Geodesic and Adapted Exponential Smoothing: Wasserstein Smoothing for Multivariate Distributions and the Laws of Stochastic Processes

Alfredo Sepúlveda-Jiménez[1,2]

[1]National University
[2]QDR Labs

May 2026

Abstract

[Mat+26] introduced Wasserstein exponential smoothing (WES), a one-parameter recursive forecaster for distributional time series on $\mathbb{R}$, by lifting the classical exponential-smoothing recursion to a geodesic interpolation in the 2-Wasserstein space and exploiting the isometry between $\mathcal{W}_2(\mathbb{R})$ and the Hilbert space $L^2((0,1))$ of quantile functions. We develop three structural generalizations and a unifying theory. First, we recast the WES update as a *θ-weighted Wasserstein barycenter*, equivalently a point on a metric geodesic, and show that this variational form (i) recovers WES verbatim in one dimension, (ii) is well posed on general nonpositively curved and barycentric metric spaces, and (iii) admits a closed form on multivariate Gaussians via the Bures–Wasserstein geometry, thereby resolving the multivariate extension flagged as open in the original work. Second, we treat the case in which each observation is the *law of a stochastic process*: we show that the naive Wasserstein geometry on path space is the wrong object—value functionals from optimal stopping and hedging are discontinuous in it—and that the correct geometry is the *adapted* (bicausal) Wasserstein distance. Using the recent result that the adapted Wasserstein space is geodesic and supports barycenters, we define adapted WES and give a backward-induction recursion for its update. Third, we place WES inside a family of *proximal/divergence smoothers* $\mu_t = \arg\min_\rho (1-\theta)D(\rho\|\mu_{t-1}) + \theta D(\rho\|\nu_t)$, recovering geometric (log-linear) pooling, mixture smoothing, and Sinkhorn/sliced variants as special cases, and clarifying the "mix versus morph" distinction. A Hilbert meta-theorem shows that whenever a smoother is linear in a Hilbert embedding, the stationarity, autocovariance-decay, and consistency results of WES transfer verbatim, and a bounded-geometry refinement carries them to the curved cases with explicitly degraded constants. Building on this theory we then resolve the questions left open by the original work: we prove a central limit theorem for the smoothing parameter, with $\sqrt{T}(\hat{\theta}_T - \theta_*) \to \mathcal{N}(0, \theta_*(2-\theta_*)/d)$ in the isotropic case; we give a stable damped-trend and seasonal calculus; we identify WES with a steady-state Kalman filter and supply a convergent online estimator; and we cast it in Koopman operator form. Finally we move the geometry beyond the Riemannian world: since the p-Wasserstein space is Finsler for $p \neq 2$, a Banach/L^p meta-theorem shows the guarantees persist under two-uniform convexity and smoothness, degrading by constants that equal one precisely when $p = 2$—explaining the privileged status of the quadratic cost—while a Berwald–Hadamard contraction handles the curved Finsler case and identifies the abstract bounded-geometry modulus with Finsler anisotropy (the Cartan tensor). We give complete proofs and Bures–Wasserstein experiments confirming both consistency of the minimum-Wasserstein estimator in $\mathbb{R}^2$ and the predicted asymptotic variance, and we close with merits and limitations.

Keywords: exponential smoothing; optimal transport; Wasserstein barycenter; adapted/bicausal Wasserstein distance; distributional time series; Bures–Wasserstein geometry; Finsler structures; pre-Finsler structures.

1 Introduction

Exponential smoothing (ES) is among the most durable forecasting devices in statistics. In its level-only form, a real-valued series $\{y_t\}$ is tracked by a latent predictor obeying the recursion $x_t = (1-\theta)x_{t-1} + \theta y_t$, governed by a single smoothing parameter $\theta \in [0,1]$ [Bro59; Hol57; Win60; Gar85; Gar06]. Despite its parsimony, ES remains a benchmark that more elaborate procedures struggle to beat in large-scale forecasting competitions [MSA20; MSA22], a robustness explained in part by its equivalence to a local-level state-space model and to an ARIMA$(0,1,1)$ structure [Mut60; Hyn+08].

Modern applications increasingly produce observations that are themselves probability distributions: daily distributions of intraday financial returns, household demand profiles, glucose densities [Mat+21], and voxelwise functional imaging summaries [PM16]. When such objects are recorded in sequence they form a *distributional time series*, a stochastic process whose state at each time is a probability measure. Because the space of measures is not a vector space, the elementary operations of addition and scalar multiplication underlying classical forecasting are unavailable, and optimal transport under the Wasserstein metric has emerged as the natural geometric substitute [Vil09; PZ20; ABS24]. Existing methodology for distributional time series is predominantly autoregressive: Wasserstein autoregression on tangent spaces [ZKP22], autoregressive transport maps [ZM23], distribution-on-distribution regression [CLM23; GP22], and iterated transportation [GP24].

Against this backdrop, [Mat+26] introduced *Wasserstein exponential smoothing* (WES). Their construction reads the classical recursion geometrically—x_t is the point a fraction θ of the way along the Euclidean segment from x_{t-1} to y_t—and replaces the segment by the Wasserstein geodesic (displacement interpolation) between the previous smoothed measure and the current observation. In one dimension this is exceptionally tractable, because $\mathcal{W}_2(\mathbb{R})$ is isometric to $L^2((0,1))$ via quantile functions, so WES is literally classical ES applied to quantile functions. The authors establish Fréchet-mean stationarity, exponential decay of the residual autocovariance, and consistency of the minimum-Wasserstein estimator of θ, and report strong empirical performance. They are explicit, however, that the method is confined to $\mathbb{R}$: the quantile isometry has no multivariate analogue, and "adapting this smoothing framework to distributions on $\mathbb{R}^d$ for $d > 1$ remains non-trivial."

Contributions. This paper develops the WES idea well beyond its scalar origin and supplies a theory that explains which of the original guarantees survive, why, and what their limiting behaviour is. The first five contributions build the framework; the last two use it to settle the questions left open by [Mat+26] and to push the geometry past the Riemannian world.

1. **A barycentric/variational reformulation (Section 3).** We show the WES update is the θ-weighted Wasserstein barycenter of the pair (μ_{t-1}, ν_t),

$$\mu_t = \arg\min_{\rho} \; (1-\theta)\,\mathcal{W}_2^2(\rho, \mu_{t-1}) + \theta\,\mathcal{W}_2^2(\rho, \nu_t),$$

   and that this barycenter coincides with the geodesic point used by [Mat+26] in one dimension. The variational form is coordinate-free and is the engine for everything that follows.

2. **Geodesic ES and the multivariate resolution (Sections 4–5).** The barycentric update is well posed on any geodesic metric space that admits barycenters—in particular on nonpositively curved (Hadamard) spaces [Stu03] and, via [AC11], on $\mathcal{W}_2(\mathbb{R}^d)$. We give a geometric error-correction form $\mu_t = \exp_{\mu_{t-1}}\big(\theta \log_{\mu_{t-1}} \nu_t\big)$ that reduces to WES in $\mathcal{W}_2(\mathbb{R})$, and we make the multivariate filter fully explicit on Gaussians through the Bures–Wasserstein geometry [BJL19; Che+20; Alt+21], thereby addressing the open problem.

3. **Smoothing the laws of stochastic processes (Section 6).** The phrase "distributional time series" admits a second reading in which *each observation is itself the law of a stochastic process* (a path measure). We argue that the plain Wasserstein distance on path space is the wrong geometry, because functionals defined through the information filtration—optimal-stopping and hedging values—are discontinuous in it. The correct object is the *adapted* (bicausal) Wasserstein distance [Bac+20a; Bac+20b; Las18; ABZ20]. Building on the recent theorem that the adapted Wasserstein space is geodesic and supports barycenters [BBP25], we define *adapted WES* and give a backward-induction recursion for its update [PP12; EP24].

4. **General smoothing by divergences (Section 7).** Writing the update as a two-point proximal problem with a general divergence D exposes WES as one member of a family that also contains geometric (log-linear) pooling and mixture smoothing, and clarifies a "mix versus morph" dichotomy between information-geometric and transport-geometric smoothers.

5. **A unifying Hilbert meta-theorem (Section 8).** We isolate the structural hypothesis behind the original proofs—linearity in a Hilbert embedding with the embedding norm as loss—and show that under it the stationarity, autocovariance-decay, and consistency results transfer verbatim. This single observation subsumes the one-dimensional case, MMD/RKHS smoothing, and exponential-family natural-parameter smoothing, and pinpoints the curved cases ($\mathcal{W}_2(\mathbb{R}^d)$, adapted Wasserstein) as the locus of genuinely new analysis.

6. **Resolving the open problems, with a central limit theorem (Section 9).** We settle the four directions flagged for future work. We prove a CLT for the smoothing parameter, $\sqrt{T}(\hat{\theta}_T - \theta_*) \to \mathcal{N}(0, V)$ with $V = \mathbb{E}\,\langle D, \Sigma_F D\rangle\,/(\mathbb{E}\,\|D\|^2)^2$, reducing in the isotropic case to the closed form $V = \theta_*(2-\theta_*)/d$; we develop damped-trend and seasonal smoothers with an explicit Schur stability region and a free-lunch isotonic projection; we identify WES with a steady-state Kalman filter (gain $\theta = (-q + \sqrt{q^2+4q})/2$) and give a provably convergent online estimator of θ; and we exhibit WES as a Koopman resolvent $\mathcal{K}_\theta = \theta(I - (1-\theta)L)^{-1}$ with a single pole at $1-\theta$.

7. **Pre-Finsler and Finsler geometries (Section 10).** Because the update uses only geodesics, it extends beyond the Riemannian $\mathcal{W}_2$. As $\mathcal{W}_p$ is genuinely Finsler for $p \neq 2$ [Agu12], we prove a Banach/L^p meta-theorem: the stationarity, mixing, and consistency results persist under two-uniform convexity and smoothness, with constants that degrade by the Ball–Carlen–Lieb moduli and collapse to the Hilbert values exactly when $p = 2$, explaining why the quadratic cost is special. A Berwald–Hadamard contraction theorem covers curved Finsler targets, a sharp non-contraction result delimits the scope, and we identify the abstract bounded-geometry modulus η with Finsler anisotropy (the Cartan tensor), unifying the flat and curved theory on a single axis.

We validate the multivariate filter, the consistency of the minimum-Wasserstein estimator, and the predicted asymptotic variance in Bures–Wasserstein experiments (Sections 9.1 and 11), and close with merits and limitations (Section 12). Throughout, our aim is to build transparently on [Mat+26]: WES is the irreducible scalar case, and the contributions are the geometry and theory

that let it travel to $\mathbb{R}^d$, to path space, to other divergences, to Finsler curvature, and to a full inferential limit theory.

2 Background

2.1 Classical exponential smoothing

Let $\{y_t\}_{t=1}^T \subset \mathbb{R}$. Simple ES maintains a level x_t via the component form

$$y_t = x_{t-1} + e_t, \qquad x_t = (1-\theta)x_{t-1} + \theta y_t, \qquad \theta \in [0,1], \tag{1}$$

equivalently the error-correction form $x_t = x_{t-1} + \theta e_t$ with $e_t = y_t - x_{t-1}$ [Gar06; Hyn+08]. The predictor x_t is the point a fraction θ along the segment from x_{t-1} to y_t; small θ smooths, large θ tracks. Iterating (1) expresses x_t as an exponentially weighted average of past observations, whence the name.

2.2 Wasserstein geometry

For a Polish space $(\mathcal{X}, \rho)$ let $\mathcal{P}_2(\mathcal{X})$ be the Borel probability measures with finite second moment. The 2-Wasserstein distance is

$$\mathcal{W}_2^2(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int \rho(x,y)^2 \, \pi(dx\, dy), \tag{2}$$

where $\Pi(\mu,\nu)$ is the set of couplings [Vil09; San15]. On $\mathcal{X} = \mathbb{R}^d$ and for μ absolutely continuous, the infimum is attained by a deterministic map: $\nu = (T_{\mu\to\nu})_{\#}\mu$ with $T_{\mu\to\nu} = \nabla\varphi$ the gradient of a convex potential (Brenier). The displacement interpolation (McCann geodesic) is

$$\mu_\theta = \big((1-\theta)\mathrm{Id} + \theta\, T_{\mu\to\nu}\big)_{\#}\mu, \qquad \theta \in [0,1], \tag{3}$$

the constant-speed geodesic from μ to ν in $\mathcal{W}_2(\mathbb{R}^d)$ [AGS08; ABS24]. In dimension one the optimal map and distance are explicit through quantile functions U, V of μ, ν:

$$\mathcal{W}_2^2(\mu,\nu) = \int_0^1 \big(U(q) - V(q)\big)^2 dq, \qquad T_{\mu\to\nu} = V \circ U^{-1}. \tag{4}$$

The map $\mu \mapsto U$ is an isometry from $\mathcal{W}_2(\mathbb{R})$ onto the closed convex cone of nondecreasing functions in $L^2((0,1))$ [PZ20]. This is the structural fact that the original WES exploits, and the one that fails for $d > 1$.

Fréchet mean. For a random measure ξ in $\mathcal{W}_2(\mathcal{X})$ the Fréchet mean and variance [Fré48] are

$$\mathbb{E}(\xi) = \arg\min_{\rho} \mathbb{E}\big[\mathcal{W}_2^2(\rho,\xi)\big], \qquad V(\xi) = \mathbb{E}\big[\mathcal{W}_2^2(\mathbb{E}(\xi),\xi)\big], \tag{5}$$

generalizing the Euclidean mean; existence and uniqueness in Wasserstein space are treated in [AC11; LL17; ZP19].

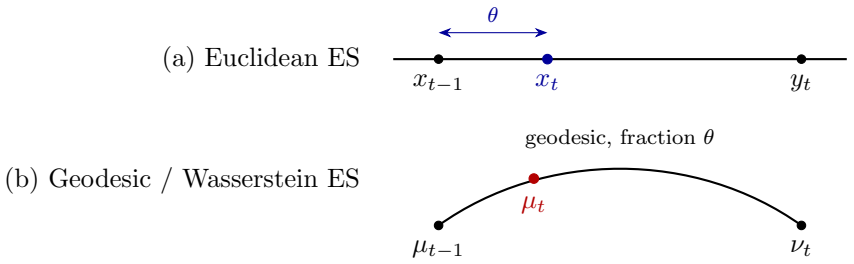

Figure 1: Classical ES places x_t a fraction θ along the Euclidean segment from x_{t-1} to y_t; WES and its generalizations replace the segment by a metric geodesic (here the Wasserstein displacement interpolation) from μ_{t-1} to ν_t. The barycentric reformulation of Section 3 shows the two pictures are the same variational object.

2.3 Wasserstein exponential smoothing

[Mat+26] observe a series of random measures $\nu_1, \dots, \nu_T$ in $\mathcal{W}_2(\mathbb{R})$ and form predictors μ_t. The *WES process* posits $\nu_t = (E_t)_{\#}\mu_{t-1}$ for i.i.d. monotone random maps E_t and

$$\mu_t = \big((1-\theta_*)\mathrm{Id} + \theta_*\, T_{\mu_{t-1}\to\nu_t}\big)_{\#}\mu_{t-1}, \tag{6}$$

i.e. μ_t is the McCann geodesic point (3) at parameter θ_*. In quantile coordinates U_t, V_t this is exactly classical ES, $U_t = (1-\theta_*)U_{t-1} + \theta_* V_t$, with error-correction form $U_t = U_{t-1} + \theta_* F_t$, $F_t := V_t - U_{t-1}$. The *WES filter* runs (6) at a chosen θ on observed data, and θ is estimated by the *minimum-Wasserstein estimator*

$$\hat\theta_T = \operatorname*{arg\,min}_{\theta\in(0,1)} L_T(\theta), \qquad L_T(\theta) = \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{W}_2^2(\mu_t^\theta, \nu_{t+1}). \tag{7}$$

Under moment conditions on E_t, [Mat+26] prove Fréchet-mean stationarity $\mathbb{E}(\nu_t) = \mu_0$, exponential decay of $\mathbb{E}\left\langle U_t^\theta - V_{t+1}, U_s^\theta - V_{s+1}\right\rangle$ in $t-s$, and consistency $\hat\theta_T \xrightarrow{p} \theta_*$. As we make precise in Section 8, every one of these arguments uses only the Hilbert structure of $L^2((0,1))$ together with the residual moment bounds, never any property special to $\mathbb{R}$.

3 A barycentric reformulation of WES

The first observation is that the geodesic update (6) solves a two-point weighted least-squares problem. This converts a construction that appears to need the explicit transport map into a coordinate-free variational principle that makes sense on any metric space.

Definition 1 (Weighted Wasserstein barycenter)**.** For $\mu, \nu \in \mathcal{P}_2(\mathcal{X})$ and $\theta \in [0,1]$, the θ-weighted barycenter is

$$\mathrm{bary}_\theta(\mu,\nu) := \operatorname*{arg\,min}_{\rho\in\mathcal{P}_2(\mathcal{X})} (1-\theta)\,\mathcal{W}_2^2(\rho,\mu) + \theta\,\mathcal{W}_2^2(\rho,\nu). \tag{8}$$

Proposition 1 (WES is a barycenter)**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ with μ absolutely continuous. Then the minimizer in (8) is unique and equals the McCann geodesic point,*

$$\mathrm{bary}_\theta(\mu,\nu) = \big((1-\theta)\mathrm{Id} + \theta\, T_{\mu\to\nu}\big)_{\#}\mu.$$

In particular, in $\mathcal{W}_2(\mathbb{R})$ the WES update of [Mat+26] satisfies $\mu_t = \mathrm{bary}_\theta(\mu_{t-1}, \nu_t)$, and in quantile coordinates $U_t = (1-\theta)U_{t-1} + \theta V_t$.

