ONLINE CORESETS FOR CONDITIONAL FLOW MATCHING: LEARNT SUPPORT AND TRANSPORT TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow matching trains generative models by learning a velocity field that transports a source law to a data law. In the conditional case, each data endpoint fixes an interpolation path and its velocity target, so the rule deciding which endpoints a streaming learner retains also decides the population on which the generator trains. Keeping every endpoint wastes memory under stationarity and can preserve obsolete regimes under drift, while a fixed or marginal buffer can erase rare events and weak conditional modes. We cast endpoint memory as a condition-wise model-order problem and introduce Online Memory Compression and Identification with Maximum Mean Discrepancy (OMCI–MMD). For each condition, OMCI–MMD approximates the RKHS mean of the endpoint law with a sparse weighted subset of observed endpoints and selects the first support size whose MMD falls below the sampling floor, retaining only structure the data can resolve. For stationary streams, this floor scales as $n_c(t)^{-1/2}$ up to logarithms; the computed order never understates the oracle observed-support order, a small path slack yields a conditional upper bracket, and Frank–Wolfe reaches any MMD radius ε with $O(\varepsilon^{-2})$ atoms. Under drift, a recent-window path separates sampling, stale-law, candidate-availability, and numerical error. We characterize when the MMD bound transfers to Wasserstein, Sinkhorn, and endpoint-indexed flow-matching losses, which may refine the endpoint law only after its cardinality is frozen. Experiments recover support-size plateaus, track a regime change, expose a conditional failure hidden by exact marginal agreement, and show that aggregating at statistical resolution avoids both replaying stale modes and dropping current ones, two upstream sources of hallucination.

1 INTRODUCTION

Generative modeling asks us to sample from a complex data law (distribution), often after conditioning on a regime or class. Score matching uses log-density gradients (Hyvärinen, 2005); denoising supplies a surrogate (Vincent, 2011); score-based diffusion learns the corresponding reverse-time dynamics (Song et al., 2021). Flow Matching instead trains a velocity field along a prescribed source-to-data probability path (Lipman et al., 2023; Liu et al., 2023). Optimal-transport conditional Flow Matching (OT-CFM) changes how source and endpoint samples define that path (Tong et al., 2024); C^2OT shows that condition-blind minibatch transport can distort the conditional training prior (Cheng and Schwing, 2025). Both lines suppose oracle access to endpoint population distribution. Static or marginal endpoint summaries create a separate failure: they can erase rare conditional structure even when the marginal endpoint law appears accurate. This is one source of the fabled "confident but wrong" phenomenon of generative modeling. Sequential retention adds a second problem because full-history memory grows linearly and becomes stale after distributional change.

We therefore design memory at the endpoint layer. Retaining more observations does not solve both failures: after drift, full history approximates a stale mixture, while a fixed or marginal buffer can delete condition-specific modes. We seek the smallest active endpoint bank that matches the current conditional law up to its *sampling resolution*, a notion we define as the support size. In a stationary finite-effective population, this learnt support size should eventually flatline; when new conditional structure becomes statistically resolvable, it should grow.

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

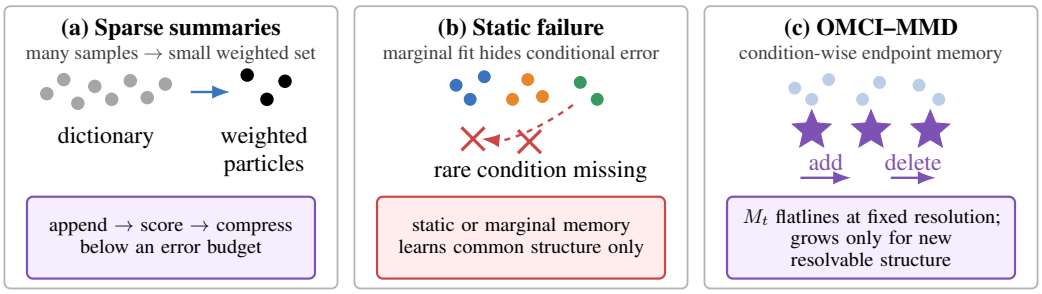


Figure 1: **Endpoint memory as condition-wise model order.** **(a)** Monte Carlo and Bayesian inference methods often replace a growing empirical measure by a small weighted summary. **(b)** A static or marginal summary can match the global endpoint law while deleting rare conditional structure. **(c)** OMCI–MMD constructs a sparse path for each condition and changes support only when its MMD error crosses the sampling radius.

Recent rectified-flow analyses establish order-optimal rates after the endpoint law has been supplied (Sahoo et al., 2026). Stein discrepancies assess a fixed target (Gorham and Mackey, 2015; Liu et al., 2016); Stein points select representative samples (Chen et al., 2018), while kernelized thinning compresses a fixed chain (Riabiz et al., 2022; Koppel et al., 2024). These methods all pre-suppose access to the empirical endpoint law that supplies paths and labels, which is assumed identical to the population. Thus, we study the upstream question that this fixed-law approaches leave open: which conditional endpoint law should a streaming learner use now?

We use the maximum mean discrepancy (MMD) (Gretton et al., 2012), an integral probability metric that is computable in closed form under kernel parameterizations (Berlinet and Thomas-Agnan, 2004), as this turns endpoint compression into a Hilbert-valued sparse approximation problem. This choice is buttressed by the observation that Wasserstein critics, Sinkhorn first-variation potentials, and endpoint-indexed flow losses inherit the endpoint certificates of a RKHS-based method when the RKHS approximates them with a controlled error, as we establish in Section 4; however, this happens quite often, and is akin to “error-bound conditions" in numerical optimization that indicate when non-convex landscape is approximately benign (Polyak, 1963b; Lojasiewicz, 1963; Chen and Tseng, 2003). Witness transfer defines which downstream quantities inherit the resolution and memory selected by the MMD crossing. Thus, our main contributions are the following:

- **Conditional model order.** We define the conditional model order as the first crossing of a sparse-MMD curve below a condition-specific resolution (Sec. 2).
- **OMCI–MMD.** We introduce OMCI–MMD, whose fully corrective path returns both a weighted endpoint law and the memory–error curve needed to locate that crossing (Sec. 3).
- **Stationary crossing guarantees.** Feasibility prevents the computed path from understating the oracle observed-support order at the same radius, and a small oracle path slack yields a conditional upper bracket (Sec. 4).
- **Tracking under drift.** A staleness envelope separates the sampling, stale-law, and candidate-availability error terms and recovers the error floor under non-stationarity (Sec. 4).
- **Witness transfer.** We give conditions under which Wasserstein, Sinkhorn, and endpoint-indexed flow losses inherit MMD certificates, together with a plug-in refinement principle at a frozen cardinality (Sec. 4).
- **Experiments.** We recover finite-support plateaus, track a programmed regime change, expose a conditional failure hidden by marginal agreement, and test how the selected endpoint law changes neural-teacher fidelity (Sec. 5).

2 Conditional Endpoint Memory

We formalize endpoint memory as a per-condition sparse-MMD model-selection problem. We first define the endpoint-induced flow objective, then the oracle crossing that an online algorithm must estimate. Exact marginal compression still cannot protect a conditional flow objective; Appendix A places this distinction in the related literature.

For the rectified path used by Flow Matching (Liu et al., 2023; Lipman et al., 2023) and optimal-transport conditional Flow Matching (OT-CFM) (Tong et al., 2024), let x_0 denote a source draw, Y the data endpoint, and $s \in [0,1]$ the interpolation time:

$$X_s = (1-s)x_0 + sY, \qquad u_s^\star = Y - x_0, \tag{1}$$

In this path, the data endpoint Y changes both the regression input and its velocity label. We therefore hold the path rule and flow architecture fixed and study how a streaming weighted endpoint law replaces the full conditional population. See Appendix A for additional positioning.

Let $\mathcal{Y}$ denote the endpoint space, and let $k(y,y') = \langle \phi(y), \phi(y') \rangle_{\mathcal{H}}$ be a kernel with feature map ϕ into RKHS $\mathcal{H}$, where we assume $\|\phi(y)\|_{\mathcal{H}} \le K$. Samples $(Y_t, c_t) \in \mathcal{Y} \times \mathcal{C}$ arrive sequentially, where $\mathcal{C}$ is a finite condition set, $\mathbf{1}\{\cdot\}$ denotes an indicator, and δ_y denotes a Dirac mass at y. For each condition with $n_c(t) > 0$, the count, empirical law, and empirical kernel mean are

$$n_c(t) = \sum_{i \le t} \mathbf{1}\{c_i = c\}, \qquad \widehat{\nu}_t^c = \frac{1}{n_c(t)} \sum_{i \le t : c_i = c} \delta_{Y_i}, \qquad \widehat{\mu}_t^c = \frac{1}{n_c(t)} \sum_{i \le t : c_i = c} \phi(Y_i). \tag{2}$$

For any probability law R, write $\mu_R := \int \phi(y)\, dR(y)$ and $Rf := \int f(y)\, dR(y)$. Maximum mean discrepancy (MMD) satisfies, for probability laws P and Q,

$$\mathrm{MMD}_k(P,Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}} = \sup_{\|f\|_{\mathcal{H}} \le 1} |Pf - Qf|, \tag{3}$$

so a weighted memory simultaneously approximates the empirical mean and bounds expectation error over the RKHS unit ball (Gretton et al., 2012; Sejdinovic et al., 2013).

For a retained index set A^c and simplex weights $a \in \mathsf{S}(A^c) := \{a \ge 0 : \sum_{j \in A^c} a_j = 1\}$, write $\nu_{A,a}^c = \sum_{j \in A^c} a_j \delta_{Y_j}$ and

$$F_t^c(A^c, a) = \frac{1}{2} \left\| \sum_{j \in A^c} a_j \phi(Y_j) - \widehat{\mu}_t^c \right\|_{\mathcal{H}}^2. \tag{4}$$

The best observed-support error at cardinality m and its norm-scale curve are

$$F_{t,m}^{c,\star} = \inf_{\substack{A^c \subseteq \{i \le t : c_i = c\},\ |A^c| \le m \\ a \in \mathsf{S}(A^c)}} F_t^c(A^c, a), \qquad e_{t,m}^{c,\star} = \sqrt{2 F_{t,m}^{c,\star}}. \tag{5}$$

Fix a confidence level $\delta \in (0,1)$, let $C_{\mathrm{H}} > 0$ denote the universal concentration constant used in Lemma 3, and choose a predeclared safety factor $\kappa \ge 1$. Rather than prescribe m, we compare this curve with the condition-specific sampling radius

$$E_t^c(\delta) = C_{\mathrm{H}} K \sqrt{\frac{\log(4|\mathcal{C}|t^2/\delta)}{n_c(t)}}, \qquad \varepsilon_t^c = \kappa E_t^c(\delta), \qquad m_t^{\star,c}(\rho) = \min\{m : e_{t,m}^{c,\star} \le \rho\}, \tag{6}$$

with the convention $\min \varnothing := +\infty$. The oracle model order is $m_t^{\star,c} = m_t^{\star,c}(\varepsilon_t^c)$. Below this crossing, the memory misses structure visible at the declared resolution; beyond it, additional atoms fit differences smaller than that resolution.

Proposition 1 (Marginal agreement can hide conditional failure)**.** *Write P^c and Q^c for the endpoint laws conditional on $C = c$, and let W_1 denote the 1-Wasserstein distance. Let $C \in \{0,1\}$ be uniform. Under P, set $Y = -1$ for $C = 0$ and $Y = 1$ for $C = 1$; under Q, swap these conditionals. Then $P_Y = Q_Y$, while $\max_c W_1(P^c, Q^c) = 2$.*

Thus a condition-agnostic coreset can be exact marginally while supplying the wrong endpoint and velocity law to every condition. OMCI–MMD therefore constructs one sparse curve and one sampling radius per condition. The crossing in (6) is still combinatorial; Section 3 replaces it by a computable fully corrective path. Before continuing, we now introduce a fully worked toy numerical instantiation.

Toy Example. We begin with a toy example with ten fixed endpoints in $\mathbb{R}^2$. Before time $t_0 = 2200$, six endpoints carry mass 0.16 each and four weak endpoints carry mass 0.01 each. At t_0, two central weak endpoints rise to mass 0.145, the original six fall to 0.115, and two outer endpoints remain weak. With a recent window of $L = 500$, the intended statistically resolvable order is therefore six

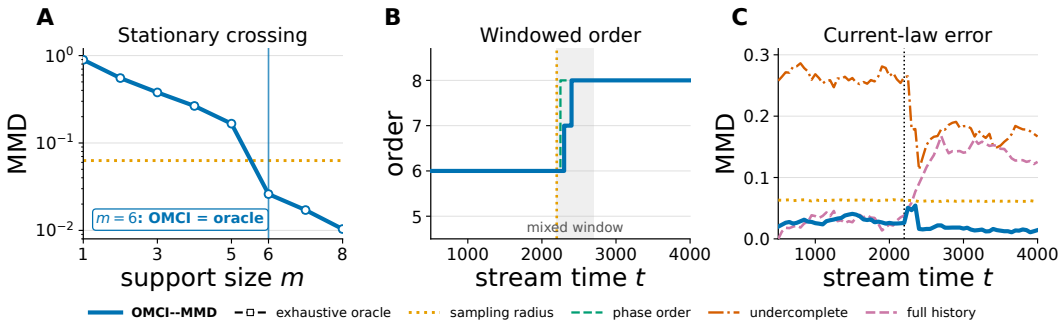


Figure 2: The running toy links model-order identification to drift. **A:** OMCI and exhaustive search cross the stationary radius at six atoms. **B:** after $t_0 = 2200$, the target drifts, and OMCI detects the drift in its windowed estimate transitioning to eight required points. **C:** OMCI recovers its MMD below the sampling radius, whereas undercomplete and full-history memories retain information that harms their representation after drift.

before the change and eight after it. This instance asks three questions: does the stationary rule stop at six rather than fit the four weak atoms below the sampling floor; does a windowed rule move to eight after the change; and does full-history memory lag because it averages the two phases? Section 3 works through the first crossing and uses its failure after t_0 to derive the nonstationary algorithm.

We visualize the results in Figure 2. Finite support lets us compute the sparse-MMD path, the exhaustive observed-support oracle, and the current-window discrepancy exactly. We use a radial basis function (RBF) kernel of bandwidth 1.15, a window of $L = 500$, an independently calibrated 95% sampling floor, and fully corrective weights. The support-selection stage (Stage I) sees only the MMD path and its radius; teacher, Wasserstein, and Sinkhorn errors are evaluated after the support size is frozen. Appendix D gives the protocol and Appendix E specifies the hindsight-only memory–accuracy frontier.

At $t = 2000$, MMD falls from 0.1665 at $m = 5$ to 0.0260 at $m = 6$, below radius 0.0629, so OMCI and exhaustive search both select six points. After the change, the windowed order passes through seven and stabilizes at eight. OMCI attains current-window MMD 0.0143; the six-atom law attains 0.1669 and full history 0.1255, both above radius 0.0615. Across forty independent post-change windows, exact scans again select eight with no observed-support oracle miss. For analytic Gaussian-path teachers, the selected law reduces teacher-field mean-squared error from 15.51 to 0.850 and paired terminal ordinary differential equation (ODE) rollout error from 1.728 to 0.504 relative to the undercomplete law. It also lowers median MMD from 0.194 to 0.040, W_1 from 0.182 to 0.036, and Sinkhorn divergence from 0.0278 to 0.0010; finite-RKHS interpolants verify the corresponding witness bounds. Appendix E.3 gives the complementary continuous-mixture failure: generic RBF-MMD can cross its radius without minimizing Sinkhorn divergence, whereas a frozen witness bank improves the median but not every stream. These results separate selection from refinement. OMCI freezes the statistically justified cardinality; WCSL, Sinkhorn–CO2, EMS, or direct Sinkhorn descent can then optimize transport geometry at that fixed size.

3 Online Memory Compression and Identification with MMD

Section 2 defines model order through the best m-atom MMD curve, but evaluating that curve exactly would require a combinatorial search at every support size. OMCI–MMD instead builds a nested, fully corrective conditional-gradient path. Frank–Wolfe supplies linear-minimization geometry (Jaggi, 2013), which we apply via kernel herding to a mean embedding (Bach et al., 2012; Lacoste-Julien et al., 2015). Monte Carlo thinning and Hilbert coresets are clear inspiration here (Riabiz et al., 2022; Campbell and Broderick, 2019). OMCI–MMD changes the stopping problem: the path must reveal the first statistically calibrated crossing rather than return only one fixed-budget approximation.

Let $\mathcal{J}_t^c \subseteq \{i \le t : c_i = c\}$ be the declared candidate indices at checkpoint (t, c), and let $1 \le M_{\max} \le |\mathcal{J}_t^c|$ be the largest reported support budget. At budget m, the nested dictionary $A_{t,m}^c$ contains the selected indices, $a_{t,m}^c \in \mathsf{S}(A_{t,m}^c)$ are their fully corrected weights, $q_{t,m}^c = \sum_{j \in A_{t,m}^c} a_{t,m,j}^c \phi(Y_j)$ is the coreset mean, and $r_{t,m}^c = \widehat{\mu}_t^c - q_{t,m}^c$ is its residual. The linear minimization oracle selects the

Algorithm 1 Stationary OMCI–MMD: complete CrossingReadout path

Require: $\widehat{\mu}_t^c$, radius ε_t^c, candidates $\mathcal{J}_t^c$, maximum size $M_{\max}$
Choose $i_1 \in \arg\min_{i \in \mathcal{J}_t^c} F_t^c(\{i\}, 1)$; set $A_{t,1}^c = \{i_1\}$ and $\widehat{m}_t^c \leftarrow +\infty$.
for $m = 1, \ldots, M_{\max}$ **do**
  Fully correct $a_{t,m}^c$; record $\widehat{e}_{t,m}^c$ and $\widehat{F}_{t,m}^c = \frac{1}{2}(\widehat{e}_{t,m}^c)^2$.
  if $\widehat{m}_t^c = +\infty$ and $\widehat{e}_{t,m}^c \le \varepsilon_t^c$ **then**
    Set $\widehat{m}_t^c \leftarrow m$ and freeze the weighted law at this first crossing.
  end if
  if $m < M_{\max}$ **then**
    Compute $r_{t,m}^c$ and select $\widehat{i}_{t,m+1}^c$ by (7).
    if the exact gap (23) equals zero **then**
      Use Lemma 1; copy the terminal law and error through budgets $m+1{:}M_{\max}$; stop.
    else
      Form $A_{t,m+1}^c$ and fully correct its weights by (8).
    end if
  end if
end for
Return the complete budget-indexed path, its first crossing, and the frozen weighted law.

candidate with largest residual correlation,

$$\widehat{i}_{t,m+1}^c \in \arg\max_{i \in \mathcal{J}_t^c} \langle r_{t,m}^c, \phi(Y_i) \rangle_{\mathcal{H}}, \tag{7}$$

$$A_{t,m+1}^c := A_{t,m}^c \cup \{\widehat{i}_{t,m+1}^c\}, \qquad a_{t,m+1}^c \in \arg\min_{a \in \mathsf{S}(A_{t,m+1}^c)} F_t^c(A_{t,m+1}^c, a). \tag{8}$$

The residual points toward what memory still misses, while full correction makes the reported error depend on the selected dictionary rather than a greedy step-size schedule. Appendix B.1 derives (7) from the directional derivative in (18), records full correction in (21), proves the zero-gap optimality test (24), and fixes the complete checkpoint-path notation. The path records $\widehat{e}_{t,m}^c = \|q_{t,m}^c - \widehat{\mu}_t^c\|_{\mathcal{H}}$ and, with $\min \varnothing := +\infty$, reports

$$\widehat{m}_t^c = \min\{1 \le m \le M_{\max} : \widehat{e}_{t,m}^c \le \varepsilon_t^c\}. \tag{9}$$

The index m is a selected-dictionary budget. Full correction may assign zero weight to some selected atom, so the frozen law stores at most m nonzero atoms; implementations report that nonzero count and may prune zero weights after readout. The routine nevertheless evaluates every budget or copies a terminal law after proving zero gap in the declared range needed for the path slack in Theorem 1. Stopping at the first crossing alone would not construct the object used by that theorem.

Worked stationary crossing. In the example from Sec. 2, with pre-change checkpoint $t = 2000$, the independently calibrated radius is 0.0629. Five atoms leave MMD error 0.1665, so the data still resolve missing structure; six atoms reduce the error to 0.0260, so Algorithm 1 records $\widehat{m}_{2000}^c = 6$. Exhaustive subset search gives the same curve and crossing. OMCI retains the six substantial modes without spending memory on four 0.01-mass atoms below the current sampling resolution. The same routine fails after the designed change if it continues to target the full-history mean. At $t = 4000$, that historical law remains 0.1255 MMD units from the current window, more than twice the 0.0615 radius. A perfectly optimized stationary coreset would therefore compress the wrong law. The nonstationary repair changes the target estimator, not the sparse-path geometry: it expires stale observations, forms a recent-window mean, and enlarges the radius by a declared staleness envelope.

For window length L, define the active condition-c indices, count, and empirical mean by

$$I_{t,L}^c := \{s : \max(1, t-L+1) \le s \le t,\ c_s = c\}, \quad n_{t,L}^c := |I_{t,L}^c|, \quad \widehat{\mu}_{t,L}^c := \frac{1}{n_{t,L}^c} \sum_{s \in I_{t,L}^c} \phi(Y_s). \tag{10}$$

whenever $n_{t,L}^c > 0$. Let $E_{t,L}^c$ denote the window sampling term specified in (73); let $\overline{S}_{t,L}^c$ be a declared upper envelope on stale-law bias; set $\varepsilon_{t,L}^c := \kappa(E_{t,L}^c + \overline{S}_{t,L}^c)$; and declare a nonempty set $\mathcal{J}_{t,L}^c \subseteq I_{t,L}^c$ with $1 \le M_{\max} \le |\mathcal{J}_{t,L}^c|$. Appendix B.7 provides further details.

The analysis uses each design choice directly. The nested path supplies one error per support budget; full correction removes step-size artifacts; the declared candidate set exposes oracle miss; and the

Algorithm 2 Windowed OMCI–MMD for a drifting stream

Require: stream (Y_t, c_t), window length L, δ, κ, envelopes $\overline{S}^c_{t,L}$, candidates $\mathcal{J}^c_{t,L}$, $M_{\max}$

for each arrival (Y_t, c_t) **do**
  Insert (Y_t, c_t) and remove every expired pair (Y_s, c_s) with $s < t - L + 1$.
  Update the queue, count, and mean for every condition in $\{c_t\} \cup \{c_s : s \text{ expired at } t\}$.
  if t is a reporting checkpoint **then**
    for each reported c with $n^c_{t,L} > 0$ **do**
      Set $\varepsilon^c_{t,L} = \kappa(E^c_{t,L} + \overline{S}^c_{t,L})$ and declare $\mathcal{J}^c_{t,L}$.
      Invoke Algorithm 1 with $(\widehat{\mu}^c_{t,L}, \varepsilon^c_{t,L}, \mathcal{J}^c_{t,L}, M_{\max})$.
      Relabel the returned crossing and path as $\widehat{m}^c_{t,L}$ and $m \mapsto (\widehat{F}^c_{t,L,m}, \widehat{e}^c_{t,L,m})$.
      Return $\widehat{m}^c_{t,L}$, the frozen weighted window law, the complete path, and the reported pool and oracle-miss terms.
    end for
  end if
end for

path slack (30) enters the sandwich (31) that yields the stationary crossing bracket. Appendix D details the online admission/deletion implementation, while Appendix B.7 defines the full-window and candidate-restricted paths used in the nonstationary regime (Algorithm 2). Next, we shift to analyzing the stationary crossing, its failure under drift, and the additional witness assumptions needed downstream.

4 Theoretical Analysis

In this section, we establish when the first sparse-MMD crossing identifies a support size based upon Pinelis' Hilbert-space concentration inequality, (Pinelis, 1994), and a specific usage of Frank–Wolfe rate analysis (Jaggi, 2013). These prior works do not select the budget obtained by inverting a path, which is our specific analytical contribution, under stationarity and a recent-window target under drift. Then, we state the witness assumptions needed for transport and neural-risk statements. Appendix B develops the model-order proof, while Appendix C develops witness transfer and neural risk.

Stationary streams. Appendix B.5 conditions on the realized labels and assumes independent draws from a fixed endpoint law P^c for each condition c. Pinelis' Hilbert-space concentration inequality (Pinelis, 1994) then yields the simultaneous event (39). For a predeclared path prefix $\mathcal{I} = \{1, \dots, m_{\max}\}$, define the norm-scale path slack

$$s^c_{t,\mathcal{I}} := \sup_{m \in \mathcal{I}} [\widehat{e}^c_{t,m} - e^{c,\star}_{t,m}]_+, \qquad [x]_+ := \max\{x, 0\}. \tag{11}$$

The oracle and correction steps in (18) and (21) are standard; the departure lies in the complete-path sandwich (31) and its support-size inversion (33).

Assumptions and stationary departure. Appendix B.4 fixes the stationary law, bounded kernel, confidence allocation, checkpoint rule, candidate universe, and solver tolerances before inverting the path. Concentration controls the empirical kernel mean; conditional-gradient analysis controls approximation at each prescribed budget. The additional step is pathwise: the sandwich (31) compares the complete computed and oracle curves, and (33) converts that comparison into a support-size bracket. The selected support may depend on the same sample because the sandwich holds uniformly over the declared path; data-dependent kernel or radius tuning still requires predictability, splitting, or simultaneous control.

Theorem 1 (Stability of the support-size crossing). *(Informal) Set $e^{c,\star}_{t,0} := +\infty$. If $m^{\star,c}_t(\varepsilon^c_t)$ and $m^{\star,c}_t(\varepsilon^c_t - s^c_{t,\mathcal{I}})$ are finite and belong to $\mathcal{I}$, and $s^c_{t,\mathcal{I}} < \varepsilon^c_t$, then*

$$m^{\star,c}_t(\varepsilon^c_t) \le \widehat{m}^c_t \le m^{\star,c}_t(\varepsilon^c_t - s^c_{t,\mathcal{I}}). \tag{12}$$

Moreover, if $e^{c,\star}_{t,m-1} > \varepsilon^c_t$ and $e^{c,\star}_{t,m} \le \varepsilon^c_t - s^c_{t,\mathcal{I}}$, then $\widehat{m}^c_t = m$.

Feasibility proves the lower bracket; only an exhaustive search or an independent slack bound proves the upper bracket. Thus a flat oracle curve supports an interval, not an invented integer. The separate

Frank–Wolfe argument (Jaggi, 2013; Lacoste-Julien et al., 2015) proves the fixed-radius bound (60). At the Monte Carlo radius $\widetilde{O}(n_c(t)^{-1/2})$, that inherited bound still permits $\widetilde{O}(n_c(t))$ atoms; it does not prove generic compression. Hilbert coresets (Campbell and Broderick, 2019) and kernelized thinning (Riabiz et al., 2022; Koppel et al., 2024) optimize fixed-target summaries. Our departure is the statistically calibrated *first crossing* of a complete condition-wise path. Appendix B.6 proves the rate interpretation and finite-support stabilization conditions. Drift breaks the target itself: exact compression of full history can still approximate the wrong current law, as detailed in Figure 2(C), and formalized next.

Nonstationary streams. For every time-condition pair, Appendix B.7 defines an endpoint law P_t^c and mean μ_t^c before forming the population-window and empirical-window means in (72). The identity (90) separates sampling fluctuation from stale-law bias. A current-law bound therefore requires a deterministic or separately calibrated envelope $S_{t,L}^c \le \overline{S}_{t,L}^c$ and the radius in (75); the full-window oracle curve (76) defines the target order, and (81) gives exact-scan reachability. This couples the classical local-estimation bias–variance tradeoff (Dahlhaus, 1997; Basseville and Nikiforov, 1993) to support-size selection.

Finite pools add errors that windowing cannot erase. The Frank–Wolfe recurrence is standard; the departures are the explicit availability split (99), the regret-to-search interface (110), and the current-law composition in (135) and (137). Proposition 6 and Corollary 3 retain the resulting path and crossing terms separately. Related RKHS-based supervised learning seeks to track dynamic regret when governed by comparator path length and projection tolerance (Bedi et al., 2020), but these quantities neither define a conditional endpoint law nor guarantee uniform residual-gain ranking on the searched pool. We invoke it only through the explicit interface (110); the availability term remains because no predictor can select an endpoint that the pool omitted. A valid current-law crossing still bounds only the functions that the kernel represents.

Current-law conclusion. Path regret alone does not show that the returned law tracks P_t^c. Appendix B.11 separates the two required steps: candidate and solver errors determine whether the path crosses by budget M, while sampling, stale-law bias, and output feasibility determine the current-law error after a crossing. Equation (135) bounds the fraction of checkpoints at which either requirement fails. DynaPOLK controls predictive dynamic regret for a drifting RKHS predictor (Bedi et al., 2020); local-stationarity and concept-drift analyses motivate the window bias–variance split (Dahlhaus, 1997; Gama et al., 2014). Neither supplies candidate availability, path-wide crossing separation, or stale-window control, which Proposition 7 keeps explicit.

Witness transfer and neural risk. Appendix C.1 defines the downstream witness discrepancy and its RKHS approximation profile in (138). The approximation inequality (140) is standard; the departures are its learnt-crossing specialization (143)–(144), transport composition (162), and neural-risk substitution (235). For Sinkhorn divergence, (176) supplies the first-variation identity and (184) evaluates it at the selected cardinality. Universality alone does not make the bound useful: the required RKHS norm can grow until the bound becomes vacuous. Fixed-law rectified-flow rates analyze neural estimation after one supplies an endpoint law (Sahoo et al., 2026); they do not bound replacement by a sparse, stale, or condition-mismatched law. Appendix C.4 defines endpoint-indexed Flow-Matching risk in (189)–(190), separates endpoint-memory from path or coupling error in (206), and proves value transfer in (230) without a Polyak–Łojasiewicz (PL) condition. Gradient transfer uses PL only under the additional assumptions in (231).

