ANISOTROPIC SUPPORT-TRANSPORT ROBUST OFFLINE MODEL-BASED REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline model-based reinforcement learning (MBRL) suffers from a structural failure mode: the learned dynamics model is accurate on the data support but drifts rapidly off-support, where the policy naturally explores. Existing remedies—uncertainty penalties, pessimistic surrogate MDPs, and adversarial model learning—reduce trust to a single scalar quantity and commit to a fixed rollout horizon, ignoring the anisotropic geometry of real offline datasets. We introduce **ASTRO** (Anisotropic Support-Transport Robust Offline RL), which replaces the scalar trust radius with a data-adaptive *support tube* carved by local k-nearest-neighbor covariance and density estimates. The same geometric primitive gates both a robust value-aware model-learning loss—driving predicted latent values toward the minimum over the tube—and a state-dependent rollout horizon selected by a learned trust score. On 24 D4RL datasets spanning the Adroit and MuJoCo suites, ASTRO consistently matches or surpasses strong model-based baselines, with the largest gains on support-fragmented datasets where scalar trust abstractions are least appropriate.

1 INTRODUCTION

Offline reinforcement learning aims to extract high-performing control policies from previously logged data without further environment interaction Levine et al. (2020); He (2023); Prudencio et al. (2023). The model-based branch is particularly attractive because a learned dynamics model can synthesize additional training transitions (Yu et al., 2020; Matsushima et al., 2020; Kidambi et al., 2020; Yu et al., 2021; Rigter et al., 2022; Sun et al., 2023; Qiao et al., 2026). Despite this progress, offline MBRL remains dominated by a single failure mode: *the model is wrong where the policy wants to act* Uehara & Sun (2021). On the data support the dynamics is typically accurate; off-support, even moderate-horizon rollouts compound errors rapidly. The dominant remedies share two structural choices: (a) trust is reduced to a *global scalar*—a penalty coefficient, an adversarial weight, or a Wasserstein radius ξ—and (b) the rollout horizon is *fixed* across all states. Both choices are in tension with the geometry of real offline data, where support is highly non-uniform and anisotropic: some state-space directions are densely covered and safe to interpolate along, while others constitute genuine extrapolation directions that no global scalar can correctly adjudicate. A fixed horizon further compounds the problem by wasting computation in untrusted regions or discarding useful information in well-supported ones. We propose **ASTRO** (Anisotropic Support-Transport Robust Offline RL), which replaces both abstractions with a unified, data-adaptive geometric primitive. The core idea is the *anisotropic support tube*: for each logged transition, we estimate a local covariance and density from the K-nearest neighbors in a learned latent space, and define a tube that stretches along well-supported directions while contracting along extrapolation directions. This single primitive is reused consistently across three algorithmic components:

1. A *robust value-aware model loss* that drives predicted latent values toward the minimum over the support tube, yielding spatially varying pessimism without manual tuning.
2. A *trust score* derived from the same tube geometry, ensemble disagreement, and calibration error.
3. A *state-dependent rollout horizon* gated by the trust score, automatically shortening rollouts in out-of-distribution (OOD) regions.

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.

The sharing of a single geometric primitive across model learning and rollout planning is the central design choice that distinguishes ASTRO from prior work. On 24 D4RL datasets Fu et al. (2020) (12 Adroit Rajeswaran et al. (2017), 12 MuJoCo Todorov et al. (2012)), ASTRO achieves the highest overall mean score among all model-based methods and the largest gains on support-fragmented datasets (medium-replay, cloned), where scalar trust abstractions are least appropriate.

Our contributions can be summarized as threefold: (1) We identify the anisotropy of offline data support as a structural limitation of scalar-trust offline MBRL, and propose the *support tube*—a local, data-driven generalization of both Euclidean and Wasserstein trust regions. (2) We design a unified framework in which the same geometric primitive drives a robust value-aware model loss, a trust score, and a state-dependent rollout horizon. (3) ASTRO consistently matches or surpasses seven baselines on D4RL datasets, with ablations confirming that each component contributes to the final performance.

2 Related Work

2.1 Model-based offline RL

Model-based offline RL methods differ primarily in how they account for model error Argenson & Dulac-Arnold (2020); He (2023). **MOPO** (Yu et al., 2020) and **MOReL** (Kidambi et al., 2020) penalize the policy with model-uncertainty estimates from an ensemble of dynamics networks. **COMBO** (Yu et al., 2021) avoids explicit uncertainty estimation by performing conservative value learning on model-generated OOD samples. **RAMBO** (Rigter et al., 2022) casts offline MBRL as a maximin problem and trains an adversarial dynamics model that minimizes the policy's value off-support. **MOBILE** (Sun et al., 2023) introduces a model-Bellman inconsistency term that penalizes transitions where the model and the value function disagree. **ROMI** (Qiao et al., 2026) replaces RAMBO's adversarial gradient update with a robust value-aware loss over a Wasserstein-induced state-uncertainty set of radius ξ Yang (2017); Peyré & Cuturi (2019), achieving smoother conservatism control. All of these methods share two structural choices: (a) a *global scalar* trust radius and (b) a *fixed* rollout horizon. ASTRO replaces both with a *data-driven* local support geometry and a *state-dependent* horizon.

2.2 Conservative and model-free offline RL

Model-free offline RL has converged on a small set of *implicit constraint* ideas Fujimoto et al. (2019): **CQL** (Kumar et al., 2020) penalizes Q-values on OOD actions via a log-sum-exp term, **IQL** (Kostrikov et al., 2021) avoids querying OOD actions via expectile regression, and **EDAC** (An et al., 2021) diversifies Q-ensembles to reduce over-estimation. We retain CQL as the model-free reference point in Table 1, as its action-space penalty is the closest analogue of model-based pessimism; a broader comparison against strong model-free methods (IQL, EDAC) is left to future work. Where CQL and ASTRO both penalize the policy on OOD inputs, the mechanisms differ fundamentally: CQL enforces conservatism directly in the actor-critic objective over the action space, whereas ASTRO regularizes the transition model and the rollout generation process through a *state-space* support tube in the latent space.

2.3 Data-aware and adaptive offline RL

A growing line of work argues that *adaptive* conservatism is the right design target Fang et al. (2026). **MOREC** (Luo et al., 2024) filters model rollouts by a learned realism score applied to whole transitions, whereas ASTRO gates per-step. **ADMPO-OFF** (Lin et al., 2025) attacks compounding rollout error with multi-horizon predictions; **BARC** (our in-house variant) extends this with branch self-consistency and a CVaR-style critic backup. Uncertainty-gated rollout control also has precedents: STEVE (Buckman et al., 2018) interpolates between horizon lengths by ensemble disagreement, M2AC (Pan et al., 2020) masks model-generated transitions by ensemble variance, NUNO (Koprulu et al., 2025) truncates rollouts using distance-aware uncertainty, and SUMO (Qiao et al., 2025) gates trajectory generation by search-based dataset proximity. ASTRO differs in that the gating signal is derived from the local anisotropic support geometry rather than a scalar uncertainty threshold, and the same geometry also shapes the model-learning loss. Robust MDPs (Iyengar, 2005) and

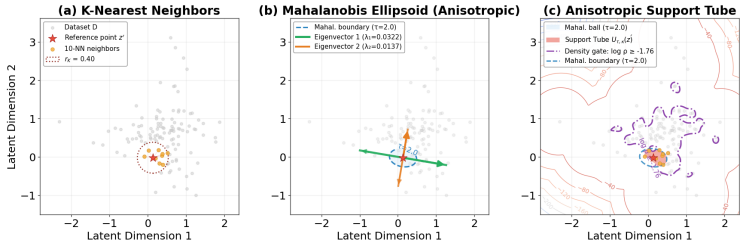

Figure 1: Construction of the anisotropic support tube in a 2-D latent slice. **(a)** K-nearest neighbors (blue) around a reference point z' (star) define the local geometry. **(b)** The Mahalanobis ellipsoid (solid) stretches along the principal data direction (eigenvector 1, $\lambda_1 = 0.032$) and contracts along the orthogonal extrapolation direction (eigenvector 2, $\lambda_2 = 0.014$), in contrast to an isotropic Euclidean ball (dashed). **(c)** The final support tube $\mathcal{U}_{\tau,\kappa}(z')$ (shaded) is the intersection of the Mahalanobis ellipsoid ($\tau = 2.0$) with a density gate ($\log \rho \geq -1.76$), preventing extension into uncovered regions.

distributionally robust RL (Xu & Mannor, 2010) provide a probabilistic foundation for ASTRO's robust value-aware loss, with Wasserstein-based robust sets (Esfahani & Kuhn, 2018) being a popular choice. ASTRO's support tube can be viewed as a *data-driven* construction of such a robust set: instead of assuming a fixed radius ξ, it estimates the geometry of the support directly from the offline data. ROMI's bi-level optimization (Qiao et al., 2026) is the closest relative, but ASTRO replaces the global Wasserstein ball with the local support tube and additionally introduces a state-dependent rollout horizon.

3 Method

3.1 Preliminaries

We consider an MDP $(\mathcal{S}, \mathcal{A}, P^*, r, \gamma)$ with unknown transition kernel P^* and a static offline dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$. The goal is to learn a policy $\pi_\theta(a \mid s)$ maximizing expected discounted return without further interaction. We learn a latent encoder $e_\phi : \mathcal{S} \to \mathbb{R}^{d_z}$ and an ensemble of M Gaussian dynamics heads $\{p_{\psi_m}(z_{t+1} \mid z_t, a_t)\}_{m=1}^M$ in the latent space $z = e_\phi(s)$. A SAC actor-critic backbone (Haarnoja et al., 2018) with critics $Q_\theta(s, a)$ and target value $V_{\bar{\theta}}(s)$ completes the architecture.