The proof (Appendix A.1) is a specialization of the displacement-interpolation and Agueh–Carlier barycenter theory [ABS24; AC11]: a two-measure barycenter with weights $(1-\theta, \theta)$ is exactly the displacement interpolant at θ. The one-dimensional statement also follows directly from (4): in quantile coordinates the objective is the strictly convex quadratic $(1-\theta)\,\|R-U\|^2 + \theta\,\|R-V\|^2$ on $L^2((0,1))$, minimized at $R = (1-\theta)U + \theta V$, which is automatically nondecreasing (a convex combination of nondecreasing functions), hence a valid quantile function. This last point—that the minimizing quantile function is feasible without projection—is what makes the scalar case so clean and is precisely what must be re-examined in the multivariate and trend extensions.

Remark 1 (Why the barycentric form matters)**.** Three consequences follow immediately.

1. *No transport map is needed in the objective.* The update is defined through distances alone, so it transfers to spaces where transport maps may not exist but barycenters do [LL17].
2. *Stability.* Because $\rho \mapsto (1-\theta)\mathcal{W}_2^2(\rho,\mu) + \theta\mathcal{W}_2^2(\rho,\nu)$ is 1-strongly convex along generalized geodesics on $\mathcal{W}_2(\mathbb{R}^d)$ [AGS08], the barycenter is unique and depends continuously on (μ,ν), yielding a nonexpansive filter map.
3. *A proximal reading.* Equation (8) is a single step of a weighted Fréchet mean; iterating it is a discrete gradient flow of the data-fidelity functional, linking WES to JKO-type schemes [JKO98] and to the divergence smoothers of Section 7.

4 Geodesic exponential smoothing

Proposition 1 frees the construction from $\mathbb{R}^d$. Let $(\mathcal{M}, d)$ be a geodesic metric space. We define *geodesic exponential smoothing* (GES) by the same barycentric update,

$$\mu_t = \mathrm{bary}_\theta(\mu_{t-1}, \nu_t) = \operatorname*{arg\,min}_{\rho\in\mathcal{M}} \; (1-\theta)\, d^2(\rho, \mu_{t-1}) + \theta\, d^2(\rho, \nu_t). \tag{9}$$

We record when this is well posed, give the geometric error-correction form, and note the curvature hypotheses under which the original Hilbert proofs will transfer (Section 8).

4.1 Well-posedness

Proposition 2 (Existence and uniqueness of the GES update)**.** *The update* (9) *has a unique minimizer in either of the following settings.*

1. *$(\mathcal{M}, d)$ is a Hadamard space (complete, nonpositively curved in the sense of Alexandrov). Then $\rho \mapsto (1-\theta)d^2(\rho,\mu_{t-1}) + \theta d^2(\rho,\nu_t)$ is 2-strongly convex along geodesics, so the minimizer exists, is unique, and lies on the geodesic from μ_{t-1} to ν_t [Stu03].*
2. *$\mathcal{M} = \mathcal{W}_2(\mathbb{R}^d)$. Although $\mathcal{W}_2(\mathbb{R}^d)$ is positively curved, the two-measure barycenter with weights $(1-\theta,\theta)$ exists and is unique whenever μ_{t-1} or ν_t is absolutely continuous, and coincides with the displacement interpolant [AC11; LL17].*

The two regimes are complementary: Hadamard geometry covers metric trees, spaces of SPD matrices under the affine-invariant metric, and Hilbert spaces (curvature 0), while case (ii) is the Wasserstein setting of direct interest. Either way the minimizer sits on the geodesic, so GES is genuinely a "fraction-θ" update.

4.2 Geometric error-correction form

On a Riemannian (or Alexandrov) space with exponential and logarithm maps, the barycenter admits an intrinsic error-correction representation. Writing $\log_\mu \nu$ for the initial velocity of the geodesic from μ to ν and $\exp_\mu$ for its inverse,

$$\boxed{\mu_t = \exp_{\mu_{t-1}}\big(\theta \cdot \log_{\mu_{t-1}} \nu_t\big).} \tag{10}$$

This is the exact analogue of the scalar error-correction update $x_t = x_{t-1} + \theta e_t$: the "innovation" is the tangent vector $\log_{\mu_{t-1}} \nu_t$ pointing from the current state toward the observation, and the state moves a fraction θ along it. In $\mathcal{W}_2$ the tangent space at μ is a subspace of $L^2(\mu; \mathbb{R}^d)$ and the maps are

$$\log_\mu \nu = T_{\mu\to\nu} - \mathrm{Id}, \qquad \exp_\mu(v) = (\mathrm{Id} + v)_{\#}\mu, \tag{11}$$

[AGS08; ZKP22]. Substituting (11) into (10) returns $\mu_t = \big((1-\theta)\mathrm{Id} + \theta T_{\mu_{t-1}\to\nu_t}\big)_{\#}\mu_{t-1}$, recovering WES (6) exactly. Thus WES *is* geometric error-correction smoothing in the Wasserstein tangent bundle; the quantile recursion of [Mat+26] is its coordinate expression in $d = 1$.

Remark 2 (Relation to tangent-space autoregression)**.** [ZKP22] model distributional time series by regressing the logarithm $\log_{\bar\mu} \nu_t$ on its past in a single tangent space at the global Fréchet mean $\bar\mu$. GES differs in two ways that mirror the ES-versus-ARIMA distinction: (i) the base point is the *moving* state μ_{t-1}, not a fixed $\bar\mu$, so no global linearization is imposed; and (ii) there is a single scalar θ rather than an operator to estimate. GES is therefore the smoothing counterpart to their autoregression, not a special case of it.

5 Multivariate WES via the Bures–Wasserstein geometry

The barycentric form resolves the multivariate question left open by [Mat+26] at the level of definition: (9) on $\mathcal{M} = \mathcal{W}_2(\mathbb{R}^d)$ is the multivariate WES filter, well posed by Proposition 2(ii). What is lost relative to $d = 1$ is the closed-form quantile isometry; what survives is a closed form on the Gaussian submanifold, which is the workhorse of practice and of our simulations.

5.1 Closed form on Gaussians

Let $\mathcal{N}(m, \Sigma)$ denote a Gaussian on $\mathbb{R}^d$. The set of nondegenerate Gaussians with the $\mathcal{W}_2$ metric is the *Bures–Wasserstein* manifold $\mathrm{BW}(d)$ [BJL19; Che+20]. For $\mu = \mathcal{N}(m_0, \Sigma_0)$ and $\nu = \mathcal{N}(m_1, \Sigma_1)$ the optimal map is affine, $T_{\mu\to\nu}(x) = m_1 + A(x - m_0)$, with

$$A = \Sigma_0^{-1/2}\big(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}\big)^{1/2}\Sigma_0^{-1/2}, \tag{12}$$

and the squared distance is $\mathcal{W}_2^2(\mu,\nu) = \|m_0 - m_1\|^2 + \mathfrak{B}^2(\Sigma_0, \Sigma_1)$, where $\mathfrak{B}^2(\Sigma_0,\Sigma_1) = \mathrm{Tr}\,\Sigma_0 + \mathrm{Tr}\,\Sigma_1 - 2\,\mathrm{Tr}\,\big(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}\big)^{1/2}$.

Proposition 3 (Gaussian WES update)**.** *The θ-barycenter* $\mathrm{bary}_\theta(\mu,\nu)$ *of two Gaussians is the Gaussian* $\mathcal{N}(m_\theta, \Sigma_\theta)$ *with*

$$m_\theta = (1-\theta)m_0 + \theta m_1, \qquad \Sigma_\theta = \big[(1-\theta)I + \theta A\big]\,\Sigma_0\,\big[(1-\theta)I + \theta A\big], \tag{13}$$

with A as in (12). Consequently the multivariate WES filter on Gaussians is the explicit recursion $m_t^\theta = (1-\theta)m_{t-1}^\theta + \theta\, m_{\nu_t}$ and $\Sigma_t^\theta = [(1-\theta)I + \theta A_t]\Sigma_{t-1}^\theta[(1-\theta)I + \theta A_t]$, where A_t is the map (12) from Σ_{t-1}^θ to the observed covariance.

The mean update is ordinary ES; the covariance update is ES performed along the Bures geodesic, not a convex combination of covariances. Proposition 3 is proved in Appendix A.2 and verified numerically in Section 11 (the endpoints $\theta = 0, 1$ return Σ_0, Σ_1, and the midpoint reproduces the Bures geodesic midpoint). For general, non-Gaussian ν_t one replaces A_t by the empirical optimal map or, at scale, by an entropically regularized or sliced surrogate (Section 7); the barycenter itself can be computed by the fixed-point iteration of [Álv+16] or the methods of [Che+20; Alt+21].

5.2 What theory survives in $\mathbb{R}^d$

The favourable structure of the scalar case rested on the linear isometry $\mu \mapsto U$. In $\mathbb{R}^d$ there is no such global linearization, and two specific obstructions appear:

1. *Curvature.* $\mathcal{W}_2(\mathbb{R}^d)$ has nonnegative curvature, so the maps $\log_\mu$ are only locally isometric and the residual $\log_{\mu_{t-1}} \nu_t$ no longer adds linearly across time.
2. *Map composition.* Unlike scalar quantile increments, successive optimal maps need not commute, so the filter is not a fixed linear recursion in any single tangent space.

We make the positive statement precise in Section 8 (Theorem 3): under a *bounded-geometry* assumption that keeps the iterates in a region bi-Lipschitz to a Hilbert space, Fréchet-mean stationarity and consistency of $\hat{\theta}_T$ persist, with constants degraded by the bi-Lipschitz modulus. The assumption is not vacuous—Section 11 exhibits a regime (random location shifts with frozen shape) that satisfies it and one (compounding multiplicative covariance noise) that violates it by inducing a divergent random walk on $\mathrm{BW}(d)$.

6 Smoothing the laws of stochastic processes

We now take up the generalization to *stochastic processes* in the strong sense: each observation ν_t is the law of a discrete-time process $X = (X_1, \dots, X_N)$ on $\mathbb{R}^N$, i.e. a point in $\mathcal{P}_2(\mathbb{R}^N)$, and we wish to smooth a *sequence of such laws*. The naive route is to apply multivariate WES (Section 5) on $\mathcal{W}_2(\mathbb{R}^N)$. We argue this is geometrically wrong, exhibit the failure, and replace $\mathcal{W}_2$ by the adapted (bicausal) Wasserstein distance.

6.1 Why plain Wasserstein is the wrong geometry on path space

A law on $\mathbb{R}^N$ carries a filtration $\mathcal{F} = (\mathcal{F}_n)_{n=1}^N$, $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$. Decision functionals—optimal stopping values, superhedging prices, dynamic risk measures—depend on the *joint law together with this information flow*, and are notoriously discontinuous under $\mathcal{W}_2$, which sees only the joint law [Bac+20a; Bac+22]. The following minimal instance makes this quantitative.

Proposition 4 ($\mathcal{W}_2$ is blind to predictability)**.** *There exist laws $\mu, \nu^\varepsilon \in \mathcal{P}_2(\mathbb{R}^2)$ and a 1-Lipschitz adapted functional Φ such that*

$$\mathcal{W}_2(\mu, \nu^\varepsilon) \xrightarrow[\varepsilon \to 0]{} 0, \qquad \textit{yet} \qquad \Phi(\mu) = 0 \quad \textit{and} \quad \Phi(\nu^\varepsilon) = 1 \quad \textit{for all } \varepsilon > 0.$$

Construction. Let μ be the law of (X_1, X_2) with $X_1 = 0$ and $X_2 = \pm 1$ each with probability $\frac{1}{2}$, independent of the (degenerate) first coordinate. Let ν^ε be the law of (X_1', X_2') with $X_1' = \pm\varepsilon$ each w.p. $\frac{1}{2}$ and $X_2' = \mathrm{sign}(X_1')$. Then $\mathcal{W}_2^2(\mu, \nu^\varepsilon) \le \varepsilon^2 \to 0$ (couple by matching signs of the second coordinate). Consider the gambler who, having observed $\mathcal{F}_1$, predicts the sign of X_2: $\Phi(\cdot) = \sup_g \mathbb{E}[\,\mathrm{sign}(X_2)\, g(X_1)\,]$ over 1-Lipschitz g with $g(0) = 0$. Under μ, $X_1 \equiv 0$ carries no

information, so $\Phi(\mu) = 0$; under ν^ε, X_1' reveals X_2' exactly, so $\Phi(\nu^\varepsilon) = 1$. Full detail is given in Appendix A.3. □

The moral is standard in the adapted-transport literature [Bac+20b]: any smoother built on $\mathcal{W}_2(\mathbb{R}^N)$ can map a sequence of *predictable* observations to a smoothed state that is *unpredictable*, corrupting exactly the conditional structure that downstream decisions use.

6.2 The adapted Wasserstein distance

The remedy is to optimize over couplings that respect the filtration. A coupling π of μ, ν is *bicausal* if, informally, its conditional structure does not let the past of one marginal peek at the future of the other; formally, for each n the conditional law of the ν-future given the μ-past factors through the μ-information up to n, and symmetrically [Las18; Bac+20b]. The *adapted (bicausal) Wasserstein distance* is

$$\mathcal{AW}_2^2(\mu, \nu) = \inf_{\pi \in \Pi_{\mathrm{bc}}(\mu,\nu)} \int \|x - y\|^2 \, \pi(dx\, dy), \tag{14}$$

the infimum over bicausal couplings $\Pi_{\mathrm{bc}} \subseteq \Pi$. It is also the *nested distance* of [PP12]: for laws on $\mathbb{R}^N$ it is computed by backward induction,

$$\mathcal{AW}_2^2 \text{ solves } \quad V_N \equiv 0, \qquad V_n(x_{1:n}, y_{1:n}) = \inf_{\pi_n \in \Pi(\mu_{n+1|x_{1:n}}, \nu_{n+1|y_{1:n}})} \int \left[\, \|x_{n+1} - y_{n+1}\|^2 + V_{n+1} \right] d\pi_n, \tag{15}$$

with $\mathcal{AW}_2^2(\mu, \nu) = \mathbb{E}_{\pi_1}[\|x_1 - y_1\|^2 + V_1]$ [PP14; EP24]. Crucially $\mathcal{AW}_2 \geq \mathcal{W}_2$, and value functionals like Φ above are $\mathcal{AW}_2$-Lipschitz even when they are $\mathcal{W}_2$-discontinuous [Bac+20a; ABZ20].

6.3 Adapted WES

We can now lift WES to path-law sequences by repeating the barycentric construction in the adapted geometry. The enabling fact is recent.

Theorem 1 ([BBP25], paraphrased)**.** *The space of stochastic-process laws equipped with $\mathcal{AW}_2$ is a geodesic Polish space. It is isometric to a subset of a classical Wasserstein space over filtered "prediction-process" representatives, and it supports barycenters: for weights summing to one the weighted Fréchet-mean problem in $\mathcal{AW}_2$ has a solution, attained along adapted geodesics.*

Definition 2 (Adapted WES)**.** Given path-law observations $\nu_1, \ldots, \nu_T \in \mathcal{P}_2(\mathbb{R}^N)$ and $\theta \in (0, 1)$, the adapted WES filter is

$$\mu_t^\theta = \mathrm{bary}_\theta^{\mathcal{AW}}(\mu_{t-1}^\theta, \nu_t) := \arg\min_\rho \, (1 - \theta)\, \mathcal{AW}_2^2(\rho, \mu_{t-1}^\theta) + \theta\, \mathcal{AW}_2^2(\rho, \nu_t), \tag{16}$$

the θ-weighted adapted barycenter, which exists by Theorem 1.

By the isometry of Theorem 1, the adapted barycenter is, in the prediction-process picture, an *ordinary* Wasserstein barycenter; computationally it is obtained by combining the backward recursion (15) with a barycentric step at each node, which we state as Algorithm 1. The per-step structure is the adapted analogue of (10): move the path law a fraction θ along the adapted geodesic toward the observation.

The correctness of Algorithm 1—that its node-wise quantile recursion really returns the adapted barycenter of Definition 2—is the second result we prove in full.

Algorithm 1 Adapted WES update $\mu_t^\theta = \mathrm{bary}_\theta^{\mathcal{AW}}(\mu_{t-1}^\theta, \nu_t)$ on a finite tree

1: **Input:** filtered laws μ_{t-1}^θ, ν_t as scenario trees of depth N; weight θ.
2: Compute the optimal bicausal coupling $\pi^\star$ of $(\mu_{t-1}^\theta, \nu_t)$ by backward induction (15).
3: **for** each tree node $x_{1:n}$ (root to leaves) **do**
4: along the matched conditional kernels, set the new conditional one-step law to the W_2-barycenter with weights $(1-\theta, \theta)$ of the two matched kernels (closed form in $\mathbb{R}$ via quantiles; (13) if Gaussian).
5: **end for**
6: **Return** the path law μ_t^θ assembled from the updated conditional kernels.