Once OMCI freezes the cardinality, WCSL optimizes a fixed-size Sinkhorn-loss coreset (Yin et al., 2025); CO2 builds a divergence-adapted MMD for fixed-target selection (Kokot and Luedtke, 2025); and EMS optimizes weighted free-support representatives (Yin et al., 2026). Compressed Online Sinkhorn instead compresses the representation of a dual-potential (Wang et al., 2023). Proposition 10 formalizes the complementary select–freeze–refine step in (185)–(186). Next, we shift to experimenting with this endpoint-law replacement in neural teachers.

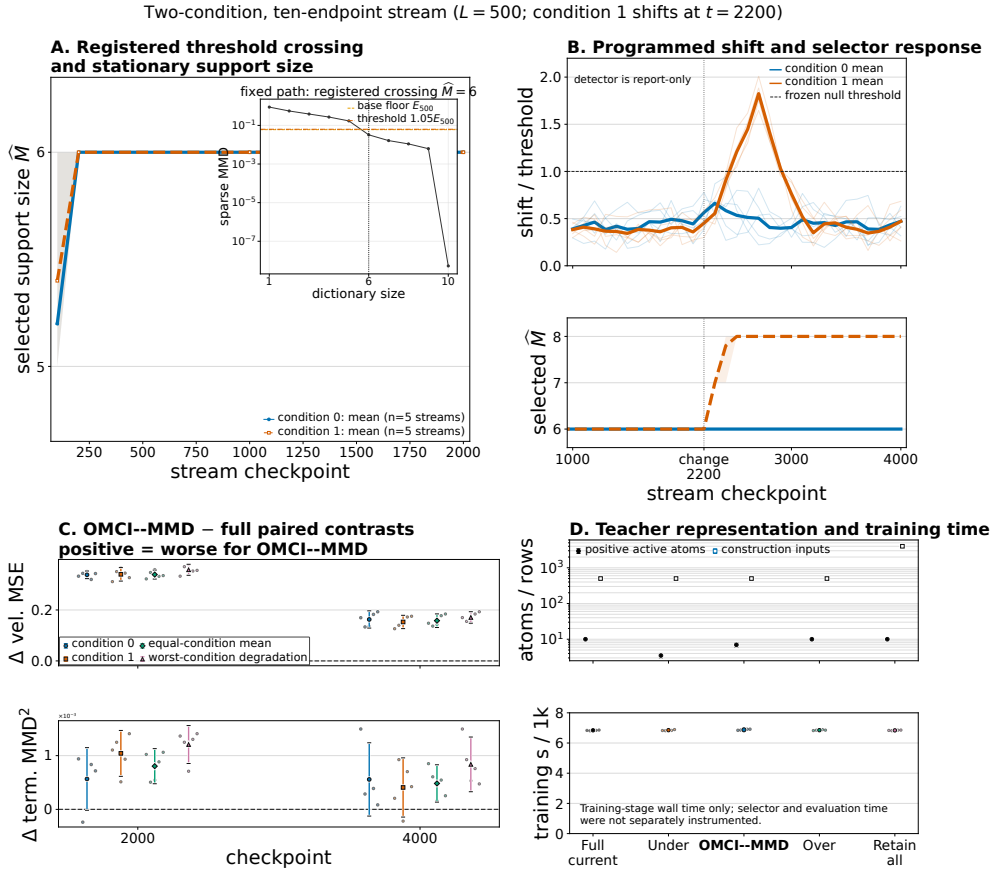


Figure 3: **Neural teachers preserve the selected model order. A:** the registered crossing stabilizes at six atoms. **B:** only the shifted condition moves from six to eight. **C:** paired differences from full current remain small in held-out velocity MSE and terminal MMD^2; positive values favor full current. **D:** OMCI–MMD averages seven active atoms under fixed construction rows and training time.

5 Experiments

Setup. We return to the two-condition, ten-endpoint stream in Section 2. Condition 0 remains stationary; condition 1 changes at $t_0 = 2200$, making two rare atoms resolvable. At $t \in \{2000, 4000\}$, Algorithm 2 builds the latest $L = 500$ condition-wise path, applies (9) with (75), and freezes the selected endpoint teacher. We fit the same 56,354-parameter conditional rectified-flow network for 4000 Adam updates under (1) on five independent streams, with three nested fits per stream. Equation (254) excludes neural outcomes from selection; Appendix E.1.1 gives the full setup.

Figure 3 shows the main effect. OMCI–MMD keeps the stationary condition at six atoms and expands the shifted condition to eight. With seven active atoms on average, it nearly matches the full-current teacher: worst-condition terminal MMD^2 is 0.00414 versus 0.00350, and velocity MSE is 3.212 versus 3.049. A half-size teacher *collapses newly resolvable conditional mass* and raises terminal MMD^2 to 0.1660; retain-all history trains on a stale pre/post-change mixture and remains at 0.01773. Shared architectures, optimization, inputs, and training budgets isolate the endpoint-law term in (206). OMCI–MMD therefore avoids *both* under-selection and intrinsic bias of full retention after drift.

Figure 4 explains why the frozen choice stabilizes neural training. OMCI–MMD stays in the same low-error cluster as full-current and over-selected teachers across the 15 fits, while under-selection and retain-all form distinct high-error groups (Panel A). Its selected sizes sit at the onset of the post hoc error plateau (Panel B), consistent with (185), but never use neural error for selection. Holding all ten atom types fixed, retain-all still loses current-law fidelity (Panel C), so stale data—not inadequate capacity—cause the failure. Panel D shows why training cannot repair that mistake: the under-selected teacher attains low own-teacher loss while producing high held-out error. OMCI–MMD

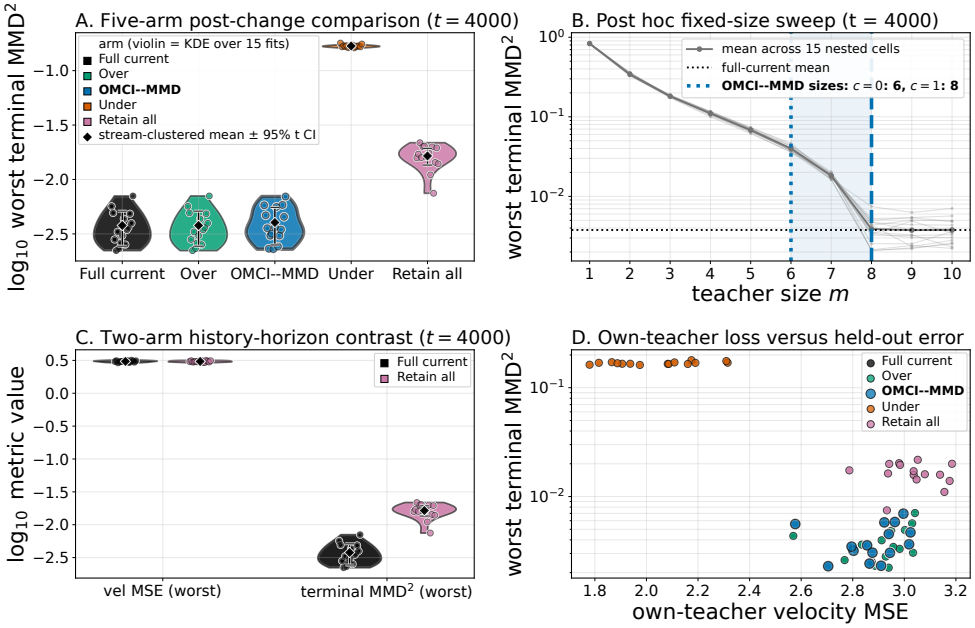


Figure 4: **Post-selection effect of the frozen teacher at $t = 4000$. A:** Across 15 fits, OMCI–MMD remains with the low-error full-current and over-selected teachers; half-size and retain-all teachers fail. **B:** The frozen sizes, six and eight, lie where terminal error flattens. **C:** With all ten atom types fixed, full history performs worse than the current window, isolating stale-law bias. **D:** Low own-teacher loss can coexist with high current-law error. Panels B and D are post-selection analyses only.

gives the network a current, sufficiently rich endpoint law before fitting it. Appendix E.3 reports full tables, paired contrasts, sensitivity checks, and transport-aware results.

Generated stale outputs. In Appendix E.4, we showcase how the proposed mechanism can actually mitigate hallucinations in two practical generation tasks: image and text generation. In particular, Figure 6 shows the four endpoint laws before neural fitting. This same setting is investigated in Figure 7, where we separate the image-flow result in its top row from the table-to-text result in its bottom row. Appendix E.4 trains two small generators after freezing those laws. A 13,302-parameter latent rectified flow produces obsolete scenes in 20.1% of retain-all samples, versus 4.7% under OMCI–MMD and 3.0% under the full-current reference; fixed-small memory covers only 75% of current scene modes. A 30,224-parameter structured table-to-text model produces obsolete facts in 17.4% of retain-all samples, while OMCI–MMD and full current produce none; fixed-small memory again covers 75% of current facts. Neither model influences the support decision.

6 Conclusion

OMCI–MMD selects condition-wise endpoint memory by comparing sparse MMD error with the sampling radius. Under stationarity, the first crossing identifies the support that the data justify; under drift, the recent-window target prevents the algorithm from compressing an obsolete full-history law. Witness transfer states when this endpoint approximation bounds transport or flow losses, and fixed-cardinality refinement lets transport-specific optimizers improve the law at the frozen size. The neural-teacher, latent-flow, and structured-text experiments show the two complementary failures: fixed-small memory drops current modes, while retain-all history reproduces stale outputs. The main lesson is simple: endpoint memory should track the current conditional law, not the clock.

AI USE STATEMENT

AI tools assisted with technical writing, running experiments, and aspects of literature review. The authors independently verified every citation, mathematical claim, code path, and reported result before submission. The authors take responsibility for the final content, including all text, claims, and artifacts produced with AI assistance.

ETHICS STATEMENT

The current experiments use controlled synthetic endpoint laws and do not involve human subjects, private data, or deployment decisions. The method targets one specific source of conditional-generation error—a stale or undercomplete retained endpoint law—and should not be interpreted as a general solution to hallucination or distribution shift in deployed generative systems.

REPRODUCIBILITY STATEMENT

Algorithms 1 and 2 specify the method, Section 4 states the guarantees, Appendices B and C provide complete proofs, Appendix D describes implementation details, and Appendix E records the experimental protocols and additional results. The supplementary source archive includes the figure scripts and experiment runners used by the manuscript. The source archive also includes the exact frozen laws, seeds, model code, checkpoints, generated samples, and raw metrics for the dynamic-scene flow and structured table-to-text pilots in `experiments/experiment3_dynamic_scene_flow` and `experiments/experiment4_table_to_text`.

REFERENCES

Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1359–1366, 2012. URL https://icml.cc/2012/papers/577.pdf.

Michèle Basseville and Igor V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, Englewood Cliffs, New Jersey, 1993.

Amrit Singh Bedi, Alec Koppel, Ketan Rajawat, and Brian M. Sadler. Trading dynamic regret for model complexity in nonstationary nonparametric optimization. In *2020 American Control Conference (ACC)*, pages 321–326, 2020. doi: 10.23919/ACC45564.2020.9148028. Archival conference version of the DynaPOLK analysis.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, New York, 2004. doi: 10.1007/978-1-4419-9096-9.

Trevor Campbell and Tamara Broderick. Automated scalable Bayesian inference via Hilbert coresets. *Journal of Machine Learning Research*, 20(15):1–38, 2019. URL https://jmlr.org/papers/v20/17-613.html.

Wilson Ye Chen, Lester Mackey, Jackson Gorham, François-Xavier Briol, and Chris J. Oates. Stein points. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 844–853, 2018. URL https://proceedings.mlr.press/v80/chen18f.html.

Xin Chen and Paul Tseng. Non-interior continuation methods for solving semidefinite complementarity problems. *Mathematical Programming*, 95(3):431–474, 2003.

Ho Kei Cheng and Alexander Schwing. The curse of conditions: Analyzing and improving optimal transport for conditional flow-based generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15875–15884, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Cheng_The_Curse_of_Conditions_Analyzing_and_Improving_Optimal_Transport_for_ICCV_2025_paper.html.

William G. Cochran. *Sampling Techniques*. John Wiley & Sons, New York, 3 edition, 1977.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pages 2292–2300, 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Rainer Dahlhaus. Fitting time series models to nonstationary processes. *The Annals of Statistics*, 25 (1):1–37, 1997. doi: 10.1214/AOS/1034276620.

Ofer Dekel, Shai Shalev-Shwartz, and Yoram Singer. The Forgetron: A kernel-based perceptron on a budget. *SIAM Journal on Computing*, 37(5):1342–1372, 2008. doi: 10.1137/060666998.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006. doi: 10.1111/j.1467-9868.2006.00553.x.

Alex Delalande. Nearly tight convergence bounds for semi-discrete entropic optimal transport. In *Proceedings of the Twenty-Fifth International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 1619–1642, 2022. URL https://proceedings.mlr.press/v151/delalande22a.html.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690, 2019. URL https://proceedings.mlr.press/v89/feydy19a.html.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014. doi: 10.1145/2523813.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1574–1583, 2019. URL https://proceedings.mlr.press/v89/genevay19a.html.

Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/698d51a19d8a121ce581499d7b701668-Abstract.html.

Siegfried Graf and Harald Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer, Berlin, 2000. doi: 10.1007/BFb0103945.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Jonathan H. Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable Bayesian logistic regression. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/2b0f658cbffd284984fb11d90254081f-Abstract.html.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL https://www.jmlr.org/papers/v6/hyvarinen05a.html.

Andrew Jacobsen, Alessandro Rudi, Francesco Orabona, and Nicolò Cesa-Bianchi. Dynamic regret reduces to kernelized static regret. *Advances in Neural Information Processing Systems*, 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/fc7f90ed14864bef5cafa89ba38f1993-Abstract-Conference.html.

Martin Jaggi. Revisiting Frank–Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 427–435, 2013. URL https://proceedings.mlr.press/v28/jaggi13.html.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. doi: 10.1109/CVPR.2017.215.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak–łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

Alex Kokot and Alex Luedtke. Coreset selection for the Sinkhorn divergence and generic smooth divergences, 2025. URL https://arxiv.org/abs/2504.20194. arXiv preprint, version 2, revised May 2025.

Alec Koppel, Garrett Warnell, Ethan Stump, and Alejandro Ribeiro. Parsimonious online learning with kernels via sparse projections in function space. *Journal of Machine Learning Research*, 20 (3):1–44, 2019. URL https://jmlr.org/papers/v20/16-585.html.

Alec Koppel, Hrusikesha Pradhan, and Ketan Rajawat. Consistent online Gaussian process regression without the sample complexity bottleneck. *Statistics and Computing*, 31(6):76, 2021. doi: 10.1007/s11222-021-10051-5.

Alec Koppel, Joe Eappen, Sujay Bhatt, Cole Hawkins, and Sumitra Ganesh. Online MCMC thinning with kernelized stein discrepancy. *SIAM Journal on Mathematics of Data Science*, 6(1):51–75, 2024. doi: 10.1137/22M1510108.

Simon Lacoste-Julien, Fredrik Lindsten, and Francis Bach. Sequential kernel herding: Frank–Wolfe optimization for particle filtering. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 544–552, 2015. URL https://proceedings.mlr.press/v38/lacoste-julien15.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Qiang Liu, Jason Lee, and Michael I. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 276–284, 2016. URL https://proceedings.mlr.press/v48/liub16.html.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

Simon Mak and V. Roshan Joseph. Support points. *The Annals of Statistics*, 46(6A):2562–2592, 2018. doi: 10.1214/17-AOS1629.

Arthur Mensch and Gabriel Peyré. Online Sinkhorn: Optimal transport distances from sample streams. In *Advances in Neural Information Processing Systems*, volume 33, pages 1657–1667, 2020.

Francesco Orabona, Joseph Keshet, and Barbara Caputo. Bounded kernel-based online learning. *Journal of Machine Learning Research*, 10(92):2643–2666, 2009. URL https://jmlr.org/papers/v10/orabona09a.html.

Xiaxue Ouyang, Hao Zheng, Haoxian Liang, Jingyi Zhang, Yixuan Qiu, Cheng Meng, and Mengyu Li. Sparsification techniques for large-scale optimal transport problems. *WIREs Computational Statistics*, 18(1):e70056, 2026. doi: 10.1002/wics.70056. URL https://doi.org/10.1002/wics.70056.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1186. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.89.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*, volume 11 of *Foundations and Trends in Machine Learning*. Now Publishers, 2019. doi: 10.1561/2200000073.

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994. doi: 10.1214/aop/1176988477.

Boris T. Polyak. Gradient methods for minimizing functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963a.

Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*, 3(4):643–653, 1963b.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://proceedings.neurips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Marina Riabiz, Wilson Ye Chen, Jon Cockayne, Pawel Swietach, Steven A. Niederer, Lester Mackey, and Chris J. Oates. Optimal thinning of mcmc output. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(4):1059–1081, 2022. doi: 10.1111/rssb.12503.

Hari Krishna Sahoo, Mudit Gaur, and Vaneet Aggarwal. Order-optimal sample complexity of rectified flows, 2026. URL https://arxiv.org/abs/2601.20250. arXiv preprint, version 1, submitted January 2026.

Dino Sejdinovic, Bharath K. Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013. doi: 10.1214/13-AOS1140.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010. URL https://jmlr.org/papers/v11/sriperumbudur10a.html.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410, 2011. URL https://jmlr.org/papers/v12/sriperumbudur11a.html.

Sophia Tang and Pranam Chatterjee. Expanding flow maps, 2026. URL https://arxiv.org/abs/2607.21585. arXiv preprint, version 1, submitted July 2026.

Zihao Tang and Yixuan Qiu. Safe and sparse Newton method for entropic-regularized optimal transport. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4127. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ea8620683340facbd5f754dd169e0980-Abstract-Conference.html.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=CD9Snc73AW.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 2009. doi: 10.1007/978-3-540-71050-9.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Chenrui Wang and Yixuan Qiu. The sparse-plus-low-rank quasi-Newton method for entropic-regularized optimal transport. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 64202–64221, 2025. URL https://proceedings.mlr.press/v267/wang25ck.html.

Fengpei Wang, Clarice Poon, and Tony Shardlow. Compressed online Sinkhorn, 2023. URL https://arxiv.org/abs/2310.05019. arXiv preprint, version 1.

Xiao Wang, Zihua She, and Jianxi Su. Coreset-induced conditional velocity flow matching, 2026. URL https://arxiv.org/abs/2605.12951. arXiv preprint, version 2, revised June 2026.

Zhuang Wang, Koby Crammer, and Slobodan Vucetic. Breaking the curse of kernelization: Budgeted stochastic gradient descent for large-scale SVM training. *Journal of Machine Learning Research*, 13(100):3103–3131, 2012. URL https://jmlr.org/papers/v13/wang12b.html.

Xinyang Wen. ForgettingOT: Certified speculative batching from Sinkhorn's projective forgetting, 2026. URL https://arxiv.org/abs/2607.24741. arXiv preprint, version 2, revised August 2026.

Haoyun Yin, Yixuan Qiu, and Xiao Wang. Wasserstein coreset via Sinkhorn loss. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=DrMCDS88IL. Published 12 February 2025.

Haoyun Yin, Chuanhui Liu, and Xiao Wang. EMS coreset: An efficient expectation–maximization algorithm for Sinkhorn coreset, 2026. URL https://arxiv.org/abs/2608.16101. arXiv preprint, version 1, submitted August 2026.

A Related Work

Endpoint memory as a distinct statistical layer. Flow Matching and rectified flows explain how a supplied endpoint law induces interpolation paths and regression targets (Lipman et al., 2023; Liu et al., 2023). Optimal-transport conditional Flow Matching changes the source–target coupling (Tong et al., 2024); C^2OT shows how condition-blind transport can distort the condition prior (Cheng and Schwing, 2025). These methods begin with an endpoint population already in hand. OMCI–MMD addresses the preceding decision: which conditional endpoint law should training retain from a stream? Under stationarity, extra endpoints mainly increase replay and coupling cost. Under drift, full history estimates a historical mixture and can average incompatible displacement targets. A fixed buffer creates the opposite error by deleting rare conditions or weak modes. Local-estimation and change-point theory explain the bias introduced by stale observations (Dahlhaus, 1997; Basseville and Nikiforov, 1993). Concept-drift and stratified-sampling results explain why a marginal summary need not protect a current or rare conditional stratum (Gama et al., 2014; Cochran, 1977).

Sparse summaries of probability laws. Early budgeted online kernels impose explicit dictionary limits (Dekel et al., 2008; Orabona et al., 2009); budgeted stochastic gradients maintain a fixed dictionary (Wang et al., 2012), while POLK and POG use projection to trade approximation against model growth (Koppel et al., 2019; 2021). Frank–Wolfe supplies the sparse convex-optimization step (Jaggi, 2013); kernel herding applies it to Hilbert-space residual geometry (Bach et al., 2012; Lacoste-Julien et al., 2015). Bayesian and Hilbert coresets preserve posterior or log-likelihood summaries (Huggins et al., 2016; Campbell and Broderick, 2019). Support points target distributional representation (Mak and Joseph, 2018), while Monte Carlo thinning preserves selected test-function expectations (Riabiz et al., 2022; Koppel et al., 2024). OMCI–MMD borrows the weighted-atomic approximation principle but changes the statistical question. Classical quantization and sequential Monte Carlo likewise replace a distribution by a finite weighted measure, usually at a prescribed distortion or particle count (Graf and Luschgy, 2000; Del Moral et al., 2006). Kernel mean embeddings supply the probability representation (Berlinet and Thomas-Agnan, 2004; Sriperumbudur et al., 2010); MMD turns its empirical approximation into an RKHS distance and connects it to distance-based statistics (Gretton et al., 2012; Sejdinovic et al., 2013). This geometry supports an online mean recursion, Hilbert-space concentration, and convex reweighting on a fixed support. OMCI uses those tools to estimate a conditional, time-varying law and reads the support budget from the first sparse-MMD crossing below a sampling-and-drift radius. DynaPOLK and related nonstationary kernel methods instead track a time-varying predictor with a projection-controlled dictionary and dynamic-regret guarantees (Bedi et al., 2020); they do not define a conditional endpoint-law path or calibrate its first crossing to statistical resolution.

Transport compression acts at several layers. Recent transport work compresses several different objects: endpoint measures, dual-potential representations, curvature matrices, reusable solver state, and generated state dimension. Table 1 separates these layers. This distinction mirrors the broader difference between measure sparsification and computational sparsification in large-scale optimal transport (Ouyang et al., 2026).

Method family	Object changed	Main objective or guarantee	Relation to OMCI–MMD
Flow Matching (Lipman et al., 2023), optimal-transport conditional Flow Matching (OT-CFM) (Tong et al., 2024), conditional optimal transport (C^2OT) (Cheng and Schwing, 2025)	Path, velocity target, or source–target coupling	Defines or improves flow-training geometry for a supplied endpoint sample	Operates after the endpoint population has been chosen.
Direct Sinkhorn coresets: WCSL (Yin et al., 2025), CO2 (Kokot and Luedtke, 2025), and EMS (Yin et al., 2026)	Weighted atomic approximation of a fixed empirical measure	Direct Sinkhorn optimization, divergence-adapted second-order MMD, or weighted-centroid refinement	They refine transport geometry after OMCI selects and freezes the cardinality.
Coreset-Induced Conditional Velocity Flow Matching (CCVFM) preprint (Wang et al., 2026)	Prescribed-cardinality entropic coreset, Gaussian-mixture surrogate, and correction flow	Reduces the residual learned by a hierarchical flow after choosing the coreset size	Complements OMCI when that size is supplied by the online crossing.
Online Sinkhorn and Compressed Online Sinkhorn (COS) (Mensch and Peyré, 2020; Wang et al., 2023)	Kernel expansions representing entropic dual potentials	Learns transport potentials from streams and compresses their representation	The stored nodes parameterize a dual estimator rather than the endpoint law used for flow training.
Safe sparse Newton (Tang and Qiu, 2024), sparse-plus-low-rank quasi-Newton (Wang and Qiu, 2025), and ForgettingOT (Wen, 2026)	Hessian, quasi-Newton, or reusable solver state	Accelerates one solve or a sequence of related solves	Reduces transport computation without selecting the incoming endpoint population.
Expanding Flow Maps (Tang and Chatterjee, 2026)	Number of generated coordinates or tokens	Learns controllable output dimension or sequence length	Changes state dimension; OMCI keeps the state space fixed and changes retained endpoint cardinality.

Table 1: Neighboring methods place sparsity at different layers of the transport-and-generation pipeline. Their guarantees become complementary once endpoint-cardinality selection is separated from fixed-size transport refinement and solver acceleration.

Online conditional cardinality. Direct Sinkhorn coresets make strong fixed-size comparators. WCSL optimizes a Sinkhorn-loss coreset for a fixed dataset (Yin et al., 2025). CO2 derives a divergence-specific second-order MMD and discusses tolerance-based fixed-target size selection (Kokot and Luedtke, 2025). EMS learns weighted free-support representatives (Yin et al., 2026), while CCVFM starts from a fixed-K entropic coreset (Wang et al., 2026). These methods optimize a supplied size or a fixed target. They do not define a condition-wise crossing for an evolving endpoint law whose threshold reflects sampling error and stale-data bias. OMCI, in turn, does not replace their transport objectives. The resulting pipeline assigns one role to each stage: OMCI selects and freezes the cardinality; a fixed-size transport optimizer then refines weights and, when permitted, atom locations. A separate support-size sweep measures the hindsight memory–accuracy frontier but never influences deployment.

Why the model-order rule uses MMD. MMD turns endpoint compression into sparse approximation of a Hilbert-space mean and links kernel probability metrics to classical distance-based statistics (Berlinet and Thomas-Agnan, 2004; Sejdinovic et al., 2013). This geometry supplies a time-uniform concentration radius, exact residual correlations, convex reweighting, and a linear support oracle. Raw Wasserstein distance uses the Lipschitz dual class and has dimension-sensitive empirical behavior (Villani, 2009; Peyré and Cuturi, 2019). Entropic regularization makes the transport problem computationally tractable (Cuturi, 2013). Sinkhorn divergences supply smooth first-variation potentials (Feydy et al., 2019), whose empirical accuracy still depends on the active law, entropy parameter, and sample size (Genevay et al., 2019). CO2 shows how a second-order, data-dependent MMD can recover local smooth-divergence geometry for a fixed target (Kokot and Luedtke, 2025). Recomputing such a kernel from the same stream would require a new calibration argument, so the present model-order theorem uses a fixed or independently calibrated bounded kernel and transfers to transport through explicit witnesses.

Observed support, free support, and transport refinement. The OMCI theorem selects observed endpoints. A free-support coreset may attain lower error with fewer atoms, but it solves a larger and generally nonconvex problem. If the endpoint space has metric $d_{\mathcal{Y}}$, a candidate dictionary $\mathcal{D} \subseteq \mathcal{Y}$ has covering radius $h := \sup_{y \in \mathcal{Y}} \inf_{z \in \mathcal{D}} d_{\mathcal{Y}}(y, z)$, and ϕ is L_ϕ-Lipschitz into $\mathcal{H}$, snapping a free atom to its nearest candidate perturbs its feature vector by at most $L_\phi h$; without such coverage, observed- and free-support orders need not agree. EMS, for example, learns weighted centroids rather than restricting every atom to an observation (Yin et al., 2026). The matched-size experiment therefore reports observed-support, uniform-weight, and free-support methods in separate strata while allowing every method to benefit from the same frozen cardinality.

Witnesses connect endpoint memory to the trained flow. An endpoint MMD bound applies only to functions that the chosen reproducing-kernel Hilbert space (RKHS) represents with controlled norm and approximation error. Kantorovich duality identifies the W_1 critics (Villani, 2009); entropic first variations supply Sinkhorn witnesses (Peyré and Cuturi, 2019; Feydy et al., 2019). Flow Matching contributes endpoint-indexed loss witnesses. A frozen witness bank can insert application-facing quantities before calibration, such as prompt, identity, motion, payoff, tail-event, or no-arbitrage functions. The analysis therefore keeps three effects separate: replacing the endpoint law, changing the path or coupling, and optimizing a neural velocity field. Equation (192) addresses the first for a fixed path; optimization, generalization, and ordinary differential equation (ODE) or stochastic differential equation (SDE) stability govern the latter two.

Windowing as local estimation. The nonstationary construction uses a recent window as a local estimator of the current endpoint law. A longer window reduces variance but averages a less local population, reproducing the classical bias–variance tradeoff in locally stationary processes and change-point analysis (Dahlhaus, 1997; Basseville and Nikiforov, 1993). OMCI adds support selection to that local-estimation problem: the window defines the target, the sampling-and-staleness radius defines its resolution, and the sparse path defines the memory required to reach it. Drift detection and candidate coverage remain separate modules, which lets the analysis state exactly which part of the error each module must control.

B Model-order construction and analysis

This appendix follows the quantities used by OMCI–MMD. The stationary analysis derives the support oracle, constructs the complete path, and proves concentration, crossing stability, and fixed-radius reachability. The nonstationary analysis applies the same path to a recent window, separates sampling, staleness, candidate availability, path construction, and numerical error, and then proves current-law convergence.

B.1 Stationary checkpoint path

Fix a reporting time t and condition c with at least one observation. Recall from (2) that $\widehat{\mu}_t^c$ denotes the empirical condition-c kernel mean. Define

$$H_t^c := \{i \in \{1, \dots, t\} : c_i = c\}, \qquad n_c(t) := |H_t^c|, \qquad \varnothing \neq \mathcal{J}_t^c \subseteq H_t^c, \tag{13}$$

where H_t^c is the complete observed index set and $\mathcal{J}_t^c$ is the declared set searched by the checkpoint linear minimization oracle. Exact observed-support search sets $\mathcal{J}_t^c = H_t^c$.

