiszár (1975) yields global convergence; the proof of Theorem L.1 in Appendix N works directly in the W-geometry at the feasible point, which is what a rate for Kullback–Leibler projections requires, and needs neither convexity in Bregman coordinates nor the Lewis–Luke–Malick framework. That framework is retained (Theorem M.6) for two purposes: the transversality constant $c_{\mathcal{ABD}}$ of Lemma M.5, which is computable from the data but is not claimed to be the Kullback–Leibler contraction factor; and the inequality-bounded variant of the projection, in which $\mathcal{D}$ is replaced by a one-sided polytope — as required when the predictable drift is constrained to dominate or be dominated pathwise — whose feasible set is no longer affine but remains super-regular, and for which the affine argument of Appendix N would have to be replaced.

ROLE OF THE GLOBAL-BALANCE CORRECTION.

The shift δ_k in (12) restores the *necessary* mean balance, and it is what the tower property requires; it does not by itself give $\mathcal{A}\cap\mathcal{B}\cap\mathcal{D} \neq \emptyset$. The exact criterion is one of convex order: with $m_i = x_i + \tilde{b}_k[i]$ and $\eta = \sum_i \mu_k[i]\delta_{m_i}$, a coupling with the prescribed marginals and row means exists if and only if $\eta \leq_{\mathrm{cx}} \nu_k$, by Strassen's theorem; on the line it suffices to check $\sum_i \mu_k[i](m_i - t)_+ \leq \sum_j \nu_k[j](x_j - t)_+$ at the knots of the two supports. Row-wise reachability together with mean balance is necessary and not sufficient, since it ignores the capacity of the destination marginal: with $\mu = (0.45, 0.10, 0.45)$, $\nu = (0.05, 0.90, 0.05)$ upon $(-1, 0, 1)$ and $m = (-0.9, 0, 0.9)$, every target is interior and the means balance, yet $\sum_i \mu_i m_i^2 = 0.729 > 0.1 = \sum_j \nu_j x_j^2$, which conditional Jensen forbids. Independent clipping can moreover destroy the balance it was meant to preserve. The correct repair is to project the proposed row means onto the barycentric feasible set, a convex quadratic programme; we note this rather than claim it, since the fixed-point analysis of Section 6 is derived for the uniform shift and would have to be redone. The shift is the empirical analogue of the tower identity $\mathbb{E}[r_{k+1} - r_k] = \mathbb{E}[\mathbb{E}[r_{k+1} - r_k \mid \mathcal{G}_k]]$: without it, $\mathcal{D}$ cannot meet the marginal slice $\mathcal{A} \cap \mathcal{B}$ (for instance when the trial parameter θ is inconsistent with the empirical mean increment), in which case the alternating projection oscillates between an infeasible region and a meaningless "best-effort" compromise. To state separately what each device guarantees: the clipping secures row-wise interior reachability, which is (T2); the uniform shift δ_k restores the aggregate mean identity; neither, separately or jointly, guarantees $\mathcal{A} \cap \mathcal{B} \cap \mathcal{D} \neq \emptyset$, for which the convex-order condition of Remark G.1 is needed; and when that condition holds, the shifted and clipped targets are those to which Theorem L.1 applies.

EXACT VERSUS APPROXIMATE PROJECTION ONTO $\mathcal{D}$.

The argument of Theorem L.1 requires the alternating step to be a true projection up to a controlled tolerance. The Newton update of Γ in `code/core.py` computes the exact KL projection onto $\mathcal{D}$ row by row to floating-point tolerance, by exploiting the strict monotonicity and convexity of the one-dimensional residual. A naive single-step gradient update yields only an averaged projection with a weaker constant; the Newton variant is therefore the correct choice for the rate of Theorem L.1 to apply.

RELATION TO THE NON-CONVERGENCE OF THREE-BLOCK DIRECT-EXTENSION ADMM.