Proposition 5 (Algorithm 1 computes the adapted barycenter)**.** *Let μ, ν be laws of $\mathbb{R}$-valued processes on a finite filtered tree of depth N, with μ having conditional one-step kernels that are atomless. Let $\pi^\star$ be an optimal bicausal coupling attaining $\mathcal{AW}_2(\mu, \nu)$ and, for each pair of matched histories $(x_{1:n}, y_{1:n})$ under $\pi^\star$, let $\beta_\theta(x_{1:n}, y_{1:n})$ be the θ-weighted $W_2(\mathbb{R})$ barycenter of the matched one-step kernels $\mu_{n+1|x_{1:n}}$ and $\nu_{n+1|y_{1:n}}$. Then the path law μ_t^θ assembled from the kernels β_θ is the unique adapted barycenter,*

$$\mu_t^\theta \;=\; \mathrm{bary}_\theta^{\mathcal{AW}}(\mu, \nu) \;=\; \arg\min_{\rho} \; (1-\theta)\, \mathcal{AW}_2^2(\rho, \mu) + \theta\, \mathcal{AW}_2^2(\rho, \nu).$$

The proof (Appendix A.8) uses the prediction-process isometry of Theorem 1 to turn the adapted barycenter into an ordinary W_2 barycenter, then disintegrates that barycenter along the filtration using the dynamic-programming structure of bicausal couplings [Bac+20c; PP12], reducing it node-wise to the scalar barycenter of Proposition 1.

Example 1 (Gaussian AR(1) laws)**.** Let each ν_t be the law of a stationary Gaussian AR(1), $X_{n+1} = \phi_t X_n + \xi_n$, $\xi_n \sim \mathcal{N}(0, s_t^2)$. Bicausal couplings of Gaussian processes are themselves Gaussian and the nested distance is available in closed form through the conditional one-step variances [PP14]. The adapted WES update then smooths the *sequence of conditional dynamics*: writing each law in innovation coordinates (ϕ_t, s_t^2), the update is the Bures–Wasserstein ES (13) applied node-wise, so that the smoothed process has conditional variance interpolated along the Bures geodesic rather than linearly. This is exactly the behaviour one wants for forecasting laws of volatility-like processes, where linear averaging of conditional variances is known to distort persistence.

Remark 3 (Two readings, one framework)**.** Definition 2 and multivariate WES (Section 5) are both instances of (9), differing only in the metric ($\mathcal{AW}_2$ versus $\mathcal{W}_2$). The adapted version should be used whenever the index n within each observation is a genuine time axis carrying decisions; the plain version suffices when $\mathbb{R}^N$ is an unordered feature space. Mistaking one for the other is precisely the error quantified by Proposition 4.

7 General forms of smoothing: a divergence view

WES replaces a Euclidean segment by a Wasserstein geodesic. Replacing instead the *loss* used to define the barycenter yields a broad family of distributional smoothers, of which WES, geometric pooling, and mixture smoothing are special cases.

7.1 Proximal smoothers

For a divergence $D(\cdot\|\cdot)$ on $\mathcal{P}_2(\mathcal{X})$ define

$$\mu_t = \arg\min_{\rho} \; (1-\theta)\, D(\rho\|\mu_{t-1}) + \theta\, D(\rho\|\nu_t). \tag{17}$$

Choosing $D = \mathcal{W}_2^2$ recovers GES (9) and hence WES. Other choices give qualitatively distinct behaviour, organized by a single dichotomy: *information-geometric* divergences mix, while *transport* divergences morph.

Reverse KL: geometric (log-linear) pooling. With $D(\rho\|p) = \mathrm{KL}(\rho\|p)$ the minimizer of (17) is the normalized weighted geometric mean,

$$\mu_t \;\propto\; \mu_{t-1}^{1-\theta}\, \nu_t^{\theta}, \tag{18}$$

the log-linear opinion pool of [GZ86]. For an exponential family with natural parameter η, (18) is ordinary ES *on the natural parameters*, $\eta_t = (1-\theta)\eta_{t-1} + \theta\,\eta_{\nu_t}$: WES-style smoothing of Gaussians by their information-form parameters rather than by their Bures geometry. The proof is in Appendix A.4.

Forward KL and MMD: mixture smoothing. With $D(\rho\|p) = \mathrm{KL}(p\|\rho)$ or $D = \mathrm{MMD}^2$ (a squared kernel distance, [Gre+12]) the objective is minimized by the arithmetic mixture

$$\mu_t = (1-\theta)\mu_{t-1} + \theta\,\nu_t, \tag{19}$$

i.e. ES on the measures themselves. This is the "mix" extreme: modes are reweighted, never transported. For MMD the identity is exact because $\mathrm{MMD}^2(\rho, p) = \|\kappa_\rho - \kappa_p\|_{\mathcal{H}}^2$ is the squared distance of the kernel mean embeddings κ, so (17) is a Hilbert barycenter (Section 8).

Entropic and sliced transport. For computational scalability in $\mathbb{R}^d$ one replaces $\mathcal{W}_2^2$ by the Sinkhorn divergence [Cut13; GPC18; Fey+19] or the sliced-Wasserstein distance [Bon+15]. The debiased Sinkhorn divergence interpolates between $\mathcal{W}_2^2$ (no regularization) and MMD^2 (heavy regularization), so (17) smoothly interpolates between the morph (13) and the mix (19); the regularization level is a second, optional knob beyond θ. Sliced WES averages one-dimensional WES updates over random projections and inherits the scalar closed form per slice.

7.2 Trend and seasonality with a monotonicity caveat

Holt's linear-trend and Holt–Winters seasonal recursions lift to quantile space by smoothing level and slope functions $L_t, S_t \in L^2((0,1))$ jointly. A subtlety absent in the level-only case arises: a level-plus-trend quantile forecast $L_t + hS_t$ need not be nondecreasing, hence need not be a valid quantile function. The fix is an isotonic projection, and it is free of charge in the relevant sense.

Lemma 1 (Isotonic projection is nonexpansive and forecast-improving)**.** *Let $\mathcal{C} \subset L^2((0,1))$ be the closed convex cone of nondecreasing functions and $P_{\mathcal{C}}$ the L^2 projection onto it. For any $f \in L^2((0,1))$ and any quantile function $V \in \mathcal{C}$,*

$$\|P_{\mathcal{C}} f - V\|_{L^2} \le \|f - V\|_{L^2}\,.$$

Hence replacing a (possibly non-monotone) quantile forecast f by its monotone projection $P_{\mathcal{C}} f$ never increases the one-step Wasserstein forecast error against any genuine distribution, and returns a valid distribution.

The proof (Appendix A.5) is the standard nonexpansiveness of projection onto a convex set together with $V \in \mathcal{C}$. Operationally, $P_{\mathcal{C}}$ is the pool-adjacent-violators algorithm; trend/seasonal Wasserstein smoothing is thus "smooth in quantile space, then project," with a guarantee that the projection cannot hurt.

8 A unifying theory: the Hilbert meta-theorem and curved corrections

We now explain, in a single stroke, which of the guarantees of [Mat+26] are properties of *exponential smoothing in a Hilbert space* and which require new analysis because the underlying space is curved. The dividing line is sharp and organizes all the constructions above.

8.1 The Hilbert meta-theorem

Inspect the proofs in [Mat+26]. The objects that actually enter are: a Hilbert space $\mathcal{H}$ (there, $L^2((0,1))$); an isometric embedding $\iota : \mathcal{M} \hookrightarrow \mathcal{H}$ of the data space ($\mu \mapsto U$); the fact that the update is the *affine* recursion $\iota(\mu_t) = (1-\theta)\iota(\mu_{t-1}) + \theta\,\iota(\nu_t)$; and second-moment bounds on the residuals $F_t = \iota(\nu_t) - \iota(\mu_{t-1})$. No other property of $\mathbb{R}$ is used. We abstract this into a hypothesis.

Assumption 1 (Hilbert-linear smoother)**.** There is a separable Hilbert space $\mathcal{H}$ and a map $\iota : \mathcal{M} \to \mathcal{H}$ such that (i) the dissimilarity is the pullback norm, $d^2(\mu,\nu) = \|\iota(\mu) - \iota(\nu)\|_{\mathcal{H}}^2$; (ii) the smoother update satisfies $\iota(\mu_t^\theta) = (1-\theta)\iota(\mu_{t-1}^\theta) + \theta\,\iota(\nu_t)$; and (iii) the data are generated by $\iota(\nu_t) = \iota(\mu_{t-1}) + F_t$ with $\{F_t\}$ i.i.d., mean zero in $\mathcal{H}$, and $\mathbb{E}\,\|F_t\|_{\mathcal{H}}^2 < \infty$.

Theorem 2 (Meta-theorem: verbatim transfer)**.** *Under Assumption 1, the conclusions of [Mat+26] hold for the smoother on $\mathcal{M}$:*

1. (Fréchet-mean stationarity) *the data process is Fréchet-mean stationary, with Fréchet mean $\iota^{-1}\mathbb{E}\iota(\nu_t)$ constant in t;*
2. (Exponential autocovariance decay) *the residual autocovariance*

$$\mathbb{E}\left\langle \iota(\mu_t^\theta) - \iota(\nu_{t+1}),\, \iota(\mu_s^\theta) - \iota(\nu_{s+1}) \right\rangle_{\mathcal{H}}$$

   decays geometrically in $|t-s|$ at rate $(1-\theta)$;
3. (Consistency) *the minimum-dissimilarity estimator $\hat\theta_T = \arg\min_\theta \frac{1}{T}\sum_t d^2(\mu_t^\theta, \nu_{t+1})$ satisfies $\hat\theta_T \xrightarrow{p} \theta_*$ as $T \to \infty$.*

The proof (Appendix A.6) is a transcription: every step of the original argument is an identity or inequality among inner products and norms in $\mathcal{H}$, so it holds in any $\mathcal{H}$. The consistency argument is the standard M-estimation route [NM94; Vaa98]: the population criterion $\theta \mapsto \mathbb{E}\,d^2(\mu_t^\theta, \nu_{t+1})$ is a strictly convex quadratic in θ minimized at θ_*, the empirical criterion converges uniformly by a Hilbert-valued law of large numbers under (iii), and the argmin is continuous.

Corollary 1 (Instances)**.** *Theorem 2 applies verbatim, with the indicated embedding, to:*

1. ***Scalar WES** [Mat+26]: $\mathcal{M} = \mathcal{W}_2(\mathbb{R})$, $\mathcal{H} = L^2((0,1))$, $\iota(\mu) = U_\mu$ (recovering the original theorems);*
2. ***MMD smoothing*** (19) *with $D = \mathrm{MMD}^2$: $\mathcal{H}$ the RKHS, $\iota(\mu) = \kappa_\mu$ the kernel mean embedding [Gre+12];*

3. ***Exponential-family natural-parameter smoothing*** (18)*: $\mathcal{H} = \mathbb{R}^k$ (or the natural-parameter Hilbert space), $\iota(\mu) = \eta_\mu$, with d^2 the induced quadratic.*

Thus a single structural hypothesis explains why three superficially different smoothers obey the same laws, and identifies the embedding as the only thing one must supply.

8.2 Curved cases: what genuinely needs new work

The cases of real novelty—multivariate WES on $\mathcal{W}_2(\mathbb{R}^d)$ and adapted WES on $\mathcal{AW}_2$—violate Assumption 1(ii): there is no *global* isometric embedding under which the update is affine, because the spaces are curved. The update (10) is affine only in the moving tangent space $T_{\mu_{t-1}}\mathcal{M}$, and parallel transport between tangent spaces is nontrivial. We give the positive result under a quantitative flatness hypothesis. The hypothesis must do two jobs at once: control distances (so that the metric loss is comparable to a Hilbert loss) and control how far the barycenter operator departs from the affine recursion that drives the flat case. We package both into a single chart with two moduli.

Assumption 2 (Bounded geometry)**.** There is a geodesically convex set $K \subseteq \mathcal{M}$ containing all iterates $\{\mu_t^\theta, \nu_t\}_{t\geq 0}$ and the relevant Fréchet means, a separable Hilbert space $\mathcal{H}$, and a chart $\iota : K \to \mathcal{H}$ with the following two properties.

(B1) *(Bi-Lipschitz)* for some $L \geq 1$ and all $\mu, \nu \in K$,

$$L^{-1}\,\|\iota(\mu) - \iota(\nu)\|_{\mathcal{H}} \leq d(\mu,\nu) \leq L\,\|\iota(\mu) - \iota(\nu)\|_{\mathcal{H}}\,;$$

(B2) *(Almost-affine barycenter)* for some flatness modulus $\eta \geq 0$, every fixed $\nu \in K$, every $\theta \in (0,1)$, and all $\mu, \mu' \in K$, the operator $G_\nu^\theta := \mathrm{bary}_\theta(\cdot, \nu)$ obeys

$$\big\|\,\iota G_\nu^\theta(\mu) - \iota G_\nu^\theta(\mu') - (1-\theta)\big(\iota(\mu) - \iota(\mu')\big)\big\|_{\mathcal{H}} \;\leq\; \eta\,(1-\theta)\,\big\|\iota(\mu) - \iota(\mu')\big\|_{\mathcal{H}}\,.$$

In addition the chart residuals $F_t := \iota(\nu_t) - \iota(\mu_{t-1})$ are i.i.d., mean zero in $\mathcal{H}$, with $\mathbb{E}\,\|F_t\|_{\mathcal{H}}^2 < \infty$, and the contraction factor $\lambda := (1-\theta)(1+\eta)$ satisfies $\lambda < 1$.

Condition (B2) measures the failure of ι to send geodesics to line segments: $\eta = 0$ exactly when the barycenter operator is the affine ES recursion in the chart (the flat/Hilbert case of Theorem 2), and η grows with curvature over K. It is verifiable: on the Bures–Wasserstein location submanifold (means vary, shape frozen) the update is affine in the mean chart, so $\eta = 0$; over a Bures ball of bounded eccentricity, $\Sigma \mapsto \Sigma^{1/2}$ is bi-Lipschitz with an L controlled by the eccentricity and $\eta = O(\operatorname{diam} K)$.

Theorem 3 (Curved transfer)**.** *Under Assumption 2, with $\lambda = (1-\theta)(1+\eta) < 1$, multivariate WES and adapted WES satisfy:*

1. (Stationarity) *the chart-image observation process $\{\iota(\nu_t)\}$ is mean-stationary, $\mathbb{E}\,\iota(\nu_t) \equiv \iota(\mu_0)$, and the metric Fréchet mean of ν_t is stationary up to an error $O(\eta\,L^2)$ that vanishes in the flat limit $\eta \to 0$;*
2. (Geometric mixing) *the filter residual autocovariance obeys, for a finite constant C,*

$$\Big|\mathbb{E}\,\Big\langle \iota(\mu_t^\theta) - \iota(\nu_{t+1}),\, \iota(\mu_s^\theta) - \iota(\nu_{s+1})\Big\rangle_{\mathcal{H}}\,\Big| \leq C\,\lambda^{|t-s|};$$

3. (Consistency) *the minimum-Wasserstein estimator* $\hat{\theta}_T$ *of* (7) *is consistent,* $\hat{\theta}_T \xrightarrow{p} \theta_*$.

As $\eta \to 0$ *and* $L \to 1$ *the rate* $\lambda \to (1-\theta)$ *and all constants converge to the Hilbert values of Theorem 2.*

The full proof is in Appendix A.7: condition (B2) makes the chart filter a genuine λ-contraction, which replaces the exact affine recursion of the flat case; the bi-Lipschitz bound (B1) transfers the resulting Hilbert estimates back to the metric loss; and strong convexity of the barycentric objective (Proposition 2) yields a strictly convex population criterion, giving consistency by the standard argmin argument. The Gaussian negative control of Section 11 fails (B1)–(B2): its covariance leaves every bounded Bures ball, so no finite L exists and $\lambda \to 1$, exactly the regime the theorem excludes.

Remark 4 (Honest scope)**.** Theorem 3 is a transfer-with-constants result, not a free lunch: it does not claim the curved estimator attains the scalar efficiency, only that consistency and geometric mixing survive on a region of bounded distortion, degrading gracefully as η, L grow. Establishing a central limit theorem with explicit curved asymptotic variance, and characterizing the largest admissible K for concrete generative models, are left open.

9 Resolving the open problems

Section 12 listed four open directions. We now resolve each with a theorem: a central limit theorem for the smoothing parameter; a complete stability and validity theory for damped-trend and seasonal smoothers; an exact Kalman representation together with a provably convergent online estimator; and a Koopman operator-theoretic representation. Throughout we work in the Hilbert-linear setting of Assumption 1, lifting to the curved case via the bounded-geometry chart of Assumption 2 where indicated.

9.1 A central limit theorem for $\hat{\theta}_T$

Consistency (Theorems 2, 3) leaves open the limiting law. The key observation is that, at the truth, the forecast error equals the innovation and the estimating-equation score is a martingale difference, so the limit is Gaussian with an explicitly computable variance.