For a finite active set $A \subseteq \mathcal{J}_t^c$, define its simplex

$$\mathsf{S}(A) := \left\{ a \in \mathbb{R}^{|A|} : a_j \geq 0 \text{ for all } j \in A, \ \sum_{j \in A} a_j = 1 \right\}. \tag{14}$$

For $a \in \mathsf{S}(A)$, define the weighted feature mean, residual, and squared-MMD objective

$$q(A, a) := \sum_{j \in A} a_j \phi(Y_j), \qquad r(A, a) := \widehat{\mu}_t^c - q(A, a), \qquad F_t^c(A, a) := \frac{1}{2} \|q(A, a) - \widehat{\mu}_t^c\|_{\mathcal{H}}^2. \tag{15}$$

Thus $q(A, a)$ is the coreset mean, $r(A, a)$ is the feature-space structure that the coreset has not represented, and $F_t^c(A, a)$ is one half of the squared MMD from the empirical condition-c law.

Why residual correlation is the linear oracle. Write $F(q) := \frac{1}{2}\|q - \widehat{\mu}_t^c\|_{\mathcal{H}}^2$. Take a current feasible mean q, its residual $r = \widehat{\mu}_t^c - q$, a candidate feature $s = \phi(Y_i)$, and the feasible segment

$$q_\gamma := (1-\gamma)q + \gamma s = q + \gamma(s-q), \qquad 0 \le \gamma \le 1. \tag{16}$$

Expanding the squared norm gives

$$\begin{aligned} F(q_\gamma) &= \frac{1}{2}\|q - \widehat{\mu}_t^c + \gamma(s-q)\|_{\mathcal{H}}^2 \\ &= F(q) - \gamma\langle r, s-q\rangle_{\mathcal{H}} + \frac{\gamma^2}{2}\|s-q\|_{\mathcal{H}}^2, \end{aligned} \tag{17}$$

and hence

$$\begin{aligned} \frac{d}{d\gamma}F(q_\gamma)\Big|_{\gamma=0} &= -\langle r, s-q\rangle_{\mathcal{H}} \\ &= -\langle r, s\rangle_{\mathcal{H}} + \langle r, q\rangle_{\mathcal{H}}. \end{aligned} \tag{18}$$

The final term is common to all candidates. At path size m, write $q_{t,m}^c := q(A_{t,m}^c, a_{t,m}^c)$ and $r_{t,m}^c := \widehat{\mu}_t^c - q_{t,m}^c$. Therefore minimizing the directional derivative is equivalent to

$$i_{t,m+1}^{c,\star} \in \arg\max_{i \in \mathcal{J}_t^c} \langle r_{t,m}^c, \phi(Y_i)\rangle_{\mathcal{H}}, \tag{19}$$

which is the residual-correlation step in (7). This is the standard Frank–Wolfe/kernel-herding oracle (Bach et al., 2012; Jaggi, 2013); OMCI–MMD uses it to construct an entire support-indexed error path rather than only one fixed-budget iterate.

If the implementation uses an approximate oracle, it returns $\widehat{i}_{t,m+1}^c \in \mathcal{J}_t^c \setminus A_{t,m}^c$ whenever the exact gap is positive. Define its additive miss inside the declared set by

$$\omega_{t,m}^c := \max_{i \in \mathcal{J}_t^c} \langle r_{t,m}^c, \phi(Y_i)\rangle_{\mathcal{H}} - \langle r_{t,m}^c, \phi(Y_{\widehat{i}_{t,m+1}^c})\rangle_{\mathcal{H}} \ge 0. \tag{20}$$

This quantity measures optimization inside $\mathcal{J}_t^c$. It does not measure whether $\mathcal{J}_t^c$ omitted a useful observed endpoint; Appendix B.7 treats that availability error separately.

After adding an atom, full correction solves

$$a_{t,m}^c \in \arg\min_{a \in \mathsf{S}(A_{t,m}^c)} F_t^c(A_{t,m}^c, a). \tag{21}$$

The selected dictionary $A_{t,m}^c$ remains nested even if the optimizer assigns zero weight to a selected index. Its positive support is

$$A_{t,m}^{c,+} := \{j \in A_{t,m}^c : a_{t,m,j}^c > 0\}, \qquad |A_{t,m}^{c,+}| \le m. \tag{22}$$

Thus m is a support budget and the deployed law stores no more than m nonzero atoms; implementations may prune zero weights after they record the complete path. For the corrected mean $q_{t,m}^c := q(A_{t,m}^c, a_{t,m}^c)$ and residual $r_{t,m}^c := r(A_{t,m}^c, a_{t,m}^c)$, define the declared-hull Frank–Wolfe gap

$$g_{\mathrm{FW}}(q_{t,m}^c; \mathcal{J}_t^c) := \max_{i \in \mathcal{J}_t^c} \langle r_{t,m}^c, \phi(Y_i) - q_{t,m}^c\rangle_{\mathcal{H}}. \tag{23}$$

Lemma 1 (Zero-gap certificate on the declared hull)**.** *Let* $\mathcal{D}_t^c := \mathrm{conv}\{\phi(Y_i) : i \in \mathcal{J}_t^c\}$. *Then*

$$g_{\mathrm{FW}}(q_{t,m}^c; \mathcal{J}_t^c) = 0 \iff q_{t,m}^c \in \arg\min_{q \in \mathcal{D}_t^c} \frac{1}{2}\|q - \widehat{\mu}_t^c\|_{\mathcal{H}}^2. \tag{24}$$

Proof. **Proof map.** We use convex first-order optimality in (25) and linear maximization over the declared convex hull. Equation (24) permits exact path termination on that hull without claiming optimality over omitted atoms. For a differentiable convex function F on a convex set $\mathcal{D}_t^c$, a point q is optimal if and only if

$$\langle \nabla F(q), z - q\rangle_{\mathcal{H}} \ge 0 \quad \text{for every } z \in \mathcal{D}_t^c. \tag{25}$$

Here $\nabla F(q) = q - \widehat{\mu}_t^c = -r$. Thus (25) is equivalent to $\langle r, z-q\rangle_{\mathcal{H}} \le 0$ for every $z \in \mathcal{D}_t^c$. A linear functional attains its maximum over a convex hull at one of its generators, so this condition is equivalent to $g_{\mathrm{FW}}(q; \mathcal{J}_t^c) \le 0$. The gap is always nonnegative because q is a convex combination of the generators and the corresponding weighted average of the inner products equals zero. Hence $g_{\mathrm{FW}} \le 0$ if and only if $g_{\mathrm{FW}} = 0$. □

The active-atom shortcut used by an exact implementation follows from the simplex optimality conditions, but only after full correction has solved (21) exactly. The derivative of the weight-space objective with respect to coordinate a_j is

$$\frac{\partial}{\partial a_j} F_t^c(A_{t,m}^c, a_{t,m}^c) = \langle q_{t,m}^c - \widehat{\mu}_t^c, \phi(Y_j)\rangle_{\mathcal{H}} = -\langle r_{t,m}^c, \phi(Y_j)\rangle_{\mathcal{H}}. \tag{26}$$

Introduce a multiplier λ for $\sum_j a_j = 1$ and multipliers $\nu_j \geq 0$ for $a_j \geq 0$. Stationarity and complementary slackness give

$$-\langle r_{t,m}^c, \phi(Y_j)\rangle_{\mathcal{H}} + \lambda - \nu_j = 0, \qquad \nu_j a_{t,m,j}^c = 0. \tag{27}$$

Consequently,

$$\langle r_{t,m}^c, \phi(Y_j)\rangle_{\mathcal{H}} = \lambda \quad \text{for } j \in A_{t,m}^{c,+}, \qquad \langle r_{t,m}^c, \phi(Y_j)\rangle_{\mathcal{H}} \leq \lambda \quad \text{for } j \in A_{t,m}^c \setminus A_{t,m}^{c,+}. \tag{28}$$

Multiply the equality in (28) by $a_{t,m,j}^c$ and sum over the positive support. Because the positive weights sum to one, this gives $\lambda = \langle r_{t,m}^c, q_{t,m}^c\rangle_{\mathcal{H}}$. At least one positive-weight atom has correlation λ, so the maximum correlation over $\mathcal{J}_t^c$ is at least λ. If the exact oracle in (19) returns an already-selected index, (28) also bounds that returned correlation by λ. Exact maximality forces equality, and hence

$$g_{\mathrm{FW}}(q_{t,m}^c; \mathcal{J}_t^c) = \max_{i \in \mathcal{J}_t^c} \langle r_{t,m}^c, \phi(Y_i)\rangle_{\mathcal{H}} - \langle r_{t,m}^c, q_{t,m}^c\rangle_{\mathcal{H}} = 0.$$

Conversely, if the gap is positive, every exact maximizer lies outside the selected dictionary. This argument certifies optimality only on the declared hull, not over omitted endpoints. It also fails after an inexact corrective solve: then the KKT equalities need not hold, and an oracle returning an active index does not prove zero gap. An approximate implementation must compute (23) directly or report a valid upper bound on it; it may carry the terminal law forward as exact only when that certificate equals zero.

B.2 Complete CrossingReadout and the path sandwich

Let $1 \leq M_{\max} \leq |\mathcal{J}_t^c|$ be a predeclared maximum support size. CrossingReadout builds nested supports and, at every constructed size m, records

$$q_{t,m}^c := q(A_{t,m}^c, a_{t,m}^c), \qquad \widehat{F}_{t,m}^c := F_t^c(A_{t,m}^c, a_{t,m}^c), \qquad \widehat{e}_{t,m}^c := \sqrt{2\widehat{F}_{t,m}^c}. \tag{29}$$

The routine stores the first crossing when it occurs but continues until $M_{\max}$ or an exact zero-gap certificate. If the gap vanishes at budget $m < M_{\max}$, it copies the certified terminal law and error to every larger budget. The copied law remains feasible because each path index denotes an *at-most-m support budget*, not a requirement of exactly m positive weights; it is optimal over the declared hull at every later budget. Deployment uses the stored crossing, whereas Theorem 1 needs a value at every budget in its predeclared analysis range.

Algorithm 3 Complete CrossingReadout at checkpoint (t, c)

Require: empirical mean $\widehat{\mu}_t^c$, declared candidates $\mathcal{J}_t^c$, radius ρ, maximum size $M_{\max}$

Choose the best singleton $i_1 \in \arg\min_{i \in \mathcal{J}_t^c} \frac{1}{2}\|\phi(Y_i) - \widehat{\mu}_t^c\|_{\mathcal{H}}^2$; set $A_{t,1}^c = \{i_1\}$ and $\widehat{m} = +\infty$.
for $m = 1, \ldots, M_{\max}$ **do**
  Fully correct $a_{t,m}^c$ by (21); record $(\widehat{F}_{t,m}^c, \widehat{e}_{t,m}^c)$ from (29).
  if $\widehat{m} = +\infty$ and $\widehat{e}_{t,m}^c \le \rho$ **then**
    Set $\widehat{m} = m$ and store $(A_{t,m}^c, a_{t,m}^c)$ as the deployable crossing law; continue the complete path.
  end if
  if $m = M_{\max}$ **then**
    Stop.
  end if
  Compute $r_{t,m}^c$ and the exact gap (23); use (19) or a declared approximate oracle with miss (20).
  if the exact gap (23) is zero **then**
    Copy the terminal law and error through budgets $m+1{:}M_{\max}$; record the declared-hull certificate; stop.
  end if
  Require the returned approximate-oracle index to lie outside $A_{t,m}^c$, then set $A_{t,m+1}^c = A_{t,m}^c \cup \{\widehat{i}_{t,m+1}^c\}$.
end for
Return the constructed path, first crossing $\widehat{m}$, stored weighted law, oracle misses, and any zero-gap certificate.

Let $\mathcal{I} = \{m_{\min}, m_{\min}+1, \ldots, m_{\max}\} \subseteq \{1, \ldots, M_{\max}\}$ be a predeclared contiguous interval of budgets in the complete path, and write $[x]_+ := \max\{x, 0\}$. Recall the path slack from (11),

$$s_{t,\mathcal{I}}^c = \sup_{m \in \mathcal{I}} [\widehat{e}_{t,m}^c - e_{t,m}^{c,\star}]_+. \tag{30}$$

Proposition 2 (Computed/oracle path sandwich)**.** *For every $m \in \mathcal{I}$,*

$$e_{t,m}^{c,\star} \le \widehat{e}_{t,m}^c \le e_{t,m}^{c,\star} + s_{t,\mathcal{I}}^c. \tag{31}$$

Proof. **Proof map.** Feasibility gives the lower path inequality. The uniform slack in (30) gives the matching upper inequality (31), which the crossing argument later inverts. The computed weighted law uses at most m observed endpoints and feasible simplex weights, so it is feasible for the infimum defining $e_{t,m}^{c,\star}$ in (5). This gives the left inequality. The definition (30) gives $[\widehat{e}_{t,m}^c - e_{t,m}^{c,\star}]_+ \le s_{t,\mathcal{I}}^c$. The left inequality makes the expression inside the positive part nonnegative; removing the positive part and rearranging gives the right inequality. □

The two inequalities play different roles. Feasibility alone prevents the computed path from crossing before the best observed-support path at the same radius. The upper inequality requires a bound on $s_{t,\mathcal{I}}^c$ and prevents a much later crossing. Exhaustive verification can evaluate the slack exactly; a scalable implementation must derive or report a separate upper bound. CrossingReadout does not infer that bound from its own output.

B.3 Algorithm-to-theorem map

Each design choice has one analytical role.

1. The empirical kernel mean turns the endpoint law into a Hilbert-space target whose sampling error admits Lemma 3 (Pinelis, 1994; Gretton et al., 2012).
2. The derivative (18) justifies the support oracle, while full correction (21) makes each path value the best weighting on its selected support (Jaggi, 2013; Lacoste-Julien et al., 2015).
3. The complete path in Algorithm 3 supplies one recorded value per support budget. The sandwich (31) converts its approximation error into the crossing bracket (12).

4. The declared candidate set makes the within-set miss (20) observable. Under drift, Appendix B.7 further separates candidate availability from within-pool optimization.
5. Standard Frank–Wolfe geometry supplies the inherited fixed-radius rate. The novel statistical object is the first condition-wise crossing below a sampling-and-staleness radius and its stability under path error.
6. The endpoint MMD bound remains an RKHS statement. Appendix C supplies the additional witness assumptions needed before Wasserstein, Sinkhorn, or Flow-Matching losses inherit it.

This chain distinguishes OMCI–MMD from fixed-budget posterior coresets or Monte Carlo thinning: the sparse approximation primitive is inherited, while the conditional sampling floor, complete crossing readout, windowed target, and witness-transfer interface change what that primitive must bound.

B.4 Stationary assumptions and crossing inversion

The proof uses standard concentration and conditional-gradient facts, but it uses them for a different target. Fixed-budget analyses ask how accurately m atoms approximate a fixed mean. OMCI–MMD asks for the first m at which a complete, data-dependent approximation path reaches a data-dependent statistical radius. We therefore state every conditioning and solver requirement before we invert the path.

Stationary assumption package. Fix a condition c and a reported checkpoint set $\mathcal{T}^c$. Work on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. Let Y_s denote the endpoint observed at time s, and let P^c denote the fixed conditional law of Y_s on the reported stationary interval whenever the observed condition equals c. We impose the following requirements.

- **S1.** *Stationary conditional sampling.* Conditional on the condition labels and the predictable checkpoint schedule, the condition-c endpoints have conditional mean embedding $\mu^c := \mathbb{E}_{Y\sim P^c}[\phi^c(Y)]$. Independence is sufficient, but a bounded Hilbert-valued martingale-difference condition also suffices for the cited concentration inequality.
- **S2.** *Bounded kernel chosen before evaluation.* The kernel k^c has RKHS $\mathcal{H}^c$, feature map ϕ^c, and diagonal bound $k^c(y,y) \leq K_c^2$. The algorithm fixes k^c before evaluation, chooses it predictably from past data, or pays a simultaneous-validity penalty over a declared kernel family. Tuning k^c on the same held-out transport loss that later evaluates the coreset violates this clause.
- **S3.** *Simultaneous radius.* For every $t \in \mathcal{T}^c$, the reported radius ε_t^c satisfies $\|\widehat{\mu}_t^c - \mu^c\|_{\mathcal{H}^c} \leq \varepsilon_t^c$ on one event of the advertised probability. Here $\widehat{\mu}_t^c$ is the empirical condition-c kernel mean at checkpoint t.
- **S4.** *Declared path domain.* At checkpoint t, the algorithm declares a candidate universe $\mathcal{J}_t^c$ and a finite support interval $\mathcal{I}_t^c \subset \{1, \dots, |\mathcal{J}_t^c|\}$ before it reads the crossing. The candidate universe may depend on the observed endpoints; the theorem measures coverage relative to that declared universe, while the availability term in (99) compares it with the full history.
- **S5.** *Complete path with measured solver error.* For each $m \in \mathcal{I}_t^c$, the algorithm records a feasible weighted law $\widehat{Q}_{t,m}^c$ with at most m distinct atoms, together with the actual MMD error. The measured path slack absorbs approximate linear-oracle and fully corrective solves.
- **S6.** *Separation for exact integer recovery.* A claim that the computed and oracle crossings equal the same integer requires an oracle value at the crossing that lies below the radius by more than the path slack. Without this margin, the theorem gives a bracket, not exact model-order recovery.

For every $m \in \mathcal{I}_t^c$, we reuse the oracle error $e_{t,m}^{c,\star}$ from (5), the computed error $\widehat{e}_{t,m}^c$ from (29), and the path slack $s_{t,\mathcal{I}}^c$ from (30). All three quantities use the same checkpoint, candidate universe, and support interval. Proposition 2 proves the deterministic comparison (31); approximate linear-oracle and corrective solves affect the crossing only through this measured slack.

For a radius $r > 0$, set

$$m_t^{c,\star}(r) := \min\{m \in \mathcal{I}_t^c : e_{t,m}^{c,\star} \leq r\}, \qquad \widehat{m}_t^c(r) := \min\{m \in \mathcal{I}_t^c : \widehat{e}_{t,m}^c \leq r\}, \tag{32}$$

with the value $+\infty$ when the relevant set is empty.

Proposition 3 (Crossing inversion and exact-recovery margin)**.** *If $0 \le s^c_{t,\mathcal{I}} < r$ and the crossings below are finite, then*

$$m^{c,\star}_t(r) \le \widehat{m}^c_t(r) \le m^{c,\star}_t\big(r - s^c_{t,\mathcal{I}}\big). \tag{33}$$

In particular, if $m_\star = m^{c,\star}_t(r)$ and $e^{c,\star}_{t,m_\star} \le r - s^c_{t,\mathcal{I}}$, then $\widehat{m}^c_t(r) = m_\star$.

Proof. **Proof map.** We invert pointwise path inequalities at a common threshold. Applying that argument to the complete-path sandwich (31) over one declared support range gives the model-order bracket (33). Take any $m < m^{c,\star}_t(r)$. Definition (32) gives $e^{c,\star}_{t,m} > r$. The left side of (31) then yields $\widehat{e}^c_{t,m} > r$, so the computed path cannot cross before the oracle path. This proves the first inequality in (33).

Now take $m = m^{c,\star}_t(r - s^c_{t,\mathcal{I}})$. The oracle definition gives $e^{c,\star}_{t,m} \le r - s^c_{t,\mathcal{I}}$. The right side of (31) gives $\widehat{e}^c_{t,m} \le r$, so the computed path crosses no later than m. This proves the second inequality. Under the stated margin at $m_\star$, the upper bound becomes $\widehat{m}^c_t(r) \le m_\star$ while the lower bound gives the reverse inequality. □

Lemma 2 (Uniform empirical-to-population curve transfer)**.** *On the event in S3, define*

$$g^{c,\star}_{t,m} := \inf_{Q \in \mathcal{Q}_m(\mathcal{J}^c_t)} \mathrm{MMD}_{k^c}(P^c, Q).$$

Then, simultaneously for every $m \in \mathcal{I}^c_t$,

$$\big|g^{c,\star}_{t,m} - e^{c,\star}_{t,m}\big| \le \varepsilon^c_t. \tag{34}$$

The same inequality holds for any computed, data-dependent $\widehat{Q}^c_{t,m}$ after replacing the two infima by the corresponding MMD values.

Proof. **Proof map.** We use the reverse triangle inequality in the RKHS. The uniform statement (34) covers the entire feasible coreset class, so the selected support may depend on the same sample; adaptive kernels or radii still need separate control. For every probability law Q, the reverse triangle inequality for the RKHS norm gives

$$\Big|\mathrm{MMD}_{k^c}(P^c, Q) - \mathrm{MMD}_{k^c}(\widehat{P}^c_t, Q)\Big| \le \mathrm{MMD}_{k^c}(P^c, \widehat{P}^c_t) = \|\mu^c - \widehat{\mu}^c_t\|_{\mathcal{H}^c} \le \varepsilon^c_t.$$

This bound holds for every Q on the same event, so it also holds after taking an infimum over the common feasible class $\mathcal{Q}_m(\mathcal{J}^c_t)$. Uniformity over Q is why the computed support may depend on the evaluation sample without requiring a second union bound. Kernel or radius selection from the same data remains subject to S2–S3 because it changes the event itself. □

Consequence for model-order language. Equation (34) gives the statistical meaning of the empirical crossing: it locates population structure only up to the sampling radius. Exact equality with a population-defined integer requires an additional separation margin on the population curve. We use "learnt support size" for the sampling-calibrated empirical decision and reserve exact recovery for the explicit margin case.

Standard ingredients and OMCI composition. Hilbert-space concentration supplies S3 under S1–S2 Pinelis (1994). Frank–Wolfe, kernel-herding, and Hilbert-coreset analyses supply fixed-target approximation recursions and, under full observed-support access, the familiar $O(r^{-2})$ worst-case support bound Jaggi (2013); Lacoste-Julien et al. (2015). Those results hold with a budget fixed before the approximation. Proposition 3 instead inverts a complete, approximate path at a data-dependent statistical radius. The path sandwich, common-domain requirement, and separation condition are the additional logical steps that make that inversion valid. They also explain why Algorithm 1 must continue the complete path after freezing the first deployable crossing.

B.5 Stationary concentration and sparse-path reachability

Let P^c denote the stationary endpoint law for condition c, and define its kernel mean

$$\mu^c := \int \phi(y)\, dP^c(y). \tag{35}$$

Condition on the complete realized label sequence $(c_i)_{i\geq 1}$. Assume that the endpoints are conditionally independent and that $Y_i \sim P^{c_i}$ under this conditional law. Equivalently, for each condition c, the subsequence $\{\phi(Y_i) : c_i = c\}$ is independent with common mean μ^c. Assume also that $\|\phi(Y_i)\|_{\mathcal{H}} \leq K$ almost surely. The labels themselves need not be independent or balanced; their realized counts enter the radius through $n_c(t)$. Because Lemma 3 holds simultaneously for every time t, the reporting schedule may depend on previously observed data without an additional union bound. The proof does require conditional independence and a common condition-wise law. If dependence remains after conditioning on the labels, one must replace the Hoeffding step by a martingale, mixing, or other dependence-aware concentration inequality and enlarge the radius accordingly.

We separate four logical layers. First, Hilbert-space concentration calibrates the empirical endpoint mean. Second, conditional-gradient geometry bounds one computed sparse path. Third, the path sandwich compares that path with the best observed-support path at every declared budget. Fourth, the crossing theorem turns the uniform path comparison into a support-size statement. The first two layers are standard; the last two create the model-order interface used by OMCI–MMD.

Fix one condition c and one reporting time t. Let

$$\mathcal{X}_t^c := \{\phi(Y_i) : i \leq t,\ c_i = c\} \subset \mathcal{H}, \qquad \widehat{\mu}_t^c := \frac{1}{n_c(t)} \sum_{i\leq t : c_i = c} \phi(Y_i), \tag{36}$$

and let $\mathcal{D}_t^c := \operatorname{conv}(\mathcal{X}_t^c)$. For support A and weights $a \in \mathsf{S}(A)$, write

$$q(A,a) := \sum_{j\in A} a_j \phi(Y_j), \qquad F_t^c(A,a) := \frac{1}{2}\|q(A,a) - \widehat{\mu}_t^c\|_{\mathcal{H}}^2. \tag{37}$$

The optimum over the complete observed convex hull is

$$F_t^{c,\mathrm{full}} := \min_{q\in\mathcal{D}_t^c} \frac{1}{2}\|q - \widehat{\mu}_t^c\|_{\mathcal{H}}^2 = 0, \tag{38}$$

because $\widehat{\mu}_t^c$ is itself the empirical convex combination of the observed features. Equation (5) defines the best m-atom curve, and Algorithm 1 returns the computed path defined in (29).

Proof of Proposition 1. **Proof map.** We calculate W_1 directly for point masses. The example keeps the model order in (6) condition-specific even when the marginal endpoint laws agree exactly. Under P, the marginal endpoint law is $P_Y = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$. Swapping the two conditional laws leaves the same marginal under Q, so $P_Y = Q_Y$. Conditional on $C = 0$, however, the two laws are δ_{-1} and δ_1. Transporting the unit mass from -1 to 1 costs two, hence $W_1(P^0, Q^0) = 2$. The same argument with the roles reversed gives $W_1(P^1, Q^1) = 2$, and therefore $\max_c W_1(P^c, Q^c) = 2$.

Lemma 3 (Time-uniform endpoint radius)**.** *With probability at least* $1-\delta$*, every reported pair* (t,c) *with* $n_c(t) > 0$ *satisfies*

$$\|\widehat{\mu}_t^c - \mu^c\|_{\mathcal{H}} \leq E_t^c(\delta). \tag{39}$$

Proof of Lemma 3. **Proof map.** We combine the fixed-time Hilbert-valued Hoeffding bound (40) with a summable union bound. Equation (41) makes (39) simultaneous over all reported times and conditions; OMCI departs from this standard step when it inverts the sparse path at that radius. The radius uses an inherited Hilbert-space concentration step; the crossing argument starts only after we establish this event. Conditional on the labels, each centered vector $\phi(Y_i) - \mu^c$ has mean zero. Jensen's inequality gives $\|\mu^c\|_{\mathcal{H}} \leq \int \|\phi(y)\|_{\mathcal{H}}\, dP^c(y) \leq K$, so every centered vector has norm at most $2K$. We absorb this factor into a universal constant $C_{\mathrm{H}} > 0$. A Hoeffding inequality for bounded independent Hilbert-space-valued random variables (Pinelis, 1994) then gives, for fixed (t,c) and $\eta \in (0,1)$,

$$\Pr\left(\|\widehat{\mu}_t^c - \mu^c\|_{\mathcal{H}} > C_{\mathrm{H}} K \sqrt{\frac{\log(2/\eta)}{n_c(t)}} \right) \leq \eta. \tag{40}$$

Apply (40) to the condition-c subsequence under the conditional sampling model stated after (35). Assign failure probability

$$\eta_{t,c} := \frac{\delta}{2|\mathcal{C}|t^2} \tag{41}$$

to each reported pair (t, c). A union bound gives total failure probability at most

$$\sum_{c\in\mathcal{C}}\sum_{t=1}^{\infty}\eta_{t,c} = \frac{\delta}{2}\sum_{t=1}^{\infty}t^{-2} < \delta. \tag{42}$$

Moreover,

$$\log\frac{2}{\eta_{t,c}} = \log\frac{4|\mathcal{C}|t^2}{\delta}. \tag{43}$$

Substituting this identity into (40) proves (39) simultaneously over all times and conditions with positive count under the conditional law. The bound holds for every realized label sequence that satisfies the stated conditional model, so integrating over the labels gives the same unconditional probability bound. Any data-dependent reporting subset inherits the event without another union bound. This proof calibrates the radius; it does not select a support size.

Lemma 4 (Empirical-mean motion)**.** *If the observation at time t has condition c and $n_c(t) \geq 2$, then*

$$\|\widehat{\mu}_t^c - \widehat{\mu}_{t-1}^c\|_{\mathcal{H}} \leq \frac{2K}{n_c(t)}. \tag{44}$$

Proof. **Proof map.** We apply the running-mean identity and the triangle inequality. Equation (44) supplies the checkpoint-motion term used by the warm-start analysis; this step adds no new concentration argument. Because the new observation belongs to condition c,

$$\widehat{\mu}_t^c = \frac{n_c(t)-1}{n_c(t)}\widehat{\mu}_{t-1}^c + \frac{1}{n_c(t)}\phi(Y_t). \tag{45}$$

Subtracting $\widehat{\mu}_{t-1}^c$ gives

$$\widehat{\mu}_t^c - \widehat{\mu}_{t-1}^c = \frac{\phi(Y_t) - \widehat{\mu}_{t-1}^c}{n_c(t)}. \tag{46}$$

The bounded-feature assumption gives $\|\phi(Y_t)\|_{\mathcal{H}} \leq K$. Since $\widehat{\mu}_{t-1}^c$ is a convex combination of vectors of norm at most K, it also has norm at most K. The triangle inequality therefore proves (44).

For any convex combination q of observed features, set $u := q - \widehat{\mu}_{t-1}^c$ and $v := \widehat{\mu}_{t-1}^c - \widehat{\mu}_t^c$. Then $\|u\|_{\mathcal{H}} \leq 2K$, and

$$\left|\frac{1}{2}\|u+v\|_{\mathcal{H}}^2 - \frac{1}{2}\|u\|_{\mathcal{H}}^2\right| = \left|\langle u, v\rangle_{\mathcal{H}} + \frac{1}{2}\|v\|_{\mathcal{H}}^2\right| \tag{47}$$

$$\leq \|u\|_{\mathcal{H}}\|v\|_{\mathcal{H}} + \frac{1}{2}\|v\|_{\mathcal{H}}^2 \tag{48}$$

$$\leq \frac{4K^2}{n_c(t)} + \frac{2K^2}{n_c(t)^2}. \tag{49}$$

Thus one stationary update changes every fixed-support objective by $O(K^2/n_c(t))$. □

Proposition 4 (Conservative Frank–Wolfe path bound)**.** *Fix (t, c) and take the declared oracle domain to be the complete observed set, $\mathcal{J}_t^c = H_t^c$. Start from one feasible atom and let q_m be the fully corrected iterate supported on at most m observed features. At the transition from m to $m+1$ atoms, let ω_m^c be the oracle miss in (20); because $\mathcal{J}_t^c = H_t^c$ here, it is a miss relative to the complete observed set. Then, for $m \geq 2$,*

$$\widehat{F}_{t,m}^c - F_t^{c,\mathrm{full}} \leq \frac{8K^2}{m} + \frac{2}{m(m-1)}\sum_{\ell=1}^{m-1}\ell\,\omega_\ell^c. \tag{50}$$

For $m = 1$, $\widehat{F}_{t,1}^c - F_t^{c,\mathrm{full}} \leq 8K^2$.