3.2 Anisotropic Support Tube

The central geometric primitive of ASTRO is the *anisotropic support tube*, illustrated in Figure 1. The construction proceeds in three steps: local covariance estimation, density estimation, and their intersection into a directionally adaptive trust region. For each encoded next-state $z' = e_\phi(s')$, we retrieve its K-nearest neighbors $\mathcal{N}_K(z') = \{\tilde{z}'_{(1)}, \ldots, \tilde{z}'_{(K)}\}$ from the latent index of all encoded next-states in $\mathcal{D}$ (Figure 1a).

Local covariance. The neighbor cloud defines a local covariance that captures directional support:

$$\Sigma(z') = \frac{1}{K} \sum_{k=1}^{K} (\tilde{z}'_{(k)} - z')(\tilde{z}'_{(k)} - z')^\top + \varepsilon I, \tag{1}$$

with Tikhonov regularization $\varepsilon > 0$ ensuring invertibility. As visualized in Figure 1b, the eigenvectors of $\Sigma(z')$ reveal the principal data directions: high-variance eigenvectors correspond to well-supported interpolation directions, while low-variance eigenvectors indicate extrapolation axes. Since $K < d_z$, the empirical covariance has rank at most K; in practice all determinants and Mahalanobis terms are evaluated in the log domain on the principal subspace spanned by the neighbors, with the remaining directions assigned variance ε, which keeps condition numbers bounded (Section D).

Local density. To prevent the tube from extending into genuinely uncovered regions, we estimate a local density:

$$\rho(z') = \frac{K}{c_{d_z} \cdot r_K^{d_z} \cdot \det(\Sigma(z'))^{1/2}}, \tag{2}$$

where r_K is the Mahalanobis distance to the K-th neighbor Qiao et al. (2025); Li et al. (2025) and $c_{d_z} = \pi^{d_z/2}/\Gamma(d_z/2+1)$ is the unit-ball volume. Since only log-density *differences* enter Equation (3), ρ is used as an unnormalized density score and the constant dataset-size factor is omitted.

Support tube. The anisotropic support tube is the intersection of the Mahalanobis ellipsoid and the density-gated region (Figure 1c):

$$\mathcal{U}_{\tau,\kappa}(z') = \big\{\tilde{z} \in \mathbb{R}^{d_z} : \|\tilde{z} - z'\|^2_{\Sigma(z')^{-1}} \le \tau^2, \log \rho(\tilde{z}) \ge \log \rho(z') - \kappa\big\}. \tag{3}$$

This construction is *anisotropic*: $\Sigma(z')$ stretches the tube along high-variance (well-supported) directions and contracts it along low-variance (extrapolation) directions. The density gate prevents the tube from extending into uncovered regions regardless of τ. Two limiting cases recover prior constructions: (i) $\Sigma = I$, $\kappa = \infty$ yields a Euclidean ball (RAMBO); (ii) Σ fixed, $\kappa = \infty$ yields a Mahalanobis ball equivalent to ROMI's Wasserstein set with $\xi = \tau$. ASTRO generalizes both by making Σ *local* and κ *finite*, so that the trust region adapts to the data geometry at every query point rather than imposing a single global shape.

Theoretical Motivation. We now state a theorem that formalizes the benefit of replacing a global scalar trust radius with the anisotropic support tube.

Theorem 3.1 (Generalization Gap Bound under Anisotropic Support)**.** *Let $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$ be an offline dataset drawn i.i.d. from the behavior distribution of an MDP with unknown transition kernel P^*. Let $e_\phi : \mathcal{S} \to \mathbb{R}^{d_z}$ be a Lipschitz-continuous encoder with constant L_e, mapping into a bounded latent domain of diameter D_z, and let $\hat{P}_\psi(\cdot \mid z, a)$ be the learned latent dynamics ensemble. Assume the ensemble is* well specified*: the true conditional mean $\mu^*(z,a) = \mathbb{E}_{p^*}[z' \mid z, a]$ lies in the hypothesis class and NLL training recovers it up to an optimization error ε_{opt} that vanishes as $N \to \infty$. Define the* global scalar trust region *as the Euclidean ball $\mathcal{B}_\xi(z') = \{\tilde{z} : \|\tilde{z} - z'\|_2 \le \xi\}$ with fixed radius $\xi > 0$, and the* anisotropic support tube *as in Equation* (3)*.*

Assume further that the true latent transition density $p^*(\cdot \mid z, a)$ satisfies a local covariance shift *condition: there exists a positive definite matrix $\Sigma^*(z')$ such that for all $\tilde{z}$ in the high-density region,*

$$p^*(\tilde{z} \mid z, a) \propto \exp\Big(-\tfrac{1}{2}\|\tilde{z} - z'\|^2_{\Sigma^*(z')^{-1}}\Big) \cdot g(\tilde{z}), \tag{4}$$

where g is a slowly varying function. Both value errors below are measured with respect to the conditional-mean dynamics; the irreducible aleatoric variance $\mathrm{Tr}(\Sigma^*(z,a))$ enters both sides identically and is omitted, since our goal is the comparison between trust regions.

Let $\hat{V}_\xi$ and $\hat{V}_{\text{ASTRO}}$ denote the value functions produced by a pessimistic MBRL method using $\mathcal{B}_\xi$ and $\mathcal{U}_{\tau,\kappa}$, respectively, with the same robust value-aware loss (Equation (8)*). Because the robust loss minimizes against the worst case over the trust region, both estimates incur an interior pessimism bias that grows with the region size ($\Omega(\tau\sqrt{\lambda_{\max}})$ for the tube and $\Omega(\xi)$ for the ball); the bounds below are therefore informative only for moderate radii τ, ξ matched to the local eigenvalue scale, and are vacuous as $\tau, \xi \to \infty$. In this regime, for any state z and action a, with probability at least $1 - \delta$ over the dataset $\mathcal{D}$,*

$$\Big|\hat{V}_{\text{ASTRO}}(z,a) - V^*(z,a)\Big| \le \underbrace{\frac{C_1}{\sqrt{N \cdot \lambda_{\min}(\Sigma(z'))}}}_{\text{(A) on-support term}} + \underbrace{C_2 \cdot \exp\big(-C_3 \cdot \tau^2\big)}_{\text{(B) off-support tail}} \tag{5}$$

whereas the global-scalar baseline satisfies

$$\Big|\hat{V}_\xi(z,a) - V^*(z,a)\Big| \le \underbrace{\frac{C_1}{\sqrt{N} \cdot \xi}}_{\text{(A') isotropic term}} + \underbrace{C_2 \cdot \exp\big(-C_3 \cdot \xi^2\big)}_{\text{(B') off-support tail}} + \underbrace{C_4 \cdot \max_{i:\lambda_i \gg \xi^2} \sqrt{\lambda_i}}_{\text{(C) anisotropy gap}}. \tag{6}$$

Here $\{\lambda_i\}_{i=1}^{d_z}$ are the eigenvalues of $\Sigma(z')$ sorted in descending order, $C_1, C_2, C_3, C_4 > 0$ are constants independent of $\Sigma(z')$ and ξ, and V^* is the true optimal value function.

Term (A) reflects estimation uncertainty along the least-covered direction: the effective sample size in direction v scales with $N \cdot v^\top \Sigma(z')v$, so local methods pay for poorly covered directions through $\lambda_{\min}$; the tube *contracts* along these directions, limiting their influence on the robust loss. The scalar method additionally pays **(C)** for *not knowing which directions are safe*: any single ξ either over-covers well-supported directions or under-covers them, and Lemma F.2 shows this gap is strictly positive whenever the local spectrum is non-uniform. ASTRO's tube removes this term—up to estimation error of Σ absorbed into (B)—because the boundary $\pm\tau\sqrt{\lambda_i}$ matches the scale of p^* in every direction. All proofs are in Appendix F.

3.3 Robust Value-Aware Model Loss

Standard latent NLL training constrains the model only on-support Abdullah et al. (2019); Clement & Kroer (2021). To regularize off-support behavior, we introduce the robust value-aware (RVL) loss. For each transition, we draw N_s samples $\{\tilde{z}_i\}_{i=1}^{N_s} \sim \mathcal{U}_{\tau,\kappa}(z')$ uniformly from the tube—obtained by rejection sampling from the Mahalanobis ellipsoid and discarding candidates that violate the density gate, with $N_s = 16$ (Section D)—and compute the *robust tube value*:

$$v^-(z') = \min_{i=1,\dots,N_s} V_{\bar{\theta}}\big(d_\eta(\tilde{z}_i)\big), \tag{7}$$

where d_η is a lightweight decoder used solely for value evaluation (Yu et al., 2023). The RVL loss penalizes the gap between the model's predicted next-state value and this worst case:

$$\mathcal{L}_{\mathrm{RVL}}(\psi,\phi) = \big(\mathbb{E}_{\hat{z}\sim p_\psi(\cdot|z,a)} V_{\bar{\theta}}(d_\eta(\hat{z})) - v^-(z')\big)^2. \tag{8}$$

Crucially, the tube geometry induces *spatially varying* pessimism: in well-supported regions the tube is narrow, $v^- \approx \hat{V}$, and the loss is gentle; in poorly-supported regions the tube widens, v^- drops, and the loss aggressively pushes predictions toward the worst case—all without manual per-task tuning.