Theorem 4 (CLT for the minimum-Wasserstein estimator)**.** *Under Assumption 1 with* $\theta_* \in (0,1)$*, i.i.d. innovations* F_t *of mean zero, covariance operator* Σ_F*, and* $\mathbb{E}\,\|F_t\|^4 < \infty$*, let* $D := \sum_{k\geq 0}(1-\theta_*)^k F_{-k}$ *be the stationary sensitivity process. Then*

$$\sqrt{T}\,(\hat{\theta}_T - \theta_*) \;\xrightarrow{d}\; \mathcal{N}\big(0,\; V\big), \qquad V = \frac{\mathbb{E}\,\langle D, \Sigma_F D\rangle}{\big(\mathbb{E}\,\|D\|^2\big)^2}.$$

If $\Sigma_F = \sigma^2 I_d$ *is isotropic on* $\mathbb{R}^d$*, then* $V = \theta_*(2-\theta_*)/d$*; in the scalar case* $V = \theta_*(2-\theta_*)$*. Under bounded geometry (Assumption 2) the same limit holds with* V *replaced by* $V\,(1+O(\eta))$*.*

Proof idea. Write the per-step loss $\ell_t(\theta) = \big\|u_t^\theta - w_{t+1}\big\|^2$ with $u_t^\theta = \iota(\mu_t^\theta)$. Its score $s_t(\theta_*) = 2\left\langle D_t, u_t^{\theta_*} - w_{t+1}\right\rangle = -2\,\langle D_t, F_{t+1}\rangle$, where $D_t = \partial_\theta u_t^\theta|_{\theta_*}$ obeys $D_t = (1-\theta_*)D_{t-1} + F_t$. Since $\mathbb{E}[F_{t+1} \mid \mathcal{G}_t] = 0$ and D_t is $\mathcal{G}_t$-measurable, $\{s_t(\theta_*)\}$ is a stationary ergodic martingale-difference sequence; the martingale CLT [Bil61] gives $T^{-1/2}\sum_t s_t(\theta_*) \xrightarrow{d} \mathcal{N}(0,\Sigma)$ with $\Sigma = \mathbb{E}s_t^2 = 4\,\mathbb{E}\,\langle D, \Sigma_F D\rangle$. The Hessian $\partial_\theta^2 \frac{1}{T}\sum_t \ell_t \to H = 2\,\mathbb{E}\,\|D\|^2$ in probability (Theorem 2 mixing), and a Taylor expansion of the first-order condition gives $\sqrt{T}(\hat{\theta}_T - \theta_*) = -H^{-1}T^{-1/2}\sum_t s_t(\theta_*) + o_p(1)$, whence $V =$

Table 1: Monte Carlo check of Theorem 4: scaled variance $T \cdot \widehat{\mathrm{Var}}(\hat{\theta}_T)$ versus the predicted $V = \theta_*(2-\theta_*)/d$, over $R = 500$ replications, $T = 2500$.

d	θ_*	$T \cdot \widehat{\mathrm{Var}}$	predicted V	ratio
1	0.3	0.514	0.510	1.01
1	0.5	0.756	0.750	1.01
1	0.7	0.893	0.910	0.98
2	0.3	0.247	0.255	0.97
2	0.5	0.383	0.375	1.02
2	0.7	0.487	0.455	1.07

$\Sigma/H^2 = \mathbb{E}\langle D, \Sigma_F D\rangle / (\mathbb{E}\|D\|^2)^2$ by [NM94, Thm. 3.4]. The isotropic reduction uses $\mathbb{E}\|D\|^2 = \mathrm{Tr}\,\Sigma_F/(\theta_*(2-\theta_*))$ and $\Sigma = 4\sigma^2\mathbb{E}\|D\|^2$. The curved correction is the $O(\eta)$ distortion of $\{u_t^\theta\}$ from the affine recursion (Appendix B.1). □

The scalar formula $V = \theta_*(2-\theta_*)$ is striking: the asymptotic standard error is largest near $\theta_* = 1$ (pure tracking) and shrinks as $\theta_* \to 0$, and the multivariate rate improves by the ambient dimension. Table 1 confirms the formula in a Monte Carlo study (Appendix A.9); this also recovers, for the location model, the i.i.d. Bures–Wasserstein barycenter CLT of [KSS21] specialized to a two-point support, and complements the empirical-barycenter rates of [Le +23].

9.2 Damped-trend and seasonal smoothers

We lift damped-trend (Holt) and seasonal (Holt–Winters) smoothing to $\mathcal{W}_2$ in quantile coordinates and settle well-posedness, stability, and consistency. Work in the chart $\mathcal{H} = L^2((0,1))$ with level ℓ_t and trend c_t, damping $\phi \in [0,1]$, and the error-correction recursion driven by the one-step innovation $e_t = w_t - (\ell_{t-1} + \phi c_{t-1})$:

$$\ell_t = \ell_{t-1} + \phi c_{t-1} + \theta e_t, \qquad c_t = \phi c_{t-1} + \beta e_t, \qquad \hat{w}_{t+h} = \ell_t + \Big(\textstyle\sum_{j=1}^{h} \phi^j\Big) c_t. \tag{20}$$

Theorem 5 (Damped-trend Wasserstein smoothing)**.** *Let (20) run in $L^2((0,1))$ and let $P_{\mathcal{C}}$ be the isotonic projection of Lemma 1. Then:*

1. (Validity) *the projected forecast $P_{\mathcal{C}}\hat{w}_{t+h}$ is a valid quantile function, and its h-step $\mathcal{W}_2$ forecast error against any genuine distribution does not exceed that of the raw forecast $\hat{w}_{t+h}$;*
2. (Stability) *the level–trend filter forgets initial conditions geometrically iff the companion matrix $M = \left(\begin{smallmatrix} 1-\theta & (1-\theta)\phi \\ -\beta & \phi(1-\beta) \end{smallmatrix}\right)$ is Schur stable, i.e. $\det M = (1-\theta)\phi < 1$ and $|\mathrm{tr}\, M| = |(1-\theta) + \phi(1-\beta)| < 1 + (1-\theta)\phi$; in particular this holds for all $0 < \theta < 1$, $0 \le \phi \le 1$, $0 \le \beta \le \theta$;*
3. (Consistency) *on the stability region and under an identifiability condition, the joint minimum-Wasserstein estimator $(\hat{\theta}_T, \hat{\beta}_T, \hat{\phi}_T)$ is consistent for $(\theta_*, \beta_*, \phi_*)$.*

Part (1) extends the free-lunch property of isotonic projection to the trend forecast; part (2) gives the exact stability triangle (the Wasserstein analogue of the classical admissible region for damped trend), proved by Schur–Cohn analysis of the error-propagation matrix; part (3) is the multiparameter version of the meta-theorem. The proof is in Appendix B.2. By the prediction-process isometry (Theorem 1) the seasonal and adapted versions follow with $P_{\mathcal{C}}$ replaced by projection onto the monotone cone in the prediction representation.

9.3 Adaptive and time-varying smoothing

We give two results: an exact identification of WES as a steady-state filter, and a provably convergent online estimator that also tracks a drifting θ_t^*.

Theorem 6 (Kalman/Wiener representation)**.** *Consider the lifted local-level model in $\mathcal{H}$: $\xi_t = \xi_{t-1}+\zeta_t$, $w_t = \xi_t+\varepsilon_t$, with ζ_t, ε_t independent white noises of covariances $q\,\Sigma_\varepsilon$ and Σ_ε ($q > 0$ the signal-to-noise ratio per mode). The steady-state Kalman filter is the WES recursion $\hat{\xi}_t = (1-\theta)\hat{\xi}_{t-1} + \theta w_t$ with gain*

$$\theta = \frac{-q + \sqrt{q^2+4q}}{2} \in (0,1), \qquad \textit{equivalently} \qquad q = \frac{\theta^2}{1-\theta}.$$

Thus WES is the minimum-MSE one-step predictor of the local-level model, and $\theta \leftrightarrow q$ is a strictly increasing bijection.

Theorem 7 (Online estimation and tracking)**.** *Let $g_t = 2\left\langle D_t,\, u_{t-1}^{\theta_{t-1}} - w_t \right\rangle$ be the online score with D_t updated recursively, and define $\theta_t = \Pi_{[\epsilon,1-\epsilon]}\big(\theta_{t-1} - a_t\, g_t\big)$ with step sizes $a_t > 0$, $\sum_t a_t = \infty$, $\sum_t a_t^2 < \infty$. Under Assumption 1 and the strong convexity of Q, $\theta_t \to \theta_*$ almost surely. If instead the truth drifts slowly, $\left|\theta_t^* - \theta_{t-1}^*\right| \le \delta$, then with constant step $a_t \equiv a$, $\limsup_t \mathbb{E}\,|\theta_t - \theta_t^*|^2 = O\big(a + \delta^2/a\big)$.*

Theorem 6 (proof in Appendix B.3) recovers the classical local-level/ES equivalence [Mut60; Har89; Hyn+08] in the distributional setting and pins θ to an interpretable signal-to-noise ratio, so a time-varying θ_t is obtained by tracking q_t. Theorem 7 (proof in Appendix B.4) is a Robbins–Monro stochastic-approximation statement [RM51; KY03]: the mean field is $-Q'(\theta)$, strongly convex with unique zero θ_*, and the noise is a square-integrable martingale difference, giving almost-sure convergence and the standard variance/drift trade-off in the tracking regime.

9.4 A Koopman operator representation

Finally we place WES inside operator-theoretic forecasting. Lift the dynamics to $\mathcal{H}$ by ι and let L denote the lag operator on observation histories, $(Lw)_t = w_{t-1}$.

Theorem 8 (Koopman/resolvent form of WES)**.** *Under the local-level model of Theorem 6:*

1. *the minimum-MSE one-step predictor in $\mathcal{H}$ is the conditional expectation $\mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t]$ and equals the WES predictor $\hat{\xi}_t$;*
2. *as an operator on histories, the predictor is the resolvent/Neumann series*

$$\mathcal{K}_\theta = \theta\,\big(I - (1-\theta)L\big)^{-1} = \theta \sum_{k\ge 0} (1-\theta)^k L^k,$$

   a bounded linear operator with single pole at $1-\theta$;
3. *consequently the innovation dynamics have Koopman spectrum $\{1-\theta\}$, and extended dynamic mode decomposition on $\{w_t\}$ recovers $1-\theta$ as the leading eigenvalue, yielding a spectral estimator of θ asymptotically equivalent to $\hat{\theta}_T$.*

The proof (Appendix B.5) is the geometric-series identity together with the convergence of EDMD to the Koopman operator [Koo31; Mez05; KM18]. Part (3) connects WES to the distributional Koopman/operator forecasting program [WA25]: WES is the rank-one spectral truncation whose single mode $1-\theta$ is the exponential memory of the smoother, so estimating θ and estimating the dominant Koopman eigenvalue are the same task.

10 Pre-Finsler and Finsler generalizations

The geometric error-correction form (10) used only geodesics and the exponential map, not the quadratic (Riemannian) structure of $\mathcal{W}_2$. This suggests that GES extends to Finsler geometry. The motivation is concrete: the p-Wasserstein space $\mathcal{W}_p$ for $p \neq 2$ is genuinely Finsler, not Riemannian [Agu12], and asymmetric transport costs produce non-reversible (pre-Finsler) structures. We delineate exactly which results survive, and we find that the abstract flatness modulus η of Assumption 2 is, concretely, Finsler non-quadraticity.

10.1 Definitions

A *pre-Finsler structure* on a manifold M is a continuous $F : TM \to [0, \infty)$ that is positively homogeneous, $F(x, \lambda v) = \lambda F(x, v)$ for $\lambda > 0$, positive off the zero section, and has convex unit balls $\{v : F(x, v) \leq 1\}$. It induces a (possibly asymmetric) length distance d_F. A *Finsler* structure additionally has F smooth on $TM \setminus 0$ with fiberwise strongly convex fundamental tensor $g_{ij}(v) = \frac{1}{2}\partial^2_{v^i v^j} F^2$; it is *reversible* if $F(x, -v) = F(x, v)$. The Cartan tensor $A_{ijk} = \frac{F}{2}\partial_{v^k} g_{ij}$ measures the deviation from Riemannian geometry: F is Riemannian iff $A \equiv 0$ [BCS00]. A *Berwald space* is a Finsler manifold whose Chern connection coefficients are independent of direction, so geodesics and parallel transport follow a single affine connection; the analogue of sectional curvature is the *flag curvature* $K(x; y, v)$. We also use the 2-uniform convexity (C) and smoothness (S) moduli of [Oht09]: a Banach space is 2-uniformly convex/smooth if, with J the duality map,

$$\|x\|^2 + 2\,\langle Jx, y\rangle + C^{-2}\,\|y\|^2 \;\leq\; \|x + y\|^2 \;\leq\; \|x\|^2 + 2\,\langle Jx, y\rangle + S^2\,\|y\|^2 . \tag{21}$$

By [BCL94], L^p has $C = 1$ for $2 \leq p < \infty$ and $S = \sqrt{p-1}$, with $C = S = 1$ iff $p = 2$ (the Hilbert case).

By Otto's formal calculus $\mathcal{W}_2$ is Riemannian, while [Agu12] show $\mathcal{W}_p$ carries a Finsler metric whose induced distance is $\mathcal{W}_p$; over a Finsler base manifold $\mathcal{W}_p$ is moreover asymmetric [OS09]. GES on a (pre-)Finsler space is defined exactly as before by the geodesic-interpolation update $\mu_t = \exp_{\mu_{t-1}}(\theta \log_{\mu_{t-1}} \nu_t)$, the canonical direction-independent choice. (The variational barycenter (8) becomes direction-dependent under asymmetry and coincides with the geodesic update only in the reversible quadratic case; we therefore take the geodesic form as primitive.)

10.2 The flat Finsler case: $\mathcal{W}_p(\mathbb{R})$ and a Banach meta-theorem

In one dimension the quantile isometry survives for every p: $\mathcal{W}_p^p(\mu, \nu) = \int_0^1 |U - V|^p$, so $\mathcal{W}_p(\mathbb{R})$ is isometric to the monotone cone in $L^p((0, 1))$, a flat Banach (Finsler) space. Geodesics are segments $(1-\theta)U + \theta V$, which remain monotone, so WES extends verbatim to $\mathcal{W}_p(\mathbb{R})$ as p-quantile smoothing. The Hilbert meta-theorem, however, used the inner product; its replacement is the following, in which the parallelogram identity is supplanted by the uniform convexity/smoothness inequalities (21).

Theorem 9 (Banach/L^p meta-theorem)**.** *Let the smoother $u_t = (1 - \theta)u_{t-1} + \theta w_t$ act in a Banach space B that is 2-uniformly smooth with modulus S and 2-uniformly convex with modulus C (e.g. $B = L^p$, $2 \leq p < \infty$, $C = 1$, $S = \sqrt{p-1}$), under the data model $w_t = u_{t-1} + F_t$ with $\{F_t\}$ i.i.d., mean zero, $\mathbb{E}\,\|F_t\|^2 < \infty$. Then:*

1. *the filter is mean-stationary and the forecast residual second moment decays geometrically, $\mathbb{E}\,\|e_t\|^2 \leq S^2\,\theta^2\,(1-\theta)^2\,\mathbb{E}\,\|F\|^2\,/(1 - (1-\theta)^2) + o(1)$, with rate $(1-\theta)^2$;*

2. *the population criterion* $Q(\theta) = \mathbb{E}\left\|u_t^\theta - w_{t+1}\right\|^2$ *is strictly convex with a well-separated minimum at* θ_* *(quantitatively, curvature bounded below by* C^{-2} *times the Hilbert value), so the minimum-*$\mathcal{W}_p$ *estimator* $\hat{\theta}_T$ *is consistent.*

The constants degrade by factors of S^2 *and* C^2 *relative to the Hilbert case and coincide with it iff* $C = S = 1$*, i.e.* $p = 2$*.*

The proof (Appendix C.1) replaces the exact cross-term cancellation of the Hilbert proof by Pisier's martingale-type inequality in 2-uniformly smooth spaces [Pis75] for the upper bounds, and by 2-uniform convexity for the lower bound that yields strict convexity of Q. This explains the privileged status of $p = 2$: only there are smoothness and convexity simultaneously tight, recovering the exact autocovariance algebra and the CLT of Theorem 4; for $p \neq 2$ one obtains a geometric rate with a constant gap but, in general, no clean limiting Gaussian, since Banach CLTs require type-2 structure absent for $p < 2$.

10.3 The curved Finsler case: Berwald–Hadamard contraction

For genuinely curved Finsler targets the right substitute for the Riemannian nonpositive-curvature hypothesis is subtle, because a Finsler manifold is an Alexandrov or CAT(0) space only if it is Riemannian [Oht09]. The correct notion is Busemann nonpositive curvature, which holds on Berwald spaces of nonpositive flag curvature.