Proof. **Proof map.** We start from smooth convex Frank–Wolfe descent through the curvature definition (51). We retain the additive oracle miss (53) through the recurrence (56) and weighted telescoping (57), which yields (50). Write $F(q) := \frac{1}{2}\|q - \widehat{\mu}_t^c\|_{\mathcal{H}}^2$ for $q \in \mathcal{D}_t^c$, and let $q^\star$ minimize F over $\mathcal{D}_t^c$. The gradient is $\nabla F(q) = q - \widehat{\mu}_t^c$ and is 1-Lipschitz. The Frank–Wolfe curvature is the smallest C_F such that

$$F(q + \gamma(s - q)) \le F(q) + \gamma\langle \nabla F(q), s - q\rangle_{\mathcal{H}} + \frac{\gamma^2 C_F}{2} \tag{51}$$

for all $q, s \in \mathcal{D}_t^c$ and $\gamma \in [0, 1]$. The diameter of $\mathcal{D}_t^c$ is at most $2K$, because every generator has norm at most K; therefore $C_F \le 4K^2$ (Jaggi, 2013).

Let s_ℓ denote the feature returned by the approximate linear oracle at the exactly corrected iterate q_ℓ, and abbreviate its residual by $r_\ell := \widehat{\mu}_t^c - q_\ell$. For any comparison step size $\gamma_\ell \in [0, 1]$, the curvature inequality gives

$$F(q_\ell + \gamma_\ell(s_\ell - q_\ell)) \le F(q_\ell) + \gamma_\ell\langle \nabla F(q_\ell), s_\ell - q_\ell\rangle + \frac{\gamma_\ell^2 C_F}{2}. \tag{52}$$

The additive miss in (20) is stated in residual-correlation form. Since $\nabla F(q_\ell) = -r_\ell$, subtracting the common term $\langle \nabla F(q_\ell), -q_\ell\rangle$ converts that definition exactly into

$$\langle \nabla F(q_\ell), s_\ell - q_\ell\rangle \le \min_{s \in \mathcal{X}_t^c} \langle \nabla F(q_\ell), s - q_\ell\rangle + \omega_\ell^c. \tag{53}$$

Because $q^\star$ belongs to the convex hull of $\mathcal{X}_t^c$, linearity implies

$$\min_{s \in \mathcal{X}_t^c} \langle \nabla F(q_\ell), s - q_\ell\rangle \le \langle \nabla F(q_\ell), q^\star - q_\ell\rangle. \tag{54}$$

Convexity of F gives

$$\langle \nabla F(q_\ell), q^\star - q_\ell\rangle \le F(q^\star) - F(q_\ell). \tag{55}$$

Define $h_\ell := F(q_\ell) - F(q^\star)$. The dictionary at step $\ell + 1$ contains both q_ℓ and s_ℓ, so every point on their segment is feasible for the corrective simplex problem. Exact full correction therefore attains an objective no larger than the right side of (52). Combining (52)–(55) yields

$$h_{\ell+1} \le (1 - \gamma_\ell)h_\ell + \gamma_\ell\omega_\ell^c + \frac{\gamma_\ell^2 C_F}{2}. \tag{56}$$

Choose $\gamma_\ell = 2/(\ell + 1)$. Multiplying (56) by $\ell(\ell + 1)$ yields a curvature term $2C_F\ell/(\ell + 1)$; since $\ell/(\ell + 1) \le 1$,

$$\ell(\ell + 1)h_{\ell+1} \le (\ell - 1)\ell h_\ell + 2\ell\omega_\ell^c + 2C_F. \tag{57}$$

Summing (57) from $\ell = 1$ to $m - 1$ cancels every intermediate weighted gap and gives

$$(m - 1)mh_m \le 2\sum_{\ell=1}^{m-1} \ell\omega_\ell^c + 2C_F(m - 1). \tag{58}$$

Dividing by $m(m - 1)$ yields

$$h_m \le \frac{2C_F}{m} + \frac{2}{m(m - 1)} \sum_{\ell=1}^{m-1} \ell\omega_\ell^c. \tag{59}$$

Substituting $C_F \le 4K^2$ proves (50). For one atom, both q_1 and $\widehat{\mu}_t^c$ have norm at most K, so $F(q_1) \le \frac{1}{2}(2K)^2 = 2K^2$; the stated looser bound $8K^2$ follows from $2K^2 \le 8K^2$ and keeps the same constant convention as the recurrence. □

Proposition 4 is inherited conditional-gradient geometry. It compares the computed path with the full convex-hull optimum and may be loose relative to the best m-atom value. The support-size novelty therefore uses the separate oracle path and its norm slack rather than relabeling the standard $O(1/m)$ objective rate as a model-order theorem.

Proof of Theorem 1. **Proof map.** We start from the generic crossing inversion (33). Substituting the condition-specific sampling radius and measured path slack gives (12); the stated margin then yields exact integer recovery. Apply Proposition 3 with radius $r = \varepsilon_t^c$, support interval $\mathcal{I}$, and the paths defined in (5) and (29). The left side of (31) prevents the computed curve from crossing before $m_t^{\star,c}(\varepsilon_t^c)$; the right side forces a crossing by $m_t^{\star,c}(\varepsilon_t^c - s_{t,\mathcal{I}}^c)$. These are exactly the two inequalities in (12). If the oracle curve remains above ε_t^c at budget $m-1$ and lies below $\varepsilon_t^c - s_{t,\mathcal{I}}^c$ at budget m, monotonicity excludes every smaller budget and the upper sandwich forces the computed crossing at m. Proposition 3 proves the generic inversion line by line; this theorem specializes it to OMCI's reported radius.

Theorem 2 (Fixed-radius reachability)**.** *For $\rho > 0$, define $\widehat{m}_t^c(\rho) := \min\{1 \le m \le n_c(t) : \widehat{e}_{t,m}^c \le \rho\}$. If CrossingReadout scans the complete observed support exactly through $M_{\max} = n_c(t)$, then, for $1 \le m \le n_c(t)$,*

$$\widehat{F}_{t,m}^c \le \frac{8K^2}{m}, \qquad \widehat{e}_{t,m}^c \le \frac{4K}{\sqrt{m}}, \qquad \widehat{m}_t^c(\rho) \le \min\left\{ n_c(t), \max\left(1, \left\lceil \frac{16K^2}{\rho^2} \right\rceil\right)\right\}. \tag{60}$$

Proof of Theorem 2. **Proof map.** We apply the fixed-radius Frank–Wolfe rate (50). Equation (60) converts a prescribed tolerance into a sufficient support count; model-order identification instead comes from (12). Equation (38) gives $F_t^{c,\mathrm{full}} = 0$. Exact observed-support linear minimization sets every $\omega_\ell^c = 0$ in Proposition 4. The theorem also requires exact full correction, which justifies the line-search comparison used in that proposition. If the exact gap vanishes before the requested budget, CrossingReadout extends the path with the terminal zero-gap law as specified after (29); because path indices are at-most-support budgets, the copied law remains feasible and the same bound holds at every larger budget. Hence

$$\widehat{F}_{t,m}^c \le \frac{8K^2}{m}. \tag{61}$$

Using $\widehat{e}_{t,m}^c = \sqrt{2\widehat{F}_{t,m}^c}$ gives

$$\widehat{e}_{t,m}^c \le \frac{4K}{\sqrt{m}}. \tag{62}$$

Therefore $m \ge 16K^2/\rho^2$ suffices for $\widehat{e}_{t,m}^c \le \rho$. Independently, the empirical frequency vector represents $\widehat{\mu}_t^c$ exactly with at most $n_c(t)$ observed atoms. If no earlier zero-gap certificate occurs, every positive-gap step adds a new observed index, so the dictionary contains the complete observed set by budget $n_c(t)$; exact correction can then choose the empirical frequency vector and attain zero error. If a zero-gap certificate occurs earlier, the copied terminal law already attains the complete-hull optimum. Taking the smaller of these two sufficient support sizes proves (60).

B.6 From fixed-budget approximation to model-order selection

The preceding theorem combines two familiar scales but uses them for a different statistical decision. Lemma 3 gives, up to logarithms,

$$E_t^c(\delta) = \widetilde{O}\left(n_c(t)^{-1/2}\right), \qquad \widehat{e}_{t,m}^c = O\left(m^{-1/2}\right) \quad \Longrightarrow \quad m = \widetilde{O}(n_c(t)) \text{ at } \rho = \varepsilon_t^c. \tag{63}$$

The first rate is the Hilbert-space analogue of Monte Carlo mean estimation. The second is the square root of the standard $O(1/m)$ Frank–Wolfe objective rate. Equation (63) shows that worst-case conditional-gradient reachability at the sampling radius does not, on its own, prove that the coreset is asymptotically smaller than the observed sample. Any sharper compression statement must use additional structure, such as finite effective support, a separated sparse curve, spectral decay, or stronger geometry.

Kernel herding analyses normally fix a support budget or approximation tolerance and then bound its discrepancy (Bach et al., 2012; Lacoste-Julien et al., 2015). Bayesian and Hilbert coresets optimize weighted summaries at a fixed budget (Campbell and Broderick, 2019). Kernelized thinning likewise fixes the target sample and discrepancy before compression (Riabiz et al., 2022; Koppel et al., 2024). OMCI–MMD asks a different question: at which integer budget does an entire sparse path first fall below a condition-specific statistical radius? Three steps answer it.

1. The complete path in Algorithm 3 records a feasible error at every declared budget. A single terminal iterate cannot identify a first crossing.
2. Proposition 2 compares the computed and oracle paths uniformly in norm scale. Feasibility gives the lower side for free; a separate path-slack bound gives the upper side.
3. Theorem 1 converts that pathwise comparison into a crossing bracket. A separated crossing yields one integer; a flat crossing yields an interval because the data do not distinguish neighboring orders.

These steps do not follow from a fixed-budget discrepancy rate. They expose the exact algorithmic quantity—the path slack—that a scalable implementation must bound before claiming oracle-level model-order recovery. The algorithm does not observe this slack automatically; exhaustive search computes it only in controlled experiments, while a scalable deployment needs a separate upper bound.

The selected order is initially empirical. Lemma 3 and Remark 1 connect the empirical and population-target curves, but neither concentration nor universality creates a separated population crossing. The next corollary gives one transparent sufficient case.

Corollary 1 (Finite-support stabilization)**.** *Suppose condition c has population support of size r_c and every support point has appeared by time T_0. At every reported checkpoint $t \geq T_0$, assume $\{1, \dots, r_c\} \subseteq \mathcal{I}$, $s^c_{t,\mathcal{I}} < \varepsilon^c_t$, and $e^{c,\star}_{t,m} > \varepsilon^c_t$ for every integer $1 \leq m < r_c$. Then $\widehat{m}^c_t = r_c$ at every such checkpoint.*

Proof. **Proof map.** Once every population atom appears, the empirical law has an exact finite representation. Combining that fact with (31) and the lower-budget separation forces the first crossing to equal the minimal effective support in (6). Once every population support point has appeared, its empirical frequency vector represents $\widehat{\mu}^c_t$ using at most r_c atoms. Hence $e^{c,\star}_{t,r_c} = 0$. For every integer $1 \leq m < r_c$, feasibility and the assumed separation give

$$\widehat{e}^c_{t,m} \geq e^{c,\star}_{t,m} > \varepsilon^c_t, \tag{64}$$

so no smaller support crosses. At r_c, the path sandwich gives

$$\widehat{e}^c_{t,r_c} \leq e^{c,\star}_{t,r_c} + s^c_{t,\mathcal{I}} = s^c_{t,\mathcal{I}} < \varepsilon^c_t. \tag{65}$$

Thus the first crossing equals r_c. □

The separation assumption is essential. The literal support size r_c need not equal the RKHS model order: affine dependence among feature vectors, repeated atoms, or a radius larger than the weakest branch can allow fewer than r_c atoms to represent the mean at the declared resolution. The corollary claims stabilization only when the oracle curve itself shows that no smaller support crosses.

Remark 1 (Empirical and population sparse curves)**.** *The selected path approximates the empirical law because that is the object available online. To compare it with the stationary population mean, define*

$$F^{c,\star}_{\mu,t,m} := \inf_{\substack{A \subseteq \{i \leq t : c_i = c\} \\ |A| \leq m}} \inf_{a \in \mathsf{S}(A)} \frac{1}{2} \left\| \sum_{j \in A} a_j \phi(Y_j) - \mu^c \right\|^2_{\mathcal{H}}. \tag{66}$$

For fixed (A, a), the reverse triangle inequality gives

$$\left| \left\| \sum_j a_j \phi(Y_j) - \widehat{\mu}^c_t \right\|_{\mathcal{H}} - \left\| \sum_j a_j \phi(Y_j) - \mu^c \right\|_{\mathcal{H}} \right| \leq \|\widehat{\mu}^c_t - \mu^c\|_{\mathcal{H}}. \tag{67}$$

For a fixed budget m, let $u_m(A, a)$ and $v_m(A, a)$ denote the two norm objectives in the preceding display. The elementary inequality $|\inf_x u_m(x) - \inf_x v_m(x)| \leq \sup_x |u_m(x) - v_m(x)|$ then applies because both infima range over the same support-and-weight class. Hence

$$\max_{1 \leq m \leq n_c(t)} \left| \sqrt{2F^{c,\star}_{t,m}} - \sqrt{2F^{c,\star}_{\mu,t,m}} \right| \leq \|\widehat{\mu}^c_t - \mu^c\|_{\mathcal{H}}. \tag{68}$$

On the event (39), the entire empirical oracle curve lies within $E^c_t(\delta)$ of the population-target curve in norm scale. A unique population model order still requires separation around the relevant radius; concentration alone cannot create such a gap.

Rare-condition sampling limit. Assume the branch indicators are independent across the n condition-c observations and each equals one with probability $p_b \in (0,1)$. Then

$$\Pr\{\text{branch } b \text{ remains unseen after } n \text{ condition-}c \text{ observations}\} = (1-p_b)^n. \tag{69}$$

If $(1-p_b)^n \geq \delta$, equivalently $n \leq \log(\delta)/\log(1-p_b)$, the sample contains no observation from the branch with probability at least δ. For small p_b, this threshold has the familiar scale $np_b \lesssim \log(1/\delta)$. Without additional structural assumptions, the observations therefore cannot certify absence of the branch at confidence $1-\delta$. Under dependent sampling the exact probability changes and requires a dependence-aware bound. OMCI–MMD reports both $n_c(t)$ and the corresponding radius E_t^c for each condition; a small estimated order under sparse data is not evidence that rare structure is absent.

B.7 Windowed path under drift

For a window length $L \geq 1$, define

$$I_{t,L} := \{\max(1, t-L+1), \ldots, t\}, \qquad I_{t,L}^c := \{s \in I_{t,L} : c_s = c\}, \qquad n_{t,L}^c := |I_{t,L}^c|. \tag{70}$$

For every stream time s and condition c, let P_s^c denote the endpoint law that would govern an observation with label c at time s, and define

$$\mu_s^c := \int \phi(y)\, dP_s^c(y). \tag{71}$$

Condition on the complete realized label sequence, as in Appendix B.5, and assume that $Y_s \sim P_s^{c_s}$ independently under this conditional law. This formulation defines P_s^c and μ_s^c for every condition, including conditions not realized at time s; it avoids conditioning on a zero-probability label event. For $n_{t,L}^c > 0$, define

$$\bar{\mu}_{t,L}^c := \frac{1}{n_{t,L}^c} \sum_{s \in I_{t,L}^c} \mu_s^c, \qquad \widehat{\mu}_{t,L}^c := \frac{1}{n_{t,L}^c} \sum_{s \in I_{t,L}^c} \phi(Y_s). \tag{72}$$

The average population mean $\bar{\mu}_{t,L}^c$ centers the empirical window, whereas μ_t^c is the current target. Their separation creates stale-data bias. This local-estimation viewpoint follows the variance–bias logic for nonstationary processes and change-point methods (Dahlhaus, 1997; Basseville and Nikiforov, 1993).

Define the sampling term, latent staleness, declared staleness envelope, and operational radius by

$$E_{t,L}^c := C_{\mathrm{H}} K \sqrt{\frac{\log(4|\mathcal{C}|t^2/\delta)}{n_{t,L}^c}}, \tag{73}$$

$$S_{t,L}^c := \frac{1}{n_{t,L}^c} \sum_{s \in I_{t,L}^c} \|\mu_s^c - \mu_t^c\|_{\mathcal{H}}, \qquad S_{t,L}^c \leq \overline{S}_{t,L}^c, \tag{74}$$

$$\varepsilon_{t,L}^c := \kappa(E_{t,L}^c + \overline{S}_{t,L}^c). \tag{75}$$

The full-window observed-support oracle is

$$F_{t,L,m}^{c,\star} := \inf_{\substack{A \subseteq I_{t,L}^c \\ |A| \leq m}} \inf_{a \in \mathsf{S}(A)} \frac{1}{2} \left\| \sum_{j \in A} a_j \phi(Y_j) - \widehat{\mu}_{t,L}^c \right\|_{\mathcal{H}}^2, \qquad e_{t,L,m}^{c,\star} := \sqrt{2F_{t,L,m}^{c,\star}}. \tag{76}$$

Declare a nonempty candidate set $\varnothing \neq \mathcal{J}_{t,L}^c \subseteq I_{t,L}^c$ and a scanned budget $1 \leq M_{\max} \leq |\mathcal{J}_{t,L}^c|$. The candidate-restricted oracle is

$$F_{t,L,m}^{c,\mathcal{J},\star} := \inf_{\substack{A \subseteq \mathcal{J}_{t,L}^c \\ |A| \leq m}} \inf_{a \in \mathsf{S}(A)} \frac{1}{2} \left\| \sum_{j \in A} a_j \phi(Y_j) - \widehat{\mu}_{t,L}^c \right\|_{\mathcal{H}}^2. \tag{77}$$

For the nested dictionary $A_{t,L,m}^c \subseteq \mathcal{J}_{t,L}^c$ and corrected weights $a_{t,L,m}^c \in \mathsf{S}(A_{t,L,m}^c)$ returned by Algorithm 2, define the computed path

$$\widehat{F}_{t,L,m}^c := \frac{1}{2} \left\| \sum_{j \in A_{t,L,m}^c} a_{t,L,m,j}^c \phi(Y_j) - \widehat{\mu}_{t,L}^c \right\|_{\mathcal{H}}^2, \qquad \widehat{e}_{t,L,m}^c := \sqrt{2\widehat{F}_{t,L,m}^c}. \tag{78}$$

With the convention $\min \varnothing := +\infty$, its operational crossing is

$$\widehat{m}^c_{t,L} := \min\left\{1 \le m \le M_{\max} : \widehat{F}^c_{t,L,m} \le \tfrac{1}{2}(\varepsilon^c_{t,L})^2\right\} = \min\{1 \le m \le M_{\max} : \widehat{e}^c_{t,L,m} \le \varepsilon^c_{t,L}\}. \tag{79}$$

Thus a finite value of $\widehat{m}^c_{t,L}$ shows that the scanned path reaches the declared window radius. These definitions separate three comparisons: computed versus candidate-restricted path, candidate-restricted versus full-window observed-support path, and empirical-window versus current-population mean. In particular,

$$[\widehat{F}^c_{t,L,m} - F^{c,\star}_{t,L,m}]_+ \le [\widehat{F}^c_{t,L,m} - F^{c,\mathcal{J},\star}_{t,L,m}]_+ + [F^{c,\mathcal{J},\star}_{t,L,m} - F^{c,\star}_{t,L,m}]_+. \tag{80}$$

The first term measures path construction and numerical correction inside the declared pool; the second measures candidate availability. Appendix B.7 expands them into the quantities used in Proposition 6.

Corollary 2 (Windowed fixed-radius reachability)**.** *If $\varepsilon^c_{t,L} > 0$ and CrossingReadout scans the complete window support $\mathcal{J}^c_{t,L} = I^c_{t,L}$ through $M_{\max} = n^c_{t,L}$ with exact linear minimization and full correction, then*

$$\widehat{m}^c_{t,L} \le \min\left\{n^c_{t,L}, \max\left(1, \left\lceil \frac{16K^2}{(\varepsilon^c_{t,L})^2} \right\rceil\right)\right\}. \tag{81}$$

Proof. **Proof map.** We apply the stationary fixed-radius reachability argument at one frozen checkpoint. Using the empirical window mean in (72) gives (81); staleness, candidate availability, and crossing error enter only in the later nonstationary proofs. Fix a checkpoint and reindex its $n^c_{t,L}$ window observations as a finite sample. The proof of Theorem 2 uses only bounded convex-hull geometry at that checkpoint; it does not use the observations' chronological positions. Because $\widehat{\mu}^c_{t,L}$ is a convex combination of those window features, the optimum over their complete convex hull is zero. Apply Theorem 2 to the fixed target $\widehat{\mu}^c_{t,L}$ with sample count $n^c_{t,L}$ and radius $\rho = \varepsilon^c_{t,L}$. The first term in the minimum follows because the empirical window law itself uses at most $n^c_{t,L}$ atoms. □

The corollary guarantees a crossing only under a complete exact scan. With a finite pool, Proposition 6 instead keeps within-pool search, candidate availability, and numerical tracking separate; Appendix B.7 defines and proves those terms line by line.

This appendix proves the recent-window radius and expands every term in Proposition 6. Windowing is inherited from local estimation under nonstationarity: shorter windows reduce stale-data bias but increase variance, while longer windows do the opposite (Dahlhaus, 1997; Basseville and Nikiforov, 1993). OMCI–MMD adds a different object to that tradeoff: a complete support-indexed path whose first crossing determines endpoint cardinality. The proof keeps within-pool search, candidate availability, corrective optimization, warm-start tracking, and population staleness separate because a different algorithmic module controls each term.

Equations (70)–(79) define all window objects in dependency order. In particular, (71) defines the current condition-c mean μ^c_t from the time-indexed endpoint law P^c_t; $\bar{\mu}^c_{t,L}$ averages those population means over the realized condition-c indices in the window; and $\widehat{\mu}^c_{t,L}$ is the corresponding empirical mean. Condition on the complete realized label sequence and assume the conditional-kernel sampling model stated after (71). For each reported pair (t, c), the centered variables $\{\phi(Y_s) - \mu^c_s : s \in I^c_{t,L}\}$ are then independent, mean zero, and bounded by $2K$. We predeclare L and the reporting rule. Selecting a window after inspecting endpoint values requires a simultaneous bound over the candidate windows or fresh calibration data; the single-window radius below does not cover that reuse.

B.8 Sampling–staleness tradeoff

Assume reporting begins only after a full window is available, so every reported checkpoint satisfies $t \ge L$. For a constant $p_c \in (0, 1]$ and every reported window up to horizon T, assume the condition count satisfies $n^c_{t,L} \ge p_c L$. Define

$$a_c(T, \delta) := C_{\mathrm{H}} K \sqrt{\frac{\log(4|\mathcal{C}|T^2/\delta)}{p_c}}. \tag{82}$$

Then the sampling term in (73) obeys

$$E^c_{t,L} \le a_c(T,\delta)L^{-1/2}. \tag{83}$$

Assume also that the current conditional mean changes by at most $v_c \ge 0$ per time step:

$$\|\mu^c_{s+1} - \mu^c_s\|_{\mathcal{H}} \le v_c \quad \text{for every relevant } s. \tag{84}$$

For $s \in I^c_{t,L}$, repeated application of the triangle inequality gives

$$\|\mu^c_s - \mu^c_t\|_{\mathcal{H}} \le \sum_{u=s}^{t-1} \|\mu^c_{u+1} - \mu^c_u\|_{\mathcal{H}} \le v_c(t-s) \le v_c(L-1). \tag{85}$$

Averaging (85) over the condition-specific window yields

$$S^c_{t,L} \le v_c(L-1). \tag{86}$$

Consequently, (83) and (86) bound the sampling-plus-staleness part of the implementable radius, up to the calibration factor κ, by

$$a_c(T,\delta)L^{-1/2} + v_c(L-1). \tag{87}$$

Abbreviate $a_c := a_c(T,\delta)$. Treating L temporarily as a positive real and differentiating the looser expression $a_cL^{-1/2} + v_cL$ gives

$$-\frac{a_c}{2}L^{-3/2} + v_c = 0, \tag{88}$$

so an order-optimal balance is $L \asymp (a_c/v_c)^{2/3}$ when $v_c > 0$. The implementation must still round to an admissible integer and respect the minimum condition count. This calculation motivates window choice; it does not estimate v_c or certify the envelope $\overline{S}^c_{t,L}$.

Lemma 5 (Windowed endpoint radius)**.** *Fix and predeclare L. Suppose the conditional-independence and bounded-feature assumptions above hold, and suppose the envelope $S^c_{t,L} \le \overline{S}^c_{t,L}$ holds deterministically at every reported pair. Then, with probability at least $1-\delta$, every reported (t,c) with $n^c_{t,L} > 0$ satisfies*

$$\|\widehat{\mu}^c_{t,L} - \mu^c_t\|_{\mathcal{H}} \le E^c_{t,L} + S^c_{t,L} \le E^c_{t,L} + \overline{S}^c_{t,L}. \tag{89}$$

Proof of Lemma 5. Proof map. We combine Hilbert-space concentration with the triangle inequality. Equation (90) separates window sampling from temporal bias, and (92) keeps staleness explicit in the current-law bound (89). Start from the exact decomposition

$$\widehat{\mu}^c_{t,L} - \mu^c_t = (\widehat{\mu}^c_{t,L} - \bar{\mu}^c_{t,L}) + (\bar{\mu}^c_{t,L} - \mu^c_t). \tag{90}$$

Conditional on the label sequence, the centered vectors $\phi(Y_s) - \mu^c_s$, $s \in I^c_{t,L}$, are independent and have norm at most $2K$. Applying the Hilbert-space concentration argument from (40) at each reported pair and the failure-probability allocation in (41), with the common constant convention absorbed into C_{H}, gives

$$\|\widehat{\mu}^c_{t,L} - \bar{\mu}^c_{t,L}\|_{\mathcal{H}} \le E^c_{t,L} \tag{91}$$

on an event of probability at least $1-\delta$, simultaneously over the reported pairs. Next,

$$\begin{aligned}\|\bar{\mu}^c_{t,L} - \mu^c_t\|_{\mathcal{H}} &= \left\| \frac{1}{n^c_{t,L}} \sum_{s \in I^c_{t,L}} (\mu^c_s - \mu^c_t) \right\|_{\mathcal{H}} \\ &\le \frac{1}{n^c_{t,L}} \sum_{s \in I^c_{t,L}} \|\mu^c_s - \mu^c_t\|_{\mathcal{H}} = S^c_{t,L},\end{aligned} \tag{92}$$

where the inequality is convexity of the norm. Apply the triangle inequality to (90), then substitute (91) and (92):

$$\|\widehat{\mu}^c_{t,L} - \mu^c_t\|_{\mathcal{H}} \le E^c_{t,L} + S^c_{t,L}. \tag{93}$$

Finally, substitute the deterministic envelope $S^c_{t,L} \le \overline{S}^c_{t,L}$ to obtain the second inequality in (89). For a random envelope, split the failure budget as $\delta_{\mathrm{samp}} + \delta_{\mathrm{stale}} \le \delta$, replace δ by δ_{samp} in the sampling radius, and require the envelope to hold jointly over reported pairs with probability at least $1 - \delta_{\mathrm{stale}}$. A union bound then restores confidence $1-\delta$. Without a deterministic or separately calibrated envelope, the algorithm bounds compression error relative to the empirical window but not proximity to the current population law; estimating the envelope from the same data without paying for that selection invalidates the stated coverage.

B.9 POOL, OBSERVED-SUPPORT, AND FREE-SUPPORT ORACLES

At checkpoint t, sparse step m, and condition c, let $A^c_{t,L,m}$ denote the dictionary prefix selected by the implementation. The analysis replays that fixed prefix against the current window objective and computes exact corrective weights

$$a^{c,\mathrm{aud}}_{t,L,m} \in \arg\min_{a \in \mathsf{S}(A^c_{t,L,m})} \frac{1}{2} \left\| \sum_{j \in A^c_{t,L,m}} a_j \phi(Y_j) - \widehat{\mu}^c_{t,L} \right\|^2_{\mathcal{H}}. \tag{94}$$

Define the corresponding replayed residual and endpoint gain by

$$r^{c,\mathrm{aud}}_{t,L,m} := \widehat{\mu}^c_{t,L} - \sum_{j \in A^c_{t,L,m}} a^{c,\mathrm{aud}}_{t,L,m,j} \phi(Y_j), \qquad g^{c,\mathrm{aud}}_{t,L,m}(y) := \langle r^{c,\mathrm{aud}}_{t,L,m}, \phi(y) \rangle_{\mathcal{H}}. \tag{95}$$

The selected indices remain exactly those produced online; replay changes only the weights and the gains used for replay. If the implementation selected a proposal using an inexact warm state, the proof evaluates that fixed proposal ex post against $g^{c,\mathrm{aud}}_{t,L,m}$. Any gain-prediction theorem invoked below must therefore control prediction error relative to these replayed gains, not merely relative to the implementation's approximate residual. The complete observed oracle ranges over the index set $I^c_{t,L}$. At a transition with positive Frank–Wolfe gap, let $\varnothing \neq \mathcal{P}^c_{t,m} \subseteq \mathcal{J}^c_{t,L} \setminus A^c_{t,L,m} \subseteq I^c_{t,L} \setminus A^c_{t,L,m}$ be the index pool that the implementation actually searches, and let $\widehat{i}^c_{t,m} \in \mathcal{P}^c_{t,m}$ be its proposal. Using indices rather than endpoint values preserves multiplicities and makes the requirement that a positive-gap step add a new dictionary element explicit. Let $\mathcal{Y}^c$ denote the ambient endpoint domain. Define

$$\omega^{c,\mathrm{pool}}_{t,m} := \max_{i \in \mathcal{P}^c_{t,m}} g^{c,\mathrm{aud}}_{t,L,m}(Y_i) - g^{c,\mathrm{aud}}_{t,L,m}(Y_{\widehat{i}^c_{t,m}}), \tag{96}$$

$$\omega^{c,\mathrm{obs}}_{t,m} := \max_{i \in I^c_{t,L}} g^{c,\mathrm{aud}}_{t,L,m}(Y_i) - g^{c,\mathrm{aud}}_{t,L,m}(Y_{\widehat{i}^c_{t,m}}), \tag{97}$$

$$\omega^{c,\mathrm{free}}_{t,m} := \sup_{y \in \mathcal{Y}^c} g^{c,\mathrm{aud}}_{t,L,m}(y) - g^{c,\mathrm{aud}}_{t,L,m}(Y_{\widehat{i}^c_{t,m}}). \tag{98}$$

The observed-support miss decomposes exactly as

$$\omega^{c,\mathrm{obs}}_{t,m} = \omega^{c,\mathrm{pool}}_{t,m} + \omega^{c,\mathrm{avail}}_{t,m}, \qquad \omega^{c,\mathrm{avail}}_{t,m} := \max_{i \in I^c_{t,L}} g^{c,\mathrm{aud}}_{t,L,m}(Y_i) - \max_{i \in \mathcal{P}^c_{t,m}} g^{c,\mathrm{aud}}_{t,L,m}(Y_i). \tag{99}$$

The first term measures ranking error inside the declared pool. The second measures whether the pool exposed the best observed endpoint at all. The free-support miss requires a separate covering or smoothness assumption and does not enter Proposition 6. If the computed gap is zero, the algorithm extends the terminal law instead of defining another pool or proposal.