A natural objection arises in view of the well-known result of Chen et al. (2016): the direct extension of two-block ADMM to three (or more) primal blocks coupled through a single linear constraint and a single dual variable is *not* guaranteed to converge, even on convex problems with $f_i \equiv 0$. Concretely, for

$$\min_{x_1,x_2,x_3} \sum_{i=1}^{3} f_i(x_i) \quad \text{s.t.} \quad A_1x_1 + A_2x_2 + A_3x_3 = b,$$

they exhibit an instance for which the joint primal–dual iteration matrix of the direct-extension scheme has spectral radius strictly greater than one. The result is sharp: additional structure (strong convexity in a subset of blocks, prox-regularisation, Gaussian back-substitution, smaller dual step) is required to recover convergence.

Two structural features of our procedure ensure that this obstruction does not apply.

1. *The outer iteration is two-block.* The decision variables split as $(\theta, \{P_k\})$ – the parametric Markovian-projection drift on one side, the family of couplings on the other – with the multipliers $\{\Gamma_k\}$ on the predictable-compensator constraints. The plans $\{P_k\}_{k\in\mathcal{P}}$ are *not* separate ADMM blocks coupled through a single dual variable: given θ, the inner sub-problems decouple completely across k, each carrying its own multiplier Γ_k, and the M-step couples them only through the shared parameter θ. Two-block ADMM/BCD converges under standard hypotheses Gabay & Mercier (1976); Boyd et al. (2011), and our outer rate in Theorem O.2 relies on this two-block structure, not on a direct multi-block extension.
2. *The inner iteration is cyclic Bregman projection, not three-block ADMM.* Inside each E-step we project a single coupling P onto three smooth submanifolds $\mathcal{A}, \mathcal{B}, \mathcal{D}$ in the KL geometry. This is the classical alternating-projection / Dykstra setting (Bauschke & Borwein, 1996; Lewis et al., 2009), structurally distinct from a three-block augmented-Lagrangian scheme: there is no shared linear constraint over three primal variables and no single dual, only one primal iterate and three projection operators. The Chen–He–Ye–Yuan counter-example exhibits a spectral radius > 1 for the joint primal–dual iteration matrix of the direct extension; the analogous quantity in our setting is the contraction factor $c_E < 1$ of Theorem L.1, the W-norm of a product of tangent projections restricted to the complement of the common tangent space, which is strictly below one whenever the intersection is non-empty.

Where the analogy *would* bite is a hypothetical variant in which each transport plan P_k is treated as a separate ADMM block coupled to θ through a single augmented Lagrangian – a $(\theta, P_{k_1}, \dots, P_{k_m})$ direct extension with one shared dual variable. By Chen et al. (2016), that scheme would not be safe in general. The estimator deliberately avoids this route: θ is the only second block, and the $\{P_k\}$ are solved as independent inner sub-problems, each as its own constrained Schrödinger bridge.

W THE COUPLINGS AS TRANSITION KERNELS

Proposition 6.1 shows that the *estimate* uses the couplings only through their first moments, and Section 7 confirms it numerically. That leaves a separate question untouched: whether the P_k are any good as kernels in their own right, which is what they would have to be for path interpolation or any downstream use of the fitted dynamics. The first-moment analysis says nothing either way, and it would be an error to read "inert at the level of the estimate" as "inert".

The generating model being known, the true row-conditional law of $X_{t_{k+1}}$ given $X_{t_k} = x_i$ is $N\big(x_i + b_k(x_i)\Delta,\ \sigma^2\Delta\big)$, which we project upon the grid and compare, row by row and weighted by μ_k, against $P_k(\cdot \mid x_i)$ — in total variation and in W_1. As a reference we compute the same distances for the *unconstrained* entropic plan between the same marginals, that is for the coupling a method which does not impose (8) would return.