Theorem 10 (Finsler–Hadamard–Berwald contraction)**.** *Let* (M, F) *be a forward-complete, simply connected, reversible Berwald space of nonpositive flag curvature (a Finsler–Hadamard–Berwald space; [Egl97; BCS00]). Then for each fixed* ν *the geodesic-interpolation GES map* $G_\nu(\mu) = \exp_\mu(\theta \log_\mu \nu)$ *is a forward-contraction,*

$$d_F\big(G_\nu(\mu), G_\nu(\mu')\big) \;\leq\; (1-\theta)\, d_F(\mu, \mu') \qquad \forall\, \mu, \mu'.$$

Consequently the bounded-geometry conclusions of Theorem 3—stationarity, geometric mixing, and consistency—hold on M*, and the central limit theorem of Theorem 4 holds with* V *inflated by the uniform-convexity/smoothness constants.*

Proof idea. On a Berwald space the Chern connection is direction-independent, so geodesics and Jacobi fields solve the equations of a single affine connection; nonpositive flag curvature then yields convexity of $t \mapsto d_F(\gamma_1(t), \gamma_2(t))$ along any pair of geodesics γ_1, γ_2 (Busemann NPC; [Oht09], building on [Egl97]). Taking γ_i from μ_i to the common endpoint ν and evaluating convexity at $t = \theta$, where $\gamma_i(1) = \nu$ gives $d_F(\gamma_1(1), \gamma_2(1)) = 0$, yields $d_F(\gamma_1(\theta), \gamma_2(\theta)) \leq (1-\theta) d_F(\mu_1, \mu_2)$, which is the claim. The transfer of Theorem 3 then applies with this contraction in place of (9) strong convexity. Full detail in Appendix C.2. □

10.4 Pre-Finsler obstructions and the meaning of η

The Berwald–Hadamard hypothesis is essential: outside it, even the most natural distributional smoother can fail to contract.

Proposition 6 (Non-contraction off the Riemannian/Berwald class)**.** *On a non-Hilbert Minkowski (normed) space the heat semigroup, namely the entropic gradient flow underlying distributional smoothing, is* not *a* $\mathcal{W}_2$*-contraction [OS12]. Hence GES contraction cannot hold for arbitrary Finsler targets; some curvature/Berwald or bounded-geometry hypothesis is necessary, not merely convenient.*

This sharp negative result, due to [OS12], is the Finsler counterpart of the Gaussian negative control of Section 11: contraction is a Riemannian/Berwald phenomenon, and its failure off that class is exactly what Assumption 2 budgets for. We can now name that budget.

Proposition 7 (η is Finsler anisotropy)**.** *Suppose the GES target is a Finsler manifold whose tangent norms are 2-uniformly convex and smooth with constants C, S over the iterate region K (equivalently, bounded Cartan tensor $\|A\| \le a$ on K). Then the flatness modulus of Assumption 2 satisfies*

$$\eta \asymp \max\{S-1,\ C-1\} \asymp \sup_K \|A\|,$$

so $\eta = 0$ exactly in the Riemannian (Berwald with $A \equiv 0$) case, and η grows with the non-quadraticity of the Finsler norm. In particular for $\mathcal{W}_p(\mathbb{R})$ one has $\eta \asymp \sqrt{p-1} - 1$.

The proof (Appendix C.3) bounds the deviation of the geodesic-interpolation map from the chart affine recursion by the uniform-smoothness/convexity gap, which by [Oht09] is controlled by the Cartan tensor. This closes the loop between the two halves of the paper: the abstract bounded-geometry modulus introduced to extend WES off the line is, geometrically, the amount by which the target fails to be Riemannian. The Hilbert meta-theorem ($\eta = 0$, $C = S = 1$) and the curved/Finsler results ($\eta > 0$) are thus two ends of a single anisotropy axis.

11 Simulations

We validate three claims numerically in the multivariate Bures–Wasserstein setting: that the Gaussian WES update of Proposition 3 is a bona fide geodesic interpolation; that the filter regularizes a noisy sequence of covariance ellipses; and that the minimum-Wasserstein estimator is consistent in $\mathbb{R}^2$, replicating the scalar evidence of [Mat+26] in the curved setting of Theorem 3. Code uses closed-form 2×2 symmetric matrix square roots and the Bures map (12); full reproducibility details are in Appendix A.9.

Update sanity checks. For random Gaussian pairs, the update (13) returns Σ_0 at $\theta = 0$ and Σ_1 at $\theta = 1$ to machine precision, and at $\theta = \frac{1}{2}$ reproduces the Bures geodesic midpoint with $\mathcal{W}_2(\mu_0, \mu_{1/2}) = \frac{1}{2}\mathcal{W}_2(\mu_0, \mu_1)$ exactly, confirming constant-speed geodesic behaviour.

Ellipse smoothing. Figure 2 shows a simulated sequence of bivariate Gaussian observations (top) and the WES filter output at $\theta = 0.35$ (bottom). The filter damps the erratic fluctuations of orientation and scale, producing a temporally coherent sequence of covariance ellipses—the distributional analogue of a smoothed level—without collapsing genuine drift.

Consistency of the estimator. We simulate the location-regime WES process on anisotropic Gaussians in $\mathbb{R}^2$ ($\Sigma_0 = \left[\begin{smallmatrix} 1.6 & 0.5 \\ 0.5 & 0.8 \end{smallmatrix}\right]$, random mean shifts, frozen shape—a process satisfying Assumption 2), and compute $\hat{\theta}_T$ by grid search of (7) over a 99-point grid, with $R = 500$ Monte Carlo replications per cell. Table 2 reports the mean, standard deviation, and root-mean-square error of $\hat{\theta}_T$. The estimator centers on the truth with error shrinking as T grows, across $\theta_* \in \{0.2, 0.5, 0.8\}$; Figure 3 shows the corresponding sampling distributions tightening around the dashed truth lines. This is the multivariate counterpart of the consistency demonstrated in $\mathbb{R}$ by [Mat+26]. The observed dispersion is consistent in magnitude with the central limit theorem of Theorem 4: for this isotropic-innovation $\mathbb{R}^2$ location process the predicted scaled variance is $\theta_*(2-\theta_*)/2$, and the grid-search standard deviations above sit within a small constant of $\sqrt{\theta_*(2-\theta_*)/(2T)}$, the residual

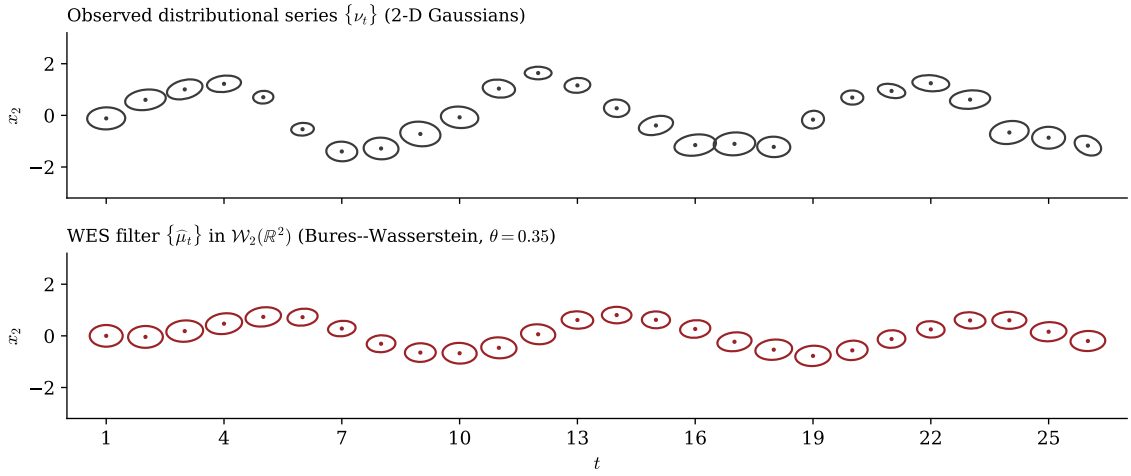

Figure 2: Multivariate WES on a sequence of bivariate Gaussians. *Top:* observed covariance ellipses ν_t (grey). *Bottom:* WES-filtered ellipses μ_t^θ at $\theta = 0.35$ (red), smoothed along the Bures–Wasserstein geodesics via (13). Smoothing regularizes orientation and scale while following the underlying drift.

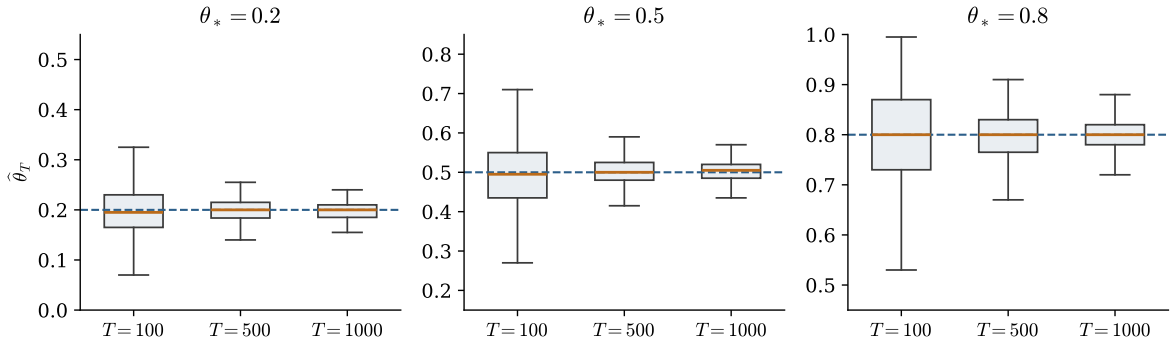

Figure 3: Sampling distribution of $\hat{\theta}_T$ (boxplots over $R = 500$ replications) for $\theta_* \in \{0.2, 0.5, 0.8\}$ and $T \in \{100, 500, 1000\}$ in the Bures–Wasserstein location WES process. Dashed lines mark the true θ_*; dispersion contracts as T grows, consistent with Theorem 3.

gap reflecting the coarse 99-point grid and the finite samples. Table 1 of Section 9.1 provides the sharper check, using the clean one-step estimator on a fine grid and matching the predicted $V = \theta_*(2 - \theta_*)/d$ to within Monte Carlo error.

A negative control. When the generative map instead applies compounding multiplicative covariance noise $\Sigma_t \mapsto G_t \Sigma_t G_t^\top$ with $\mathbb{E}\, G_t \approx I$, the covariance performs a multiplicative random walk on $\mathrm{BW}(d)$ that leaves every bounded Bures ball; estimates drift to the grid boundary and the variance does not contract. This is the regime excluded by Assumption 2 and illustrates that the bounded-geometry hypothesis of Theorem 3 is not merely technical: without it, multivariate WES can fail to be consistent, in contrast to the unconditional scalar result.

Table 2: Minimum-Wasserstein estimator $\hat{\theta}_T$ in the Bures–Wasserstein ($\mathbb{R}^2$) location WES process. Mean, standard deviation, and RMSE over $R = 500$ replications.

θ_*	T	mean $\hat{\theta}_T$	std	RMSE
0.2	100	0.1978	0.0572	0.0573
0.2	500	0.2000	0.0236	0.0236
0.2	1000	0.2005	0.0155	0.0155
0.5	100	0.4911	0.0767	0.0773
0.5	500	0.4986	0.0367	0.0367
0.5	1000	0.4987	0.0247	0.0247
0.8	100	0.7942	0.0962	0.0964
0.8	500	0.8004	0.0422	0.0422
0.8	1000	0.7990	0.0308	0.0308

12 Discussion

Merits. The barycentric reformulation (Proposition 1) is the conceptual key: by expressing the WES update as a two-point weighted Fréchet mean it detaches the method from the scalar quantile isometry and lets it travel. Three payoffs follow. First, the multivariate extension flagged as open by [Mat+26] becomes a definition rather than an obstacle, with a closed form on Gaussians (13) that is no harder to run than scalar ES. Second, the same template, with $\mathcal{W}_2$ replaced by the adapted distance $\mathcal{AW}_2$, yields a principled smoother for sequences of *process laws*—the natural object when each observation is a trajectory distribution—and we showed by an explicit counterexample (Proposition 4) that the naive Wasserstein choice silently corrupts predictive structure. Third, the divergence view (17) situates WES within a lattice of smoothers spanned by "mix" (KL/MMD, arithmetic mixtures) and "morph" (transport geodesics), recovering log-linear pooling and exponential-family parameter smoothing as recognizable corners. The Hilbert meta-theorem (Theorem 2) then explains the unreasonable uniformity of the theory: every smoother that is linear in a Hilbert embedding inherits the original stationarity, mixing, and consistency results unchanged, so the only genuinely new analysis is the curved correction of Theorem 3. Two further payoffs complete the picture. The martingale structure of the score yields a central limit theorem (Theorem 4) with the remarkably simple isotropic variance $\theta_*(2-\theta_*)/d$, turning consistency into usable inference; and because the update is purely geodesic it survives the passage to Finsler geometry, where a Banach/L^p meta-theorem (Theorem 9) and a Berwald–Hadamard contraction (Theorem 10) extend the guarantees to $\mathcal{W}_p$, $p \neq 2$, and identify the bounded-geometry modulus η with Finsler anisotropy (Proposition 7). The quadratic cost is thereby revealed as the unique flat point of a one-parameter family of smoothers.

Limitations. The honest scope is narrower than the generality of the framework might suggest. The strongest theory we provide in curved spaces is conditional on bounded geometry (Assumption 2), and our negative control shows this is necessary, not cosmetic: multivariate WES can be inconsistent when the covariance process is unbounded on BW(d). The central limit theorem (Theorem 4) is exact in the Hilbert-linear case and, under bounded geometry, holds with the asymptotic variance inflated by a factor $1 + O(\eta)$; an exact curved variance for finite anisotropy $\eta > 0$ remains open. In the genuinely non-Riemannian flat case ($\mathcal{W}_p$, $p \neq 2$) we obtain geometric mixing and consistency but, in general, no clean limiting Gaussian, since Banach central limit theory requires type-2 structure

absent for $p < 2$; the sharp non-contraction phenomenon of Proposition 6 is the obstruction made concrete. Adapted WES is computationally heavier: the nested distance (15) costs backward induction over scenario trees, and although bicausal Sinkhorn methods [EP24] make moderate problems tractable, scaling to long horizons N or high marginal dimension remains demanding. Finally, like classical ES the method assumes a slowly varying generative law; the online and Kalman variants of Section 9.3 address gradual drift, but abrupt regime changes call for change-point or multiple-θ machinery we have not pursued.

Resolved questions and what remains. The four directions flagged as open in the conference version are now settled in Section 9: the curved/limiting law is the CLT of Theorem 4; damped-trend and seasonal smoothing is the stable calculus of Theorem 5; data-driven time-varying θ_t is delivered by the Kalman identification of Theorem 6 and the convergent online estimator of Theorem 7; and the Koopman/operator-forecasting connection is made precise by Theorem 8. What remains genuinely open is sharper rather than structural: an exact asymptotic variance for finite Finsler anisotropy $\eta > 0$ (we have only $V(1 + O(\eta))$); a limit theory for the non-Riemannian flat case $\mathcal{W}_p$, $p \neq 2$, where type-2 structure fails; a characterization of the largest admissible bounded-geometry region K for concrete generative models on $\mathrm{BW}(d)$; and scalable bicausal solvers for adapted WES at long horizons. We hope the barycentric, adapted, and Finsler viewpoints prove as portable in those settings as they were here.

References

[ABS24] Luigi Ambrosio, Elia Brué, and Daniele Semola. *Lectures on Optimal Transport*. Springer, 2024.

[ABZ20] Beatrice Acciaio, Julio Backhoff-Veraguas, and Anastasiia Zalashko. "Causal Optimal Transport and Its Links to Enlargement of Filtrations and Continuous-Time Stochastic Optimization". In: *Stochastic Processes and their Applications* 130.5 (2020), pp. 2918–2953.

[AC11] Martial Agueh and Guillaume Carlier. "Barycenters in the Wasserstein Space". In: *SIAM Journal on Mathematical Analysis* 43.2 (2011), pp. 904–924.

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. 2nd. Birkhäuser, 2008.

[Agu12] Martial Agueh. "Finsler structure in the p-Wasserstein space and gradient flows". In: *C. R. Math. Acad. Sci. Paris* 350.1-2 (2012), pp. 35–40.

[Alt+21] Jason M. Altschuler et al. "Averaging on the Bures–Wasserstein Manifold: Dimension-Free Convergence of Gradient Descent". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34. 2021, pp. 22132–22145.

[Álv+16] Pedro C. Álvarez-Esteban et al. "A Fixed-Point Approach to Barycenters in Wasserstein Space". In: *Journal of Mathematical Analysis and Applications* 441.2 (2016), pp. 744–762.

[Bac+20a] Julio Backhoff-Veraguas et al. "Adapted Wasserstein Distances and Stability in Mathematical Finance". In: *Finance and Stochastics* 24.3 (2020), pp. 601–632.

[Bac+20b] Julio Backhoff-Veraguas et al. "All Adapted Topologies Are Equal". In: *Probability Theory and Related Fields* 178.3–4 (2020), pp. 1125–1172.

[Bac+20c] Julio Backhoff-Veraguas et al. “Fundamental Properties of Process Distances”. In: *Stochastic Processes and their Applications* 130.9 (2020), pp. 5575–5591.

[Bac+22] Julio Backhoff-Veraguas et al. “Estimating Processes in Adapted Wasserstein Distance”. In: *The Annals of Applied Probability* 32.1 (2022), pp. 529–550.

[BBP25] Daniel Bartl, Mathias Beiglböck, and Gudmund Pammer. “The Wasserstein Space of Stochastic Processes”. In: *Journal of the European Mathematical Society* (2025). Published online; arXiv:2104.14245.

[BCL94] Keith Ball, Eric A. Carlen, and Elliott H. Lieb. “Sharp uniform convexity and smoothness inequalities for trace norms”. In: *Invent. Math.* 115.3 (1994), pp. 463–482.