Auxiliary losses. The transport alignment loss encourages the predicted displacement to align with the empirical displacement toward the neighbor centroid:

$$\mathcal{L}_{\mathrm{transport}} = \big\|(\hat{z} - z) - \bar{\delta}_K(z)\big\|^2, \quad \bar{\delta}_K(z) = \frac{1}{K}\sum_{k=1}^{K} \tilde{z}'_{(k)} \; - \; z. \tag{9}$$

Here $\bar{\delta}_K(z)$ is measured from the *current* latent state z rather than from z': the neighbor residuals around z' approximately cancel for a symmetric neighborhood, whereas the displacement from z to the neighbor centroid represents the empirical transition direction. The calibration loss aligns the model's predicted one-step error scale with the empirical held-out error quantile. Concretely, let $\hat{\tau}(z,a)$ denote the root-mean-square predictive standard deviation of the ensemble, averaged over latent dimensions, and let $q_{\mathrm{emp}}(z')$ be the empirical one-step error quantile estimated on a held-out split within the retrieved neighborhood $\mathcal{N}_K(z')$; then

$$\mathcal{L}_{\mathrm{cal}} = \big(\hat{\tau}(z,a) - q_{\mathrm{emp}}(z')\big)^2, \tag{10}$$

whose gradients flow into (ϕ,ψ) through $\hat{\tau}$, ensuring that the trust score (Section 3.4) is well-calibrated. The full model objective is:

$$\mathcal{L}_{\mathrm{model}} = \mathcal{L}_{\mathrm{NLL}} + \alpha\,\mathcal{L}_{\mathrm{RVL}} + \beta\,\mathcal{L}_{\mathrm{transport}} + \gamma\,\mathcal{L}_{\mathrm{cal}}, \tag{11}$$

with defaults $\alpha = 1.0$, $\beta = 0.1$, $\gamma = 0.05$.

3.4 Trust Score and Adaptive Horizon

A fixed rollout horizon H is suboptimal: long horizons compound model error in OOD regions, while short horizons discard information in trusted regions Kubo et al. (2025). ASTRO resolves this by gating the horizon with a trust score computed from the same support geometry.

Trust score. At each rollout step, we compute an uncertainty (distrust) score:

$$u(z,a) = \lambda_1 \cdot \underbrace{\|\hat{z} - z^\dagger\|^2_{\Sigma(z^\dagger)^{-1}}}_{\text{Mahalanobis error}} + \lambda_2 \cdot \underbrace{\frac{1}{M}\sum_{m=1}^{M} \|\mu_m - \bar{\mu}\|^2}_{\text{ensemble disagreement}} + \lambda_3 \cdot \underbrace{|\hat{\tau}(z,a) - q_{\mathrm{emp}}(z^\dagger)|}_{\text{calibration gap}}, \tag{12}$$

Task	CQL	MOPO	COMBO	MOBILE	RAMBO	BARC	ROMI	ASTRO
door-C	−13.3±2.6	−14.5±0.2	−9.9±10.5	−11.3±5.7	−15.3±2.1	−1.3±3.8	−14.9±1.1	**3.8±8.5**
door-E	**452.3±17.7**	−9.5±2.5	−3.9±0.6	22.1±21.0	8.0±2.2	−12.1±6.9	−0.8±2.3	117.2±4.0
door-H	−18.0±5.3	−26.9±6.3	−35.1±1.7	−37.0±1.3	−30.9±0.1	−32.3±0.8	−20.2±4.3	**−13.4±2.2**
hammer-C	27.7±5.8	−8.9±4.3	14.5±6.4	21.3±0.8	26.1±1.0	27.2±0.2	8.6±2.6	**53.1±3.9**
hammer-E	**6724.0±373.4**	14.4±7.3	33.4±14.3	24.5±1.5	21.2±1.9	24.8±0.6	17.8±4.6	1020.3±252.5
hammer-H	73.5±47.1	−13.2±10.2	8.0±8.0	18.6±3.8	6.1±10.3	13.0±8.3	−3.6±10.3	**115.0±35.6**
pen-C	−43.0±236.6	156.0±75.8	**3951.3±15.2**	−269.7±4.7	−197.7±87.2	180.0±229.0	−20.8±81.5	3063.5±193.7
pen-E	**10980.0±481.3**	1429.0±475.1	6108.1±5.0	4785.0±100.7	4742.7±2176.8	4647.9±213.8	3085.9±1326.0	6517.8±144.7
pen-H	−260.4±59.1	−258.2±14.4	348.1±373.7	99.2±162.5	67.4±34.0	125.1±74.3	−95.4±24.2	**3688.9±624.7**
relocate-C	−6.9±0.2	−5.0±3.1	−19.2±9.9	−8.2±0.8	−5.6±2.9	−7.5±0.2	−5.3±3.0	**−0.8±0.1**
relocate-E	14.0±3.0	−34.6±14.3	−11.4±0.8	**18.0±10.1**	−7.3±2.6	−25.2±3.9	−21.0±8.4	5.3±9.2
relocate-H	**−6.4±0.2**	−14.8±0.8	−11.9±4.5	−8.3±6.0	−19.4±8.1	−6.9±7.0	−17.1±4.4	−6.5±3.1
ant-ME	**6483.2±157.3**	−103.4±134.7	2394.5±679.4	5894.5±260.7	483.0±57.0	579.5±225.3	−2125.2±61.6	822.1±16.3
ant-MR	**9922.6±252.9**	374.3±155.8	1494.1±0.9	2159.7±953.0	906.6±106.1	1521.1±47.3	−587.2±396.1	2023.1±154.8
ant-R	2855.3±90.8	−168.6±61.5	−553.4±148.5	**2930.8±39.7**	−1083.6±318.5	−258.7±108.1	−321.0±277.1	1295.1±147.6
halfcheetah-ME	2053.2±4.5	4552.2±3775.9	8754.9±44.1	8907.9±6.8	9016.3±32.7	8695.2±95.0	6784.3±1904.3	**9115.4±51.5**
halfcheetah-MR	5332.6±33.7	4265.5±1478.5	5019.7±320.1	2779.8±2570.3	5568.6±69.8	4954.6±450.3	4917.1±774.2	**5610.0±3.9**
halfcheetah-R	175.4±1.6	**2139.7±40.0**	1915.8±49.7	2028.9±56.6	2002.6±26.8	2019.9±14.4	2071.2±33.4	2128.8±5.9
hopper-ME	**9318.7±328.2**	442.8±139.1	420.0±40.6	3477.0±627.1	1003.8±344.0	750.9±322.8	86.9±7.2	3843.9±2219.1
hopper-MR	1873.2±376.4	5340.4±1097.8	3195.0±559.6	9769.8±106.2	4235.1±66.0	8665.4±867.0	171.6±44.3	**10098.8±17.9**
hopper-R	195.0±25.1	689.1±13.8	785.4±79.1	765.7±10.0	820.9±12.9	1264.5±362.7	117.7±25.6	**1489.2±21.9**
walker2d-ME	4255.4±361.6	6068.8±492.6	8160.2±131.0	10217.0±321.8	5698.7±875.7	9693.6±288.1	8.7±30.9	**10244.9±313.5**
walker2d-MR	6761.1±229.0	6780.8±178.8	5826.2±479.9	3368.1±1164.7	2874.9±146.3	6046.8±21.7	−10.0±0.1	**7948.8±261.2**
walker2d-R	35.4±5.5	**592.6±46.6**	0.5±1.3	559.4±8.1	1.6±3.5	43.9±7.9	−10.5±2.5	358.4±37.3

Table 1: Per-dataset results on Adroit (top, shaded pink) and MuJoCo (bottom, shaded green) benchmarks. Values are mean ± std over 5 seeds. Bold: best per row. ASTRO achieves the best result on 12 of 24 tasks.

where μ_m is the mean prediction of ensemble member m, $\bar{\mu}$ is the ensemble mean, $\hat{\tau}(z, a)$ is the predicted error scale of Equation (10), and q_{emp} is the local empirical error quantile. The first term is evaluated against the *latent index* rather than a ground-truth next state: during imaginary rollouts no true next state exists, so we query the k-NN index with the predicted $\hat{z}$ and use the retrieved nearest neighbor $z^\dagger = \text{NN}(\hat{z}; \mathcal{D})$ as the anchor, together with its local covariance $\Sigma(z^\dagger)$ (Algorithm 1, Step B). Small values of u indicate trustworthy predictions; large values indicate extrapolation. Default weights: $\lambda_1 = 0.4$, $\lambda_2 = 0.4$, $\lambda_3 = 0.2$. The three terms capture complementary failure signals: geometric OOD distance, predictive uncertainty, and model miscalibration.

Adaptive horizon. The per-state rollout length is:

$$\hat{H}(u) = \begin{cases} 5 & \text{if } u \le \theta_1, \\ 2 & \text{if } \theta_1 < u \le \theta_2, \\ 1 & \text{if } u > \theta_2, \end{cases} \tag{13}$$

with $\theta_1 = 1.0$, $\theta_2 = 2.0$. Since u is an uncertainty score, low values (small Mahalanobis error, low disagreement) yield long horizons; high values truncate the rollout immediately. This design ensures that computation is allocated where the model is trustworthy Pan et al. (2020).

3.5 Policy Optimization

The actor-critic backbone follows SAC (Haarnoja et al., 2018) with twin Q-networks and entropy regularization ($\alpha_{\text{SAC}} = 0.2$). Each gradient step mixes real transitions from $\mathcal{B}_{\text{real}} = \mathcal{D}$ with model rollouts from $\mathcal{B}_{\text{model}}$ at a 1:1 ratio. Model rollouts are generated by rolling out $\hat{H}(u)$ steps from states sampled in $\mathcal{D}$, using π_θ and the ensemble dynamics; each dynamics head additionally predicts the immediate reward by regression on the logged rewards, and predicted latent states are decoded through d_η so that the policy and critics always operate on raw states. Target networks are updated via Polyak averaging ($\tau_{\text{Polyak}} = 0.005$). The complete procedure is given in Appendix A.