Table 6: Fidelity of the row-conditional laws to the true transition kernel; medians over five seeds, at $\epsilon = 0.05$ and at $\epsilon^\star$.

	total variation		W_1	
coupling	$N = 14$	$N = 100$	$N = 14$	$N = 100$
drift-constrained, $\epsilon = 0.05$	0.033	0.091	0.023	0.022
drift-constrained, $\epsilon = \epsilon^\star$	0.034	0.091	0.024	0.022
unconstrained, $\epsilon = 0.05$	0.173	0.203	0.168	0.167
unconstrained, $\epsilon = \epsilon^\star$	0.135	0.134	0.094	0.065

The constraint earns its place here, and by a wide margin: it improves the kernel by a factor of three to eight in W_1 and of one and a half to five in total variation, uniformly over seeds, grids and both

regularisations — the W_1 comparison being the more meaningful of the two here, for the reason given below, and also the more favourable. This is the complement of Proposition 6.1 rather than a contradiction of it. The drift constraint fixes the first conditional moment of every row, which is exactly the feature of the true kernel that an unconstrained optimal-transport plan gets wrong — such a plan being chosen for cheapness of transport and not for fidelity to a diffusion — whereas it is also, by (29), the one feature the M-step subsequently discards. The transport layer is therefore idle for the parameter and useful for the kernel, and a practitioner who wants only $\hat{\theta}$ may take the two-line regression of (30), while one who wants the couplings should pay for them.

Two smaller points may be read from Table 6. The unconstrained plan improves markedly upon moving from ϵ to $\epsilon^\star$, as the Föllmer reading would suggest, whereas the constrained plan is insensitive to ϵ — once the first moment is pinned, the entropic parameter has little left to do. And the constrained total variation degrades from 0.03 to 0.09 as the grid is refined while W_1 holds at 0.022: total variation between laws upon a finer grid is the more demanding comparison, and W_1, which respects the geometry of the line, is the more meaningful of the two here.

X THE GRID, AND WHAT WOULD REPLACE IT

The estimator is Eulerian: it carries a coupling $P_k \in \mathbb{R}^{N\times N}$ upon an atomic grid, and every statement above is made for a scalar observed process. Partitioning a d_x-dimensional observation space into n bins per axis gives $N = n^{d_x}$ atoms and couplings with n^{2d_x} entries, so the representation is confined in practice to $d_x \leq 2$ or 3. This is a limitation of the implementation rather than of the formulation, and it is worth saying which parts would survive a change of representation.

What is intrinsically low-dimensional is the object the estimator actually uses. By Proposition 6.1 the fixed point depends upon the couplings only through the pairs $(\bar{x}_k, \gamma_k)$, and by (15) the M-step has $2d + 2$ parameters and K residuals whatever N may be. Neither scales with the grid at all. It is the E-step alone that carries the N^2 burden, and it does so to produce a quantity which (29) then discards down to its first moment.

Three replacements suggest themselves, in increasing order of ambition. A *particle* or Lagrangian representation would carry the M samples themselves and solve each pairwise entropic problem by Sinkhorn upon the $M \times M$ cost, which is dimension-free in d_x though quadratic in M; the drift constraint remains one linear constraint per source particle, and the global-balance shift (12) and the reachability clipping carry over verbatim. A *low-rank* or tensor-train factorisation of P_k would exploit the fact that the entropic kernel $\exp(-C/\epsilon)$ is, for a separable cost, a product across coordinates, so that Sinkhorn may be run upon factors; the drift constraint couples coordinates only through a scalar per row and is compatible with such a factorisation. A *continuous* formulation would dispense with the grid entirely, parameterising the Schrödinger potentials by neural networks as in the data-driven bridging of Appendix U and imposing the conditional first moment as a penalty; this is the least conservative option and would forfeit the certified global solves of Section 5, which are what the present paper is able to offer.

The convergence theory is indifferent to which is chosen: Theorem L.1 requires only that the three sets be affine in the coupling and pairwise transversal, and Theorem R.10 refers to the grid solely through the term $\mathcal{O}(\Delta^{-1}h^r)$, which a grid-free representation removes. What would need re-examination is the feasibility analysis of Appendix G, since the reachable range $[x_k^- - x_i, x_k^+ - x_i]$ is a statement about the support of an atomic marginal.