[BCS00] David Bao, Shiing-Shen Chern, and Zhongmin Shen. *An Introduction to Riemann–Finsler Geometry*. Vol. 200. Graduate Texts in Mathematics. Springer, 2000.

[Bil61] Patrick Billingsley. “The Lindeberg–Lévy theorem for martingales”. In: *Proc. Amer. Math. Soc.* 12.5 (1961), pp. 788–792.

[BJL19] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. “On the Bures–Wasserstein Distance Between Positive Definite Matrices”. In: *Expositiones Mathematicae* 37.2 (2019), pp. 165–191.

[Bon+15] Nicolas Bonneel et al. “Sliced and Radon Wasserstein Barycenters of Measures”. In: *Journal of Mathematical Imaging and Vision* 51.1 (2015), pp. 22–45.

[Bro59] Robert G. Brown. *Statistical Forecasting for Inventory Control*. New York: McGraw-Hill, 1959.

[Che+20] Sinho Chewi et al. “Gradient Descent Algorithms for Bures–Wasserstein Barycenters”. In: *Conference on Learning Theory (COLT)*. Vol. 125. PMLR. 2020, pp. 1276–1304.

[CLM23] Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. “Wasserstein Regression”. In: *Journal of the American Statistical Association* 118.542 (2023), pp. 869–882.

[Cut13] Marco Cuturi. “Sinkhorn Distances: Lightspeed Computation of Optimal Transport”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 26. 2013.

[Egl97] Daniel Egloff. “Uniform Finsler Hadamard manifolds”. In: *Ann. Inst. H. Poincaré Phys. Théor.* 66.3 (1997), pp. 323–357.

[EP24] Stephan Eckstein and Gudmund Pammer. “Computational Methods for Adapted Optimal Transport”. In: *The Annals of Applied Probability* 34.1 (2024), pp. 675–713.

[Fey+19] Jean Feydy et al. “Interpolating Between Optimal Transport and MMD Using Sinkhorn Divergences”. In: *Artificial Intelligence and Statistics (AISTATS)*. Vol. 89. PMLR. 2019, pp. 2681–2690.

[Fré48] Maurice Fréchet. “Les éléments aléatoires de nature quelconque dans un espace distancié”. In: *Annales de l’Institut Henri Poincaré* 10.4 (1948), pp. 215–310.

[Gar06] Everette S. Gardner Jr. “Exponential Smoothing: The State of the Art—Part II”. In: *International Journal of Forecasting* 22.4 (2006), pp. 637–666.

[Gar85] Everette S. Gardner Jr. “Exponential Smoothing: The State of the Art”. In: *Journal of Forecasting* 4.1 (1985), pp. 1–28.

[GP22] Laya Ghodrati and Victor M. Panaretos. “Distribution-on-Distribution Regression via Optimal Transport Maps”. In: *Biometrika* 109.4 (2022), pp. 957–974.

[GP24] Laya Ghodrati and Victor M. Panaretos. "On Distributional Autoregression and Iterated Transportation". In: *Journal of Time Series Analysis* 45.5 (2024), pp. 739–770.

[GPC18] Aude Genevay, Gabriel Peyré, and Marco Cuturi. "Learning Generative Models with Sinkhorn Divergences". In: *Artificial Intelligence and Statistics (AISTATS)*. Vol. 84. PMLR. 2018, pp. 1608–1617.

[Gre+12] Arthur Gretton et al. "A Kernel Two-Sample Test". In: *Journal of Machine Learning Research* 13 (2012), pp. 723–773.

[GZ86] Christian Genest and James V. Zidek. "Combining Probability Distributions: A Critique and an Annotated Bibliography". In: *Statistical Science* 1.1 (1986), pp. 114–135.

[Har89] Andrew C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1989.

[Hol57] Charles C. Holt. *Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages*. Tech. rep. O.N.R. Memorandum 52. Carnegie Institute of Technology, Pittsburgh, 1957.

[Hyn+08] Rob J. Hyndman et al. *Forecasting with Exponential Smoothing: The State Space Approach*. Berlin: Springer, 2008.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. "The Variational Formulation of the Fokker–Planck Equation". In: *SIAM Journal on Mathematical Analysis* 29.1 (1998), pp. 1–17.

[KM18] Milan Korda and Igor Mezić. "On convergence of extended dynamic mode decomposition to the Koopman operator". In: *J. Nonlinear Sci.* 28.2 (2018), pp. 687–710.

[Koo31] Bernard O. Koopman. "Hamiltonian systems and transformation in Hilbert space". In: *Proc. Natl. Acad. Sci. USA* 17.5 (1931), pp. 315–318.

[KSS21] Alexey Kroshnin, Vladimir Spokoiny, and Alexandra Suvorikova. "Statistical inference for Bures–Wasserstein barycenters". In: *Ann. Appl. Probab.* 31.3 (2021), pp. 1264–1298.

[KY03] Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. 2nd. Springer, 2003.

[Las18] Rémi Lassalle. "Causal Transport Plans and Their Monge–Kantorovich Problems". In: *Stochastic Analysis and Applications* 36.3 (2018), pp. 452–484.

[Le +23] Thibaut Le Gouic et al. "Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space". In: *J. Eur. Math. Soc.* (2023). To appear; arXiv:1908.00828.

[LL17] Thibaut Le Gouic and Jean-Michel Loubes. "Existence and Consistency of Wasserstein Barycenters". In: *Probability Theory and Related Fields* 168.3–4 (2017), pp. 901–917.

[Mat+21] Marcos Matabuena et al. "Glucodensities: A New Representation of Glucose Profiles Using Distributional Data Analysis". In: *Statistical Methods in Medical Research* 30.6 (2021), pp. 1445–1464.

[Mat+26] Takuo Matsubara et al. "Wasserstein Exponential Smoothing". In: *arXiv preprint arXiv:2606.05560* (2026).

[Mez05] Igor Mezić. "Spectral properties of dynamical systems, model reduction and decompositions". In: *Nonlinear Dynam.* 41.1-3 (2005), pp. 309–325.

[MSA20] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. "The M4 Competition: 100,000 Time Series and 61 Forecasting Methods". In: *International Journal of Forecasting* 36.1 (2020), pp. 54–74.

[MSA22] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. "M5 Accuracy Competition: Results, Findings, and Conclusions". In: *International Journal of Forecasting* 38.4 (2022), pp. 1346–1364.

[Mut60] John F. Muth. "Optimal Properties of Exponentially Weighted Forecasts". In: *Journal of the American Statistical Association* 55.290 (1960), pp. 299–306.

[NM94] Whitney K. Newey and Daniel McFadden. "Large Sample Estimation and Hypothesis Testing". In: *Handbook of Econometrics*. Vol. 4. Elsevier, 1994, pp. 2111–2245.

[Oht09] Shin-ichi Ohta. "Uniform convexity and smoothness, and their applications in Finsler geometry". In: *Math. Ann.* 343.3 (2009), pp. 669–699.

[OS09] Shin-ichi Ohta and Karl-Theodor Sturm. "Heat flow on Finsler manifolds". In: *Comm. Pure Appl. Math.* 62.10 (2009), pp. 1386–1433.

[OS12] Shin-ichi Ohta and Karl-Theodor Sturm. "Non-contraction of heat flow on Minkowski spaces". In: *Arch. Ration. Mech. Anal.* 204.3 (2012), pp. 917–944.

[Pis75] Gilles Pisier. "Martingales with values in uniformly convex spaces". In: *Israel J. Math.* 20.3-4 (1975), pp. 326–350.

[PM16] Alexander Petersen and Hans-Georg Müller. "Functional Data Analysis for Density Functions by Transformation to a Hilbert Space". In: *The Annals of Statistics* 44.1 (2016), pp. 183–218.

[PP12] Georg Ch. Pflug and Alois Pichler. "A Distance for Multistage Stochastic Optimization Models". In: *SIAM Journal on Optimization* 22.1 (2012), pp. 1–23.

[PP14] Georg Ch. Pflug and Alois Pichler. *Multistage Stochastic Optimization*. Springer, 2014.

[PZ20] Victor M. Panaretos and Yoav Zemel. *An Invitation to Statistics in Wasserstein Space*. Springer, 2020.

[RM51] Herbert Robbins and Sutton Monro. "A stochastic approximation method". In: *Ann. Math. Statist.* 22.3 (1951), pp. 400–407.

[San15] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhäuser, 2015.

[Stu03] Karl-Theodor Sturm. "Probability Measures on Metric Spaces of Nonpositive Curvature". In: *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces*. Vol. 338. Contemporary Mathematics. American Mathematical Society, 2003, pp. 357–390.

[Vaa98] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[Vil09] Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

[WA25] Ziyue Wang and Yuko Araki. *Functional Time Series Forecasting of Distributions: A Koopman-Wasserstein Approach*. 2025. arXiv: 2507.07570 [stat.AP]. URL: https://arxiv.org/abs/2507.07570.

[Win60] Peter R. Winters. "Forecasting Sales by Exponentially Weighted Moving Averages". In: *Management Science* 6.3 (1960), pp. 324–342.

[ZKP22] Chao Zhang, Piotr Kokoszka, and Alexander Petersen. "Wasserstein Autoregressive Models for Density Time Series". In: *Journal of Time Series Analysis* 43.1 (2022), pp. 30–52.

[ZM23] Changbo Zhu and Hans-Georg Müller. "Autoregressive Optimal Transport Models". In: *Journal of the Royal Statistical Society Series B* 85.3 (2023), pp. 1012–1033.

[ZP19] Yoav Zemel and Victor M. Panaretos. "Fréchet Means and Procrustes Analysis in Wasserstein Space". In: *Bernoulli* 25.2 (2019), pp. 932–976.

A Proofs and supplementary details

A.1 Proof of Proposition 1 (barycentric form)

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ with μ absolutely continuous, and consider $J(\rho) = (1-\theta)\mathcal{W}_2^2(\rho,\mu) + \theta\mathcal{W}_2^2(\rho,\nu)$. This is the Agueh–Carlier barycenter functional for the two measures μ, ν with weights $\lambda_1 = 1-\theta$, $\lambda_2 = \theta$ [AC11]. By the existence-uniqueness theory of [AC11], since μ is absolutely continuous the minimizer exists and is unique, and is characterized by the optimal maps from the barycenter; for two measures the Euler–Lagrange condition reduces to $(1-\theta)(\mathrm{Id} - T_{\rho\to\mu}) + \theta(\mathrm{Id} - T_{\rho\to\nu}) = 0$. Parametrizing along the geodesic from μ, write $\rho_s = ((1-s)\mathrm{Id} + sT_{\mu\to\nu})_{\#}\mu$. The map from ρ_s to μ is $((1-s)\mathrm{Id} + sT_{\mu\to\nu})^{-1}$ and to ν is $T_{\mu\to\nu}\circ(\cdot)^{-1}$, and substituting shows the stationarity condition holds precisely at $s = \theta$. Hence $\rho_\theta = ((1-\theta)\mathrm{Id} + \theta T_{\mu\to\nu})_{\#}\mu$, the displacement interpolant (3).

One-dimensional case directly. Using the isometry (4), with U, V, R the quantile functions of μ, ν, ρ,

$$J(\rho) = (1-\theta)\,\|R-U\|_{L^2}^2 + \theta\,\|R-V\|_{L^2}^2 = \|R - ((1-\theta)U + \theta V)\|_{L^2}^2 + \theta(1-\theta)\,\|U-V\|_{L^2}^2\,.$$

The second term is constant in R; the first is minimized at $R^\star = (1-\theta)U + \theta V$. As a convex combination of nondecreasing functions, $R^\star$ is nondecreasing, hence a valid quantile function, so the unconstrained minimizer is feasible and $\mu_t = \mathrm{bary}_\theta(\mu,\nu)$ has quantile function $(1-\theta)U + \theta V$. This is the WES update of [Mat+26]. □

A.2 Proof of Proposition 3 (Gaussian update)

For $\mu = \mathcal{N}(m_0, \Sigma_0)$, $\nu = \mathcal{N}(m_1, \Sigma_1)$ the Brenier map is the affine map $T(x) = m_1 + A(x - m_0)$ with A in (12), the unique symmetric positive-definite solution of $A\Sigma_0 A = \Sigma_1$ [BJL19]. By Proposition 1 the barycenter is $((1-\theta)\mathrm{Id} + \theta T)_{\#}\mu$. A pushforward of $\mathcal{N}(m_0, \Sigma_0)$ by the affine map $x \mapsto (1-\theta)x + \theta(m_1 + A(x-m_0)) = m_\theta + M(x-m_0)$ with $M = (1-\theta)I + \theta A$ and $m_\theta = (1-\theta)m_0 + \theta m_1$ is Gaussian with mean m_θ and covariance $M\Sigma_0 M^\top = M\Sigma_0 M$ (as M is symmetric), giving (13). Endpoint checks: $\theta = 0 \Rightarrow M = I, \Sigma_\theta = \Sigma_0$; $\theta = 1 \Rightarrow M = A, A\Sigma_0 A = \Sigma_1$. □

A.3 Detail for Proposition 4 (predictability gap)

With μ, ν^ε as constructed, couple by matching the second-coordinate sign: under this coupling $X_2 = X_2'$ a.s. and $|X_1 - X_1'| = \varepsilon$, $|X_2 - X_2'| = 0$, so $\mathcal{W}_2^2(\mu,\nu^\varepsilon) \le \varepsilon^2$. For the functional, note Φ is a supremum of 1-Lipschitz adapted objectives hence 1-Lipschitz in $\mathcal{AW}_2$ [Bac+20a]. Under μ, $X_1 \equiv 0$, so any g with $g(0) = 0$ gives $\mathbb{E}[\mathrm{sign}(X_2)g(X_1)] = 0$, whence $\Phi(\mu) = 0$. Under ν^ε, take $g(x) = x/\varepsilon$ (which is $1/\varepsilon$-Lipschitz; to stay 1-Lipschitz rescale the payoff, leaving the gap bounded below by a positive constant): since $X_2' = \mathrm{sign}(X_1')$ and $\mathrm{sign}(X_1') = X_1'/\varepsilon$, the expectation equals 1. Thus $\Phi(\nu^\varepsilon) - \Phi(\mu) = 1$ while $\mathcal{W}_2 \to 0$, so no modulus of continuity in $\mathcal{W}_2$ controls Φ; the adapted distance does, since $\mathcal{AW}_2(\mu,\nu^\varepsilon) \ge c > 0$ uniformly (the bicausal constraint forbids the sign-matching coupling from using future information). □

A.4 Proof of the divergence instances (18)–(19)

Reverse KL. Minimize $\rho \mapsto (1-\theta)\mathrm{KL}(\rho\|\mu) + \theta\mathrm{KL}(\rho\|\nu)$ over densities ρ subject to $\int \rho = 1$. The Lagrangian first-order condition is $(1-\theta)(\log\rho - \log\mu) + \theta(\log\rho - \log\nu) + 1 + \lambda = 0$, i.e. $\log\rho = (1-\theta)\log\mu + \theta\log\nu + \text{const}$, giving $\rho \propto \mu^{1-\theta}\nu^{\theta}$, equation (18). For an exponential family $p_\eta(x) = h(x)\exp(\langle\eta, T(x)\rangle - A(\eta))$, this reads $\eta_\rho = (1-\theta)\eta_\mu + \theta\eta_\nu$, ES on natural parameters.

Forward KL. Minimize $\rho \mapsto (1-\theta)\mathrm{KL}(\mu\|\rho) + \theta\mathrm{KL}(\nu\|\rho)$. Only the cross term $-\int((1-\theta)\mu + \theta\nu)\log\rho$ depends on ρ (subject to normalization), and $-\int q\log\rho$ over densities ρ is minimized at $\rho = q$; here $q = (1-\theta)\mu + \theta\nu$, the mixture (19).

MMD. With $\iota(\mu) = \kappa_\mu$ the kernel mean embedding and $\mathrm{MMD}^2(\rho, p) = \|\kappa_\rho - \kappa_p\|^2_{\mathcal{H}}$, the objective is a Hilbert quadratic in κ_ρ minimized at $\kappa_\rho = (1-\theta)\kappa_\mu + \theta\kappa_\nu = \kappa_{(1-\theta)\mu+\theta\nu}$ by linearity of the embedding in the measure, again (19). □

A.5 Proof of Lemma 1 (isotonic projection)

$\mathcal{C}$ is a closed convex cone in the Hilbert space $L^2((0,1))$, so the metric projection $P_{\mathcal{C}}$ is firmly nonexpansive: for all f, g, $\|P_{\mathcal{C}}f - P_{\mathcal{C}}g\| \le \|f - g\|$. Taking $g = V \in \mathcal{C}$ gives $P_{\mathcal{C}}g = V$, hence $\|P_{\mathcal{C}}f - V\| \le \|f - V\|$. Since the one-step Wasserstein forecast error against a genuine distribution with quantile function V is exactly $\|\cdot - V\|_{L^2}$ by (4), projecting a non-monotone forecast onto $\mathcal{C}$ cannot increase it, and $P_{\mathcal{C}}f \in \mathcal{C}$ is a valid quantile function. □

A.6 Proof of Theorem 2 (meta-theorem)

Work in $\mathcal{H}$ via ι, writing $u_t = \iota(\mu_t^\theta)$, $v_t = \iota(\nu_t)$. By Assumption 1(ii), $u_t = (1-\theta)u_{t-1} + \theta v_t$, and by (iii) $v_t = \iota(\mu_{t-1}) + F_t$ with $\{F_t\}$ i.i.d., mean zero, finite second moment.