Predicted gain and its error. A gain learner observes labels only at indices where the implementation evaluates the kernel residual. Let $\widehat{G}^c_{t,m} : \mathcal{Y}^c \to \mathbb{R}$ denote the score predictor that ranks the declared pool. For $i \in \mathcal{P}^c_{t,m}$, write

$$\widetilde{g}^c_{t,L,m}(Y_i) = g^{c,\mathrm{aud}}_{t,L,m}(Y_i) + \epsilon^{c,\mathrm{feat}}_{t,m}(i), \tag{100}$$

where exact RKHS evaluation gives $\epsilon^{c,\mathrm{feat}}_{t,m}(i) = 0$ and random features contribute approximation error. Define

$$\zeta^c_{t,m} := \max_{i \in \mathcal{P}^c_{t,m}} |\epsilon^{c,\mathrm{feat}}_{t,m}(i)|,$$

$$\xi^c_{t,m} := \max_{i \in \mathcal{P}^c_{t,m}} |\widehat{G}^c_{t,m}(Y_i) - \widetilde{g}^c_{t,L,m}(Y_i)|,$$

and define the total uniform error relative to the exact gain by

$$\chi^c_{t,m} := \max_{i \in \mathcal{P}^c_{t,m}} |\widehat{G}^c_{t,m}(Y_i) - g^{c,\mathrm{aud}}_{t,L,m}(Y_i)|. \tag{101}$$

The triangle inequality gives $\chi^c_{t,m} \leq \xi^c_{t,m} + \zeta^c_{t,m}$.

Lemma 6 (Uniform gain error controls the pool miss)**.** *If $\widehat{i}^c_{t,m}$ maximizes $\widehat{G}^c_{t,m}(Y_i)$ over $i \in \mathcal{P}^c_{t,m}$, then*

$$\omega^{c,\mathrm{pool}}_{t,m} \le 2\chi^c_{t,m}. \tag{102}$$

Proof. **Proof map.** We use argmax stability under uniform approximation error. Equation (102) turns the gain predictor error into the additive within-pool linear-oracle miss required by the Frank–Wolfe recurrence. Choose $i^\star \in \arg\max_{i\in\mathcal{P}^c_{t,m}} g^{c,\mathrm{aud}}_{t,L,m}(Y_i)$, and abbreviate $g := g^{c,\mathrm{aud}}_{t,L,m}$, $\widehat{G} := \widehat{G}^c_{t,m}$, $\widehat{i} := \widehat{i}^c_{t,m}$, and $\chi := \chi^c_{t,m}$. Uniform approximation gives

$$g(Y_{i^\star}) \le \widehat{G}(Y_{i^\star}) + \chi. \tag{103}$$

Maximality of $\widehat{i}$ on the same index pool gives

$$\widehat{G}(Y_{i^\star}) \le \widehat{G}(Y_{\widehat{i}}). \tag{104}$$

A second use of uniform approximation gives

$$\widehat{G}(Y_{\widehat{i}}) \le g(Y_{\widehat{i}}) + \chi. \tag{105}$$

Combining the three inequalities yields $g(Y_{i^\star}) - g(Y_{\widehat{i}}) \le 2\chi$, which is exactly (102). □

A mean-squared or pointwise prediction guarantee does not imply the finite-pool supremum bound in (102). The implementation must control $\chi^c_{t,m}$, use a simultaneous confidence bound over the pool, or report the measured pool miss.

Let $\tau_1 < \cdots < \tau_R \le T$ be the actual stream times of the $R \ge 1$ reported checkpoints up to stream horizon T, and write $\mathcal{T}_{\mathrm{rep}} := \{\tau_1, \dots, \tau_R\}$. We keep $t \in \mathcal{T}_{\mathrm{rep}}$ as the stream-time index so that objects such as $I^c_{t,L}$ retain their original meaning. For terminal path budget $2 \le M \le M_{\max}$ and declared tolerances $\eta^c_{t,m,\mathrm{gain}} \ge 0$, write $[x]_+ := \max\{x, 0\}$ and define the curvature-weighted within-pool search term

$$\mathcal{R}^c_{R,M} := \frac{2}{RM(M-1)} \sum_{t\in\mathcal{T}_{\mathrm{rep}}} \sum_{m=1}^{M-1} m \left(\eta^c_{t,m,\mathrm{gain}} + [\omega^{c,\mathrm{pool}}_{t,m} - \eta^c_{t,m,\mathrm{gain}}]_+ \right). \tag{106}$$

Dynamic-regret interface for support proposals. To state exactly what a gain learner must provide, let $G : \mathcal{Y}^c \to \mathbb{R}$ be any endpoint-scoring function and define the finite-pool uniform loss

$$\ell^c_{t,m}(G) := \max_{i\in\mathcal{P}^c_{t,m}} |G(Y_i) - g^{c,\mathrm{aud}}_{t,L,m}(Y_i)|. \tag{107}$$

For a predictor $\widehat{G}^c_{t,m}$ and a comparator sequence $G^{c,\circ}_{t,m}$, define the curvature-weighted dynamic regret and comparator approximation terms

$$\mathrm{DReg}^c_{R,M} := \sum_{t\in\mathcal{T}_{\mathrm{rep}}} \sum_{m=1}^{M-1} m \left[\ell^c_{t,m}(\widehat{G}^c_{t,m}) - \ell^c_{t,m}(G^{c,\circ}_{t,m}) \right], \tag{108}$$

$$\mathrm{App}^c_{R,M} := \sum_{t\in\mathcal{T}_{\mathrm{rep}}} \sum_{m=1}^{M-1} m \ell^c_{t,m}(G^{c,\circ}_{t,m}). \tag{109}$$

Here $R \ge 1$ is the number of reported checkpoints and $2 \le M \le M_{\max}$ is the terminal path budget; T remains the stream horizon used in the sampling radius. The comparator may vary with time; its path length or product-RKHS norm is whatever quantity the chosen dynamic-regret theorem requires.

Proposition 5 (From dynamic gain regret to within-pool search error)**.** *Suppose $\widehat{i}^c_{t,m}$ maximizes $\widehat{G}^c_{t,m}(Y_i)$ over $i \in \mathcal{P}^c_{t,m}$ at every reported transition, and set $\eta^c_{t,m,\mathrm{gain}} = 0$ in (106). Then*

$$\mathcal{R}^c_{R,M} \le \frac{4}{RM(M-1)} \left(\mathrm{DReg}^c_{R,M} + \mathrm{App}^c_{R,M} \right). \tag{110}$$

Proof. **Proof map.** We use dynamic-regret bookkeeping for finite-pool losses. The uniform loss in (107) converts predictor regret into the search error (110); candidate availability remains a separate term. By definition, $\chi^c_{t,m} = \ell^c_{t,m}(\widehat{G}^c_{t,m})$. Lemma 6 gives

$$\omega^{c,\mathrm{pool}}_{t,m} \le 2\ell^c_{t,m}(\widehat{G}^c_{t,m}). \tag{111}$$

With zero declared gain tolerance, (106) becomes

$$\mathcal{R}^c_{R,M} = \frac{2}{RM(M-1)} \sum_{t\in\mathcal{T}_{\mathrm{rep}}} \sum_{m=1}^{M-1} m\omega^{c,\mathrm{pool}}_{t,m}. \tag{112}$$

Apply the preceding pointwise inequality and use

$$\sum_{t\in\mathcal{T}_{\mathrm{rep}}} \sum_{m=1}^{M-1} m\ell^c_{t,m}(\widehat{G}^c_{t,m}) = \mathrm{DReg}^c_{R,M} + \mathrm{App}^c_{R,M}, \tag{113}$$

which follows directly from (108)–(109). This proves (110). □

Proposition 5 is an interface, not a free support-discovery theorem. DynaPOLK combines functional online gradient descent with greedy RKHS projections and obtains, for convex losses, a path-length dynamic-regret bound of order $O(1 + T\sqrt{\epsilon_{\mathrm{proj}}} + W_T)$ up to stepsize constants, where W_T denotes the comparator path length in that analysis; its projection tolerance ϵ_{proj} controls dictionary growth (Bedi et al., 2020). More recent dynamic-to-static reductions provide another route for time-varying RKHS comparators (Jacobsen et al., 2025). Neither result can be substituted mechanically here. Our support oracle ranks endpoints by the residual gain in (95); Proposition 5 needs a guarantee for the uniform finite-pool loss (107). Mean-squared prediction regret, pointwise loss regret, or a small projected dictionary does not imply that uniform ranking guarantee. Even a perfect ranking inside $\mathcal{P}^c_{t,m}$ leaves the availability gap (99) untouched because no learner can select an observed endpoint that the pool excluded.

B.10 Aggregate path terms and proof of Proposition 6

Keep the reporting-time set $\mathcal{T}_{\mathrm{rep}}$ and fix a terminal support budget $2 \le M \le M_{\max}$. If an exact zero gap occurs before budget M, extend the terminal law as in Algorithm 3 and set the subsequent oracle misses to zero. At one such checkpoint t, replay the implementation's selected endpoint prefixes against the current window objective. Using the exact corrective weights from (94), define the replay objective at budget m by

$$\widetilde{F}^c_{t,L,m} := \frac{1}{2}\left\| \sum_{j\in A^c_{t,L,m}} a^{c,\mathrm{aud}}_{t,L,m,j}\phi(Y_j) - \widehat{\mu}^c_{t,L} \right\|^2_{\mathcal{H}}. \tag{114}$$

The replay keeps the endpoint sequence fixed, evaluates every gain against the corresponding exact residual in (95), and changes only the corrective weights used for replay. The reported terminal value $\widehat{F}^c_{t,L,M}$ may use an inexact quadratic-program solve or a warm-started online state. We require explicit nonnegative bounds $\epsilon^c_{\mathrm{opt},t,M}$ and $\epsilon^c_{\mathrm{track},t,M}$ such that

$$\widehat{F}^c_{t,L,M} \le \widetilde{F}^c_{t,L,M} + \epsilon^c_{\mathrm{opt},t,M} + \epsilon^c_{\mathrm{track},t,M}. \tag{115}$$

The first term bounds incomplete corrective optimization on the reported dictionary. The second bounds any remaining discrepancy between the stored warm-start path and its current-window checkpoint replay. Recomputing the path from scratch and solving each simplex problem exactly sets both terms to zero; otherwise (115) is an assumption that the implementation must verify or upper-bound.

Equation (106) defines the within-pool term $\mathcal{R}^c_{R,M}$. Define the remaining candidate-availability and numerical terms by

$$\mathcal{A}^c_{R,M} := \frac{2}{RM(M-1)} \sum_{t\in\mathcal{T}_{\mathrm{rep}}} \sum_{m=1}^{M-1} m\,\omega^{c,\mathrm{avail}}_{t,m}, \tag{116}$$

$$\mathcal{N}^c_{R,M} := \frac{1}{R} \sum_{t\in\mathcal{T}_{\mathrm{rep}}} \left(\epsilon^c_{\mathrm{opt},t,M} + \epsilon^c_{\mathrm{track},t,M}\right). \tag{117}$$

The weights m are not cosmetic: they are exactly the telescoping weights generated by the Frank–Wolfe recurrence in (57).

Proposition 6 (Average terminal-budget excess)**.** *Let $R \geq 1$ be the number of reporting checkpoints and let $2 \leq M \leq M_{\max}$ be the terminal support budget. At each checkpoint, suppose the implementation reports a feasible at-most-M-atom law supported on the current window, and suppose the analysis constructs the exact replay through budget M and computes the oracle misses in (96)–(99) along that replay. If the window-objective Frank–Wolfe curvature obeys $C_F \leq 4K^2$ and (115) holds, then*

$$\frac{1}{R}\sum_{t\in\mathcal{T}_{\mathrm{rep}}}[\widehat{F}^{c}_{t,L,M} - F^{c,\star}_{t,L,M}]_+ \leq \frac{2C_F}{M} + \mathcal{R}^{c}_{R,M} + \mathcal{A}^{c}_{R,M} + \mathcal{N}^{c}_{R,M}. \tag{118}$$

The three additive terms have distinct operational meanings: search inside the declared pool, availability of useful observed endpoints, and numerical or warm-start error. A separate coverage theorem may upper-bound the availability term; without one, that term remains explicit. The result concerns the terminal budget M, not the uniform error of the entire path. Its proof compares the replay with the full observed convex hull, whose optimum is zero, and therefore yields a conservative excess bound relative to the best M-atom curve. A uniform crossing theorem would require pathwise control over every candidate budget, just as Theorem 1 does under stationarity.

Proof of Proposition 6. Proof map. We reuse the Frank–Wolfe recurrence (56) and weighted telescoping (57). Replaying the current window and applying (123) gives (118) with search and availability errors kept separate. Fix a checkpoint t. The replay uses the same selected endpoint sequence as the implementation but evaluates every gain against the current window objective and solves every corrective weight problem exactly. At transition m, its additive linear-oracle miss relative to the complete observed window is precisely $\omega^{c,\mathrm{obs}}_{t,m}$. Apply the approximate Frank–Wolfe recurrence (56) with those misses. The complete observed window convex hull contains $\widehat{\mu}^{c}_{t,L}$, so the full-hull optimum equals zero. Full correction can only improve on each feasible line step, and the weighted recurrence (57) therefore telescopes through terminal budget $M \geq 2$ to give

$$\widetilde{F}^{c}_{t,L,M} \leq \frac{2C_F}{M} + \frac{2}{M(M-1)}\sum_{m=1}^{M-1} m\,\omega^{c,\mathrm{obs}}_{t,m}. \tag{119}$$

Substitute the numerical interface (115):

$$\widehat{F}^{c}_{t,L,M} \leq \frac{2C_F}{M} + \frac{2}{M(M-1)}\sum_{m=1}^{M-1} m\,\omega^{c,\mathrm{obs}}_{t,m} + \epsilon^{c}_{\mathrm{opt},t,M} + \epsilon^{c}_{\mathrm{track},t,M}. \tag{120}$$

Because the best M-atom full-window objective in (76) is nonnegative,

$$[\widehat{F}^{c}_{t,L,M} - F^{c,\star}_{t,L,M}]_+ \leq \widehat{F}^{c}_{t,L,M}. \tag{121}$$

For each transition (t,m), (99) and the elementary inequality

$$x \leq \eta + [x-\eta]_+ \qquad (x,\eta \geq 0) \tag{122}$$

give

$$\omega^{c,\mathrm{obs}}_{t,m} \leq \eta^{c}_{t,m,\mathrm{gain}} + [\omega^{c,\mathrm{pool}}_{t,m} - \eta^{c}_{t,m,\mathrm{gain}}]_+ + \omega^{c,\mathrm{avail}}_{t,m}. \tag{123}$$

Insert (123) into (120), apply (121), sum over the R checkpoints, and divide by R. The three weighted sums are exactly (106)–(117), which proves (118).

Corollary 3 (From terminal-budget error to crossing frequency)**.** *For each checkpoint, define the squared window threshold*

$$\theta^{c}_{t,L} := \frac{1}{2}(\varepsilon^{c}_{t,L})^2. \tag{124}$$

Suppose there is a margin $\Delta > 0$ such that

$$F^{c,\star}_{t,L,M} \leq \theta^{c}_{t,L} - \Delta \qquad \textit{for every } t \in \mathcal{T}_{\mathrm{rep}}. \tag{125}$$

Define

$$\mathcal{B}^c_{R,M} := \frac{2C_F}{M} + \mathcal{R}^c_{R,M} + \mathcal{A}^c_{R,M} + \mathcal{N}^c_{R,M}. \tag{126}$$

Then the fraction of checkpoints at which the reported path has not crossed by budget M satisfies

$$\frac{1}{R}\sum_{t\in\mathcal{T}_{\mathrm{rep}}} \mathbf{1}\{\widehat{m}^c_{t,L} > M\} \le \frac{1}{R}\sum_{t\in\mathcal{T}_{\mathrm{rep}}} \mathbf{1}\{\widehat{F}^c_{t,L,M} > \theta^c_{t,L}\} \le \min\left\{1, \frac{\mathcal{B}^c_{R,M}}{\Delta}\right\}. \tag{127}$$

Here $\widehat{m}^c_{t,L} = +\infty$ counts as a failure to cross by M.

Proof. **Proof map.** We combine a margin argument with an indicator bound. Equation (127) converts terminal-budget path error into the frequency of failing to cross by the declared budget. If $\widehat{m}^c_{t,L} > M$, no path value at a budget $1, \dots, M$ lies below its threshold. In particular, $\widehat{F}^c_{t,L,M} > \theta^c_{t,L}$, which proves the first inequality in (127); this implication does not require numerical monotonicity of the reported path. Under the oracle margin (125), the event $\widehat{F}^c_{t,L,M} > \theta^c_{t,L}$ implies

$$[\widehat{F}^c_{t,L,M} - F^{c,\star}_{t,L,M}]_+ > \Delta. \tag{128}$$

Therefore

$$\mathbf{1}\{\widehat{F}^c_{t,L,M} > \theta^c_{t,L}\} \le \frac{[\widehat{F}^c_{t,L,M} - F^{c,\star}_{t,L,M}]_+}{\Delta}. \tag{129}$$

Sum this inequality over checkpoints, divide by R, and apply Proposition 6. The bound by one is automatic for any empirical fraction. □

Comparison with dynamic-regret analyses. DynaPOLK and related nonstationary kernel methods bound predictive dynamic regret while a projection rule controls dictionary complexity (Bedi et al., 2020; Jacobsen et al., 2025). Proposition 6 and Corollary 3 solve a different interface. They begin with a current conditional endpoint law, calibrate a windowed sampling-and-staleness threshold, propagate approximate support-oracle errors through the Frank–Wolfe path, and convert an oracle margin into a bound on failed support-size crossings. The conclusion remains conditional on the declared staleness envelope, candidate coverage, and numerical error bounds. The proof displays those dependencies instead of absorbing them into one regret term.

Optional product-RKHS gain comparator. Let $\mathcal{H}_\tau$ be an RKHS over reporting time and $\mathcal{H}_y$ an RKHS over endpoints. For each sparse index m, suppose the comparator sequence in (108) admits the product-space representation

$$G^{c,\circ}_m \in \mathcal{H}_\tau \otimes \mathcal{H}_y, \qquad \|G^{c,\circ}_m\|_{\mathcal{H}_\tau\otimes\mathcal{H}_y} \le B^c_m. \tag{130}$$

Define $G^{c,\circ}_{t,m}(y) := G^{c,\circ}_m(t, y)$. The dynamic-to-static reduction of Jacobsen et al. (2025) supplies one route to a regret bound for such time-varying comparators. To invoke Proposition 5, the chosen learner and observation scheme must produce a guarantee for the uniform finite-pool loss (107), or a confidence bound that dominates it. A mean-squared regret theorem alone is insufficient, and no gain theorem can remove the availability term because an endpoint absent from the pool cannot be selected.

Increasing M reduces sparse approximation error but raises memory and correction cost. Increasing L reduces sampling noise but includes older observations. Enlarging the candidate pool lowers $\mathcal{A}^c_{R,M}$ but raises search cost. The explicit terms in Proposition 6 identify which algorithmic module must justify each tradeoff.

B.11 From path tracking to current-law convergence

Dynamic regret compares a sequence of approximate optimization objectives with a comparator sequence. OMCI–MMD needs a stronger operational conclusion: at most checkpoints, the algorithm must both find a crossing within the memory budget and return a law close to the current endpoint law. We derive that conclusion directly and show exactly where a dynamic-regret estimate may enter.

Fix condition c, window length L, memory budget M, and the ordered checkpoint set $\mathcal{K}_T^c := \{t_1 < \cdots < t_N\}$ up to horizon T, where $N := |\mathcal{K}_T^c|$. At checkpoint $t \in \mathcal{K}_T^c$, let $\overline{P}_{t,L}^c$ and $\widehat{P}_{t,L}^c$ denote laws with mean embeddings $\bar{\mu}_{t,L}^c$ and $\widehat{\mu}_{t,L}^c$ from (72). Lemma 5 gives, on one simultaneous event,

$$\mathrm{MMD}_{k^c}(P_t^c, \overline{P}_{t,L}^c) \le S_{t,L}^c \le \overline{S}_{t,L}^c, \qquad \mathrm{MMD}_{k^c}(\overline{P}_{t,L}^c, \widehat{P}_{t,L}^c) \le E_{t,L}^c. \tag{131}$$

This reuses the staleness and sampling terms from (73)–(74); it does not introduce a second notation for the same errors. Without a valid envelope $\overline{S}_{t,L}^c$, the procedure bounds the empirical-window fit but cannot claim proximity to P_t^c.

Reuse the full-window oracle $F_{t,L,M}^{c,\star}$ from (76), the candidate-restricted oracle $F_{t,L,M}^{c,\mathcal{J},\star}$ from (77), and the computed value $\widehat{F}_{t,L,M}^c$ from (78). Define

$$\alpha_t^c(M) := F_{t,L,M}^{c,\mathcal{J},\star} - F_{t,L,M}^{c,\star} \ge 0, \qquad \beta_t^c(M) := \widehat{F}_{t,L,M}^c - F_{t,L,M}^{c,\mathcal{J},\star} \ge 0. \tag{132}$$

The first term measures candidate availability. The second collects linear-oracle, corrective, numerical, and warm-start error at budget M; equation (115) separates its numerical pieces when the implementation reports them individually. Let

$$\tau_{t,L}^c := \tfrac{1}{2}(\varepsilon_{t,L}^c)^2, \qquad \chi_t^c(M) := \mathbf{1}\Big\{F_{t,L,M}^{c,\star} + \alpha_t^c(M) + \beta_t^c(M) > \tau_{t,L}^c\Big\}. \tag{133}$$

Thus $\chi_t^c(M) = 0$ means that the computed squared-MMD path crosses by budget M. When it crosses, let $\widehat{Q}_t^c$ denote the returned law and let $\nu_t^c \ge 0$ bound any norm-scale feasibility or objective-reporting error beyond the reported radius. Define

$$D_t^c := \overline{S}_{t,L}^c + E_{t,L}^c + \varepsilon_{t,L}^c + \nu_t^c. \tag{134}$$

Candidate and optimization errors determine whether the path reaches the squared threshold $\tau_{t,L}^c$. Once it crosses, the norm-scale quantities in D_t^c bound the distance to the current law.

Proposition 7 (Bad-checkpoint frequency and current-law convergence)**.** *Assume that the two bounds in (131) hold simultaneously for every $t \in \mathcal{K}_T^c$ and that a reported crossing satisfies $\mathrm{MMD}_{k^c}(\widehat{P}_{t,L}^c, \widehat{Q}_t^c) \le \varepsilon_{t,L}^c + \nu_t^c$. Then, for every $q \ge 1$ and $\eta > 0$,*

$$\frac{1}{N}\sum_{t\in\mathcal{K}_T^c} \mathbf{1}\Big\{\widehat{m}_{t,L}^c > M \text{ or } \mathrm{MMD}_{k^c}(P_t^c, \widehat{Q}_t^c) > \eta\Big\} \le \frac{1}{N}\sum_{t\in\mathcal{K}_T^c} \chi_t^c(M) + \frac{1}{N\eta^q}\sum_{t\in\mathcal{K}_T^c} (D_t^c)^q. \tag{135}$$

When no crossing occurs, set $\widehat{Q}_t^c$ to an arbitrary probability law; the event $\widehat{m}_{t,L}^c > M$ is already true, so this convention does not change the union indicator. Therefore, if $N^{-1}\sum_t \chi_t^c(M) \to 0$ and $N^{-1}\sum_t (D_t^c)^q \to 0$, the fraction of checkpoints that exceed budget M or current-law error η converges to zero for every fixed $\eta > 0$.

Proof. **Proof map.** We use the triangle inequality and a deterministic indicator–moment bound. Combining crossing failure (133) with the current-law decomposition (137) gives the checkpoint-frequency guarantee (135). First suppose that $\chi_t^c(M) = 0$. Equations (132) and (133) give $\widehat{F}_{t,L,M}^c \le \tau_{t,L}^c$. Because $\widehat{m}_{t,L}^c$ is the first crossing of the same computed path, $\widehat{m}_{t,L}^c \le M$. This implication uses the entire support-indexed path; an objective value at a single adaptively chosen budget would not establish it.

Next suppose that the algorithm reports a crossing. Apply the triangle inequality through the window-average and empirical-window laws:

$$\mathrm{MMD}_{k^c}(P_t^c, \widehat{Q}_t^c) \le \mathrm{MMD}_{k^c}(P_t^c, \overline{P}_{t,L}^c) + \mathrm{MMD}_{k^c}(\overline{P}_{t,L}^c, \widehat{P}_{t,L}^c) + \mathrm{MMD}_{k^c}(\widehat{P}_{t,L}^c, \widehat{Q}_t^c) \tag{136}$$

$$\le \overline{S}_{t,L}^c + E_{t,L}^c + \varepsilon_{t,L}^c + \nu_t^c = D_t^c. \tag{137}$$

The first line isolates the statistical target mismatch from sparse approximation; the second invokes the staleness envelope, sampling event, and reported crossing bound.

For each checkpoint, the union indicator is at most $\chi_t^c(M) + \mathbf{1}\{D_t^c > \eta\}$. Since $D_t^c \ge 0$, $\mathbf{1}\{D_t^c > \eta\} \le (D_t^c/\eta)^q$. Sum this inequality over $\mathcal{K}_T^c$ and divide by N to obtain (135). The convergence statement follows by sending T, and hence N, to infinity. □

Relation to dynamic regret and prior work. Locally stationary and concept-drift analyses motivate the separation between $S^c_{t,L}$ and $E^c_{t,L}$ Dahlhaus (1997); Gama et al. (2014). DynaPOLK and dynamic-to-static RKHS reductions address predictive tracking under changing comparators Bedi et al. (2020); Jacobsen et al. (2025). DynaPOLK-type arguments can control an aggregate optimization or warm-start component inside $\beta^c_t(M)$ when their comparator and loss assumptions match the fully corrective subproblem. They do not control $S^c_{t,L}$, $\alpha^c_t(M)$, or the crossing margin in (133). Proposition 7 adds the missing deterministic conversion from those separate obligations to an interpretable current-law statement. In particular, a vanishing relative path-regret average cannot rescue a window whose stale-target bias stays bounded away from zero.

C Witness transfer and neural risk

This appendix first proves the generic RKHS witness inequality and then evaluates it at the law returned by the learnt crossing. It next verifies the identities and numerical terms required for Wasserstein, Sinkhorn, and endpoint-indexed Flow-Matching losses. The final results distinguish value transfer, gradient transfer, and parameter recovery.

C.1 Generic witness interface

For probability laws P and Q on a common endpoint domain and every function f integrable under both laws, write $Pf := \int f\,dP$ and $Qf := \int f\,dQ$. Let $\mathcal{F}$ be a bounded downstream witness class. For a declared RKHS norm budget $B > 0$, define the witness discrepancy and its norm-constrained RKHS approximation profile by

$$\gamma_{\mathcal{F}}(P,Q) := \sup_{f\in\mathcal{F}} |Pf - Qf|, \qquad \mathfrak{A}_{\mathcal{F},k}(B) := \sup_{f\in\mathcal{F}} \inf_{\substack{h\in\mathcal{H}\\ \|h\|_{\mathcal{H}}\le B}} \|f-h\|_\infty. \tag{138}$$

For later specializations, define the transfer envelope

$$\mathcal{T}_{\mathcal{F},k}(r;B) := Br + 2\mathfrak{A}_{\mathcal{F},k}(B). \tag{139}$$

The first quantity is the downstream integral probability metric. The second asks how accurately the radius-B RKHS ball approximates every downstream witness uniformly. This norm–approximation tradeoff, rather than universality alone, determines whether an MMD bound remains finite and informative.

Lemma 7 (Witness transfer)**.** *For every $B > 0$,*

$$\gamma_{\mathcal{F}}(P,Q) \le \mathcal{T}_{\mathcal{F},k}(\mathrm{MMD}_k(P,Q);B) = B\,\mathrm{MMD}_k(P,Q) + 2\mathfrak{A}_{\mathcal{F},k}(B). \tag{140}$$

Appendix C proves this lemma line by line and then verifies the additional identities needed for Wasserstein, Sinkhorn, and Flow-Matching witnesses. The lemma is the standard RKHS approximation step. OMCI applies it at a cardinality selected before any downstream transport loss is inspected.