Y ADDITIONAL EXPERIMENTS

The decomposition upon the instance of Section 7. The outer iterates fall by a factor of 700 at the first step, the reduced M-step depending upon the E-step through $(\bar{x}_k, \gamma_k)$ alone and so being at its fixed point after one exact E-step (Appendix Q); the E-step residual falls from $8\cdot10^{-2}$ to $8\cdot10^{-11}$ in 43 sweeps with slope $c_E = 0.62$ (Figure 2). We recover $\hat{a} = (-0.586, 0.081)$, $\hat{c} = (0.400, 0.401)$ against $(-0.6, 0.08)$, $(0.4, 0.4)$; the pipeline agrees with the closed form to 10^{-4} from $N = 20$ onwards (Appendix T).

Global versus local solution of the sub-problem, in full. Figure 4 compares the solvers of Section 5 on five instances each. On the acquiescent system they agree to within a few per cent, as Theorem P.2(i) predicts. On the non-acquiescent system two things break: branch and bound restricted to $\mathcal{B}_\varrho$ returns 0.42–0.93 against 0.005–0.064 unrestricted, so acquiescence costs two orders of magnitude and is a genuine restriction; and projected gradient, which the theory covers only inside $\mathcal{B}_\varrho$, fails to solve even its own restricted problem on two of five instances, landing at 24 and 44 against certified 0.93 and 0.42, one of which the Ho–Kalman warm start repairs. A substantial fraction of random restarts land above the certified global value. These instances use the fixed affine basis and are not in the hard family of Theorem 4.1 (Section 4.3); what the restarts exhibit is the bilinear non-convexity of the recursion, which is what the global solver is for.

What the transport layer contributes, in full. Proposition 6.1 says the fixed point uses the marginals only through their means, so the block iteration ought to coincide with a two-line regression, and it does (Figure 5(a)): the relative gap is $2 \cdot 10^{-3}$–$7 \cdot 10^{-3}$ at $N = 8$ and falls to 10^{-4} by $N = 20$ and stays below $4 \cdot 10^{-4}$ upon every finer grid we have examined (Appendix T). In the Non-Convex Setting the pipeline and an LDS fitted directly to (γ_k) differ by $1.5 \cdot 10^{-4}$–$1.3 \cdot 10^{-3}$ in parameter, with errors against the truth that agree to three decimals over five seeds. This is the identity above at work: the transport layer can move the estimate only when the drift target is infeasible, which on a fine grid it is not.

The decomposition against the complete problem, in full. Table 7 in Appendix Y puts the two side by side on instances small enough for (10) to be closed ($N = 10$, $K = 11$, $d = 2$, 1361 variables, both with $a \in \mathcal{B}_\varrho$). They are *not* the same estimator: branch and bound attains a penalised objective 14–26% lower, because the block iteration is not minimising that objective at all — by Proposition 6.1 its fixed point is a regression of mean displacements — yet the distance to the true parameter, which is what matters for identification, is the same to within 10% and orders the two the other way on one of the three instances. The decomposition is therefore a fast, consistent estimator in its own right rather than a solver for (10); conversely (10) is no dead end, closing in under a second at this size.