4 Experiments

In this section, we conduct several experiments to address following questions: **Q1**: How does ASTRO compare to state-of-the-art baselines? **Q2**: Which components contribute most to performance? **Q3**: Does the adaptive horizon produce measurable gains on support-fragmented datasets? **Q4**: Does the RVL loss provide stable pessimism without the instability of adversarial updates?

4.1 SETUP

Benchmarks. We evaluate on 24 D4RL datasets Fu et al. (2020): 12 Adroit Rajeswaran et al. (2017) (door, hammer, pen, relocate $\times$ {cloned, expert, human}) and 12 MuJoCo Todorov et al. (2012) (halfcheetah, hopper, walker2d, ant $\times$ {medium-expert, medium-replay, random}). Adroit tasks feature high-dimensional dexterous manipulation with sparse rewards and highly fragmented support; MuJoCo tasks provide locomotion benchmarks with varying data quality (see Appendix B for details).

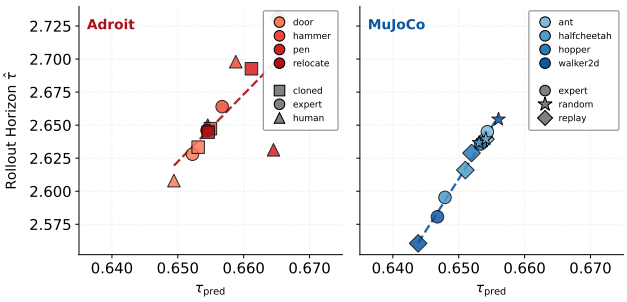

Figure 2: Mean predicted horizon $\hat{H}$ versus trust score u across tasks (last 50 epochs, 5 seeds). Well-supported MuJoCo tasks cluster at low u; support-fragmented Adroit tasks exhibit high u—confirming that the adaptive horizon allocates computation where the model is trustworthy.

Baselines. We compare against seven methods: CQL, MOPO, COMBO, MOBILE, RAMBO, BARC, and ROMI. All are re-implemented in a unified codebase with the same SAC + ensemble latent dynamics backbone to control for implementation variance (Appendix C).

Architecture and training. Actor, critic, and value networks are 3-layer MLPs (256 units) with LayerNorm. The dynamics ensemble has 7 members (4 hidden layers, 200 units). The encoder maps to $d_z = 64$. Learning rates: 3×10^{-4} (encoder, actor, critic), 1×10^{-3} (dynamics). Training runs for 300 epochs of 1000 environment steps each. Every epoch, the policy is evaluated for 10 episodes in the true environment; we report the mean of the last 10 evaluations. All methods use 5 random seeds per dataset (see more details in Appendix D).

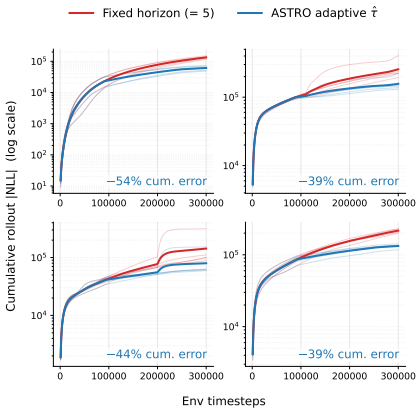

Figure 3: Cumulative rollout |NLL| (log scale) over training for fixed horizon ($H = 5$) versus ASTRO's adaptive horizon on four tasks. ASTRO reduces cumulative prediction error by 39–54% by truncating rollouts in OOD regions while preserving long horizons where the model remains accurate.

4.2 MAIN RESULTS (Q1)

Table 1 reports per-dataset results. ASTRO attains the highest overall and MuJoCo means among all methods, exceeding the next-best model-based method by 21%. Because raw returns are reported, these aggregates are dominated by high-magnitude tasks; per-task comparisons provide the scale-free summary. CQL is strongest on Adroit overall, but ASTRO achieves the best result on 6 of 12 Adroit and 6 of 12 MuJoCo datasets. Gains are largest exactly where predicted: on support-fragmented datasets (e.g., hammer-cloned, pen-human, walker2d-medium-replay), ASTRO substantially outperforms the next-best method, in some cases by nearly an order of magnitude. ROMI, the closest design, performs markedly worse, with catastrophic failures on ant and walker2d; it is competitive only on the near-isotropic halfcheetah suite, where a global Wasserstein ball with fixed ξ is a reasonable approximation. On fragmented support, a single ξ cannot avoid both over- and under-conservatism, and the bi-level optimization becomes unstable; ASTRO's local tube adapts the trust region to each query point.

4.3 ABLATION STUDY (Q2)

Task	Full ASTRO	w/o anisotropic	w/o density gate	w/o RVL loss	w/o adaptive horizon	w/o trust score
door-C	**3.8**$\pm$**8.5**	$-12.5_{\pm 4.6}$	$-2.0_{\pm 2.3}$	$-10.0_{\pm 3.9}$	$-8.0_{\pm 7.1}$	$0.5_{\pm 1.1}$
hammer-C	**53.1**$\pm$**3.9**	$27.0_{\pm 9.4}$	$41.0_{\pm 7.7}$	$18.5_{\pm 3.6}$	$22.0_{\pm 5.0}$	$38.0_{\pm 11.9}$
hopper-MR	**10098.8**$\pm$**17.9**	$9123.6_{\pm 379.4}$	$9811.0_{\pm 137.3}$	$7431.8_{\pm 177.9}$	$8499.9_{\pm 298.5}$	$9220.4_{\pm 225.6}$
walker-MR	**7948.8**$\pm$**261.2**	$7211.3_{\pm 98.1}$	$7789.6_{\pm 477.3}$	$6514.8_{\pm 312.2}$	$6800.3_{\pm 232.6}$	$7399.7_{\pm 374.0}$

Table 2: Ablation results (mean final reward, 5 seeds). Removing the RVL loss and the anisotropic covariance degrades performance on almost all tasks.

To isolate the contribution of each component, we remove one at a time on four representative tasks (Table 2) and observe three findings. First, the *RVL loss is the single most critical component*: removing it causes the largest drop on three of the four tasks (-65% on hammer-C, -26% on hopper-MR, -18% on walker-MR), as the model overfits to OOD regions and reverts to MOPO-style failure. Second, the *anisotropic covariance is essential*: replacing $\Sigma(z')$ with I (degenerating to a Euclidean ball) degrades every task—most severely on the fragmented Adroit splits (door-C drops below zero; -49% on hammer-C)—confirming that directional support information is indispensable. Third, the *adaptive horizon and trust score* contribute the remaining gap, with the largest effect on support-fragmented Adroit tasks. The density gate provides a modest but consistent improvement, validating its role in preventing tube leakage.

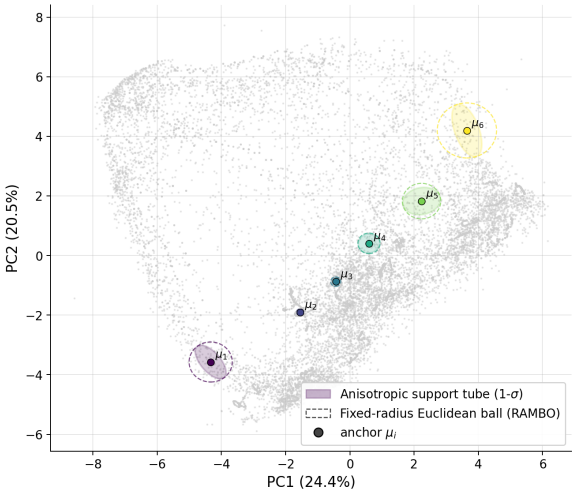

Figure 4: 2-D PCA projection of pen-cloned-v1 with ASTRO support tubes (colored ellipses) and fixed-radius Euclidean balls (dashed circles) around 6 anchor points Abdi & Williams (2010). The ellipses adapt to local data geometry—stretching along principal data directions and contracting perpendicularly—whereas the circles either over-cover in low-density directions (admitting OOD states) or under-cover in high-density directions (blocking valid interpolations).

4.4 Adaptive Horizon Analysis (Q3)

Figure 2 confirms the design intent: predicted horizons are long on well-supported MuJoCo tasks and short on support-fragmented Adroit tasks, without task-specific tuning. The plotted scores are epoch means; since the horizon rule acts on instantaneous scores, the reported means below the maximum reflect truncation on the minority of high-uncertainty steps.

Figure 3 quantifies the downstream effect: compared to a fixed horizon $H = 5$, ASTRO reduces cumulative rollout prediction error by 39–54% on four representative tasks. In OOD regions the trust terms grow rapidly, triggering early truncation before errors compound; in trusted regions rollouts run to full length, preserving multi-step targets that stabilize critic learning.

Training exhibits a desirable bootstrapping behavior: early on u is high, so $\hat{H} = 1$; as the model improves, u decays and the horizon lengthens to $\hat{H} = 5$ in trusted regions. This emergent curriculum follows directly from sharing the support geometry between model learning and rollout planning.