C.2 Witness transfer at the learnt crossing

Fix a checkpoint for which the stationary crossing in (9) is finite, $\widehat{m}^c_t < +\infty$. Let

$$\widehat{P}^c_t := \widehat{\nu}^c_t, \qquad \widehat{Q}^c_t := \nu^c_{A_{t,\widehat{m}^c_t}, a_{t,\widehat{m}^c_t}}. \tag{141}$$

The first law is the empirical condition-c endpoint law, and the second is the weighted law stored at the OMCI crossing. By (9),

$$\mathrm{MMD}_k(\widehat{P}^c_t, \widehat{Q}^c_t) = \widehat{e}^c_{t,\widehat{m}^c_t} \le \varepsilon^c_t. \tag{142}$$

Corollary 4 (Witness bound at the learnt crossing)**.** *For any bounded witness class $\mathcal{F}^c$ and any $B_c > 0$,*

$$\gamma_{\mathcal{F}^c}(\widehat{P}^c_t, \widehat{Q}^c_t) \le B_c\varepsilon^c_t + 2\mathfrak{A}_{\mathcal{F}^c,k}(B_c). \tag{143}$$

On the event in Lemma 3, if P^c denotes the stationary population law with mean embedding μ^c, then

$$\gamma_{\mathcal{F}^c}(P^c, \widehat{Q}^c_t) \le B_c\big(E^c_t(\delta) + \varepsilon^c_t\big) + 2\mathfrak{A}_{\mathcal{F}^c,k}(B_c). \tag{144}$$

Proof. **Proof map.** We combine the generic witness-transfer inequality (140) with the MMD triangle inequality. Evaluating both at the data-selected law through (142) gives (144). Apply Lemma 7 to $(\widehat{P}_t^c, \widehat{Q}_t^c)$ and use (142) to obtain (143). For the population statement, the triangle inequality for MMD and Lemma 3 give

$$\begin{aligned}\mathrm{MMD}_k(P^c, \widehat{Q}_t^c) &\le \mathrm{MMD}_k(P^c, \widehat{P}_t^c) + \mathrm{MMD}_k(\widehat{P}_t^c, \widehat{Q}_t^c) \\ &\le E_t^c(\delta) + \varepsilon_t^c. \end{aligned} \tag{145}$$

A second application of Lemma 7 proves (144). □

The corollary links model-order selection to downstream error bounds. It does not assert that a generic kernel sees every transport or flow quantity. Appendix C identifies the additional Kantorovich, Sinkhorn first-variation, and endpoint-indexed flow-loss witnesses. This first-variation route is distinct from CO2's fixed-target, second-order divergence-adapted MMD construction (Kokot and Luedtke, 2025).

MMD bounds downstream quantities only through functions that its RKHS represents. Corollary 4 first applies that comparison at the actual OMCI crossing. This appendix then spells out every additional obligation behind a Wasserstein, Sinkhorn, or Flow-Matching claim: a valid witness identity, a fixed gauge, RKHS approximation, numerical and quadrature errors, uniformity over the claimed class, and any witness staleness. The order matters. The crossing proves an endpoint MMD statement; transport duality or learning theory must supply the remaining bridge.

Throughout, P and Q are probability measures on a compact endpoint set $\mathcal{Y}_0$, k is a bounded kernel with feature map ϕ and RKHS $\mathcal{H}$, and $\mathcal{F}$ is the bounded downstream witness class defined in (138). Define the kernel means $\mu_P := \int \phi(y)\, dP(y)$ and $\mu_Q := \int \phi(y)\, dQ(y)$, and write

$$\varepsilon_{\mathrm{MMD}} := \mathrm{MMD}_k(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}. \tag{146}$$

A high-probability truncation event can replace compactness if the final bound includes its tail term. MMD is the integral probability metric over the RKHS unit ball (Gretton et al., 2012; Sriperumbudur et al., 2010). Characteristic kernels identify probability laws, but that injectivity alone does not give a useful finite norm bound for a particular downstream witness class (Sriperumbudur et al., 2011).

Proof of Lemma 7. **Proof map.** We insert an RKHS approximant and apply MMD duality in (148)–(151). The lemma is generic; OMCI enters when Corollary 4 substitutes the law selected at the learnt crossing. Fix $B > 0$. If $\mathfrak{A}_{\mathcal{F},k}(B) = +\infty$, the right side of (140) is $+\infty$, so the extended-real inequality holds; hence assume the profile is finite. Fix $f \in \mathcal{F}$ and an arbitrary $\eta > 0$. By the definition of $\mathfrak{A}_{\mathcal{F},k}(B)$ in (138), choose $h_{f,\eta} \in \mathcal{H}$ such that

$$\|h_{f,\eta}\|_{\mathcal{H}} \le B, \qquad \|f - h_{f,\eta}\|_\infty \le \mathfrak{A}_{\mathcal{F},k}(B) + \eta. \tag{147}$$

Insert this approximant and apply the triangle inequality:

$$|(P-Q)f| \le |(P-Q)h_{f,\eta}| + |P(f - h_{f,\eta}) - Q(f - h_{f,\eta})| \tag{148}$$

$$\le |(P-Q)h_{f,\eta}| + 2\|f - h_{f,\eta}\|_\infty. \tag{149}$$

The reproducing property and Cauchy–Schwarz give

$$|(P-Q)h_{f,\eta}| = |\langle h_{f,\eta}, \mu_P - \mu_Q\rangle_{\mathcal{H}}| \tag{150}$$

$$\le \|h_{f,\eta}\|_{\mathcal{H}} \|\mu_P - \mu_Q\|_{\mathcal{H}} \le B\, \mathrm{MMD}_k(P, Q). \tag{151}$$

Substitute (147) and (151) into (149) to obtain

$$|(P-Q)f| \le B\, \mathrm{MMD}_k(P, Q) + 2\mathfrak{A}_{\mathcal{F},k}(B) + 2\eta. \tag{152}$$

The bound holds for every $f \in \mathcal{F}$. Take the supremum over f and then let $\eta \downarrow 0$ to prove (140).

Proposition 8 (Frozen witness bank)**.** *Let k_0 be a bounded base kernel with feature map ϕ_0, let $\psi_1, \dots, \psi_R$ be bounded witnesses fixed before calibration, and choose $\lambda_r > 0$. Define the augmented kernel*

$$k_{\mathrm{aug}}(x, y) := k_0(x, y) + \sum_{r=1}^{R} \lambda_r \psi_r(x)\psi_r(y), \tag{153}$$

and let $\mathcal{H}_{\mathrm{aug}}$ denote its RKHS. Then, for every r,

$$|(P-Q)\psi_r| \le \lambda_r^{-1/2}\, \mathrm{MMD}_{k_{\mathrm{aug}}}(P, Q). \tag{154}$$

Proof of Proposition 8. **Proof map.** We use a direct-sum RKHS construction. Freezing the witness bank before calibration makes (155) valid for selection and yields the exact bankwise bound (154) through (156). Let $\mathcal{H}_{\mathrm{aug}}$ denote the RKHS of k_{aug}. The augmented kernel has the direct-sum feature map

$$\phi_{\mathrm{aug}}(y) = \big(\phi_0(y), \sqrt{\lambda_1}\psi_1(y), \dots, \sqrt{\lambda_R}\psi_R(y)\big). \tag{155}$$

The coordinate vector with entry $\lambda_r^{-1/2}$ in coordinate r and zero elsewhere represents ψ_r in the augmented RKHS, so $\|\psi_r\|_{\mathcal{H}_{\mathrm{aug}}} \le \lambda_r^{-1/2}$. Hence

$$\begin{aligned} |(P-Q)\psi_r| &\le \|\psi_r\|_{\mathcal{H}_{\mathrm{aug}}} \,\mathrm{MMD}_{k_{\mathrm{aug}}}(P,Q) \\ &\le \lambda_r^{-1/2}\,\mathrm{MMD}_{k_{\mathrm{aug}}}(P,Q), \end{aligned} \tag{156}$$

which proves (154). Freezing the bank before radius calibration is essential. If an algorithm chooses witnesses after inspecting $P-Q$, it changes the evaluated function class in a data-dependent way. A valid adaptive construction must use an independent selection split or prove a simultaneous bound over every witness that the selection rule could return; the coordinate argument above does not pay for that adaptivity.

C.3 Transport witnesses: Wasserstein and Sinkhorn

We start each transport claim from a witness identity. Kantorovich–Rubinstein duality supplies the identity for W_1; differentiability of entropic transport along a declared path of measures supplies it for Sinkhorn divergence (Villani, 2009); differentiability of entropic transport along a declared path supplies it for Sinkhorn divergence (Peyré and Cuturi, 2019; Feydy et al., 2019). Finite-sample and numerical errors require additional control (Genevay et al., 2019). The witness identity and the numerical error bound are different steps. A Sinkhorn objective tolerance does not, by itself, bound the dual potential in supremum norm; such a claim needs a setting-specific stability theorem or a residual-to-potential estimate (Delalande, 2022).

Proposition 9 (Integrated transport-witness verifier)**.** *Let ρ be a probability measure on an index set $\mathcal{S}$. Suppose $(s,y) \mapsto \psi_s(y)$ is measurable and integrable under $\rho \otimes P$ and $\rho \otimes Q$, and suppose a nonnegative downstream quantity admits the representation*

$$D(P,Q) = \int_{\mathcal{S}} (P-Q)\psi_s \, d\rho(s). \tag{157}$$

Let $\{(s_r, w_r)\}_{r=1}^R$ be a quadrature rule with $w_r \ge 0$, $\sum_r w_r = 1$, and deterministic identity/quadrature error

$$q_R := \left| D(P,Q) - \sum_{r=1}^R w_r (P-Q)\psi_{s_r} \right|. \tag{158}$$

For every node s_r, suppose there are nonnegative constants α_r, β_r, and B_r such that

$$\|\psi_{s_r} - \widehat{\psi}_{s_r}\|_\infty \le \beta_r, \tag{159}$$

$$\|\widehat{\psi}_{s_r} - h_r\|_\infty \le \alpha_r, \qquad \|h_r\|_{\mathcal{H}} \le B_r, \quad h_r \in \mathcal{H}. \tag{160}$$

Define the quadrature-weighted averages

$$\overline{B}_R := \sum_{r=1}^R w_r B_r, \qquad \overline{\alpha}_R := \sum_{r=1}^R w_r \alpha_r, \qquad \overline{\beta}_R := \sum_{r=1}^R w_r \beta_r. \tag{161}$$

Then

$$D(P,Q) \le \overline{B}_R\, \varepsilon_{\mathrm{MMD}} + 2(\overline{\alpha}_R + \overline{\beta}_R) + q_R. \tag{162}$$

Proof. **Proof map.** We apply triangle inequalities and RKHS duality. Equations (165)–(170) keep witness approximation, numerical error, and quadrature error separate before (162) combines them. Start from the definition of q_R in (158). Since $D(P,Q) \ge 0$,

$$D(P,Q) \le \left| \sum_{r=1}^R w_r (P-Q)\psi_{s_r} \right| + q_R \tag{163}$$

$$\le \sum_{r=1}^R w_r \, |(P-Q)\psi_{s_r}| + q_R, \tag{164}$$

where the second line is the triangle inequality and uses $w_r \geq 0$. For one quadrature node, insert the computed witness:

$$|(P-Q)\psi_{s_r}| \leq \left|(P-Q)\widehat{\psi}_{s_r}\right| + \left|P(\psi_{s_r} - \widehat{\psi}_{s_r}) - Q(\psi_{s_r} - \widehat{\psi}_{s_r})\right| \tag{165}$$

$$\leq \left|(P-Q)\widehat{\psi}_{s_r}\right| + 2\beta_r. \tag{166}$$

The last inequality follows because P and Q are probability measures and $\|\psi_{s_r} - \widehat{\psi}_{s_r}\|_\infty \leq \beta_r$. Insert the RKHS approximant:

$$\left|(P-Q)\widehat{\psi}_{s_r}\right| \leq |(P-Q)h_r| + \left|P(\widehat{\psi}_{s_r} - h_r) - Q(\widehat{\psi}_{s_r} - h_r)\right| \tag{167}$$

$$\leq |(P-Q)h_r| + 2\alpha_r. \tag{168}$$

By the reproducing-property characterization of MMD,

$$|(P-Q)h_r| = |\langle h_r, \mu_P - \mu_Q\rangle_{\mathcal{H}}| \leq \|h_r\|_{\mathcal{H}}\, \varepsilon_{\mathrm{MMD}} \leq B_r \varepsilon_{\mathrm{MMD}}. \tag{169}$$

Combining (166), (168), and (169) gives

$$|(P-Q)\psi_{s_r}| \leq B_r \varepsilon_{\mathrm{MMD}} + 2(\alpha_r + \beta_r). \tag{170}$$

Substitute (170) into (164) and collect the weighted averages. This yields (162). □

Corollary 5 (Finite-configuration W_1 verifier)**.** *Let $(\mathcal{Y}_0, d_{\mathcal{Y}})$ be a compact metric space, let W_1 use the ground metric $d_{\mathcal{Y}}$, and let $f^\star$ be an optimal Kantorovich–Rubinstein critic normalized by a fixed gauge. If a computed critic $\widehat{f}$, an RKHS function h, and nonnegative constants α, β, and B satisfy*

$$\|f^\star - \widehat{f}\|_\infty \leq \beta, \qquad \|\widehat{f} - h\|_\infty \leq \alpha, \qquad \|h\|_{\mathcal{H}} \leq B, \tag{171}$$

then

$$W_1(P,Q) \leq B\varepsilon_{\mathrm{MMD}} + 2(\alpha + \beta). \tag{172}$$

Proof. **Proof map.** Kantorovich–Rubinstein duality supplies the exact identity. Substituting it into the general verifier gives (172); no additional optimal-transport inequality is required. Kantorovich–Rubinstein duality gives $W_1(P,Q) = |(P-Q)f^\star|$ after choosing the sign of $f^\star$ (Peyré and Cuturi, 2019). Apply Proposition 9 with a singleton index set, unit quadrature weight, and $q_R = 0$. □

Let $c_{\mathrm{OT}} : \mathcal{Y}_0 \times \mathcal{Y}_0 \to [0,\infty)$ be the declared ground cost, let $\Pi(P,Q)$ denote the couplings of P and Q, let $P \otimes Q$ denote their product measure, and let KL denote Kullback–Leibler divergence. Fix the entropy convention

$$\mathrm{OT}_\varepsilon(P,Q) := \inf_{\pi \in \Pi(P,Q)} \left\{ \int c_{\mathrm{OT}}(x,y)\, d\pi(x,y) + \varepsilon\, \mathrm{KL}(\pi \,\|\, P \otimes Q) \right\}, \tag{173}$$

$$S_\varepsilon(P,Q) := \mathrm{OT}_\varepsilon(P,Q) - \tfrac{1}{2}\,\mathrm{OT}_\varepsilon(P,P) - \tfrac{1}{2}\,\mathrm{OT}_\varepsilon(Q,Q), \tag{174}$$

The entropy-regularized objective follows Cuturi (2013), and the debiased divergence follows Feydy et al. (2019). Changing the reference-measure convention changes the corresponding first variation, but leaves the witness argument unchanged after we fix that convention (Peyré and Cuturi, 2019).

Corollary 6 (Finite-configuration Sinkhorn verifier)**.** *Fix $\varepsilon > 0$ and assume the declared ground cost lies in a setting where the debiased Sinkhorn divergence in (174) is nonnegative. For $s \in [0,1]$, define $P_s = (1-s)Q + sP$. Assume $F(\Lambda) := S_\varepsilon(\Lambda, Q)$ is differentiable along this path, $F(Q) = 0$, and there exist normalized first-variation witnesses $\psi_{\varepsilon,s}$ satisfying*

$$\frac{d}{ds} F(P_s) = (P-Q)\psi_{\varepsilon,s}. \tag{175}$$

The fundamental theorem of calculus then gives the exact mixture-path identity

$$S_\varepsilon(P,Q) = \int_0^1 (P-Q)\psi_{\varepsilon,s}\, ds. \tag{176}$$

For quadrature nodes $s_1, \dots, s_R$, weights $w_1, \dots, w_R$, and computed/RKHS witnesses satisfying Proposition 9, write $B_{\varepsilon,r}$, $\alpha_{\varepsilon,r}$, and $\beta_{\varepsilon,r}$ for the nodewise norm, RKHS-approximation, and solver-error bounds. Define

$$\overline{B}_{\varepsilon,R} := \sum_{r=1}^{R} w_r B_{\varepsilon,r}, \qquad \overline{\alpha}_{\varepsilon,R} := \sum_{r=1}^{R} w_r \alpha_{\varepsilon,r}, \qquad \overline{\beta}_{\varepsilon,R} := \sum_{r=1}^{R} w_r \beta_{\varepsilon,r}, \tag{177}$$

and let $q_{\varepsilon,R}$ be the corresponding quadrature identity error in (158). Then

$$S_\varepsilon(P,Q) \le \overline{B}_{\varepsilon,R}\, \varepsilon_{\mathrm{MMD}} + 2(\overline{\alpha}_{\varepsilon,R} + \overline{\beta}_{\varepsilon,R}) + q_{\varepsilon,R}. \tag{178}$$

Proof. **Proof map.** We apply the fundamental theorem of calculus along a mixture path and the entropic first variation in (179)–(181). Equation (178) then retains the numerical and quadrature terms instead of hiding them in the witness bound. By the fundamental theorem of calculus and $F(Q) = S_\varepsilon(Q,Q) = 0$,

$$S_\varepsilon(P,Q) = F(P) - F(Q) \tag{179}$$

$$= \int_0^1 \frac{d}{ds} F(P_s)\, ds \tag{180}$$

$$= \int_0^1 (P-Q)\psi_{\varepsilon,s}\, ds. \tag{181}$$

The differentiability and first-variation interpretation are the structural inputs supplied by entropic OT theory; see Feydy et al. (2019); Peyré and Cuturi (2019). Equation (181) is exactly the witness representation in Proposition 9, with $\mathcal{S} = [0,1]$ and Lebesgue measure. Applying that proposition gives (178). □

Relation to fixed-target Sinkhorn coresets. The argument above transfers a predeclared MMD bound through Sinkhorn first variations. It does not optimize Sinkhorn geometry. WCSL instead differentiates the Sinkhorn loss with respect to a fixed-size coreset and analyzes the stability of that direct optimization (Yin et al., 2025). CO2 takes a functional second-order expansion of a smooth divergence, constructs a divergence-adapted MMD, and specializes that fixed-target selection rule to Sinkhorn divergence (Kokot and Luedtke, 2025). EMS enlarges the fixed-size search class to nonuniform weighted, free-support centroids and derives closed-form expectation–maximization updates, consistency, and perturbation stability (Yin et al., 2026). These methods can produce a sharper coreset once a target and cardinality are fixed. Equations (176)–(178) solve the complementary problem: they state what one must verify before passing an online, condition-wise MMD crossing to Sinkhorn loss without using that loss to choose the cardinality.

Corollary 7 (Sinkhorn bound at the OMCI-selected cardinality)**.** *Let P be the endpoint target and let $\bar{Q}_{\mathrm{OMCI}}$ be the weighted law frozen at the OMCI-selected cardinality $\bar{m}$. Suppose the endpoint analysis supplies*

$$\mathrm{MMD}_k(P, \bar{Q}_{\mathrm{OMCI}}) \le \bar{\varepsilon}_{\mathrm{MMD}}, \tag{182}$$

where $\bar{\varepsilon}_{\mathrm{MMD}}$ includes the empirical crossing and, when the target is a population law, its sampling and staleness allowances. Under Corollary 6, define

$$U_{\mathrm{OMCI}} := \overline{B}_{\varepsilon,R}\, \bar{\varepsilon}_{\mathrm{MMD}} + 2(\overline{\alpha}_{\varepsilon,R} + \overline{\beta}_{\varepsilon,R}) + q_{\varepsilon,R}. \tag{183}$$

Then

$$S_\varepsilon(P, \bar{Q}_{\mathrm{OMCI}}) \le U_{\mathrm{OMCI}}. \tag{184}$$

Proof. **Proof map.** We substitute the OMCI law into the Sinkhorn verifier. Equation (182) evaluates the witness term at the frozen cardinality, and (184) collects the transport, sampling, and numerical errors. Apply (178) to $(P, \bar{Q}_{\mathrm{OMCI}})$. Replace the MMD factor by (182); the remaining terms are exactly those collected in (183). □

For a stationary empirical target, Algorithm 1 supplies the crossing term in (182). For a stationary population target, the triangle inequality, (39), and (142) give the admissible choice $\bar{\varepsilon}_{\mathrm{MMD}} = E_t^c(\delta) + \varepsilon_t^c$. In the windowed setting, Lemma 5 adds the sampling and declared staleness allowances. Thus the bound keeps endpoint estimation and transport-witness approximation as visible, separate reported terms.

Proposition 10 (Fixed-cardinality transport refinement)**.** *Assume the Sinkhorn bound* $S_\varepsilon(P, \bar{Q}_{\mathrm{OMCI}}) \le U_{\mathrm{OMCI}}$ *from Corollary 7. Let* $\mathcal{Q}_{\bar{m}}$ *be a transport-specific class of probability laws with at most* $\bar{m}$ *atoms, or any other declared class that contains the weighted OMCI law* $\bar{Q}_{\mathrm{OMCI}}$*. Let* $\eta_{\mathrm{opt}} \ge 0$*, and suppose* $\widetilde{Q}_{\bar{m}} \in \mathcal{Q}_{\bar{m}}$ *satisfies the declared fixed-cardinality optimization guarantee*

$$S_\varepsilon(P, \widetilde{Q}_{\bar{m}}) \le \inf_{Q \in \mathcal{Q}_{\bar{m}}} S_\varepsilon(P, Q) + \eta_{\mathrm{opt}}. \tag{185}$$

Then

$$S_\varepsilon(P, \widetilde{Q}_{\bar{m}}) \le \inf_{Q \in \mathcal{Q}_{\bar{m}}} S_\varepsilon(P, Q) + \eta_{\mathrm{opt}} \le U_{\mathrm{OMCI}} + \eta_{\mathrm{opt}}. \tag{186}$$

Proof of Proposition 10. Proof map. We use feasibility and approximate fixed-cardinality optimization. The containment condition (187) enforces select–freeze–refine order, and (186) prevents downstream loss from choosing the support size. Because $\bar{Q}_{\mathrm{OMCI}} \in \mathcal{Q}_{\bar{m}}$, feasibility gives

$$\inf_{Q \in \mathcal{Q}_{\bar{m}}} S_\varepsilon(P, Q) \le S_\varepsilon(P, \bar{Q}_{\mathrm{OMCI}}). \tag{187}$$

Combining (185), (187), and (184) yields

$$\begin{aligned} S_\varepsilon(P, \widetilde{Q}_{\bar{m}}) &\le \inf_{Q \in \mathcal{Q}_{\bar{m}}} S_\varepsilon(P, Q) + \eta_{\mathrm{opt}} \\ &\le S_\varepsilon(P, \bar{Q}_{\mathrm{OMCI}}) + \eta_{\mathrm{opt}} \\ &\le U_{\mathrm{OMCI}} + \eta_{\mathrm{opt}}, \end{aligned} \tag{188}$$

which proves (186). The first line is an explicit fixed-cardinality optimization guarantee; a local optimizer does not satisfy it merely because its objective decreased. This optimization occurs after OMCI has frozen the cardinality and therefore does not participate in size selection. Equal-weight, exactly-$\bar{m}$-positive-atom, or otherwise restricted refiners need a separate containment or projection argument when their feasible class excludes the weighted at-most-$\bar{m}$ OMCI law. The proposition preserves the stated upper bound; it does not prove that the refined law is state of the art.

Gauge, solver, and quadrature terms. Adding a constant to a transport potential leaves $(P - Q)\psi$ unchanged because P and Q have equal mass, but it can change the RKHS norm. The verifier must therefore fix one gauge before it reports B_r or an approximation error. A Sinkhorn primal–dual gap bounds objective suboptimality; it does not automatically bound the supremum-norm potential error β_r. A uniform claim must invoke a setting-specific potential-stability theorem, derive a residual-to-potential estimate, or restrict the statement to the evaluated grid. Likewise, evaluating a few path nodes does not prove the identity integral: the quadrature remainder $q_{\varepsilon,R}$ needs its own deterministic or high-probability bound. Entropic regularization improves smoothness and stability relative to raw optimal transport, but these constants can deteriorate as $\varepsilon \downarrow 0$ (Genevay et al., 2019; Delalande, 2022).

Finite witness banks. When only finitely many critics or potentials matter, Proposition 8 gives the cleanest direct bound: fix and normalize them before calibration, insert them into the augmented kernel, and let augmented MMD bound their expectation gaps directly. If the bank comes from fitted transport potentials, an independent split or a simultaneous selection bound must cover the fitting step. Witness-estimation, numerical, and staleness errors remain separate even after augmentation.

C.4 Flow-matching risk, optimization, and nonstationary witnesses

Flow Matching turns endpoint samples into supervised regression targets (Lipman et al., 2023). Fix a condition c, a parameter class Θ, and a path or coupling rule Γ. Assume the expectations below exist. A population-uniform supremum-norm result further requires bounded losses on the declared endpoint domain; on unbounded domains one must truncate to a high-probability set and add the corresponding tail contribution. We model $\Gamma(\mathrm{d}s, \mathrm{d}x_0, \mathrm{d}\omega_{\mathrm{path}} \mid y, c)$ as a Markov kernel that, conditional on endpoint y and condition c, supplies an interpolation time $s \in [0, 1]$, a source state x_0, and any auxiliary path randomness ω_{path}. This formulation includes independent source–endpoint sampling as well as endpoint-dependent source couplings such as optimal-transport conditional Flow Matching. For $\theta \in \Theta$, let v_θ be the learned velocity field, and let $X_s^\Gamma(y, x_0, \omega_{\mathrm{path}})$ and

$u_s^{\star,\Gamma}(y, x_0, \omega_{\text{path}})$ denote the path state and supervised target velocity generated by the declared rule. Define the endpoint-indexed loss

$$\ell^c_{\theta,\Gamma}(y) := \mathbb{E}_{(s,x_0,\omega_{\text{path}})\sim\Gamma(\cdot|y,c)}\left[\|v_\theta(X_s^\Gamma(y, x_0, \omega_{\text{path}}), s, c) - u_s^{\star,\Gamma}(y, x_0, \omega_{\text{path}})\|^2\right], \tag{189}$$

and, for an endpoint law P, define the population risk

$$\mathcal{L}^c_{P,\Gamma}(\theta) = P\ell^c_{\theta,\Gamma} = \int \ell^c_{\theta,\Gamma}(y)\, P(\mathrm{d}y). \tag{190}$$

When comparing endpoint laws P and Q, we hold Γ and v_θ fixed. Equation (190) then changes only the outer endpoint integration; any endpoint dependence of the source coupling remains inside the same frozen Markov kernel. This isolates the layer that fixed-law velocity-estimation analyses omit. For example, Sahoo et al. (2026) obtain order-optimal sample complexity for squared-loss rectified-flow training under a supplied endpoint law and a linear path. Their rate does not bound the distributional error created when memory replaces that law before training. The verifier below addresses exactly that replacement term; it does not reprove neural approximation, optimization, generalization, or ordinary-differential-equation stability.

The following proposition bounds the full endpoint-indexed loss class. Its assumptions are obligations, not consequences of MMD: the constants must hold uniformly over $\theta \in \Theta$. Checking finitely many trained checkpoints covers only those checkpoints unless a covering, Lipschitz, complexity, or simultaneous concentration argument extends the check to the whole class. If the same endpoint sample also fits $\widehat{\ell}_\theta$, chooses the kernel, or calibrates the radius, these deterministic bounds must hold jointly over that data-dependent choice; otherwise the verifier must use sample splitting.