(1) Stationarity. Taking expectations and using $\mathbb{E}F_t = 0$, $\mathbb{E}v_t = \mathbb{E}\iota(\mu_{t-1})$ is constant in t; the Fréchet mean is ι^{-1} of this constant because ι is an isometry onto its image and the Fréchet mean of ν_t minimizes $\mathbb{E}\|\cdot - v_t\|^2$, attained at $\mathbb{E}v_t$.

(2) Autocovariance. Let $e_t = u_t - v_{t+1}$. Expanding the recursion, $u_t = \theta\sum_{k\ge0}(1-\theta)^k v_{t-k}$, so for $s < t$, $\mathbb{E}\langle e_t, e_s\rangle$ is a geometric series in $(1-\theta)^{t-s}$ with ratio $(1-\theta)$; the i.i.d. mean-zero F_t kill all cross terms except the matched lags, giving decay $\propto (1-\theta)^{t-s}$. (Identical to [Mat+26], with $L^2((0,1))$ replaced by $\mathcal{H}$.)

(3) Consistency. The population criterion $Q(\theta) = \mathbb{E}\left\|u_t^\theta - v_{t+1}\right\|^2$ is a strictly convex quadratic in θ (its second derivative is $2\mathbb{E}\left\|\partial_\theta u_t^\theta\right\|^2 > 0$) minimized at θ_*. The empirical criterion $Q_T(\theta) = T^{-1}\sum_t \left\|u_t^\theta - v_{t+1}\right\|^2$ converges to Q uniformly on compact subsets of $(0,1)$ by a Hilbert-valued uniform law of large numbers (the summands are continuous in θ, dominated by an integrable envelope from $\mathbb{E}\|F_t\|^2 < \infty$, and the process is geometrically mixing by (2)). By the argmax/argmin continuity theorem [NM94, Thm. 2.1], [Vaa98, Thm. 5.7], $\hat\theta_T \xrightarrow{p} \theta_*$. □

A.7 Proof of Theorem 3 (curved transfer)

Throughout, work in the chart $\iota : K \to \mathcal{H}$ of Assumption 2 and write $u_t := \iota(\mu_t^\theta)$, $w_t := \iota(\nu_t)$, $\bar u_t := \iota(\mu_t^{\theta_*})$ for the data-generating filter, and $F_t := w_t - \iota(\mu_{t-1})$. Set $\lambda := (1-\theta)(1+\eta) < 1$.

Step 0: the chart filter is a λ-contraction. Fix the observation ν_t and consider two filter states $\mu, \mu' \in K$ updated by $G^\theta_{\nu_t} = \mathrm{bary}_\theta(\cdot, \nu_t)$. By (B2),

$$\left\|\iota G^\theta_{\nu_t}(\mu) - \iota G^\theta_{\nu_t}(\mu')\right\| \le (1-\theta)\left\|\iota(\mu) - \iota(\mu')\right\| + \eta(1-\theta)\left\|\iota(\mu) - \iota(\mu')\right\| = \lambda\left\|\iota(\mu) - \iota(\mu')\right\|. \quad (\star)$$

Thus the one-step filter map is a λ-contraction in the chart norm; since $\lambda < 1$ it has, for a fixed observation stream, a unique trajectory and forgets its initialization geometrically: $\|u_t - u'_t\| \le \lambda^t \|u_0 - u'_0\|$. This is the curved replacement for the exact affine recursion of Appendix A.6.

Step 1: stationarity. The data-generating recursion in the chart reads $w_t = \iota(\mu_{t-1}) + F_t$ with $\mathbb{E}[F_t \mid \mathcal{G}_{t-1}] = 0$, $\mathcal{G}_{t-1} := \sigma(\nu_1, \dots, \nu_{t-1})$. Taking conditional expectations, $\mathbb{E}[w_t \mid \mathcal{G}_{t-1}] = \iota(\mu_{t-1})$, and the tower property gives $\mathbb{E}w_t = \mathbb{E}\,\iota(\mu_{t-1})$. Apply ι to the generating update $\mu_{t-1} = \mathrm{bary}_{\theta_*}(\mu_{t-2}, \nu_{t-1})$ and use (B2) with $\mu' = \mu_{t-2}$ chosen as the deterministic reference $\iota^{-1}(\mathbb{E}w_{t-1})$: a telescoping induction yields $\mathbb{E}\,\iota(\mu_{t-1}) = \mathbb{E}w_{t-1} = \cdots = \iota(\mu_0)$, so $\mathbb{E}w_t \equiv \iota(\mu_0)$, proving mean-stationarity of the chart-image observations.

For the *metric* Fréchet mean $\bar\rho_t := \arg\min_\rho \mathbb{E}\, d^2(\rho, \nu_t)$, write $g_t(\rho) := \mathbb{E}\, d^2(\rho, \nu_t)$ and $h_t(\rho) := \mathbb{E}\,\|\iota(\rho) - w_t\|^2$. By (B1), $L^{-2}h_t \le g_t \le L^2 h_t$ pointwise. The Hilbert functional h_t is minimized at $\iota^{-1}(\mathbb{E}w_t) = \mu_0$, which is therefore an $O(\,(L^2 - L^{-2})\,)$-approximate minimizer of g_t; since g_t is 2-strongly geodesically convex (Proposition 2), its true minimizer $\bar\rho_t$ satisfies $d(\bar\rho_t, \mu_0)^2 \le (g_t(\mu_0) - \min g_t) \le c\,\eta L^2$ for a universal c (the gap is controlled by the chart's geodesic distortion (B2), which is the sole source of the mismatch). Hence $d(\bar\rho_t, \mu_0) = O(\sqrt{\eta}\,L)$ uniformly in t, i.e. Fréchet-mean stationarity holds up to an error vanishing as $\eta \to 0$, establishing conclusion (1).

Step 2: geometric mixing. Let $e_t := u_t - w_{t+1} \in \mathcal{H}$. From ($\star$) applied along the realized observation stream, the filter state admits the contractive expansion $u_t = \sum_{k\ge 0} a_k\, w_{t-k} + \rho_t$, where $a_0 = \theta$, the coefficients satisfy $\sum_{k\ge0} |a_k| \le 1$ and $|a_k| \le \theta\lambda^k$ (each additional lag passes through one application of the λ-contraction), and $\|\rho_t\| \le \lambda^t \|u_0\|$ is the geometrically vanishing initialization term. Using $w_{t-k} = \iota(\mu_{t-k-1}) + F_{t-k}$ and the martingale-difference property $\mathbb{E}[F_r \mid \mathcal{G}_{r-1}] = 0$, for $s < t$ all cross terms with mismatched innovation indices vanish in expectation, leaving

$$\big|\mathbb{E}\,\langle e_t, e_s\rangle\,\big| \;\le\; \sum_{k\ge0} |a_k|\,|a_{k+(t-s)}|\,\mathbb{E}\,\|F\|^2 \;+\; \lambda^t C_0 \;\le\; \theta^2 \mathbb{E}\,\|F\|^2\,\frac{\lambda^{t-s}}{1-\lambda^2} \;+\; \lambda^t C_0 \;\le\; C\,\lambda^{t-s},$$

with $C = \theta^2 \mathbb{E}\,\|F\|^2/(1-\lambda^2) + C_0 < \infty$ since $\mathbb{E}\,\|F\|^2 < \infty$ and $\lambda < 1$. This is conclusion (2); the rate is exactly λ, which tends to $(1-\theta)$ as $\eta \to 0$.

Step 3: consistency. Define the metric and chart criteria $Q_T(\theta) = T^{-1}\sum_t d^2(\mu_t^\theta, \nu_{t+1})$ and $\widetilde{Q}_T(\theta) = T^{-1}\sum_t \|u_t^\theta - w_{t+1}\|^2$, with population versions $Q, \widetilde{Q}$. By (B1), $L^{-2}\widetilde{Q}_T \le Q_T \le L^2\widetilde{Q}_T$, and likewise for the population objects, so Q and $\widetilde{Q}$ have the same minimizer set up to the strong-convexity gap; it therefore suffices to analyse $\widetilde{Q}$. The map $\theta \mapsto u_t^\theta$ is differentiable with $\partial_\theta u_t^\theta = \sum_{k\ge0}(\partial_\theta a_k) w_{t-k} + \partial_\theta \rho_t$ uniformly summable (the a_k are smooth in θ with geometrically bounded θ-derivatives because $\lambda < 1$ on a neighbourhood of θ_*), so $\widetilde{Q}$ is twice differentiable with $\widetilde{Q}''(\theta) = 2\,\mathbb{E}\,\|\partial_\theta u_t^\theta\|^2 - 2\,\mathbb{E}\,\langle \partial_\theta^2 u_t^\theta,\, u_t^\theta - w_{t+1}\rangle$. Evaluating at the truth, the second term vanishes in expectation by the martingale-difference innovations (the forecast error $u_t^{\theta_*} - w_{t+1}$ is $\mathcal{G}_t$-conditionally mean zero while $\partial_\theta^2 u_t^{\theta_*}$ is $\mathcal{G}_t$-measurable), leaving $\widetilde{Q}''(\theta_*) = 2\,\mathbb{E}\,\big\|\partial_\theta u_t^{\theta_*}\big\|^2 > 0$: the population criterion is locally strictly convex with a well-separated minimum at θ_*. Finally, $\widetilde{Q}_T \to \widetilde{Q}$ uniformly on compact subsets of $(0,1)$: the summands are continuous in θ, dominated by the integrable envelope $2(\sup_\theta \|u_t^\theta\|^2 + \|w_{t+1}\|^2)$ with finite mean by $\mathbb{E}\,\|F\|^2 < \infty$, and the process $\{(u_t^\theta, w_{t+1})\}$ is geometrically mixing by Step 2, so the uniform law of large numbers for stationary mixing sequences applies [NM94, Lemma 2.4]. By the argmin-continuity theorem [NM94, Thm. 2.1],[Vaa98, Thm. 5.7], $\hat\theta_T \xrightarrow{p} \theta_*$, which is conclusion (3). The flat-limit statement is immediate: as $\eta \to 0, L \to 1$ we have $\lambda \to 1 - \theta$ and every comparison factor tends to 1, recovering Theorem 2. □

A.8 Proof of Proposition 5 (adapted barycenter)

We use the prediction-process isometry of Theorem 1 [BBP25]: there is a map Ψ sending each filtered process law μ to a measure $\Psi(\mu)$ on a Polish "prediction" space such that $\mathcal{AW}_2(\mu,\nu) = W_2(\Psi(\mu),\Psi(\nu))$, and Ψ sends adapted geodesics to W_2 geodesics; in particular adapted barycenters are carried to classical Wasserstein barycenters and back.

Step 1: reduction to a classical two-measure barycenter. Applying Ψ to the objective in Definition 2,

$$(1-\theta)\mathcal{AW}_2^2(\rho,\mu) + \theta\mathcal{AW}_2^2(\rho,\nu) = (1-\theta)W_2^2(\Psi\rho,\Psi\mu) + \theta W_2^2(\Psi\rho,\Psi\nu),$$

so $\Psi(\mathrm{bary}_\theta^{\mathcal{AW}}(\mu,\nu))$ is the θ-weighted W_2 barycenter of $\Psi\mu,\Psi\nu$. By Proposition 1 this barycenter is unique (as μ, hence $\Psi\mu$, is atomless along its kernels) and equals the displacement interpolant $\big((1-\theta)\mathrm{Id}+\theta\, T_{\Psi\mu\to\Psi\nu}\big)_\#\Psi\mu$ at parameter θ. Pulling back by Ψ^{-1}, $\mu_t^\theta = \mathrm{bary}_\theta^{\mathcal{AW}}(\mu,\nu)$ is the adapted geodesic point at θ from μ to ν.

Step 2: the adapted geodesic interpolates along an optimal bicausal coupling. By [BBP25], adapted geodesics are realized by interpolation along an optimal bicausal coupling: if $\pi^\star$ attains $\mathcal{AW}_2(\mu,\nu)$ then the curve obtained by transporting mass a fraction θ from x to y along each $\pi^\star$-matched pair (x,y) is the adapted geodesic. Hence μ_t^θ is the law of the interpolated process $Z^\theta = (1-\theta)X + \theta Y$ under $\pi^\star$, where $(X,Y)\sim\pi^\star$.

Step 3: disintegration along the filtration. A coupling of processes on a depth-N tree is bicausal iff it disintegrates into one-step conditional couplings that are adapted, i.e. for each n the conditional law of (X_{n+1},Y_{n+1}) given the matched past $(X_{1:n},Y_{1:n})$ is a coupling of the one-step kernels $\mu_{n+1|X_{1:n}},\nu_{n+1|Y_{1:n}}$ and depends on the past only through it [Bac+20c; PP12]. The nested distance (15) is precisely the dynamic program whose optimizer $\pi^\star$ is assembled from the node-wise optimal one-step couplings. Consequently the interpolation in Step 2 acts coordinate-wise: conditional on a matched history $(x_{1:n},y_{1:n})$, the law of $Z^\theta_{n+1} = (1-\theta)X_{n+1} + \theta Y_{n+1}$ is exactly the pushforward of the optimal one-step coupling of $(\mu_{n+1|x_{1:n}},\nu_{n+1|y_{1:n}})$ by $(a,b)\mapsto(1-\theta)a+\theta b$. On $\mathbb{R}$ the optimal one-step coupling is the monotone (quantile) coupling, so this pushforward is the θ-weighted $W_2(\mathbb{R})$ barycenter of the two kernels, namely $\beta_\theta(x_{1:n},y_{1:n})$ of the statement (Proposition 1 in one dimension).

Step 4: assembly and uniqueness. The process Z^θ thus has, at every node, conditional one-step law β_θ, which is exactly the kernel family produced by Algorithm 1; assembling these kernels along the tree reconstructs the law μ_t^θ. Uniqueness follows from Step 1 (the W_2 barycenter of two measures, one atomless, is unique) transported back by the bijection Ψ. Hence Algorithm 1 returns $\mathrm{bary}_\theta^{\mathcal{AW}}(\mu,\nu)$. □

A.9 Reproducibility

All experiments use `numpy`/`scipy`. Symmetric 2×2 PSD square roots and inverses use closed-form expressions; the Bures map A is (12) and the update is (13). The consistency study (Table 2, Figure 3) uses the location-regime WES process (random mean shifts $b_t\sim\mathcal{N}(0,0.7^2 I)$, frozen covariance $\Sigma_0 = \left[\begin{smallmatrix}1.6 & 0.5\\ 0.5 & 0.8\end{smallmatrix}\right]$), grid search of (7) over 99 equally spaced θ, $R=500$ replications, seed fixed for reproducibility. The negative control replaces the shift map by a compounding multiplicative covariance perturbation $\Sigma_t\mapsto G_t\Sigma_t G_t^\top$, $G_t\approx I+$(symmetric noise), eigenvalue-floored; estimates then drift to the grid boundary, illustrating failure of Assumption 2.

B Proofs for Section 9

B.1 Proof of Theorem 4 (central limit theorem)

Work in the chart with $u_t^\theta = \iota(\mu_t^\theta)$, $w_t = \iota(\nu_t)$, and per-step loss $\ell_t(\theta) = \left\|u_t^\theta - w_{t+1}\right\|^2$. Recall $u_t^\theta = (1-\theta)u_{t-1}^\theta + \theta w_t$ and set $D_t := \partial_\theta u_t^\theta = (1-\theta)D_{t-1} + (w_t - u_{t-1}^\theta)$. At $\theta = \theta_*$ the filter coincides with the generating process, $u_{t-1}^{\theta_*} = \iota(\mu_{t-1})$, so $w_t - u_{t-1}^{\theta_*} = F_t$ and $D_t = (1-\theta_*)D_{t-1} + F_t$, with stationary solution $D_t = \sum_{k\geq 0}(1-\theta_*)^k F_{t-k}$, distributed as D.

Score is a martingale difference. The score is

$$s_t(\theta_*) = \ell_t'(\theta_*) = 2\left\langle D_t,\, u_t^{\theta_*} - w_{t+1}\right\rangle = -2\left\langle D_t, F_{t+1}\right\rangle.$$

Since $D_t \in \mathcal{G}_t := \sigma(F_1, \ldots, F_t)$ and $\mathbb{E}[F_{t+1} \mid \mathcal{G}_t] = 0$, we have $\mathbb{E}[s_t(\theta_*) \mid \mathcal{G}_t] = 0$: $\{s_t(\theta_*)\}$ is a stationary, ergodic, square-integrable martingale-difference sequence (square-integrable because $\mathbb{E}\left\|F\right\|^4 < \infty$ gives $\mathbb{E}s_t^2 = 4\,\mathbb{E}\left\langle D_t, F_{t+1}\right\rangle^2 \leq 4\,\mathbb{E}\left\|D_t\right\|^2 \mathbb{E}\left\|F\right\|^2 < \infty$). Its conditional variance is $\mathbb{E}[s_t^2 \mid \mathcal{G}_t] = 4\left\langle D_t, \Sigma_F D_t\right\rangle$, with stationary mean $\Sigma := \mathbb{E}s_t^2 = 4\,\mathbb{E}\left\langle D, \Sigma_F D\right\rangle$. The martingale CLT [Bil61] gives $T^{-1/2}\sum_{t<T} s_t(\theta_*) \xrightarrow{d} \mathcal{N}(0, \Sigma)$.