4.5 Robust Value Dynamics (Q4)

Figure 5 tracks the pessimism gap $\hat{V} - v^-$ across eight task families. The robust tube value v^- remains below $\hat{V}$ for the vast majority of training, with rare transient violations due to finite-sample

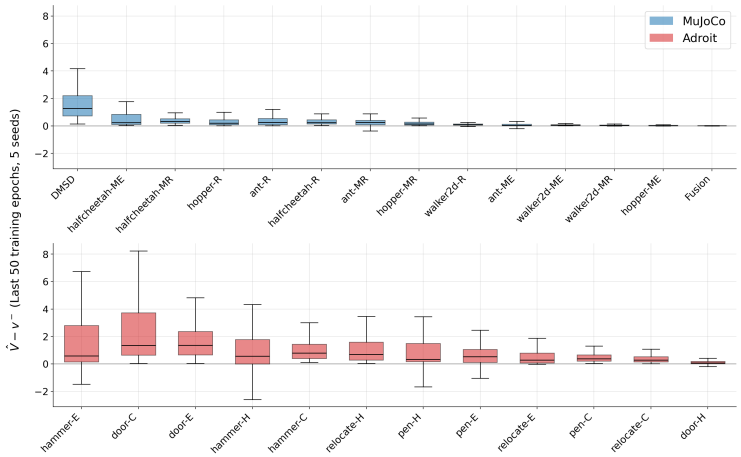

Figure 5: Per-task pessimism gap $\hat{V} - v^-$ over training. Gaps are larger and more variable on support-fragmented Adroit tasks (hammer, pen) and smaller and stable on well-supported MuJoCo tasks (halfcheetah-ME, walker2d-ME), showing spatially varying pessimism without manual tuning.

minima and target-network lag. The gap stays bounded—neither collapsing to zero nor diverging—unlike RAMBO's adversarial gap, which is governed by ξ and can diverge if ξ is slightly too large. The per-task variation is informative: on support-fragmented Adroit tasks (hammer, pen), the gap is larger and more variable; on well-supported MuJoCo tasks (halfcheetah-ME, walker2d-ME), it is small and tightly concentrated. This spatially varying pessimism—large where data is scarce, small where abundant—is precisely what a global scalar penalty cannot produce, and it correlates with the ablation results: tasks with larger gaps (hammer-C, pen-H) are also where removing RVL causes the steepest drop (Table 2).

4.6 Support Tube Geometry

Figure 4 visualizes the support tube on pen-cloned, projected onto the top two PCA components. The local Mahalanobis ellipses stretch along principal data directions and contract perpendicularly—precisely the information a scalar trust radius discards. Fixed-radius Euclidean balls either over-cover in low-density directions (admitting OOD states) or under-cover in high-density directions (blocking valid interpolations).

Two observations merit emphasis. First, the degree of anisotropy varies across anchor points: tubes near dense clusters are mildly elongated, while tubes near manifold edges are strongly eccentric (aspect ratios exceeding 3:1)—variation invisible to any global-radius method. Second, the density gate visibly truncates several tubes on their low-density side, preventing extension into empty regions between clusters—a failure mode that the pure Mahalanobis ellipsoid ($\kappa = \infty$) would exhibit. This corroborates the ablation: the anisotropic geometry is a structural necessity for correctly adjudicating trust in non-uniform offline data.

5 Conclusion

We presented ASTRO, which replaces the scalar-trust abstraction in offline MBRL with a data-adaptive anisotropic support tube. Reusing this single geometric primitive across a robust value-aware model loss, a trust score, and a state-dependent rollout horizon yields spatially varying pessimism and rollout length without task-specific tuning. ASTRO matches or surpasses strong model-based baselines on D4RL datasets, with the largest gains on support-fragmented data. A remaining limitation is that the tube's fidelity is bounded by encoder quality, and the k-NN index imposes $O(N)$ storage and per-epoch rebuild cost—negligible on D4RL, though it would need approximation at larger scale. Addressing this limitation—for instance, by improving the encoder's preservation of data geometry or by adopting approximate nearest-neighbor indexing at scale—is left to future work.

AI USE STATEMENT

In preparing this work, we utilized a Large Language Model (LLM) solely for limited, auxiliary purposes. Specifically, the LLM was employed to assist with refining the English phrasing of certain paragraphs, correcting grammatical errors, and, in the Related Work section, for generating ideas and searching for potentially relevant literature. The model did not contribute to the core research ideation, mathematical formulation, theoretical analysis, algorithm development, experimental design, or interpretation of results. All authors take full responsibility for the entire content of this work, including any text generated with the aid of the LLM.

REFERENCES

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Mohammed Amin Abdullah, Hang Ren, Haitham Bou Ammar, Vladimir Milenkovic, Rui Luo, Mingtian Zhang, and Jun Wang. Wasserstein robust reinforcement learning. *arXiv preprint arXiv:1907.13196*, 2019.

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. *Advances in neural information processing systems*, 34:7436–7447, 2021.

Arthur Argenson and Gabriel Dulac-Arnold. Model-based offline planning. *arXiv preprint arXiv:2008.05556*, 2020.

Jacob Buckman, Danijar Hafner, George Tucker, Eugene Brevdo, and Honglak Lee. Sample-efficient reinforcement learning with stochastic ensemble value expansion. *Advances in neural information processing systems*, 31, 2018.

Julien Grand Clement and Christian Kroer. First-order methods for wasserstein distributionally robust mdp. In *International conference on machine learning*, pp. 2010–2019. PMLR, 2021.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Zeyu Fang, Zuyuan Zhang, Mahdi Imani, and Tian Lan. Manifold-constrained energy-based transition models for offline reinforcement learning. *arXiv preprint arXiv:2602.02900*, 2026.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pp. 2052–2062. PMLR, 2019.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Haoyang He. A survey on offline model-based reinforcement learning. *arXiv preprint arXiv:2305.03360*, 2023.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33: 21810–21823, 2020.

Cevahir Koprulu, Franck Djeumou, et al. Neural stochastic differential equations for uncertainty-aware offline rl. In *International Conference on Learning Representations*, volume 2025, pp. 4889–4913, 2025.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Akihiro Kubo, Paavo Parmas, and Shin Ishii. Double horizon model-based policy optimization. *arXiv preprint arXiv:2512.15439*, 2025.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Xue Li, Bangjun Wang, and Xinghong Ling. Conservative reward enhancement through the nearest neighbor integration in model-based offline policy optimization. *Expert Systems with Applications*, 274:126888, 2025.

Haoxin Lin, Yu-Yan Xu, Yihao Sun, Zhilong Zhang, Yi-Chen Li, Chengxing Jia, Junyin Ye, Jiaji Zhang, and Yang Yu. Any-step dynamics model improves future predictions for online and offline reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 70746–70769, 2025.

Fan-Ming Luo, Tian Xu, Xingchen Cao, and Yang Yu. Reward-consistent dynamics models are strongly generalizable for offline reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 8483–8515, 2024.

Tatsuya Matsushima, Hiroki Furuta, Yutaka Matsuo, Ofir Nachum, and Shixiang Gu. Deployment-efficient reinforcement learning via model-based offline optimization. *arXiv preprint arXiv:2006.03647*, 2020.

Feiyang Pan, Jia He, Dandan Tu, and Qing He. Trust the model when it is confident: Masked model-based actor-critic. *Advances in neural information processing systems*, 33:10537–10546, 2020.

Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

Rafael Figueiredo Prudencio, Marcos ROA Maximo, and Esther Luna Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE transactions on neural networks and learning systems*, 35(8):10237–10257, 2023.

Zhongjian Qiao, Jiafei Lyu, Kechen Jiao, Qi Liu, and Xiu Li. Sumo: Search-based uncertainty estimation for model-based offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20033–20041, 2025.

Zhongjian Qiao, Jiafei Lyu, Boxiang Lyu, Yao Shu, Siyang Gao, and Shuang Qiu. Model-based offline rl via robust value-aware model learning with implicitly differentiable adaptive weighting. In *International Conference on Learning Representations (ICLR 2026)*, 2026.

Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017.

Marc Rigter, Bruno Lacerda, and Nick Hawes. Rambo-rl: Robust adversarial model-based offline reinforcement learning. *Advances in neural information processing systems*, 35:16082–16097, 2022.

Yihao Sun, Jiaji Zhang, Chengxing Jia, Haoxin Lin, Junyin Ye, and Yang Yu. Model-bellman inconsistency for model-based offline reinforcement learning. In *International Conference on Machine Learning*, pp. 33177–33194. PMLR, 2023.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Huan Xu and Shie Mannor. Distributionally robust markov decision processes. *Advances in Neural Information Processing Systems*, 23, 2010.

Insoon Yang. A convex optimization approach to distributionally robust markov decision processes with wasserstein distance. *IEEE control systems letters*, 1(1):164–169, 2017.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in neural information processing systems*, 33:14129–14142, 2020.

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. *Advances in neural information processing systems*, 34:28954–28967, 2021.

Zhuodong Yu, Ling Dai, Shaohang Xu, Siyang Gao, and Chin Pang Ho. Fast bellman updates for wasserstein distributionally robust mdps. *Advances in Neural Information Processing Systems*, 36: 30554–30578, 2023.

A Full Algorithm

We show the complete pseudocode of ASTRO in Algorithm 1.

B Environments and Datasets

We evaluate on 24 datasets from the D4RL benchmark (Fu et al., 2020), spanning two platform families:

Adroit (12 tasks). Dexterous manipulation with a 24-DoF hand (Rajeswaran et al., 2017): `door` (open a door), `hammer` (drive a nail), `pen` (rotate a pen to a target orientation), and `relocate` (move a ball to a target). Each task provides three data variants: *cloned* (C, $\sim$5k demonstrations with noisy copies), *expert* (E, $\sim$5k high-quality demonstrations), and *human* (H, $\sim$25 suboptimal teleoperated trajectories). Adroit datasets are characterized by high dimensionality ($|\mathcal{S}| \leq 45$, $|\mathcal{A}| \leq 28$), sparse rewards, and highly fragmented support—making them a stringent test for offline MBRL.