Proposition 11 (Uniform flow-risk verifier)**.** *Fix a path/coupling rule Γ. Suppose that every $\theta \in \Theta$ admits a computed loss witness $\widehat{\ell}_\theta$ and an RKHS function h_θ such that, for nonnegative constants α_c, β_c, and B_c, the following bounds hold uniformly over θ:*

$$\|\ell^c_{\theta,\Gamma} - \widehat{\ell}_\theta\|_\infty \le \beta_c, \qquad \|\widehat{\ell}_\theta - h_\theta\|_\infty \le \alpha_c, \qquad \|h_\theta\|_{\mathcal{H}} \le B_c. \tag{191}$$

Then

$$\delta^c_{\mathrm{FM}}(P, Q; \Gamma) := \sup_{\theta\in\Theta} |\mathcal{L}^c_{P,\Gamma}(\theta) - \mathcal{L}^c_{Q,\Gamma}(\theta)| \le B_c\varepsilon_{\mathrm{MMD}} + 2(\alpha_c + \beta_c). \tag{192}$$

Proof. **Proof map.** We use scalar witness approximation and RKHS duality. After freezing the coupling, (193) identifies the endpoint-indexed loss as the witness, and the uniform step yields (192). Fix an arbitrary $\theta \in \Theta$. By definition of the risks,

$$|\mathcal{L}^c_{P,\Gamma}(\theta) - \mathcal{L}^c_{Q,\Gamma}(\theta)| = |(P - Q)\ell^c_{\theta,\Gamma}|. \tag{193}$$

Insert the computed loss witness:

$$|(P - Q)\ell^c_{\theta,\Gamma}| \le |(P - Q)\widehat{\ell}_\theta| + |P(\ell^c_{\theta,\Gamma} - \widehat{\ell}_\theta) - Q(\ell^c_{\theta,\Gamma} - \widehat{\ell}_\theta)| \tag{194}$$

$$\le |(P - Q)\widehat{\ell}_\theta| + 2\beta_c. \tag{195}$$

Insert the RKHS approximant:

$$|(P - Q)\widehat{\ell}_\theta| \le |(P - Q)h_\theta| + |P(\widehat{\ell}_\theta - h_\theta) - Q(\widehat{\ell}_\theta - h_\theta)| \tag{196}$$

$$\le |(P - Q)h_\theta| + 2\alpha_c. \tag{197}$$

The RKHS dual bound gives

$$|(P - Q)h_\theta| \le \|h_\theta\|_{\mathcal{H}}\varepsilon_{\mathrm{MMD}} \le B_c\varepsilon_{\mathrm{MMD}}. \tag{198}$$

Combining (195), (197), and (198) yields the displayed upper bound for each fixed θ. The constants in (191) do not depend on θ, so taking the supremum over Θ proves (192). Without that uniformity, the proof establishes only a pointwise checkpoint bound. □

Proposition 12 (Approximate empirical-risk-minimization transfer)**.** *Let $\epsilon_{\mathrm{opt}} \ge 0$, and let $\widehat{\theta}_Q \in \Theta$ satisfy*

$$\mathcal{L}^c_{Q,\Gamma}(\widehat{\theta}_Q) \le \inf_{\theta\in\Theta} \mathcal{L}^c_{Q,\Gamma}(\theta) + \epsilon_{\mathrm{opt}}. \tag{199}$$

If $\delta^c_{\mathrm{FM}}(P, Q; \Gamma)$ is finite, then

$$\mathcal{L}^c_{P,\Gamma}(\widehat{\theta}_Q) - \inf_{\theta\in\Theta} \mathcal{L}^c_{P,\Gamma}(\theta) \le 2\delta^c_{\mathrm{FM}}(P, Q; \Gamma) + \epsilon_{\mathrm{opt}}. \tag{200}$$

Consequently, under Proposition 11, the right-hand side is at most

$$2B_c\varepsilon_{\mathrm{MMD}} + 4(\alpha_c + \beta_c) + \epsilon_{\mathrm{opt}}. \tag{201}$$

Proof. **Proof map.** We use the standard two-sided uniform-risk perturbation argument for approximate empirical risk minimization. Endpoint-law replacement gives (200), and the witness bound converts it to the MMD form (201). Let $\delta = \delta^c_{\mathrm{FM}}(P, Q; \Gamma)$. By definition of δ,

$$\mathcal{L}^c_{P,\Gamma}(\widehat{\theta}_Q) \le \mathcal{L}^c_{Q,\Gamma}(\widehat{\theta}_Q) + \delta. \tag{202}$$

Apply the approximate optimality condition (199):

$$\mathcal{L}^c_{P,\Gamma}(\widehat{\theta}_Q) \le \inf_{\theta\in\Theta} \mathcal{L}^c_{Q,\Gamma}(\theta) + \epsilon_{\mathrm{opt}} + \delta. \tag{203}$$

For every θ, again by the definition of δ,

$$\mathcal{L}^c_{Q,\Gamma}(\theta) \le \mathcal{L}^c_{P,\Gamma}(\theta) + \delta. \tag{204}$$

Taking the infimum over θ in (204) gives

$$\inf_{\theta\in\Theta} \mathcal{L}^c_{Q,\Gamma}(\theta) \le \inf_{\theta\in\Theta} \mathcal{L}^c_{P,\Gamma}(\theta) + \delta. \tag{205}$$

Substitute (205) into (203) and subtract $\inf_\theta \mathcal{L}^c_{P,\Gamma}(\theta)$. This proves (200). The bound (201) follows by substituting Proposition 11. □

Corollary 8 (Frozen versus recomputed coupling)**.** *Let Γ_P and Γ_Q be coupling/path rules constructed from P and Q. For any fixed θ,*

$$\begin{aligned} &|\mathcal{L}^c_{Q,\Gamma_Q}(\theta) - \mathcal{L}^c_{P,\Gamma_P}(\theta)| \\ &\quad\le \underbrace{|\mathcal{L}^c_{Q,\Gamma_Q}(\theta) - \mathcal{L}^c_{Q,\Gamma_P}(\theta)|}_{\text{coupling/path instability}} + \underbrace{|\mathcal{L}^c_{Q,\Gamma_P}(\theta) - \mathcal{L}^c_{P,\Gamma_P}(\theta)|}_{\text{endpoint-memory term}}. \end{aligned} \tag{206}$$

Proposition 11 covers the second term when the fixed-Γ_P loss witnesses are RKHS-controlled. A frozen-coupling verification sets the first term to zero. Recomputing an OT or Sinkhorn coupling requires a separate plan-stability bound.

Proof. **Proof map.** We add and subtract one intermediate risk and apply the triangle inequality. Equation (206) separates coupling or path error from endpoint-memory error, so a coupling method cannot substitute for a memory guarantee. Add and subtract $\mathcal{L}^c_{Q,\Gamma_P}(\theta)$ and apply the triangle inequality; this gives the two displayed terms exactly. OT-CFM deliberately changes the path through an estimated coupling (Tong et al., 2024), and condition-aware coupling corrections such as C^2OT target the first term (Cheng and Schwing, 2025). Neither construction, by itself, bounds the second term created when a memory policy changes the endpoint law. □

Proposition 13 (Dynamic witness verifier)**.** *For $t = 1, \dots, T$, let P_t be the current endpoint law, Q_t its active approximation, and*

$$\varepsilon_{\mathrm{MMD},t} := \mathrm{MMD}_k(P_t, Q_t). \tag{207}$$

Let $D_t(P_t, Q_t)$ be a nonnegative downstream discrepancy or risk gap, and let $q_t \ge 0$ be a declared identity or quadrature remainder. Suppose each time t admits a current true witness f_t, a numerically computed witness $\widehat{f}_t$, a stored witness $\overline{f}_t := \widehat{f}_{\tau(t)}$ last refreshed at time $\tau(t) \le t$, and an RKHS function h_t. Let the error bounds α_t, β_t, s_t and norm budget $B_{\mathrm{wit},t}$ be nonnegative and satisfy

$$D_t(P_t, Q_t) \le |(P_t - Q_t)f_t| + q_t, \tag{208}$$

$$\|f_t - \widehat{f}_t\|_\infty \le \beta_t, \tag{209}$$

$$\|\widehat{f}_t - \overline{f}_t\|_\infty \le s_t, \tag{210}$$

$$\|\overline{f}_t - h_t\|_\infty \le \alpha_t, \qquad \|h_t\|_{\mathcal{H}} \le B_{\mathrm{wit},t}. \tag{211}$$

Then, pointwise in t,

$$D_t(P_t, Q_t) \le B_{\mathrm{wit},t}\varepsilon_{\mathrm{MMD},t} + 2(\alpha_t + \beta_t + s_t) + q_t, \tag{212}$$

and consequently

$$\frac{1}{T}\sum_{t=1}^{T} D_t(P_t, Q_t) \le \left(\frac{1}{T}\sum_{t=1}^{T} B_{\mathrm{wit},t}^2\right)^{1/2} \left(\frac{1}{T}\sum_{t=1}^{T} \varepsilon_{\mathrm{MMD},t}^2\right)^{1/2} + \frac{2}{T}\sum_{t=1}^{T}(\alpha_t + \beta_t + s_t) + \frac{1}{T}\sum_{t=1}^{T} q_t. \tag{213}$$

Proof. **Proof map.** We repeat the witness-approximation argument and apply Cauchy–Schwarz. The stored-witness term (210) remains explicit in (213) rather than being folded into MMD error. Fix a time t. Start from the witness representation (208):

$$D_t(P_t, Q_t) \le |(P_t - Q_t)f_t| + q_t. \tag{214}$$

Insert the numerically computed witness:

$$|(P_t - Q_t)f_t| \le |(P_t - Q_t)\widehat{f}_t| + |P_t(f_t - \widehat{f}_t) - Q_t(f_t - \widehat{f}_t)| \tag{215}$$

$$\le |(P_t - Q_t)\widehat{f}_t| + 2\beta_t. \tag{216}$$

Insert the stored witness:

$$|(P_t - Q_t)\widehat{f}_t| \le |(P_t - Q_t)\overline{f}_t| + |P_t(\widehat{f}_t - \overline{f}_t) - Q_t(\widehat{f}_t - \overline{f}_t)| \tag{217}$$

$$\le |(P_t - Q_t)\overline{f}_t| + 2s_t. \tag{218}$$

Insert the RKHS approximant:

$$|(P_t - Q_t)\overline{f}_t| \le |(P_t - Q_t)h_t| + |P_t(\overline{f}_t - h_t) - Q_t(\overline{f}_t - h_t)| \tag{219}$$

$$\le |(P_t - Q_t)h_t| + 2\alpha_t \tag{220}$$

$$\le B_{\mathrm{wit},t}\varepsilon_{\mathrm{MMD},t} + 2\alpha_t, \tag{221}$$

where the final line uses the RKHS dual bound and $\|h_t\|_{\mathcal{H}} \le B_{\mathrm{wit},t}$. Combining (214), (216), (218), and (221) proves (212).

Sum (212) over t and divide by T:

$$\frac{1}{T}\sum_{t=1}^{T} D_t(P_t, Q_t) \le \frac{1}{T}\sum_{t=1}^{T} B_{\mathrm{wit},t}\varepsilon_{\mathrm{MMD},t} + \frac{2}{T}\sum_{t=1}^{T}(\alpha_t + \beta_t + s_t) + \frac{1}{T}\sum_{t=1}^{T} q_t. \tag{222}$$

Apply Cauchy–Schwarz to the first term:

$$\frac{1}{T}\sum_{t=1}^{T} B_{\mathrm{wit},t}\varepsilon_{\mathrm{MMD},t} \le \left(\frac{1}{T}\sum_{t=1}^{T} B_{\mathrm{wit},t}^2\right)^{1/2} \left(\frac{1}{T}\sum_{t=1}^{T} \varepsilon_{\mathrm{MMD},t}^2\right)^{1/2}. \tag{223}$$

Substitution of (223) into (222) proves (213). □

Connection to the nonstationary OMCI bound. Let $\tau_1 < \cdots < \tau_R \le T$ be the reported stream times used in Appendix B.7. Apply Proposition 13 with its horizon set to R and rename its generic time index as r. Set $P_r := P^c_{\tau_r}$, the current condition-c endpoint law, and let Q_r be the active at-most-M-atom windowed approximation at stream time τ_r. On the window-floor event of Lemma 5, the triangle inequality for MMD gives

$$\varepsilon_{\mathrm{MMD},r} \le \widehat{e}^c_{\tau_r,L,M} + E^c_{\tau_r,L} + S^c_{\tau_r,L} \le \widehat{e}^c_{\tau_r,L,M} + E^c_{\tau_r,L} + \overline{S}^c_{\tau_r,L}, \tag{224}$$

where $\widehat{e}^c_{\tau_r,L,M}$ is the MMD error between the active law and the windowed empirical law. The second inequality uses the declared envelope in (74). The reported weights remain on the simplex

and use at most M observed endpoints, so feasibility gives $\widehat{F}^c_{\tau_r,L,M} \geq F^{c,\star}_{\tau_r,L,M}$. The positive part in Proposition 6 therefore equals the ordinary difference, and (126) implies

$$\frac{1}{R}\sum_{r=1}^{R} \widehat{F}^c_{\tau_r,L,M} \leq \frac{1}{R}\sum_{r=1}^{R} F^{c,\star}_{\tau_r,L,M} + \mathcal{B}^c_{R,M}. \tag{225}$$

Write $d^c_{\tau_r,L} := E^c_{\tau_r,L} + \overline{S}^c_{\tau_r,L}$. Squaring (224) and applying $(a+b)^2 \leq 2a^2 + 2b^2$ gives

$$\varepsilon^2_{\mathrm{MMD},r} \leq 2(\widehat{e}^c_{\tau_r,L,M})^2 + 2(d^c_{\tau_r,L})^2 = 4\widehat{F}^c_{\tau_r,L,M} + 2(d^c_{\tau_r,L})^2. \tag{226}$$

Average (226) over $r = 1, \ldots, R$ and substitute (225) to obtain

$$\frac{1}{R}\sum_{r=1}^{R} \varepsilon^2_{\mathrm{MMD},r} \leq \frac{4}{R}\sum_{r=1}^{R} F^{c,\star}_{\tau_r,L,M} + 4\mathcal{B}^c_{R,M} + \frac{2}{R}\sum_{r=1}^{R} (d^c_{\tau_r,L})^2. \tag{227}$$

The bridge retains three visible sources: oracle M-atom approximation, computed-path error, and sampling plus staleness. Proposition 13 converts this endpoint bound into an average downstream bound only after the witness norm, approximation, numerical, and witness-staleness terms are also controlled. The dynamic-to-static RKHS reduction can justify the gain predictor used for support proposals (Jacobsen et al., 2025); it cannot verify the downstream witnesses.

Scope of the guarantees. The propositions support three increasingly strong claims:

1. a finite-bank bound, where all stored witnesses are inserted explicitly and their empirical expectation gaps are exact up to numerical precision;
2. a finite-configuration result, where α, β, q, and s are checked on preregistered endpoint pairs and grids;
3. a uniform theorem over a continuous population or neural class, which additionally requires compactness, covering, Lipschitz, or approximation results that make the sup-norm and RKHS-norm bounds uniform.

Only the third statement supports an unrestricted population or neural-class claim. The first two apply to the frozen bank or evaluated configurations and should be reported at that scope.

C.5 Witness transfer, PL geometry, and distinct conclusions

The MMD argument and the optimization argument solve different problems. The first compares expectations under two endpoint laws. The second converts stationarity under the full-law objective into an objective guarantee. Combining them requires either loss-value witnesses or gradient witnesses; a PL inequality cannot replace either witness condition.

Let P denote the current endpoint law, let Q denote a learnt endpoint law, let $\Theta \subseteq \mathbb{R}^p$, and let $\ell_\theta(y)$ be differentiable in θ. Define

$$R_P(\theta) := \mathbb{E}_{Y\sim P}[\ell_\theta(Y)], \qquad R_Q(\theta) := \mathbb{E}_{Y\sim Q}[\ell_\theta(Y)], \qquad R^\star_P := \inf_{\theta\in\Theta} R_P(\theta). \tag{228}$$

Assume the displayed expectations exist and justify differentiation under the expectation. For a constrained parameter set, replace the Euclidean gradient in what follows by the appropriate projected-gradient or stationarity mapping; an unconstrained gradient bound does not certify constrained stationarity. Define the uniform value and gradient perturbations

$$\Delta_0(P,Q) := \sup_{\theta\in\Theta} |R_P(\theta) - R_Q(\theta)|, \qquad \Delta_1(P,Q) := \sup_{\theta\in\Theta} \|\nabla R_P(\theta) - \nabla R_Q(\theta)\|_2. \tag{229}$$

Proposition 14 (Value transfer, PL transfer, and parameter recovery)**.** *Let $\widehat{\theta} \in \Theta$.*

1. *If $R_Q(\widehat{\theta}) \leq \inf_{\theta\in\Theta} R_Q(\theta) + \epsilon_{\mathrm{opt}}$, then*

$$R_P(\widehat{\theta}) - R^\star_P \leq 2\Delta_0(P,Q) + \epsilon_{\mathrm{opt}}. \tag{230}$$

This conclusion does not require a PL condition.

2. *Suppose R_P satisfies the PL inequality $2\mu_{\mathrm{PL}}(R_P(\theta)-R_P^\star) \le \|\nabla R_P(\theta)\|_2^2$ for every $\theta\in\Theta$, with $\mu_{\mathrm{PL}}>0$. If $\|\nabla R_Q(\widehat{\theta})\|_2 \le \epsilon_{\mathrm{stat}}$, then*

$$R_P(\widehat{\theta}) - R_P^\star \le \frac{(\epsilon_{\mathrm{stat}} + \Delta_1(P,Q))^2}{2\mu_{\mathrm{PL}}}. \tag{231}$$

3. *Let $\Theta_P^\star := \arg\min_{\theta\in\Theta} R_P(\theta)$. If, in addition, R_P satisfies quadratic growth $R_P(\theta) - R_P^\star \ge (\alpha/2)\operatorname{dist}(\theta,\Theta_P^\star)^2$ for some $\alpha>0$, then*

$$\operatorname{dist}(\widehat{\theta},\Theta_P^\star) \le \frac{\epsilon_{\mathrm{stat}} + \Delta_1(P,Q)}{\sqrt{\alpha\mu_{\mathrm{PL}}}}. \tag{232}$$

PL alone does not imply this parameter-distance claim.

Proof. **Proof map.** We combine uniform value stability, the Polyak–Łojasiewicz inequality, and quadratic growth. Equations (230)–(232) separate value, gradient, and parameter conclusions, and (235) composes each with the OMCI-selected law. For $\delta>0$, choose $\theta_{P,\delta}$ with $R_P(\theta_{P,\delta}) \le R_P^\star + \delta$. Add and subtract the Q-risk:

$$\begin{aligned} R_P(\widehat{\theta}) - R_P(\theta_{P,\delta}) &\le R_Q(\widehat{\theta}) - R_Q(\theta_{P,\delta}) + 2\Delta_0(P,Q) \\ &\le \epsilon_{\mathrm{opt}} + 2\Delta_0(P,Q). \end{aligned}$$

The second line uses $\inf_\theta R_Q(\theta) \le R_Q(\theta_{P,\delta})$. Letting $\delta\downarrow 0$ proves (230) and shows why PL plays no role in value transfer.

For the second claim, the triangle inequality gives $\|\nabla R_P(\widehat{\theta})\|_2 \le \|\nabla R_Q(\widehat{\theta})\|_2 + \Delta_1(P,Q) \le \epsilon_{\mathrm{stat}} + \Delta_1(P,Q)$. Insert this bound into the PL inequality at $\widehat{\theta}$ and divide by $2\mu_{\mathrm{PL}}$ to obtain (231). Finally, combine quadratic growth with (231) and take square roots to obtain (232). □

How MMD supplies Δ_0 and Δ_1. Let $d_k(P,Q) := \mathrm{MMD}_k(P,Q)$. Suppose that, for every θ, an RKHS function h_θ satisfies $\|h_\theta\|_{\mathcal{H}_k} \le B_0$ and $\sup_y |\ell_\theta(y) - h_\theta(y)| \le A_0$. The reproducing property and two approximation remainders give

$$\Delta_0(P,Q) \le B_0 d_k(P,Q) + 2A_0. \tag{233}$$

For gradients, suppose each coordinate $j\in\{1,\dots,p\}$ admits $g_{\theta,j}\in\mathcal{H}_k$ with $\|g_{\theta,j}\|_{\mathcal{H}_k} \le B_{1,j}$ and $\sup_y |\partial_j \ell_\theta(y) - g_{\theta,j}(y)| \le A_{1,j}$, uniformly over θ. Coordinatewise witness transfer followed by the Euclidean norm yields

$$\Delta_1(P,Q) \le \left[\sum_{j=1}^p \big(B_{1,j} d_k(P,Q) + 2A_{1,j}\big)^2\right]^{1/2}. \tag{234}$$

A vector-valued RKHS assumption can replace the coordinatewise condition when a dimension-free operator norm is available. Equations (233)–(234) are inherited witness inequalities for an IPM; the paper-specific step evaluates them at OMCI's data-selected crossing and then composes them with optimization as in Proposition 14.

Substitution at the OMCI crossing. If the Section 4 analysis supplies $d_k(P,Q) \le \rho_t$, define

$$\overline{\Delta}_{0,t} := B_0\rho_t + 2A_0, \qquad \overline{\Delta}_{1,t} := \left[\sum_{j=1}^p (B_{1,j}\rho_t + 2A_{1,j})^2\right]^{1/2}. \tag{235}$$

Then (230) and (231) hold with $\Delta_0(P,Q)$ and $\Delta_1(P,Q)$ replaced by $\overline{\Delta}_{0,t}$ and $\overline{\Delta}_{1,t}$. This substitution is the paper-specific composition: the crossing theorem chooses Q and controls $d_k(P,Q)$, while witness approximation determines whether that control reaches the teacher objective.

Scope of the PL corollary. The paper does not assert that an arbitrary neural teacher satisfies PL. The value-transfer result remains valid without PL. A neural-teacher experiment may invoke (231) only after it justifies PL or an equivalent error bound on the declared optimization region and verifies gradient-witness transfer. Otherwise, the experiment supports an empirical claim rather than the PL corollary.

Data reuse and frozen witnesses. If the algorithm fixes the kernel and witness bank before selecting Q, the preceding inequalities apply directly. If it learns witnesses from an independent split, condition on that split and apply the same argument to the evaluation split. If it learns witnesses and evaluates the bound on the same data, it must establish a uniform bound over the declared witness-selection class. A post hoc witness chosen because it maximizes the observed discrepancy does not inherit the fixed-witness bound automatically. This requirement is separate from PL geometry and remains necessary even when the optimizer reaches an exact stationary point.

Literature interface. Classical MMD/IPM theory supplies the expectation inequality. Polyak introduced a gradient-dominance condition for optimization Polyak (1963a); the Łojasiewicz inequality supplies the related analytic geometry Lojasiewicz (1963). Modern PL analyses convert stationarity into excess risk in nonconvex problems Karimi et al. (2016). The new logical interface is their composition at a learnt, condition-wise endpoint cardinality, together with the explicit separation between value witnesses, gradient witnesses, and parameter identification. Fixed-cardinality transport coresets optimize Q after a budget is supplied; they do not prove that the sampling-calibrated crossing selected that budget. Conversely, the crossing theorem does not guarantee neural-teacher performance without the witness and optimization conditions above.

D ALGORITHM IMPLEMENTATION

A support size is reproducible only under a fixed implementation protocol. This appendix specifies the exact and random-feature backends, candidate pools, fully corrective solve, admission and deletion rules, numerical tolerances, and runtime scaling.

A support size has meaning only under a fixed experimental protocol. Before inspecting downstream metrics, we fix the endpoint representation, condition map, kernel or random-feature map, sampling radius, tolerance κ, candidate search, numerical tolerances, path construction, and reporting window.

Reported configuration. Each run records:

1. the endpoint representation and condition map;
2. the exact kernel or random-feature map, including bandwidth, dimension, and seeds;
3. the minimum per-condition sample count required before reporting model order;
4. the sampling radius $E_t^c(\delta)$, compression radius $\varepsilon_t^c = \kappa E_t^c(\delta)$, confidence level, and calibration method;
5. the tolerance κ and reporting range $\mathcal{I}$;
6. whether CrossingReadout uses a nested path or a backward deletion path;
7. the candidate set searched by each linear oracle and its reported miss $\omega_{t,m}^c$;
8. the fully corrective solver, stopping rule, and stationarity residual;
9. whether the witness bank was frozen before model-order selection or used only for held-out evaluation.

Changing these choices changes the operational model order.

Exact RKHS backend. For a selected dictionary A^c,

$$F_t^c(A^c, a) = \frac{1}{2} a^\top K_{AA} a - a^\top k_{A,t}^c + \frac{1}{2} \|\widehat{\mu}_t^c\|^2, \tag{236}$$

where $(K_{AA})_{ij} = k(Y_i, Y_j)$ and

$$(k_{A,t}^c)_j = \frac{1}{n_c(t)} \sum_{i \le t : c_i = c} k(Y_j, Y_i). \tag{237}$$

Define the unnormalized condition-c Gram sum

$$G_t^c := \sum_{i,j \le t:\ c_i = c_j = c} k(Y_i, Y_j), \qquad G_0^c := 0. \tag{238}$$

When the arriving endpoint Y_t has condition c, we update it through

$$G_t^c = G_{t-1}^c + 2\sum_{i<t:c_i=c} k(Y_t, Y_i) + k(Y_t, Y_t), \qquad \|\widehat{\mu}_t^c\|^2 = \frac{G_t^c}{n_c(t)^2}. \tag{239}$$

For conditions not observed at time t, the corresponding Gram sum is unchanged. This backend computes exact objectives and residual inner products, up to the numerical solution of the simplex quadratic program (QP).

Random-feature backend. For a shift-invariant kernel, a frozen random-feature map $z : \mathcal{Y} \to \mathbb{R}^D$ defines an approximate kernel. The corresponding approximate-kernel objective is (Rahimi and Recht, 2007)

$$\bar{z}_t^c = \frac{1}{n_c(t)} \sum_{i\le t:c_i=c} z(Y_i), \qquad F_t^c(A^c, a) = \frac{1}{2}\|Z_{A^c}a - \bar{z}_t^c\|_2^2, \tag{240}$$

where $Z_{A^c} \in \mathbb{R}^{D\times|A^c|}$ has column $z(Y_j)$ for each $j \in A^c$ in the declared support order. The update costs $O(D)$ per observation, and candidate gains become Euclidean inner products. The frozen map defines an approximate kernel whose empirical mean has its own Hilbert-space radius, with K replaced by a declared bound on $\|z(y)\|_2$. Translating that approximate-kernel bound back to the original kernel additionally requires a random-feature approximation term. A scalable experiment should report either sensitivity or a separate approximation bound as D varies.

Candidate pool, admission, correction, and deletion. Between checkpoints, OMCI–MMD forms the declared index pool

$$\mathcal{J}_{t,\text{online}}^c := \mathcal{B}_t^c \cup A_{t-1}^c \cup \mathcal{R}_t^c \cup \mathcal{P}_t^c, \tag{241}$$

where $\mathcal{B}_t^c$ indexes the arriving condition-c endpoints, A_{t-1}^c is the selected dictionary retained from the preceding update, $\mathcal{R}_t^c$ indexes reservoir entries, and $\mathcal{P}_t^c$ indexes optional predictor proposals; all four sets therefore contain endpoint indices. For residual $r_{t,m}^c$, the additive linear-oracle miss of a returned index $\widehat{i}_{t,m+1}^c \in \mathcal{J}_{t,\text{online}}^c$ for the transition from budget m to $m+1$ is

$$\omega_{t,m}^c = \max_{i\in\mathcal{J}_{t,\text{online}}^c} \langle r_{t,m}^c, \phi(Y_i)\rangle_{\mathcal{H}} - \langle r_{t,m}^c, \phi(Y_{\widehat{i}_{t,m+1}^c})\rangle_{\mathcal{H}}. \tag{242}$$

For a current active set A, weights $a \in \mathsf{S}(A)$, and candidate endpoint y, let $q := q(A,a) = \sum_{j\in A} a_j\phi(Y_j)$, $s := \phi(y)$, $r := \widehat{\mu}_t^c - q$, and $u := s - q$. When $\|u\|_{\mathcal{H}} > 0$, the exact line-search step is

$$\gamma^\star = \Pi_{[0,1]}\left(\frac{\langle r, u\rangle_{\mathcal{H}}}{\|u\|_{\mathcal{H}}^2}\right), \tag{243}$$

where $\Pi_{[0,1]}(x) := \min\{1, \max\{0, x\}\}$ denotes Euclidean projection onto the unit interval; when $u = 0$, set $\gamma^\star := 0$ because the candidate leaves the coreset mean unchanged. The candidate decreases squared MMD by

$$\Delta_{t,\text{adm}}^c(y) = \frac{1}{2}\|r\|_{\mathcal{H}}^2 - \frac{1}{2}\|r - \gamma^\star u\|_{\mathcal{H}}^2. \tag{244}$$

After admission, we reoptimize every selected-dictionary weight and record a Karush–Kuhn–Tucker (KKT) or projected-gradient residual. Let A_t^c denote the current selected dictionary. The exact reoptimized deletion cost is

$$C_{t,\text{del}}^c(j) = \min_{a\in\mathsf{S}(A_t^c\setminus\{j\})} F_t^c(A_t^c \setminus \{j\}, a) - \min_{a\in\mathsf{S}(A_t^c)} F_t^c(A_t^c, a). \tag{245}$$

A lazy heap may propose a deletion, but the exact post-deletion objective must approve it.

Constructing the memory–error curve. A single active set gives one point, not a crossing. CrossingReadout returns $m \mapsto (\widehat{e}_{t,m}^c, \widehat{F}_{t,m}^c)$.

Nested path. We maintain

$$A_{t,1}^c \subset A_{t,2}^c \subset \cdots \subset A_{t,M_{\max}}^c \tag{246}$$

by adding one atom and fully correcting at every size. The exact support verifications use this path.

Algorithm 4 Operational OMCI–MMD update between checkpoints

Require: previous states $(n_c(t-1), \widehat{\mu}^c_{t-1}, A^c_{t-1}, a^c_{t-1})_{c\in\mathcal{C}}$ and arriving index batch $\mathcal{B}_t$

for each condition c represented in $\mathcal{B}_t$ **do**
  Set $\mathcal{B}^c_t := \{i \in \mathcal{B}_t : c_i = c\}$; update $n_c(t)$ and $\widehat{\mu}^c_t$; compute τ^c_t by (247).
  Form $\mathcal{J}^c_{t,\mathrm{online}}$ by (241); set $A^c_t \leftarrow A^c_{t-1}$, and if A^c_t is empty, initialize it with the best verified singleton in this nonempty pool.
  Fully correct the current weights a^c_t on A^c_t and evaluate $\overline{F}^c_{t,\mathrm{cur}}$.
  while $\overline{F}^c_{t,\mathrm{cur}} > \tau^c_t$ and some candidate has $\underline{\Delta}^c_{t,\mathrm{adm}}(y) > 0$ **do**
    Admit the candidate with the largest verified decrease; fully correct; recompute $\overline{F}^c_{t,\mathrm{cur}}$.
  end while
  if $\overline{F}^c_{t,\mathrm{cur}} \le \tau^c_t$ **then**
    while some $j \in A^c_t$ has a verified post-deletion value $\overline{F}^c_{t,-j} \le \tau^c_t$ **do**
      Delete the atom with the smallest verified post-deletion value; fully correct; update $\overline{F}^c_{t,\mathrm{cur}}$.
    end while
  else
    Report that the declared online pool did not reach the threshold; record its within-pool and availability errors.
  end if
  At a checkpoint, run the complete Algorithm 1 on the declared candidates.
end for

Deletion path. We may instead start from a larger support, delete the least costly atom, and reoptimize after each deletion. Theorem 1 then uses the oracle slack associated with that returned path; exhaustive verification can compute it, while scalable runs require a separate upper bound.

If the curve remains flat near the radius, we report the bracket from Theorem 1 rather than choose an integer by tie-breaking.