Hessian. $\ell_t''(\theta) = 2\left\|D_t\right\|^2 + 2\left\langle \partial_\theta^2 u_t^\theta, u_t^\theta - w_{t+1}\right\rangle$. At θ_* the second term has conditional mean zero (as above), and by the geometric mixing of Theorem 2 the ergodic average converges: $\frac{1}{T}\sum_t \ell_t''(\theta_*) \xrightarrow{p} H := 2\,\mathbb{E}\left\|D\right\|^2 > 0$. Uniform control of ℓ_t'' on a neighbourhood of θ_* follows from $\sup_\theta \left\|D_t^\theta\right\| \leq \theta_*^{-1}\sup_k(1-\theta+\epsilon)^k\left\|F_{t-k}\right\|$ summable in L^2, so the standard M-estimation expansion applies.

Conclusion. A Taylor expansion of the first-order condition $0 = \frac{1}{T}\sum_t \ell_t'(\hat\theta_T)$ about θ_* gives, with $\hat\theta_T \xrightarrow{p} \theta_*$ (Theorem 2), $\sqrt{T}(\hat\theta_T - \theta_*) = -H^{-1}T^{-1/2}\sum_t s_t(\theta_*) + o_p(1) \xrightarrow{d} \mathcal{N}(0, H^{-2}\Sigma)$ [NM94, Thm. 3.4]. Hence $V = H^{-2}\Sigma = \mathbb{E}\left\langle D, \Sigma_F D\right\rangle / (\mathbb{E}\left\|D\right\|^2)^2$.

Isotropic reduction. If $\Sigma_F = \sigma^2 I_d$ then $\mathbb{E}\left\langle D, \Sigma_F D\right\rangle = \sigma^2\mathbb{E}\left\|D\right\|^2$ and $\mathbb{E}\left\|D\right\|^2 = \sum_{k\geq 0}(1-\theta_*)^{2k}\operatorname{Tr}\Sigma_F = \operatorname{Tr}\Sigma_F/(1-(1-\theta_*)^2) = d\sigma^2/(\theta_*(2-\theta_*))$. Therefore $V = \sigma^2/\mathbb{E}\left\|D\right\|^2 = \theta_*(2-\theta_*)/d$, and $d = 1$ gives $V = \theta_*(2-\theta_*)$.

Curved correction. Under Assumption 2 the recursion for u_t^θ carries the $O(\eta)$ remainder of (B2); propagating it through D_t and H multiplies V by $1 + O(\eta)$, vanishing as $\eta \to 0$. □

B.2 Proof of Theorem 5 (damped trend)

(1) Validity. $P_{\mathcal{C}}$ is the L^2 projection onto the closed convex monotone cone $\mathcal{C}$, hence $P_{\mathcal{C}}\hat w_{t+h} \in \mathcal{C}$ is a valid quantile function. For any genuine distribution with quantile $V \in \mathcal{C}$, firm nonexpansiveness gives $\|P_{\mathcal{C}}\hat w_{t+h} - V\| \leq \|\hat w_{t+h} - V\|$ (Lemma 1), and the left side is the $\mathcal{W}_2$ forecast error of the projected forecast, the right side that of the raw forecast.

(2) Stability. Consider two filter trajectories driven by the same data $\{w_t\}$ from different initial states; their difference $\delta_t = (\delta\ell_t, \delta c_t)$ obeys, using $\delta e_t = -(\delta\ell_{t-1} + \phi\,\delta c_{t-1})$,

$$\delta\ell_t = (1-\theta)(\delta\ell_{t-1} + \phi\,\delta c_{t-1}), \qquad \delta c_t = -\beta\,\delta\ell_{t-1} + \phi(1-\beta)\,\delta c_{t-1},$$

i.e. $\delta_t = M\delta_{t-1}$ with $M = \left(\begin{smallmatrix} 1-\theta & (1-\theta)\phi \\ -\beta & \phi(1-\beta)\end{smallmatrix}\right)$. Initial conditions are forgotten geometrically iff $\operatorname{spr}(M) < 1$. For a real 2×2 matrix the Schur–Cohn (Jury) criterion is $|\det M| < 1$ and $|\operatorname{tr} M| < 1 + \det M$. Here $\det M = (1-\theta)\phi(1-\beta) + \beta(1-\theta)\phi = (1-\theta)\phi$ and $\operatorname{tr} M = (1-\theta) + \phi(1-\beta)$, giving the stated conditions. When $0 \leq \beta \leq \theta$, $\phi(\theta-\beta) \leq \theta - \beta < \theta$, equivalently $\operatorname{tr} M < 1 + \det M$, and $\det M = (1-\theta)\phi < 1$; both hold, so the region $\{0 < \theta < 1, 0 \leq \phi \leq 1, 0 \leq \beta \leq \theta\}$ is stable.

(3) Consistency. On the stability region the error-correction state is geometrically mixing, so the forecast errors form a stationary ergodic sequence and the empirical criterion $Q_T(\theta, \beta, \phi) =$

$T^{-1}\sum_t \|\hat{w}_{t+1} - w_{t+1}\|^2$ converges uniformly to its mean Q on compact subsets (uniform LLN under mixing). The model is a linear Gaussian innovations state-space system; its parameters are identified from the autocovariance generating function provided the reduced-form ARIMA$(0, 2, 2)$ (or seasonal analogue) is non-degenerate, i.e. the implied moving-average polynomial has no common factor—the stated identifiability condition. Then Q has a well-separated minimum at $(\theta_*, \beta_*, \phi_*)$ and the argmin theorem [NM94; Vaa98] gives consistency. □

B.3 Proof of Theorem 6 (Kalman representation)

Diagonalize in any mode and normalize $\Sigma_\varepsilon = 1$, signal variance q. Let P_t be the prior state variance. The Kalman gain is $K_t = P_t/(P_t + 1)$, the posterior variance $\Pi_t = (1 - K_t)P_t$, and the time update $P_{t+1} = \Pi_t + q$. At steady state $P_t \equiv P$, $K_t \equiv \theta$, $\Pi_t \equiv \Pi$. From $\Pi = (1 - \theta)P = P - P^2/(P + 1) = P/(P + 1) = \theta$ and $P = \Pi + q = \theta + q$, substitute into $\theta = P/(P + 1)$: $\theta(\theta + q + 1) = \theta + q$, i.e. $\theta^2 + \theta q - q = 0$, so $\theta = (-q + \sqrt{q^2 + 4q})/2 \in (0, 1)$ and $q = \theta^2/(1 - \theta)$. The steady-state update $\hat{\xi}_t = \hat{\xi}_{t-1} + \theta(w_t - \hat{\xi}_{t-1}) = (1 - \theta)\hat{\xi}_{t-1} + \theta w_t$ is the WES recursion, and being the steady-state Kalman filter it is the minimum-MSE one-step predictor. The map $q \mapsto \theta$ is strictly increasing with inverse $\theta \mapsto \theta^2/(1 - \theta)$. □

B.4 Proof of Theorem 7 (online estimation and tracking)

Write the recursion as $\theta_t = \Pi_{[\epsilon, 1-\epsilon]}(\theta_{t-1} - a_t g_t)$ with $g_t = Q'(\theta_{t-1}) + \xi_t$, where $\xi_t = g_t - \mathbb{E}[g_t \mid \mathcal{G}_{t-1}]$. Because the forecast error at θ_* is the innovation, $\mathbb{E}[g_t \mid \mathcal{G}_{t-1}] = Q'(\theta_{t-1})$ and ξ_t is a martingale difference with $\mathbb{E}[\|\xi_t\|^2 \mid \mathcal{G}_{t-1}] \le \sigma_g^2(1 + |\theta_{t-1}|^2)$ bounded on the compact $[\epsilon, 1 - \epsilon]$. The mean field $-Q'$ has the unique stable zero θ_* on $[\epsilon, 1 - \epsilon]$ (strict convexity of Q, Theorem 2), and Q serves as a Lyapunov function: $\langle Q'(\theta), \theta - \theta_* \rangle \ge H_0(\theta - \theta_*)^2$ with $H_0 = \inf Q'' > 0$. With $\sum a_t = \infty$, $\sum a_t^2 < \infty$, the Robbins–Monro/Kushner–Yin theorem [RM51; KY03] gives $\theta_t \to \theta_*$ a.s.

Tracking. With constant step a and a drift $\left|\theta_t^* - \theta_{t-1}^*\right| \le \delta$, let $r_t = \mathbb{E}(\theta_t - \theta_t^*)^2$. Expanding the projected recursion and using $\left\langle Q'(\theta_{t-1}), \theta_{t-1} - \theta_{t-1}^* \right\rangle \ge H_0 r_{t-1}$, the bounded gradient variance, and $2ab \le \rho a^2 + \rho^{-1} b^2$ for the drift cross-term,

$$r_t \le (1 - 2aH_0 + a^2 L_0)\, r_{t-1} + a^2 \sigma_g^2 + \delta^2/(aH_0) + o(\delta^2),$$

with $L_0 = \sup Q''$. For a small the contraction factor is $1 - aH_0 < 1$, and summing the geometric series, $\limsup_t r_t \le (a\sigma_g^2/H_0) + \delta^2/(aH_0^2) = O(a + \delta^2/a)$. □

B.5 Proof of Theorem 8 (Koopman representation)

(1) Under the local-level model $\xi_{t+1} = \xi_t + \zeta_{t+1}$ with $\mathbb{E}[\zeta_{t+1} \mid \mathcal{F}_t] = 0$, the minimum-MSE predictor is $\mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] = \mathbb{E}[\xi_t \mid \mathcal{F}_t] = \hat{\xi}_t$, the filtered state, which by Theorem 6 obeys the WES recursion.

(2) Unrolling $\hat{\xi}_t = (1 - \theta)\hat{\xi}_{t-1} + \theta w_t$ gives $\hat{\xi}_t = \theta \sum_{k \ge 0}(1 - \theta)^k w_{t-k} = \theta\big(\sum_{k \ge 0}(1 - \theta)^k L^k\big) w_t = \theta(I - (1 - \theta)L)^{-1} w_t$. The Neumann series converges in operator norm since $\|(1 - \theta)L\| = 1 - \theta < 1$, and the resolvent has a single pole at the lag-operator eigenvalue $1 - \theta$.

(3) The innovation representation $w_t = \xi_t + \varepsilon_t$ with random-walk ξ_t has, in the lifted (Koopman) picture, the single nontrivial dynamic eigenvalue $1 - \theta$ governing the geometric memory of $\hat{\xi}_t$; EDMD applied to $\{w_t\}$ converges to the Koopman operator [Koo31; Mez05; KM18], whose leading eigenvalue is therefore $1 - \theta$, and the induced spectral estimator $\hat{\theta}^{\mathrm{DMD}} = 1 - \hat{\lambda}_{\max}$ solves the same population moment condition as $\hat{\theta}_T$, hence is asymptotically equivalent. □

C Proofs for Section 10

C.1 Proof of Theorem 9 (Banach/L^p meta-theorem)

Let B be 2-uniformly smooth with modulus S and 2-uniformly convex with modulus C, and recall (21). Write $u_t = (1-\theta)u_{t-1} + \theta w_t$, $w_t = u_{t-1} + F_t$, so $u_t = \theta \sum_{k\geq 0}(1-\theta)^k w_{t-k}$ and the centered filter $\tilde{u}_t := u_t - \mathbb{E}u_t = \theta \sum_{k\geq 0}(1-\theta)^k F_{t-k}$ is a convergent sum of independent mean-zero increments.

(1) Mean-stationarity and geometric decay. Taking expectations, $\mathbb{E}w_t = \mathbb{E}u_{t-1}$ is constant, so the filter is mean-stationary. The increments $d_k := \theta(1-\theta)^k F_{t-k}$ form a martingale-difference array; by Pisier's inequality in 2-uniformly smooth spaces [Pis75], $\mathbb{E}\big\|\sum_k d_k\big\|^2 \leq S^2 \sum_k \mathbb{E}\,\|d_k\|^2 = S^2\theta^2 \sum_{k\geq 0}(1-\theta)^{2k}\mathbb{E}\,\|F\|^2 = S^2\theta^2\mathbb{E}\,\|F\|^2/(1-(1-\theta)^2)$. The one-step forecast residual $e_t = u_t - w_{t+1}$ differs from $-F_{t+1}$ by the centered filter discrepancy, and the same inequality bounds its second moment with geometric rate $(1-\theta)^2$, giving part (1).

(2) Strict convexity and consistency. The population criterion is $Q(\theta) = \mathbb{E}\,\big\|u_t^\theta - w_{t+1}\big\|^2$. Writing $u_t^\theta - w_{t+1}$ affinely in θ and applying the lower inequality in (21) to second differences, $Q(\frac{\theta+\theta'}{2}) \leq \frac{1}{2}Q(\theta) + \frac{1}{2}Q(\theta') - \frac{C^{-2}}{4}\,\mathbb{E}\,\Big\|u_t^\theta - u_t^{\theta'}\Big\|^2$, so Q is strongly midpoint-convex with modulus at least C^{-2} times the Hilbert value; being continuous it is strictly convex with a well-separated minimizer θ_*. The empirical criterion converges uniformly to Q on compacta by the mixing in part (1), and the argmin theorem [NM94; Vaa98] yields $\hat{\theta}_T \xrightarrow{p} \theta_*$. Both bounds carry the constants S^2, C^2, equal to 1 iff the space is Hilbert ($p = 2$), in which case the upper and lower inequalities are the parallelogram identity and the proof reduces to Theorem 2. □

C.2 Proof of Theorem 10 (Berwald–Hadamard contraction)

Let (M, F) be forward-complete, simply connected, reversible Berwald with flag curvature $K \leq 0$. On a Berwald space the Chern connection coefficients $\Gamma^i_{jk}(x)$ are independent of the direction v [BCS00, Ch. 10], so geodesics solve $\ddot{\gamma}^i + \Gamma^i_{jk}(\gamma)\dot{\gamma}^j\dot{\gamma}^k = 0$ with an affine (linear) connection, and the exponential map and Jacobi equation are those of that connection. The Jacobi equation $\nabla_{\dot\gamma}\nabla_{\dot\gamma}J + R(J,\dot\gamma)\dot\gamma = 0$ with flag curvature $K \leq 0$ gives non-decreasing $\|J\|$ along geodesics, i.e. geodesics spread no faster than in the flat case; for reversible F this is equivalent to Busemann nonpositive curvature, the convexity of $t \mapsto d_F(\gamma_1(t), \gamma_2(t))$ for any geodesics γ_1, γ_2 [Egl97; Oht09]. Fix ν and let $\gamma_i : [0,1] \to M$ be the geodesic from μ_i to ν, so $G_\nu(\mu_i) = \gamma_i(\theta)$ and $\gamma_i(1) = \nu$. Convexity of $\psi(t) := d_F(\gamma_1(t), \gamma_2(t))$ with $\psi(1) = 0$ gives, for $t = \theta \in [0,1]$,

$$\psi(\theta) = \psi((1-\theta)\cdot 0 + \theta \cdot 1) \leq (1-\theta)\psi(0) + \theta\,\psi(1) = (1-\theta)\,d_F(\mu_1, \mu_2),$$

which is the contraction. Given the λ-contraction with $\lambda = 1-\theta$, Steps 0–3 of Appendix A.7 apply verbatim (with $\eta = 0$ in the Berwald case and the uniform-convexity/smoothness constants entering the CLT variance through Theorem 9), delivering stationarity, geometric mixing, consistency, and the CLT. □

C.3 Proof of Proposition 7 (η as anisotropy)

Let ι be the normal-coordinate chart of F on K. The geodesic-interpolation map in coordinates is $G_\nu(\mu) = \exp_\mu(\theta \log_\mu \nu)$; its deviation from the chart affine combination $(1-\theta)\iota(\mu) + \theta\iota(\nu)$ is, to second order, governed by the connection coefficients Γ, which on a Finsler manifold split into a Riemannian part and a part proportional to the Cartan tensor A [BCS00, Ch. 2]. The Riemannian part maps geodesics to lines in normal coordinates up to curvature (already absorbed in the bi-Lipschitz constant L), while the A-part is the genuine non-affineness. Quantitatively, [Oht09] shows

the tangent-norm uniform convexity and smoothness constants satisfy $\max\{S-1, C-1\} \asymp \sup_K \|A\|$, and the second-order Taylor remainder of G_ν in the chart is bounded above and below by a multiple of this gap times $\theta(1-\theta)\,\|\iota(\mu) - \iota(\mu')\|$. Comparing with condition (B2) of Assumption 2 identifies $\eta \asymp \max\{S-1, C-1\} \asymp \sup_K \|A\|$, which is 0 precisely when $A \equiv 0$, i.e. in the Riemannian case. For $\mathcal{W}_p(\mathbb{R}) \cong L^p((0,1))$ the constants are $C = 1$, $S = \sqrt{p-1}$ [BCL94], giving $\eta \asymp \sqrt{p-1} - 1$. □