The entropic parameter, in full. Figure 5(b,c) sweeps ϵ over four decades around $\epsilon^\star = 2\sigma^2\Delta/S$. These runs must now be read in the light of Proposition T.1, which says that an exactly feasible plan does not depend upon ϵ at all and that its mass-weighted conditional variance is fixed by (38) before any transport problem is solved. The variation reported below is therefore a diagnostic of the solver and of the clipping, not a property of the bridge, and we retain it only as such. On a grid of spacing comparable with the diffusive scale ($h = 1.4\,\sigma\sqrt{\Delta}$ — not narrower than $\sigma\sqrt{\Delta}$, and we no longer describe it as resolving the diffusion) the implied conditional variance of the coupling is $0.36, 0.60, 1.11, 1.86, 4.00, 7.18$ times $\sigma^2\Delta$ at $\epsilon^\star/4, \dots, 32\epsilon^\star$, crossing 1 at $\epsilon^\star$ as the Föllmer reading requires; on a coarse grid ($h = 10.4\,\sigma\sqrt{\Delta}$) it is 13–$20\,\sigma^2\Delta$ at every ϵ, the bin width and not the diffusion setting the scale. The drift error is flat at 0.002–0.010 across the whole sweep. The E-step is stiffest at $\epsilon^\star$: 60 sweeps there leave a KKT residual of 43% of $\|\gamma\|_\infty$ and a drift error of 0.25, against 0.4% and 0.0033 at the 2000 sweeps used throughout (Appendix I). Taken together with Proposition T.1 this identifies what the sweep measures: at $N = 14$ the clipping is active upon a third of the pairs and the row constraint is not met exactly, so the computed plan retains an ϵ-dependence which an exactly feasible plan could not have. Upon the finer grids of Appendix T, where the drift residual is smaller, the dependence disappears — as the proposition requires. The value $\epsilon^\star$ is therefore not to be read as that at which P_k becomes the Föllmer–Schrödinger bridge: matching the unconstrained Gibbs kernel to the diffusion is a statement about a different coupling from the one the E-step computes.

Scalability, and what the comparison is between. In the Convex Setting the complete QP beats the decomposition by two to three orders of magnitude (Figure 6, Appendix Y). In the Non-Convex Setting the decomposition keeps the KN^2 transport variables out of branch and bound: at $d = 3$, as N grows from 20 to 80, the complete QCQP slows from 12 to 140 s and the decomposition from 11 to 38 s (medians over three seeds), 2.4–$3.7\times$ faster for $N \geq 30$. At $d = 2$ the QCQP stays faster, though its lead shrinks from $22\times$ to $1.4\times$; at $d = 4$ neither dominates.

Three qualifications travel with every one of those ratios, and we repeat them here rather than leave them to the caption of Figure 6. The two sides are *different estimators* solving different criteria,

Table 7: The block-coordinate decomposition next to the globally solved complete problem (10) ($N = 10$, $K = 11$, $d = 2$, $a \in \mathcal{B}_\varrho$ in both; the QCQP has 1361 variables and 684 constraints). Columns 2–4: the penalised objective of (10) at the parameter returned by each method, and the lower bound certified by spatial branch and bound. Columns 5–6: distance of the returned parameter to the truth. The two methods optimise different criteria, so the objective ordering and the accuracy ordering need not agree.

	penalised objective (10)			$\|\hat{\Theta} - \Theta_\star\|$		time (s)	
seed	decomposition	global QCQP	bound	decomp.	QCQP	decomp.	QCQP
3	1.2642	1.0047	1.0047	0.064	0.053	1.5	0.3
4	1.3741	1.2287	1.2287	0.139	0.147	1.2	0.2
5	1.7346	1.5045	1.5045	0.119	0.116	1.6	0.4

so a faster solve is not a like-for-like optimisation speed-up; a smaller value of (10) attained by the complete programme is not a certificate that the block estimator has failed to converge, since the block estimator is not minimising (10). And the two sides are not given equal resources: the complete solves use a commercial solver upon all ten cores, whereas the E-step of the decomposition is single-threaded NUMPY. The figures above are therefore what a practitioner with this hardware would observe, and not a measurement of algorithmic work.