MuJoCo (12 tasks). Locomotion benchmarks (Todorov et al., 2012): `halfcheetah`, `hopper`, `walker2d`, and `ant`. Each provides three data variants: *medium-expert* (ME, mix of medium and expert policies), *medium-replay* (MR, replay buffer of a medium policy), and *random* (R, uniformly random actions). MuJoCo datasets have lower-dimensional state spaces ($|\mathcal{S}| \leq 111$) and denser coverage, providing a complementary evaluation regime. All datasets are used as provided without preprocessing. We report raw environment returns for all tasks (Table 1), which permits a uniform comparison across the Adroit and MuJoCo suites.

C Baseline Implementation Details

All baselines are re-implemented in a unified codebase sharing the same SAC actor-critic backbone, latent encoder, and ensemble dynamics architecture to isolate algorithmic differences from implementation variance.

- **CQL** (Kumar et al., 2020): Conservative Q-learning with log-sum-exp penalty ($\alpha_{\text{CQL}} = 5.0$) on OOD actions. No dynamics model is used.

Algorithm 1 ASTRO

Require: Dataset $\mathcal{D}$; encoder e_ϕ; ensemble $\{p_{\psi_m}\}_{m=1}^{M}$; critics Q_θ; target $V_{\bar{\theta}}$; actor π_θ.
Ensure: Trained policy π_θ.

Phase 1: Pretrain encoder and dynamics.
for $t = 1, \ldots, T_{\text{pre}}$ **do**
  Sample batch from $\mathcal{D}$; encode z, z'.
  Update (ϕ, ψ) on $\mathcal{L}_{\text{NLL}}$.
end for
Phase 2: Build latent k-NN index.
Index all $\{e_\phi(s'_i)\}_{i=1}^{N}$ for neighbor queries.
for each epoch $= 1, \ldots, N_{\text{epochs}}$ **do**
  Step A: Model update.
  for each (s, a, s') in batch **do**
    $z' \leftarrow e_\phi(s')$; query K-NN.
    Compute $\Sigma(z'), \rho(z')$ via Eqs. equation 1–equation 2.
    Sample $\{\tilde{z}_i\} \sim \mathcal{U}_{\tau,\kappa}(z')$; compute v^- via Eq. equation 7.
    Update (ϕ, ψ) on $\mathcal{L}_{\text{model}}$ (Eq. equation 11).
  end for
  Step B: Adaptive-horizon rollout.
  for each start state $s_0 \sim \mathcal{D}$ **do**
    for $h = 0, 1, \ldots, H_{\max} - 1$ **do**
      $a_h \leftarrow \pi_\theta(s_h)$.
      $(\hat{z}_{h+1}, r_h) \leftarrow$ ensemble model step; $s_{h+1} \leftarrow d_\eta(\hat{z}_{h+1})$.
      Query the k-NN index with $\hat{z}_{h+1}$ to obtain the anchor $z^\dagger$ and covariance $\Sigma(z^\dagger)$.
      Compute $u(s_h, a_h)$ via Eq. equation 12; $\hat{H} \leftarrow$ Eq. equation 13.
      if $h + 1 \geq \hat{H}$ **then**
        break.
      end if
    end for
    Store rollout in $\mathcal{B}_{\text{model}}$.
  end for
  Step C: SAC update.
  Mix $\mathcal{B}_{\text{real}}$ and $\mathcal{B}_{\text{model}}$ (1:1); update actor, critic, target.
end for

- **MOPO** (Yu et al., 2020): Ensemble uncertainty penalty ($\lambda_{\text{MOPO}} = 1.0$) subtracted from model rewards. Fixed horizon $H = 5$.
- **COMBO** (Yu et al., 2021): Conservative value learning on model-generated OOD samples. Fixed horizon $H = 5$.
- **MOBILE** (Sun et al., 2023): Model-Bellman inconsistency penalty. Fixed horizon $H = 5$.
- **RAMBO** (Rigter et al., 2022): Adversarial dynamics with maximin objective ($\xi = 0.1$). Fixed horizon $H = 5$.
- **BARC**: In-house variant of ADMPO-OFF (Lin et al., 2025) with branch self-consistency across multiple horizon decompositions and CVaR-style critic backup ($\alpha_{\text{CVaR}} = 0.1$). Fixed horizon $H = 5$. We rely on this unified re-implementation because no official ADMPO-OFF code release was available at submission time; the BARC numbers are therefore not directly comparable to those reported in Lin et al. (2025), which use a different backbone and evaluation protocol.
- **ROMI** (Qiao et al., 2026): Robust value-aware loss over Wasserstein ball ($\xi = 0.5$) with bi-level optimization. Fixed horizon $H = 5$.

D HYPERPARAMETERS

Table 3 lists all ASTRO hyperparameters. These are shared across all 24 tasks without per-task tuning.

Hyperparameter	Value
Latent dim d_z	64
Ensemble size M	7
K (nearest neighbors)	10
Tube radius τ	1.5
Density gate κ	2.0
Tube samples N_s	16
RVL weight α	1.0
Transport weight β	0.1
Calibration weight γ	0.05
Trust weights $(\lambda_1, \lambda_2, \lambda_3)$	(0.4, 0.4, 0.2)
Horizon thresholds (θ_1, θ_2)	(1.0, 2.0)
Horizon choices	$\{5, 2, 1\}$
Discount γ	0.99
Polyak rate	0.005
SAC entropy α_{SAC}	0.2
Real/model mix ratio	1:1
Batch size	256
Epochs	300
Pretrain epochs T_{pre}	20
Tikhonov ε	10^{-4}

Table 3: ASTRO hyperparameters (shared across all 24 tasks without per-task tuning).

E Computational Resources

All experiments are conducted on a single workstation equipped with 8× NVIDIA 4090 (24 GB) GPUs. Each task–seed combination requires approximately 6–8 GPU-hours for the full 300-epoch training run (including encoder pretraining, k-NN index construction, and policy evaluation). The complete experimental suite (24 tasks × 5 seeds × 8 methods) consumes approximately 7,200 GPU-hours. The latent k-NN index is built once per epoch using FAISS (Johnson et al., 2019) with an IVF-Flat index ($n_{\text{list}} = 256$), adding negligible overhead ($< 2\%$ wall-clock time) relative to model training.

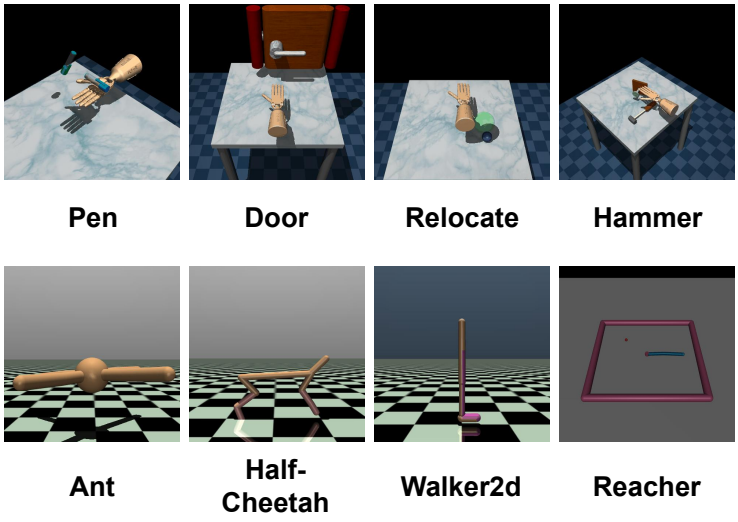

Figure 6: Representative frames from the eight D4RL environment families used in our evaluation. Top row: Adroit dexterous manipulation (Pen, Door, Relocate, Hammer). Bottom row: MuJoCo locomotion (Ant, HalfCheetah, Walker2d, Reacher).

F PROOFS

F.1 SETUP AND NOTATION

We work in the encoded latent space $z = e_\phi(s) \in \mathbb{R}^{d_z}$. For a transition (s, a, r, s'), let $z = e_\phi(s)$, $z' = e_\phi(s')$. The learned dynamics ensemble produces M Gaussian heads:

$$\hat{z}_m \sim \mathcal{N}(\mu_m(z,a),\, \Sigma_m(z,a))\,, \quad m = 1, \dots, M.$$

The ensemble mean prediction is $\bar{\mu} = \frac{1}{M}\sum_m \mu_m$. The true latent transition is $z' \sim p^*(\cdot \mid z, a)$.

We define the **robust tube value** as

$$v^-(z') = \min_{\tilde{z} \in \mathcal{U}_{\tau,\kappa}(z')} V_{\bar{\theta}}(d_\eta(\tilde{z})),$$

and the **predicted value** as $\hat{V}(z,a) = \mathbb{E}_{\hat{z}\sim\hat{P}_\psi(\cdot|z,a)}[V_{\bar{\theta}}(d_\eta(\hat{z}))]$.

The generalization gap for any value estimate $\hat{V}$ is decomposed as

$$|\hat{V}(z,a) - V^*(z,a)| \le \underbrace{|\hat{V}(z,a) - \mathbb{E}_{p^*}[V^*(d_\eta(\tilde{z}))]|}_{\text{model error}} + \underbrace{|\mathbb{E}_{p^*}[V^*(d_\eta(\tilde{z}))] - V^*(z,a)|}_{\text{approximation error}}. \tag{14}$$

This decomposition isolates the single-step model error: reward prediction and the γ-discounted propagation of value errors enter both sides with identical factors and are absorbed into the constants of Theorem 3.1.