Statistical threshold and numerical tolerances. Write $E^c_t := E^c_t(\delta)$ for the radius in (6), and define

$$\varepsilon^c_t := \kappa E^c_t, \qquad \tau^c_t := \frac{1}{2}(\varepsilon^c_t)^2. \tag{247}$$

The statistical threshold τ^c_t must not absorb an unverified optimization or oracle error. For any stored feasible weights, let $\overline{F}^c_t(A, a)$ be a certified upper bound on the exactly evaluated objective $F^c_t(A, a)$; exact kernel arithmetic permits $\overline{F}^c_t = F^c_t$, while approximate arithmetic must add a declared evaluation bound. A stored law crosses only when

$$\overline{F}^c_t(A, a) \le \tau^c_t. \tag{248}$$

For a candidate y, the exact line-search decrease in (244) lower-bounds the decrease after full correction; let $\underline{\Delta}^c_{t,\mathrm{adm}}(y)$ be a numerically certified lower bound on that quantity. For deletion of j, let $\overline{F}^c_{t,-j}$ be a certified upper bound on the fully reoptimized post-deletion objective. We admit only when $\underline{\Delta}^c_{t,\mathrm{adm}}(y) > 0$ and delete only when $\overline{F}^c_{t,-j} \le \tau^c_t$. Solver residuals, objective-evaluation tolerances, and the linear-oracle miss in (242) remain separately reported quantities; they do not become statistical resolution by being added to (247). Downstream Wasserstein, Sinkhorn, or flow errors enter the selection criterion only when a frozen witness bank was inserted into the kernel before calibration.

Reported quantities. We report the MMD residual, QP residual, oracle miss relative to the declared candidate set, witness norm and approximation error for downstream transfer, and held-out ordinary differential equation (ODE) or neural-flow errors. These quantities answer different questions and remain separate in the results.

Computational cost. Let $M^c_t := |A^c_t|$ be the selected-dictionary size, define $A^{c,+}_t := \{j \in A^c_t : a^c_{t,j} > 0\}$ and $M^c_{t,+} := |A^{c,+}_t| \le M^c_t$ as the online analogue of (22), let b^c_t be the number of candidates scored, and let D be the random-feature dimension. Candidate scoring costs $O(b^c_t D)$.

Reweighting solves a simplex quadratic program of dimension M_t^c; a generic dense solver has cubic worst-case cost, while warm-started active-set solvers are usually cheaper. After readout, pruning zero weights reduces stored endpoint count to $M_{t,+}^c$. A deletion path requires repeated reweighting.

Exact observed-support search costs $O(n_c(t))$ oracle evaluations and appears only in the exact-search experiments. A bounded-memory implementation can use a cache, reservoir, approximate maximum-inner-product search, or gain predictor. It must report both the miss within the searched pool and the availability gap to the full observed support.

E Experiments

E.1 Experimental setups

E.1.1 Neural-teacher protocol

Question and stream. The experiment asks whether the cardinality selected by the MMD crossing supplies enough endpoint information to train a conditional rectified-flow network. Both conditions share ten fixed points in $\mathbb{R}^2$. In the fixed coordinate order, the stationary and post-change endpoint laws are

$$P_0 = (0.16 \times 6,\ 0.01 \times 4), \qquad P_1 = (0.115 \times 6,\ 0.145 \times 2,\ 0.01 \times 2). \tag{249}$$

Condition 0 follows P_0 for all $T = 4000$ observations. Condition 1 follows P_0 through $t = 2200$ and P_1 thereafter. The change transfers mass to two previously rare central atoms while leaving two rare extremes unresolved. Each condition contributes one independent endpoint per stream step, and the current empirical law uses its latest $L = 500$ observations. We retain the condition index throughout: Proposition 1 shows why marginal agreement cannot protect the conditional teacher, while C^{2}OT identifies the analogous condition-prior distortion for condition-agnostic minibatch transport (Cheng and Schwing, 2025).

Selector and freeze. At checkpoints $t = 100, 200, \ldots, 4000$, Algorithm 2 scans every endpoint type observed in the active window, fully corrects its weights, and selects

$$\widehat{M}_t^c = \min\{m : \widehat{e}_{t,m}^c \le 1.05 E_{n_t}\}, \qquad n_t = \min\{t, 500\}. \tag{250}$$

The RBF bandwidth equals 1.15. Independent simulator draws calibrate E_n as the 0.95 empirical-to-population MMD quantile from 5000 replicates for each declared pure regime and sample size, augmented by four transition mixtures at $n = 500$; the selector uses the largest applicable quantile. A separate adjacent-window MMD statistic uses the 0.99 quantile of the maximum over both conditions and all checkpoints from 5000 stationary-null trajectories. That statistic appears only as a post hoc change statistic in Figure 3B. We freeze the teacher laws at $t \in \{2000, 4000\}$ before fitting any network.

Teacher arms. Table 2 changes only the endpoint teacher. The "under" and "over" arms read alternative points from the same fully corrective path, so they isolate stopping before or after the registered crossing. Retain-all keeps all observations through the checkpoint and therefore mixes P_0 and P_1 after the change.

Table 2: Endpoint teachers used for the neural comparison. Every arm trains the same network with paired initialization, source draws, time draws, minibatch order, and evaluation draws.

Teacher	Construction	Experimental role
Full current	Complete empirical law on the latest 500 observations per condition	Current-window reference
OMCI–MMD	Weighted law at the first crossing in (250)	Proposed data-dependent support rule
Under	Same path at $\max\{1, \lfloor \widehat{M}/2 \rfloor\}$	Deliberate deletion of resolvable branches
Over	Same path at $\min\{2\widehat{M}, M_{\max}\}$	Deliberate fitting past the crossing; equals full current in this ten-atom geometry
Retain all	All observations through the checkpoint	Growing-history teacher with stale pre-change mass

Neural learner. Every arm trains the same jointly conditional rectified-flow multilayer perceptron under (1). The network has 56,354 parameters: four width-128 SiLU hidden layers, a 32-dimensional sinusoidal time embedding, and a 16-dimensional condition embedding. Adam uses learning rate 10^{-3}, betas $(0.9, 0.999)$, $\epsilon = 10^{-8}$, batch size 1024 with 512 rows per condition, float32 arithmetic, and 4000 updates. The protocol uses no scheduler, early stopping, clipping, arm-specific tuning, or post-result checkpoint selection. This design isolates endpoint-law replacement from the fixed-law estimation problem studied for rectified flows (Sahoo et al., 2026). It also tests the upstream cardinality decision that coreset-induced velocity constructions such as CCVFM take as supplied (Wang et al., 2026).

Evaluation and statistical unit. At each frozen checkpoint, the evaluation draws 8192 true-current-law rectified paths per condition for held-out velocity MSE. It also integrates 8192 generated trajectories with 100-step Runge–Kutta order four (RK4) and compares their terminal points with 8192 independent true-law endpoints using the biased fixed-bandwidth RBF MMD^2 estimator. Both metrics use bandwidth 1.15 and lower values indicate better fidelity. "Overall" averages the two conditions and "worst" takes their maximum. For any condition-indexed metric q, the registered worst-condition contrast is

$$\Delta_{\mathrm{worst}}(A, B; q) := \max_c \{q_{A,c} - q_{B,c}\}. \tag{251}$$

Positive values mean that arm A performs worse than arm B. Five independent endpoint streams supply the statistical units. The analysis averages three nested network fits within each stream and reports two-sided Student-t intervals with four degrees of freedom.

Post-selection analyses. The fixed-size sweep retrains only after OMCI freezes the deployed sizes. It therefore measures where neural error flattens without entering the selection rule. The stale-history comparison keeps all ten endpoint types and the neural budget fixed while changing only the temporal horizon. At $t = 4000$, full current uses condition-1 observations 3501:4000, all from P_1, whereas retain-all uses $2200P_0 + 1800P_1$ observations; its expected mass on each emergent middle atom is 0.07075 rather than 0.145. One-factor sensitivity runs vary learning rate, batch size, update count, and κ on one declared stream/model cell; they measure robustness around the registered setting rather than select a new default.

Finite-alphabet mechanism study. The endpoint alphabet contains ten locations. We use an exact radial basis function (RBF) Gram matrix, scan the full observed support, and solve the fully corrective simplex problem at every sparse step. Independent simulator windows calibrate the radius. Exhaustive subset enumeration gives the reference curve at the final checkpoint. For $\eta = \sum_i w_i \delta_{z_i}$, draw $Z \sim \eta$ independently of $X_0 \sim \mathcal{N}(0, I_2)$, where I_2 is the 2×2 identity matrix. For $s \in [0, 1)$, the analytic Gaussian-path teacher is

$$X_s = (1 - s)X_0 + sZ, \qquad u_s^\eta(x) = \sum_i \pi_i(s, x) \frac{z_i - x}{1 - s}, \tag{252}$$

with

$$\pi_i(s, x) = \frac{w_i \exp\{-\|x - sz_i\|^2/[2(1 - s)^2]\}}{\sum_j w_j \exp\{-\|x - sz_j\|^2/[2(1 - s)^2]\}}. \tag{253}$$

This conditional-expectation field is the population squared-loss teacher induced by the linear interpolation in (1), the path used by rectified and Flow-Matching constructions (Liu et al., 2023; Lipman et al., 2023). We hold that path fixed to isolate endpoint-law replacement; fixed-law sample-complexity analysis describes the downstream estimation problem once an endpoint law has been supplied (Sahoo et al., 2026). We use common source samples and a common Runge–Kutta grid for every endpoint law.

Finite-support witness verification. We evaluate forty independent current windows. Four specified verification streams use exhaustive subset enumeration. For every OMCI path point, we compute Kantorovich and Sinkhorn witnesses. Because the endpoint alphabet is finite and the Gaussian-kernel Gram matrix on its distinct locations is strictly positive definite, their values admit exact interpolants in the finite RKHS span. The saved results include radius calibration, sparse paths, oracle misses, QP residuals, witness norms, and transfer-bound ratios.

Continuous witness stress test. The continuous target is a one-dimensional Gaussian mixture. We approximate Sinkhorn first variations by finite RBF fits on a separate grid and include witness approximation, Sinkhorn solver, grid, and quadrature errors in the bound. We implement only the entropic-anchor compression component from the Coreset-Induced Conditional Velocity Flow Matching (CCVFM) preprint; we do not implement its Gaussian-mixture lift, conditional velocity, or correction flow (Wang et al., 2026). We report Online Sinkhorn and Compressed Online Sinkhorn separately because they approximate dual potentials rather than endpoint laws (Mensch and Peyré, 2020; Wang et al., 2023).

Matched-size transport refinement. For each stream, we run OMCI once, store the selected size, and lock the selection artifact before any transport-refinement process starts. Adapter scripts then call WCSL at the frozen size (Yin et al., 2025), evaluate Sinkhorn–CO2 (Kokot and Luedtke, 2025), run the free-support EMS coreset (Yin et al., 2026), and include a direct differentiable Sinkhorn optimizer. Each adapter exposes the returned locations, weights, objective trace, stopping residual, wall-clock time, peak memory, and random seed. A separate fixed-size sweep computes a post-selection envelope, defined later in Appendix E; its outputs live in an evaluation-only namespace and never enter the deployable pipeline. The experiment reports separate tables for weighted observed-support, uniform-weight, and weighted free-support feasible classes.

E.1.2 Matched-size transport refinement

The baselines approximate different objects. Stage I uses OMCI to select the cardinality. Stage II refines a law at that fixed cardinality. We identify what each baseline stores and keep endpoint-memory error separate from coupling error.

Support selection. We choose $\widehat{m}$ only from the OMCI–MMD curve and an independently calibrated MMD radius:

$$\widehat{m} \text{ depends on the OMCI–MMD curve and its sampling radius only.} \tag{254}$$

We evaluate Sinkhorn, Wasserstein, teacher-field, rollout, image, video, and financial metrics afterward. A witness-aware kernel enters the selection rule only when we freeze its witness bank, weights, normalization, and refresh rule before calibration.

Experimental sequence. The experiments ask four linked questions: does the MMD curve select a support size; does the selected law satisfy a Wasserstein or Sinkhorn witness bound; does it preserve an analytic teacher field; and does a trained neural field preserve current-law fidelity? The controlled endpoint studies answer the first three. Section 5 answers the fourth by fitting the same conditional rectified-flow network from each frozen endpoint teacher. The matched-size benchmark then tests transport-specific refinement at the frozen OMCI cardinality against an evaluation-only size sweep.

Baseline roles. Reservoir sampling, kernel herding, k-medoids, and static pre-change coresets approximate the endpoint law. At matched size, they test whether learning the size or choosing a different sparse path matters. We implement only the entropic-anchor compression component from the recent Coreset-Induced Conditional Velocity Flow Matching (CCVFM) preprint; we do not implement its Gaussian-mixture lift, conditional velocity, or correction flow (Wang et al., 2026). Wasserstein Coreset via Sinkhorn Loss (WCSL) directly optimizes a fixed-size Sinkhorn coreset (Yin et al., 2025). Sinkhorn Coresets of Order 2 (Sinkhorn–CO2) derives a divergence-adapted surrogate (Kokot and Luedtke, 2025), while the expectation–maximization Sinkhorn (EMS) coreset optimizes weighted free-support representatives (Yin et al., 2026). They belong in the transport-refinement stage (Stage II) after OMCI has frozen the cardinality. Online and Compressed Online Sinkhorn approximate streaming transport potentials, so their support nodes and memory counts describe a dual estimator rather than an endpoint law (Mensch and Peyré, 2020; Wang et al., 2023). Safe sparse Newton sparsifies second-order solver structure (Tang and Qiu, 2024); sparse-plus-low-rank quasi-Newton methods compress related curvature information (Wang and Qiu, 2025); ForgettingOT reuses solver state across changing problems (Wen, 2026). None selects the endpoint training law. Optimal-transport conditional Flow Matching (OT-CFM) and conditional optimal transport (C^2OT) change the source–target coupling and belong in an experiment that varies the coupling rule (Tong et al., 2024; Cheng and Schwing, 2025).

Matched-size select–then–refine benchmark. For stream replicate r and condition c, let $\widehat{P}_r^c$ denote the empirical endpoint target in the selection window. The support-selection stage (Stage I) observes only that window and returns $\widehat{m}_r^c$; we serialize the integer before launching any transport-specific optimizer. For a method label A and cardinality m, let $\mathsf{T}_{A,m}$ denote that method's fixed-cardinality endpoint optimizer. Stage II then constructs

$$\widetilde{Q}_{A,r}^c = \mathsf{T}_{A,\widehat{m}_r^c}(\widehat{P}_r^c), \qquad A \in \{\text{WCSL}, \text{Sinkhorn–CO2}, \text{EMS}, \text{direct Sinkhorn}, \text{OMCI weights}\}, \tag{255}$$

using a common target law $\widehat{P}_r^c$, common entropy convention, common held-out evaluation sample, and a predeclared compute budget. The primary result is the downstream test loss of every optimizer at the same OMCI-selected cardinality. We report native initialization and, when the feasible class contains the weighted OMCI law, an OMCI warm start as separate rows.

Observed-support and free-support methods appear in separate strata. EMS optimizes weighted free-support representatives, while the OMCI theorem and Sinkhorn–CO2 construction select weighted observed points. WCSL is assigned to the stratum dictated by its published support and weight constraints. A free-support method may therefore attain lower loss at the same cardinality because it solves a larger optimization problem; that advantage is legitimate but must not be attributed to a better size rule.

Post-selection envelopes and oracle gaps. Fix a condition c, suppress its superscript in this paragraph, and write $[x]_+ := \max\{x, 0\}$. For evaluation only, each optimizer A is also run over a predeclared grid $\mathcal{M}$. Let $L_{A,r}^{\text{test}}(m)$ denote its held-out downstream loss at size m. The post-selection object is the entire memory–accuracy envelope

$$\mathcal{E}_{A,r}^{\text{test}} := \{(m, L_{A,r}^{\text{test}}(m)) : m \in \mathcal{M}\}, \tag{256}$$

not the raw minimum over m: without a memory price, transport loss commonly favors the largest available coreset. We therefore predeclare either a target loss $\tau \geq 0$ or a memory penalty $\lambda C(m)$, where $\lambda \geq 0$ and $C(m)$ is a nondecreasing, predeclared memory-cost function. The corresponding test-only oracles are

$$m_{A,r}^{\text{or}}(\tau) := \min\{m \in \mathcal{M} : L_{A,r}^{\text{test}}(m) \leq \tau\}, \tag{257}$$

$$J_{A,r}^{\text{or}}(\lambda) := \min_{m \in \mathcal{M}} \{L_{A,r}^{\text{test}}(m) + \lambda C(m)\}. \tag{258}$$

We set $m_{A,r}^{\text{or}}(\tau) = +\infty$ when the grid never reaches τ and mark the corresponding memory comparison as unresolved. For the deployed OMCI size, we report the target violation, memory overhead, and penalized oracle gap

$$\Delta_{A,r}^{\text{loss}}(\tau) := \left[L_{A,r}^{\text{test}}(\widehat{m}_r) - \tau\right]_+, \tag{259}$$

$$\Delta_{A,r}^{\text{mem}}(\tau) := \widehat{m}_r - m_{A,r}^{\text{or}}(\tau), \tag{260}$$

$$\Delta_{A,r}^{\text{pen}}(\lambda) := L_{A,r}^{\text{test}}(\widehat{m}_r) + \lambda C(\widehat{m}_r) - J_{A,r}^{\text{or}}(\lambda). \tag{261}$$

We also report runtime, peak memory, failure rate, initialization variance, and the best loss attainable by any faithful optimizer with at most $\widehat{m}_r$ atoms. None of these post hoc envelopes selects the deployed model. Validation-selected and one-standard-error sizes provide attainable comparators; the test-oracle quantities measure only how much a method could have gained by inspecting every size after the fact.

E.2 Cross-modal memory-induced hallucination pilots

Dynamic scene flow. We build ten 16×16 compositional scene atoms with exact object and relation labels, following the controlled-scene philosophy of CLEVR (Johnson et al., 2017). Condition 0 remains on six atoms. Condition 1 uses atoms $\{0, \ldots, 5\}$ before the change and $\{0, 1, 2, 3, 6, 7, 8, 9\}$ afterward, so atoms 4–5 become stale and atoms 6–9 become current. A shared convolutional autoencoder maps images to a six-dimensional latent. For each frozen memory law, a separate 13,302-parameter conditional rectified-flow network trains for 900 updates and generates 400 post-change samples for each of three independent seeds. We classify outputs by nearest latent scene prototype and report stale-scene rate, current-mode coverage, total variation from the current scene law, and off-catalog rate. A mode counts as covered when it receives at least 3% generated mass.

Structured table-to-text. We construct a temporal mission table with ten exact rows. Two entity-specific rows become obsolete after the change ("Eos launches from Florida in 2027" and "Fenix lands in Madrid in 2028") and are replaced by current California/2028 and Kyoto/2029 rows; two further rows introduce new current facts. The task follows the controlled one-sentence table-verbalization setting of ToTTo (Parikh et al., 2020), but uses a smaller exact table so every stale or current output is machine-checkable. A 30,224-parameter GRU receives the table and condition tokens, autoregressively emits (entity, relation, location, year), and passes that row to a deterministic surface realizer. Each memory law uses three independent fits, 700 updates, and 400 generated rows per fit. We report stale-fact rate, current-fact coverage, exact-current rate, unsupported-hybrid rate, and total variation from the current row law.

Frozen-law protocol. The four arms are full current window, retain all history, fixed-small support, and windowed OMCI–MMD. We freeze all four laws before fitting either neural model. Image or text outcomes never tune the kernel, radius, window, or selected support. The runnable implementations and raw checkpoints live in `experiments/experiment3_dynamic_scene_flow` and `experiments/experiment4_table_to_text`.

E.3 Additional results

Neural-teacher comparison. Table 3 reports the post-change checkpoint. OMCI–MMD uses six atoms for condition 0 and eight for condition 1, or seven positive atoms on average. It stays close to the full-current teacher while avoiding the under-selected and stale-history failure modes. Table 4 reports paired contrasts across the five independent endpoint streams.

Table 3: Neural-teacher results at $t = 4000$. Lower is better. The final column isolates the shifted condition.

Teacher	Positive atoms	Construction rows	Worst-condition velocity MSE	Worst-condition terminal MMD^2	Shifted-condition terminal MMD^2
Full current	10.0	500	3.049	0.00350	0.00309
Under	3.5	500	18.963	0.1660	0.1154
OMCI–MMD	7.0	500	3.212	0.00414	0.00349
Over	10.0	500	3.049	0.00350	0.00309
Retain all	10.0	4000	3.039	0.01773	0.01773

Table 4: Registered stream-level contrasts. The intervals average the three nested fits within each of five independent streams before applying a two-sided t_4 interval.

Contrast	Metric	Estimate	95% interval
OMCI–MMD − full current	Worst-condition velocity MSE	$+0.1706$	$[+0.1480, +0.1932]$
OMCI–MMD − full current	Worst-condition terminal MMD^2	$+0.000837$	$[+0.000328, +0.001346]$
Under − OMCI–MMD	Worst-condition terminal MMD^2	$+0.162763$	$[+0.160118, +0.165408]$
Retain all − full current	Condition-1 terminal MMD^2	$+0.014639$	$[+0.012182, +0.017095]$

Numerical integration and resource accounting. A 200-step RK4 check reused the first 1024 source and reference rows per condition; the largest single-stream absolute change from the 100-step metric was 3.77×10^{-7}. All current-window arms consume 500 construction observations. Their neural-training times lie between 6.84 and 6.88 seconds per 1000 updates; selection and evaluation are excluded from that timing. Positive atoms measure representation complexity rather than byte memory. Figure 3 reports the selector, neural fidelity, and resource quantities.

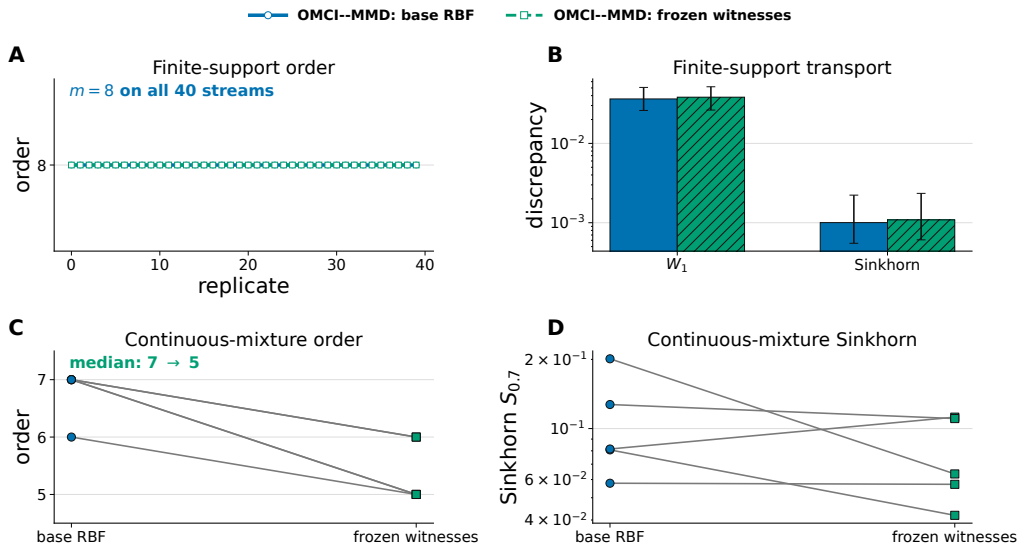


Figure 5: Witness-aware transport result. **A–B:** on finite support, the base RKHS already represents the evaluated transport witnesses, so freezing those witnesses leaves both model order and downstream discrepancy essentially unchanged. **C–D:** on the continuous mixture, the frozen bank changes the selected order and improves median held-out Sinkhorn divergence, but one stream worsens. This is the behavior predicted by (140): a witness-aware kernel helps only when the added functions expose downstream directions that the base kernel misses.

Table 5: Cross-modal pilot results (mean ± standard deviation over three independent fits). Coverage is the fraction of current scene modes or factual rows generated above the modality-specific threshold.

	Conditional latent image flow		Structured table-to-text GRU	
Memory law	stale scenes (%)	scene coverage (%)	stale facts (%)	fact coverage (%)
Full current	3.0 ± 1.1	100 ± 0	0.0 ± 0.0	100 ± 0
Retain all history	20.1 ± 1.8	91.7 ± 14.4	17.4 ± 1.2	100 ± 0
Fixed small	3.5 ± 0.8	75 ± 0	0.0 ± 0.0	75 ± 0
OMCI–MMD	4.7 ± 1.0	100 ± 0	0.0 ± 0.0	100 ± 0

E.3.1 Witness-aware transport result

E.4 Generated stale outputs and deleted current support

Figure 6 shows the intervention before either neural model is fitted. Both modalities use the same ten endpoint identities and the same four frozen memory laws. In each modality, endpoints 4–5 become obsolete after drift. Fixed-small memory deletes valid current endpoints 1 and 3, retain-all preserves positive mass on the obsolete endpoints, and OMCI–MMD freezes the same eight-endpoint support as the full-current reference. The image and language models therefore differ only in how they learn from the frozen law, not in which memory failure the experiment applies.

The downstream result in Figure 7 uses two different neural models. The top row trains a 13,302-parameter conditional latent rectified-flow image model and evaluates 400 generated scenes per fit. The bottom row trains a 30,224-parameter structured table-to-text GRU and evaluates 400 generated factual rows per fit. In both rows, the left panel counts obsolete outputs and the right panel measures coverage of the current endpoint support.

Retain-all history makes obsolete outputs probable under the training law: the image flow assigns roughly one fifth of its generated mass to stale scenes, and the row model verbalizes obsolete facts at roughly one sixth. Fixed-small memory avoids those obsolete endpoints only by deleting two valid current modes or facts. OMCI–MMD preserves the full current support and keeps stale generation near the full-current reference. The mean image off-catalog rate remains below 0.2% in every arm,

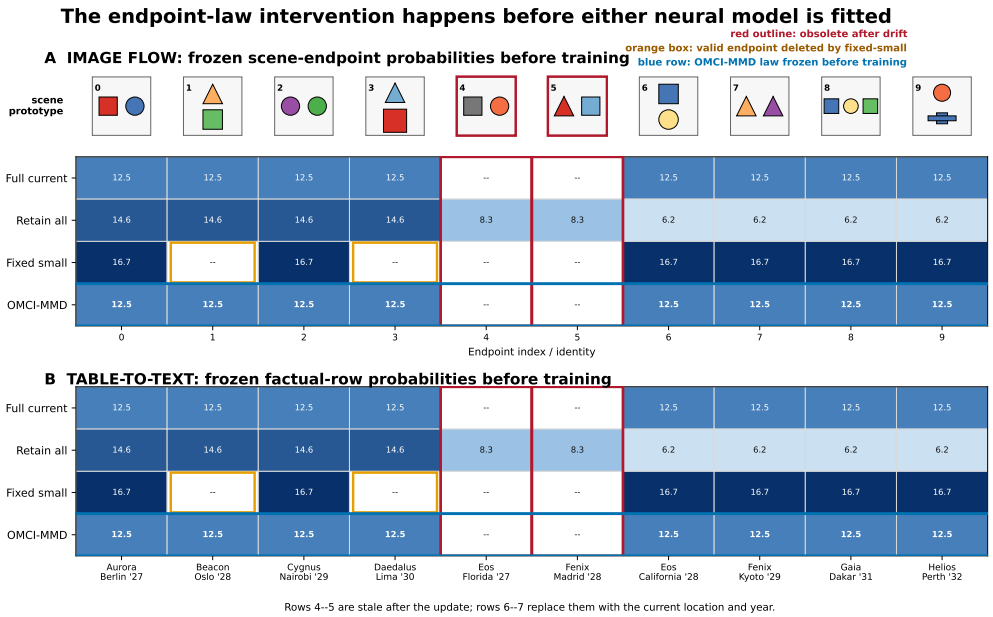


Figure 6: **Endpoint laws frozen before neural training. A:** the image experiment uses ten compositional scene endpoints; red outlines mark the two obsolete scenes, and orange boxes mark valid current scenes deleted by fixed-small memory. **B:** the language experiment uses ten factual rows; the Florida/2027 and Madrid/2028 rows become obsolete and are replaced by the California/2028 and Kyoto/2029 rows. Full current and OMCI–MMD place mass only on the eight current endpoints. Retain-all leaves mass on obsolete endpoints, whereas fixed-small omits two valid current endpoints. No image or language outcome enters this construction.

and the grammar-constrained text model generates no unsupported hybrid rows. Malformed decoding therefore does not explain the contrast.

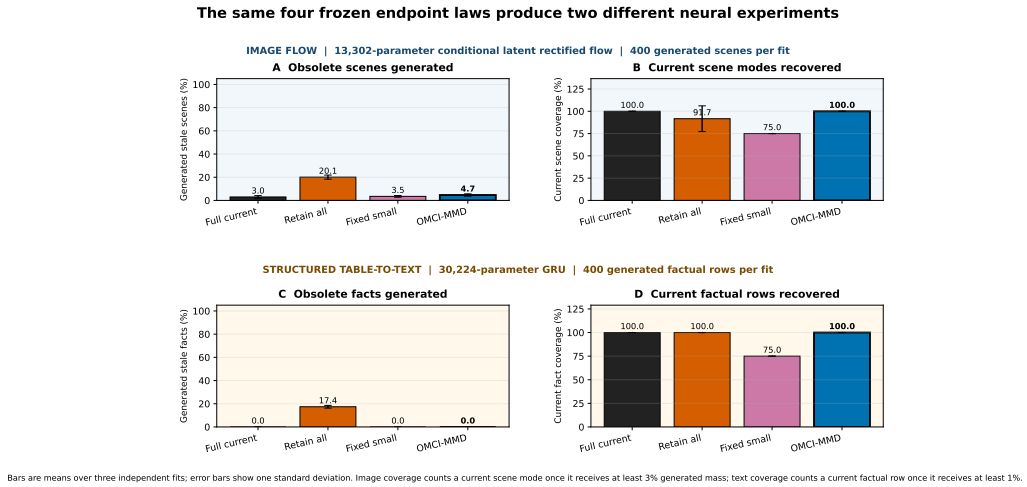


Figure 7: **Neural consequences of the frozen endpoint laws. Top row:** a 13,302-parameter conditional latent rectified-flow image model; each fit generates 400 scenes. **Bottom row:** a 30,224-parameter structured table-to-text GRU; each fit generates 400 factual rows. The left panels report obsolete outputs, while the right panels report coverage of the current endpoint support. Bars show means over three independent fits and error bars show one standard deviation. Image coverage uses a 3% generated-mass threshold; factual-row coverage uses a 1% threshold.