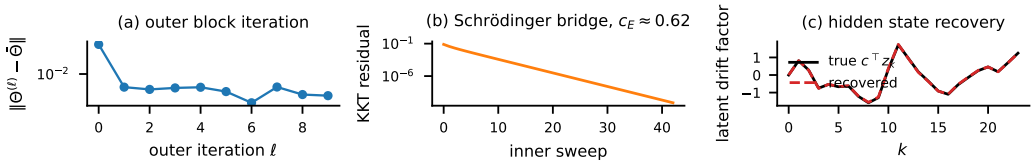

Figure 2: The Non-Convex Setting with a globally solved M-step: outer iteration; bridge projection, c_E off the slope; hidden factor and reconstruction.

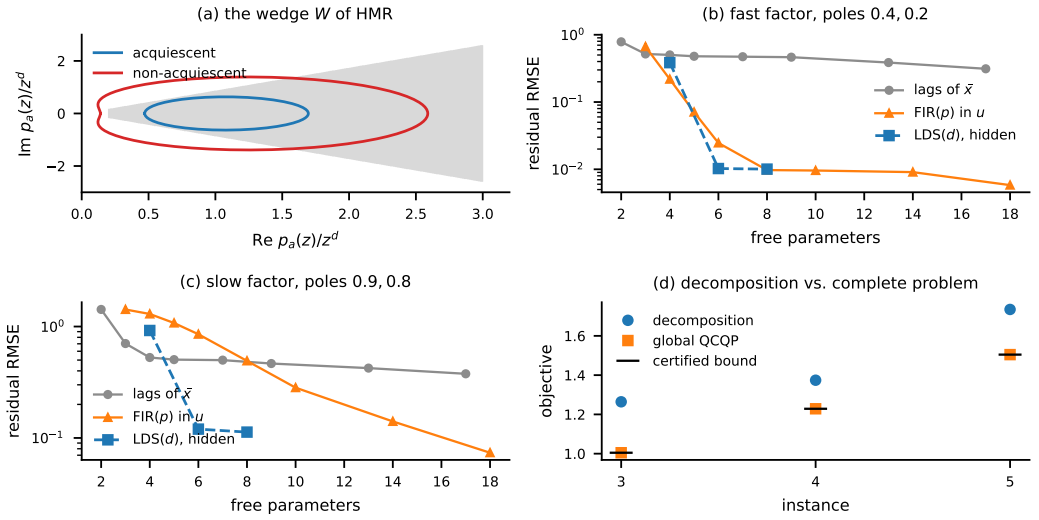

Figure 3: (a) the truncated wedge $\mathcal{W}$ of (6) and the curves $\{p_a(\zeta)/\zeta^d : |\zeta| = \varrho\}$ for an acquiescent and a non-acquiescent system; the latter is stable but leaves the wedge. (b,c) residual against number of free parameters for three model classes — lags of $\bar{x}$, a FIR filter of the input u (the memoryless case of the Non-Convex Setting, Remark 2.1), and the latent LDS — on a fast-mixing and a slow-mixing latent factor. The hidden state affords parsimony, and the gain grows with the memory of the factor; a baseline denied u (grey) is not a fair comparison. (d) objective attained by the decomposition and by global branch and bound on (10), with the certified lower bound.

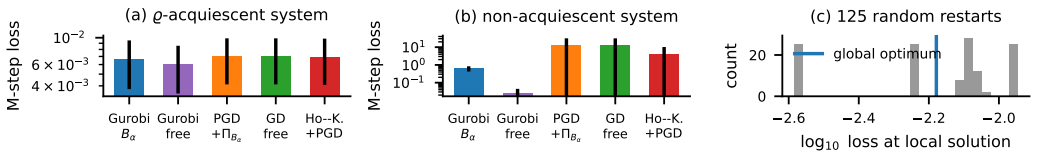

Figure 4: The drift-identification sub-problem (15) on an acquiescent and on a non-acquiescent system, and the loss reached by first-order descent from random restarts against the value certified by branch and bound.