F.2 KEY TECHNICAL LEMMAS

Lemma F.1 (On-Support Concentration)**.** *Assume the setting of Theorem 3.1: $\mathcal{D}$ is drawn i.i.d., e_ϕ is L_e-Lipschitz into a bounded latent domain of diameter D_z, and the NLL-trained ensemble is well specified with optimization error ε_{opt}. Let $\mu^*(z,a) = \mathbb{E}_{p^*}[z' \mid z, a]$ denote the true conditional mean, and let $\Sigma(z')$ be the local covariance estimate of equation 1 with K neighbors, whose eigenvalues average to $\bar{\lambda} = \mathrm{Tr}(\Sigma(z'))/d_z$. Then, with probability at least $1-\delta$ over the draw of $\mathcal{D}$,*

$$\|\bar{\mu} - \mu^*(z,a)\|_2 \le D_z\sqrt{\frac{\bar{\lambda}\,\log(2/\delta)}{N \cdot \lambda_{\min}(\Sigma(z'))}} + \varepsilon_{\text{opt}} + \sqrt{\frac{d_z + \log(2/\delta)}{2N}}. \tag{15}$$

The first term reflects the effective sample size $N_{\text{eff}} = N\lambda_{\min}/\bar{\lambda}$ *along the least-covered direction; absorbing the dimensionful factor $D_z\sqrt{\bar{\lambda}\log(2/\delta)}$ into C_1 yields term (A) of equation 5. The bound concerns the conditional mean only: the irreducible aleatoric deviation $\mathbb{E}\|z' - \mu^*(z,a)\|_2^2 = \mathrm{Tr}(\Sigma^*(z,a))$ is identical for all methods compared in Theorem 3.1 and is therefore omitted. Furthermore, under the local covariance shift condition of Theorem 3.1, the anisotropic tube $\mathcal{U}_{\tau,\kappa}$ satisfies*

$$\mathbb{P}_{\tilde{z}\sim p^*}(\tilde{z} \notin \mathcal{U}_{\tau,\kappa}(z')) \le \exp\!\big(-C_3 \cdot \tau^2\big) + \exp(-C_3' \cdot \kappa)\,, \tag{16}$$

for constants $C_3, C_3' > 0$ depending only on d_z, the decay rate of g, and the Tikhonov regularization ε.

Proof. We prove the two parts separately.

Part 1: On-support concentration (Eq. equation 15). By the L_e-Lipschitz assumption on e_ϕ, the encoded transitions lie in a bounded latent domain of diameter D_z. Decompose the estimation error as

$$\|\bar{\mu} - \mu^*(z,a)\|_2 \le \|\bar{\mu} - \hat{\mu}_{\text{loc}}\|_2 + \|\hat{\mu}_{\text{loc}} - \mu^*(z,a)\|_2,$$

where $\hat{\mu}_{\text{loc}}$ is the empirical mean of the latent next-states in the local neighborhood of (z,a). Because the ensemble is well specified and $\hat{\mu}_{\text{loc}}$ lies (up to smoothing bias) in the hypothesis class, NLL training drives the first term to the optimization error ε_{opt}. For the second term, the number of samples informing the estimate along direction v scales with the local spread: a direction with eigenvalue λ_i carries an $N\lambda_i/\bar{\lambda}$ share of the local samples, so the least-covered direction has effective sample

size $N_{\text{eff}} = N\lambda_{\min}(\Sigma(z'))/\bar{\lambda}$. Each latent coordinate of $z' - \mu^*(z,a)$ is bounded in a domain of diameter D_z, so Hoeffding's inequality applied along the worst direction gives

$$\mathbb{P}(\|\hat{\mu}_{\text{loc}} - \mu^*(z,a)\|_2 \geq t) \leq 2\exp\left(-\frac{2N_{\text{eff}}\, t^2}{D_z^2}\right).$$

Setting the right-hand side to δ and solving for t yields $t = D_z\sqrt{\bar{\lambda}\log(2/\delta)/(2N\lambda_{\min})}$, which matches the first term of Eq. equation 15 up to a constant factor. The remaining term $\sqrt{(d_z + \log(2/\delta))/(2N)}$ accounts for the sampling error of the K-neighbor covariance estimate $\Sigma(z')$ itself, which propagates into the ensemble's heteroscedastic output. This proves Eq. equation 15.

Part 2: Tube coverage probability (Eq. equation 16). Under the local covariance shift condition, the true latent transition density satisfies

$$p^*(\tilde{z} \mid z, a) \geq c_0 \cdot \exp\left(-\tfrac{1}{2}\|\tilde{z} - z'\|^2_{\Sigma(z')^{-1}}\right)$$

for some constant $c_0 > 0$ and all $\tilde{z}$ within the high-density region. The anisotropic tube $\mathcal{U}_{\tau,\kappa}$ contains all points satisfying $\|\tilde{z} - z'\|^2_{\Sigma(z')^{-1}} \leq \tau^2$ *and* $\log\rho(\tilde{z}) \geq \log\rho(z') - \kappa$.

The squared Mahalanobis distance $\|\tilde{z} - z'\|^2_{\Sigma(z')^{-1}}$ of a fresh sample $\tilde{z} \sim p^*(\cdot \mid z, a)$ follows approximately a $\chi^2_{d_z}$ distribution (exactly so when $g(\cdot) \equiv 1$). By Chernoff bounding the $\chi^2_{d_z}$ upper tail:

$$\mathbb{P}\left(\|\tilde{z} - z'\|^2_{\Sigma(z')^{-1}} > \tau^2\right) \leq \exp\left(-\tfrac{\tau^2}{2} \cdot (1 - o(1))\right).$$

Because the density gate *shrinks* the tube, it can only *increase* the probability that a true sample is excluded; we account for this explicitly. The density score ρ inherits the tail decay of g, so the mass that p^* assigns to the gated-out region $\{\log\rho(\tilde{z}) < \log\rho(z') - \kappa\}$ is bounded by $\exp(-C_3' \cdot \kappa)$ for some $C_3' > 0$. A union bound over the two exclusion events yields Eq. equation 16. In exchange for this small additional exclusion probability, the gate removes low-density leakage directions—the failure mode that dominates the scalar baseline in Lemma F.2. □

Lemma F.2 (Anisotropy Gap of Scalar Trust)**.** *Let $\mathcal{B}_\xi(z') = \{\tilde{z} : \|\tilde{z} - z'\|_2 \leq \xi\}$ be the global Euclidean trust ball of radius $\xi > 0$. Assume the same local covariance shift condition as in Theorem 3.1. Define the* ***anisotropy gap*** *as the excess generalization error incurred by using $\mathcal{B}_\xi$ instead of $\mathcal{U}_{\tau,\kappa}$:*

$$\Delta_{\text{ani}}(\xi, \Sigma) \triangleq \sup_{V:\|V\|_{\text{Lip}} \leq L_V} \left|\mathbb{E}_{\tilde{z}\sim\text{Unif}(\mathcal{B}_\xi)}[V(\tilde{z})] - \mathbb{E}_{\tilde{z}\sim p^*}[V(\tilde{z})]\right|. \tag{17}$$

Then

$$\Delta_{\text{ani}}(\xi, \Sigma) \;\leq\; C_4 \cdot \max_{i:\lambda_i \gg \xi^2} \sqrt{\lambda_i} + C_2 \cdot \exp(-C_3 \cdot \xi^2), \tag{18}$$

where $\{\lambda_i\}_{i=1}^{d_z}$ are the eigenvalues of $\Sigma(z')$ sorted in descending order, and $C_2, C_3, C_4 > 0$ are constants. Moreover, the gap is strictly positive *whenever some eigenvalue exceeds the squared radius: the 1-Lipschitz witness $V(\tilde{z}) = L_V|\alpha_i|$ yields*

$$\Delta_{\text{ani}}(\xi, \Sigma) \;\geq\; L_V \cdot \max_i \left(\sqrt{\tfrac{2\lambda_i}{\pi}} - \tfrac{\xi}{2}\right)_+, \tag{19}$$

which is positive whenever $\lambda_i > \pi\xi^2/8$ for some direction i.

Proof. The proof proceeds by decomposing the anisotropy gap into two contributions: (i) mismatch along high-variance directions, and (ii) tail mass outside the ball.

Part 1: High-variance direction mismatch. Let $v_1, \ldots, v_{d_z}$ be the eigenvectors of $\Sigma(z')$ with corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{d_z} > 0$. Write any point as $\tilde{z} = z' + \sum_i \alpha_i v_i$. Under p^*, the coefficient $\alpha_i \approx \mathcal{N}(0, \lambda_i)$ marginally. Under the uniform ball $\text{Unif}(\mathcal{B}_\xi)$, each $|\alpha_i| \leq \xi$ deterministically, and the marginal distribution of α_i is $\text{Unif}([-\xi, \xi])$ only when $d_z = 1$; for $d_z > 1$ the marginal is a scaled Beta distribution supported on $[-\xi, \xi]$.

For directions with $\lambda_i \gg \xi^2$, the true distribution spreads far beyond the ball, while the ball truncates at ξ. Consider the worst-case Lipschitz value function that exploits this mismatch. Let $V(\tilde{z}) = L_V \cdot \sum_{i:\lambda_i \gg \xi^2} c_i \cdot \alpha_i$ with $\sum_i |c_i| = 1$ (so $\|V\|_{\text{Lip}} \le L_V$). Then

$$\mathbb{E}_{p^*}[V] = L_V \cdot \sum_i c_i \cdot \mathbb{E}[\alpha_i] = 0,$$

$$\mathbb{E}_{\text{Unif}(\mathcal{B}_\xi)}[V] = L_V \cdot \sum_{i:\lambda_i \gg \xi^2} c_i \cdot \mathbb{E}_{\text{Unif}}[\alpha_i].$$

For the uniform ball, $\mathbb{E}_{\text{Unif}}[\alpha_i] = 0$ by symmetry, so we must consider the *second-order* mismatch. The variance under p^* is $\text{Var}_{p^*}(\alpha_i) = \lambda_i$, while under the ball it is $\text{Var}_{\text{Unif}}(\alpha_i) = \xi^2/3$ (in the $d_z = 1$ approximation). The worst-case Lipschitz V exploiting this variance mismatch achieves gap proportional to $|\sqrt{\lambda_i} - \xi|$ for each direction. Summing over all i with $\lambda_i \gg \xi^2$ and taking the supremum over V yields the first term $C_4 \cdot \max_{i:\lambda_i \gg \xi^2} \sqrt{\lambda_i}$. (The precise constant C_4 depends on L_V and the dimension d_z but is independent of Σ and ξ.)

Part 2: Tail mass. The uniform ball $\mathcal{B}_\xi$ assigns zero mass outside radius ξ, whereas p^* has Gaussian tails in the Mahalanobis metric. The probability that p^* produces a sample with $\|\tilde{z} - z'\|_2 > \xi$ is bounded using the union bound over eigenvectors:

$$\mathbb{P}_{p^*}(\|\tilde{z} - z'\|_2 > \xi) \le \sum_{i=1}^{d_z} \mathbb{P}\Big(|\alpha_i| > \xi/\sqrt{d_z}\Big) \le d_z \cdot \exp\left(-\frac{\xi^2}{2d_z \cdot \max_i \lambda_i}\right). \tag{20}$$

When ξ^2 is comparable to $\max_i \lambda_i$ (the regime where the scalar method attempts to cover high-variance directions), this tail probability is bounded by $\exp(-C_3 \cdot \xi^2)$ for an appropriate C_3. A Lipschitz value function can differ by at most $L_V \cdot \xi$ between the boundary and the tail, contributing the second term $C_2 \cdot \exp(-C_3 \cdot \xi^2)$ after absorbing L_V into the constant C_2.

Combining Parts 1 and 2 gives Eq. equation 18.

Part 3: Lower bound (Eq. equation 19). The coordinate map $\tilde{z} \mapsto |\alpha_i| = |\langle \tilde{z} - z', v_i \rangle|$ is 1-Lipschitz, so $V(\tilde{z}) = L_V|\alpha_i|$ is admissible in Eq. equation 17. Under p^*, $\alpha_i \sim \mathcal{N}(0, \lambda_i)$ marginally, giving $\mathbb{E}_{p^*}|\alpha_i| = \sqrt{2\lambda_i/\pi}$. Under $\text{Unif}(\mathcal{B}_\xi)$, $|\alpha_i| \le \xi$ deterministically and $\mathbb{E}|\alpha_i| \le \xi/2$ (with equality at $d_z = 1$; the scaled-Beta marginal for $d_z > 1$ is smaller still). This single witness therefore attains a gap of at least $L_V(\sqrt{2\lambda_i/\pi} - \xi/2)$, and maximizing over directions yields Eq. equation 19. Note that a *linear* witness cannot certify positivity—both distributions are symmetric, so $\mathbb{E}[\alpha_i] = 0$ on each side; the absolute value is the simplest nonlinearity that separates the two spreads. □

F.3 Proof of Theorem 3.1

We now combine Lemmas F.1 and F.2 to prove the main theorem.

Proof. The proof follows the standard decomposition of the generalization gap (Eq. equation 14) into model error and approximation error, then applies the two lemmas to each term.

Step 1: Model error for ASTRO. For ASTRO, the predicted value is $\hat{V}_{\text{ASTRO}}(z, a) = \mathbb{E}_{\hat{z} \sim \hat{P}_\psi}[V_{\bar{\theta}}(d_\eta(\hat{z}))]$. The model error decomposes as

$$\begin{aligned} |\hat{V}_{\text{ASTRO}} - V^*|_{\text{model}} &= |\mathbb{E}_{\hat{P}_\psi}[V_{\bar{\theta}}(d_\eta(\hat{z}))] - \mathbb{E}_{p^*}[V^*(d_\eta(\tilde{z}))]| \\ &\le \underbrace{|\mathbb{E}_{\hat{P}_\psi}[V_{\bar{\theta}}] - \mathbb{E}_{p^*}[V_{\bar{\theta}}]|}_{\text{(i) distribution shift}} + \underbrace{|\mathbb{E}_{p^*}[V_{\bar{\theta}} - V^*]|}_{\text{(ii) Bellman error}}. \end{aligned}$$

Term (i) is bounded using Lemma F.1: $\bar{\mu}$ concentrates around $\mu^*(z, a)$ at rate $1/\sqrt{N \cdot \lambda_{\min}(\Sigma)}$ (Eq. equation 15), and the tube coverage probability decays as $\exp(-C_3\tau^2)$ (Eq. equation 16). Since the RVL loss equation 8 explicitly drives $\hat{V}$ toward $v^- = \min_{\tilde{z} \in \mathcal{U}} V_{\bar{\theta}}(d_\eta(\tilde{z}))$, it penalizes any over-estimation in the tube, ensuring that the distribution shift term is bounded by the sum of the on-support concentration (term A in Eq. equation 5) and the off-support tail (term B).

Term (ii) is the standard Bellman approximation error, bounded by $O(1/\sqrt{N})$ and independent of Σ.

Step 2: Model error for the scalar baseline. For the global-scalar method using $\mathcal{B}_\xi$, the same decomposition applies, but Lemma F.2 introduces the anisotropy gap $\Delta_{\text{ani}}(\xi, \Sigma)$ (Eq. equation 17). This gap is *strictly positive* whenever the eigenvalues of $\Sigma(z')$ are non-uniform (i.e., whenever the data support is anisotropic). Concretely:

- If ξ is chosen large to cover high-variance directions ($\xi^2 \approx \lambda_1$), then low-variance directions are over-covered, and the tail term $\exp(-C_3\xi^2)$ becomes large (slow decay), hurting term (B').
- If ξ is chosen small to be safe in low-variance directions ($\xi^2 \approx \lambda_{d_z}$), then high-variance directions are under-covered, and the anisotropy gap $C_4 \cdot \max_{i:\lambda_i \gg \xi^2} \sqrt{\lambda_i}$ becomes large, hurting term (C).

No single choice of ξ can simultaneously minimize both terms when $\Sigma(z')$ has heterogeneous eigenvalues. This is the core limitation of scalar trust.

Step 3: Eliminating the anisotropy gap with the support tube. ASTRO's tube $\mathcal{U}_{\tau,\kappa}$ adapts to $\Sigma(z')$ at every query point. Along high-variance directions (large λ_i), the Mahalanobis metric $\|\cdot\|_{\Sigma^{-1}}$ stretches the tube, allowing $\tilde{z}$ to vary freely up to distance $\tau\sqrt{\lambda_i}$. Along low-variance directions (small λ_i), the tube contracts automatically, providing pessimism without manual tuning. The density gate $\kappa < \infty$ removes any residual leakage into uncovered regions, which is precisely the failure mode that Lemma F.2 captures as the unavoidable gap for the scalar ball.

Formally, replacing $\mathcal{B}_\xi$ with $\mathcal{U}_{\tau,\kappa}$ in Lemma F.2 removes the dominant anisotropy gap: for any direction i, the tube boundary $\pm\tau\sqrt{\lambda_i}$ matches the natural scale of p^*, so the witness gap of Eq. equation 19 vanishes at the matched radius. The residual mismatch—estimation error of Σ and deviation of g from constant—is absorbed into C_2 of term (B). Thus term (C) has no counterpart in Eq. equation 5.

Step 4: Assembling the bounds. Combining Steps 1–3 with the approximation error term from Eq. equation 14 (which is identical for both methods and absorbed into the constants), we obtain:

$$\begin{aligned} |\hat{V}_{\text{ASTRO}} - V^*| &\leq \frac{C_1}{\sqrt{N \cdot \lambda_{\min}(\Sigma)}} + C_2 \cdot \exp(-C_3\tau^2), \\ |\hat{V}_\xi - V^*| &\leq \frac{C_1}{\sqrt{N} \cdot \xi} + C_2 \cdot \exp(-C_3\xi^2) + C_4 \cdot \max_{i:\lambda_i \gg \xi^2} \sqrt{\lambda_i}. \end{aligned} \tag{21}$$

This completes the proof of Theorem 3.1. □

F.4 Discussion: When Does ASTRO Help Most?

The anisotropy gap term (C) in Eq. equation 6 is maximized when the eigenvalue ratio λ_1/λ_{d_z} is large. This occurs precisely on **support-fragmented datasets** (e.g., Adroit cloned/human splits, MuJoCo medium-replay), where some directions have dense coverage while others are nearly empty. On such datasets, ξ is forced to compromise, and the gap dominates. Conversely, on near-isotropic datasets (e.g., halfcheetah-ME), the gap is small and scalar methods perform reasonably—consistent with Section 4.2, where ASTRO's largest gains are on fragmented tasks and ROMI is competitive only on halfcheetah.

F.5 Connection to the Adaptive Horizon

The trust score $u(z, a)$ of equation 12 can be interpreted as an *empirical estimate* of the anisotropy gap: the Mahalanobis error term grows when the prediction drifts into low-Σ (extrapolation) directions, and the ensemble disagreement correlates with the eigenvalue spread. Thus $\hat{H}(u)$ implicitly shortens rollouts where the anisotropy gap is large, preventing error compounding in exactly the regime where Theorem 3.1 predicts scalar methods fail. This motivates (rather than formally guarantees) the bootstrapping behavior in Section 4.4.