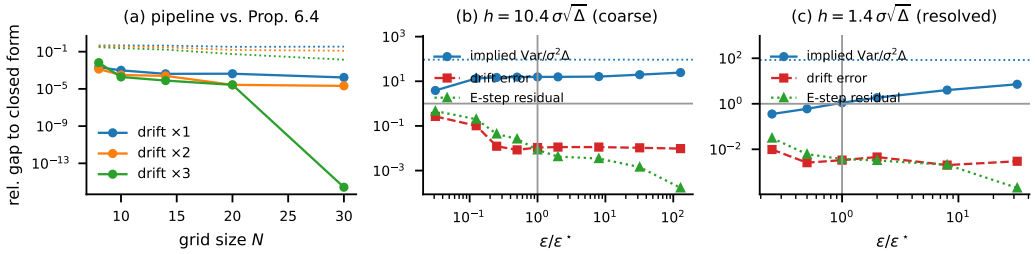

Figure 5: (a) relative gap between the full pipeline and the closed form of Proposition 6.1 against grid size (solid), with the fraction of pairs on which the drift target is clipped (dotted): the gap is the clipping and nothing else. (b,c) the conditional variance implied by the fitted couplings, in units of $\sigma^2\Delta$, and the drift-parameter error, as ϵ is swept around $\epsilon^\star = 2\sigma^2\Delta/S$; the dotted line is the variance of ν_k, which the marginals impose as a ceiling. Panels (b) and (c) are the same process on a coarse ($h = 10.4\,\sigma\sqrt{\Delta}$) and on a fine ($h = 1.4\,\sigma\sqrt{\Delta}$) grid. The second is not narrower than $\sigma\sqrt{\Delta}$; what it shows is that once h is of the order of $\sigma\sqrt{\Delta}$ the implied conditional variance tracks ϵ and crosses $\sigma^2\Delta$ at $\epsilon^\star$, whereas at ten times that spacing the bin width alone sets the scale.

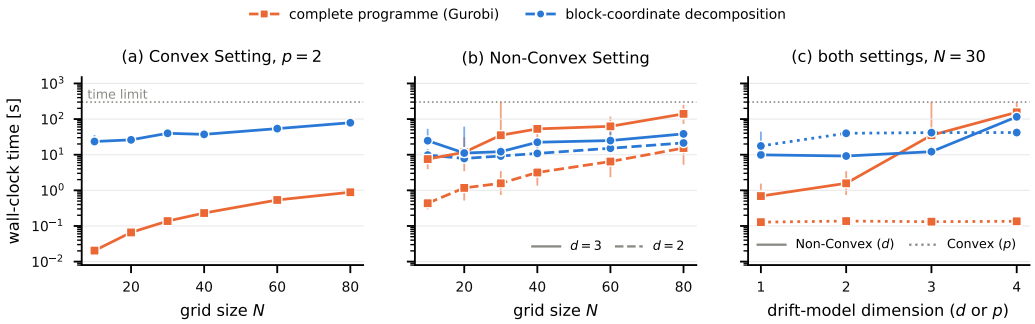

Figure 6: Wall-clock time of the complete programme — (9) as a convex QP, (10) as a non-convex QCQP, both solved by Gurobi on all ten cores of an Apple M2 Pro — and of the block-coordinate decomposition, whose E-step is single-threaded Python and whose M-step in the Non-Convex Setting is the global QCQP (15). The decomposition runs until the relative change in the parameter falls below 10^{-4}, for at most 15 outer iterations. Population scheme with $K = 11$ pairs and $M = 400$ copies, as in Table 7; medians with min–max bars over three seeds; the dotted line is the 300 s limit imposed on the complete solves. (a,b) Against the grid size N, i.e. against KN^2 transport variables, at $p = 2$ and $d \in \{2, 3\}$. (c) Against the dimension of the drift model at $N = 30$: the latent dimension d in the Non-Convex Setting and the lag order p in the Convex Setting, the latter on the same data. The complete QCQP carries all transport variables into every node of the branch and bound, whereas the M-step of the decomposition has $\mathcal{O}(Kd)$ variables whatever N is.