Locality-Aware Distributionally Robust Deep Q-Learning: Reducing Epistemic Uncertainty via Context-Conditional Ambiguity Sets

Giulio Golinelli
golinelli.giulio13@gmail.com

November 2025

Abstract

We propose an enhancement to the Distributionally Robust Deep Q-Learning (RDQN) framework that incorporates local temporal context into the reference measure estimation. By conditioning the ambiguity set on local state dynamics, we aim to reduce epistemic uncertainty while maintaining protection against distributional shifts, potentially enabling less conservative policies without sacrificing robustness guarantees. Our approach addresses the fundamental limitation of aggregation-induced conservatism in standard RDQN by distinguishing between different dynamical regimes within the state space. We further introduce a regime-conditional decision criterion that blends the worst-case (robust) and nominal (risk-neutral) Bellman targets, so the agent relaxes its hedge in favorable, high-confidence regimes while remaining defensive elsewhere — reducing conservatism without ever becoming optimistic.

1 Introduction and Motivation

1.1 Current Framework and Limitations

The current RDQN framework constructs ambiguity sets as Sinkhorn balls around a reference measure $\hat{P}(x, a)$:

$$\mathcal{B}_{\varepsilon,\delta}(\hat{P}(x,a)) = \{P \in \mathcal{M}_1(\mathcal{X}) : W_\delta(\hat{P}(x,a), P) \leq \varepsilon\} \tag{1}$$

The reference measure $\hat{P}(x, a)$ is typically estimated via the empirical distribution:

$$\hat{P}(x,a) = \frac{1}{N(x,a)} \sum_{i:(x_i,a_i)=(x,a)} \delta_{x'_i} \tag{2}$$

where $N(x, a)$ counts all historical transitions from state-action pair (x, a), and $\delta_{x'_i}$ denotes the Dirac measure at x'_i.

Critical Observation: This estimator aggregates all transitions $(x, a) \to x'$ across the entire training history, treating them as i.i.d. samples from a stationary distribution. However, in financial and dynamical systems, the transition kernel may exhibit:

1. **Regime-dependent dynamics**: Transitions during bull versus bear markets differ systematically
2. **Local momentum effects**: Recent directional trends influence short-term dynamics
3. **Volatility clustering**: Periods of high/low volatility exhibit distinct transition characteristics
4. **Mean-reversion dynamics**: Behavior near local extrema differs from typical states

By ignoring this temporal heterogeneity, we introduce **excessive epistemic uncertainty** into $\hat{P}(x, a)$, necessitating overly large ambiguity balls (large ε) and yielding overly conservative policies.

Empirical Evidence: The portfolio optimization experiments in Lu et al. (2025, Section 4.2, Table 4) show RDQN achieves superior risk-adjusted returns (Sharpe and Sortino ratios) but underperforms buy-and-hold strategies on absolute wealth accumulation, suggesting excessive conservatism that leaves profitable opportunities unexploited.

2 Locality-Aware Reference Measures

2.1 Formal Framework

We propose conditioning the reference measure on a **local context descriptor** $\tau \in \mathcal{T}$, where $\mathcal{T}$ is a space of local features capturing recent dynamics.

Definition 1 (Context-Conditional Reference Measure)**.** *Define the context-conditional reference measure as:*

$$\hat{P}(x, a, \tau) : \mathcal{X} \times \mathcal{A} \times \mathcal{T} \to \mathcal{M}_1(\mathcal{X}) \tag{3}$$

The empirical estimate becomes:

$$\hat{P}(x, a, \tau) = \frac{1}{N(x, a, \tau)} \sum_{i:(x_i, a_i, \tau_i) \approx (x, a, \tau)} \delta_{x'_i} \tag{4}$$

where "$\approx$" denotes approximate matching and $N(x, a, \tau)$ counts matching transitions.

2.2 Context Descriptors for Financial Applications

For portfolio optimization tasks, we propose a composite context descriptor $\tau = (\tau_{\text{trend}}, \tau_{\text{vol}}, \tau_{\text{pos}})$ where:

2.2.1 Local Trend Indicator

$$\tau_{\text{trend}} = \text{sign}\left(\frac{1}{k}\sum_{j=1}^{k}(x_{t-j}^{(1)} - x_{t-j-1}^{(1)})\right) \in \{-1, 0, +1\} \tag{5}$$

where $x^{(1)}$ denotes the most recent return component in the state vector, and $k \in \mathbb{N}$ defines the lookback window for trend estimation.

2.2.2 Local Volatility Regime

$$\tau_{\text{vol}} = \text{quantize}\left(\sqrt{\frac{1}{k}\sum_{j=1}^{k}(x_{t-j}^{(1)})^2}\right) \in \{\text{low}, \text{med}, \text{high}\} \tag{6}$$

where the quantization thresholds (the 33rd and 67th percentiles of window-volatilities) are computed *once* on a fixed reference period — e.g. the entire available price history — and then *frozen*. Freezing is essential: recomputing the terciles on a growing online sample would relabel the same historical day differently as training proceeds, rendering τ (and hence ϕ) non-stationary. A frozen, history-wide ruler keeps the regime definition stable throughout training and introduces no per-decision look-ahead, since τ is still computed from each day's own (past-only) return window.

2.2.3 Proximity to Local Extrema

$$\tau_{\text{pos}} = \begin{cases} +1 & \text{if } x_t^{(1)} \geq \max_{j\in[1,w]} x_{t-j}^{(1)} - \sigma \\ -1 & \text{if } x_t^{(1)} \leq \min_{j\in[1,w]} x_{t-j}^{(1)} + \sigma \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $x_t^{(1)}$ denotes the *current* return (scalar, not the full state vector), $w \in \mathbb{N}$ is a window size (default: $w = 20$), and σ is in units of the window's return standard deviation (default: $\sigma = 0.5$).

2.3 Theoretical Interpretation: State Augmentation

Remark 1 (Connection to State Augmentation)**.** *This approach can be viewed through two equivalent theoretical lenses:*

1. ***State Augmentation**: Redefine the state space as $\tilde{\mathcal{X}} = \mathcal{X} \times \mathcal{T}$, so $\tilde{x} = (x, \tau)$. Under this formulation, the Markov property holds on the augmented space $\tilde{\mathcal{X}}$.*

2. ***Context-Dependent Ambiguity***: *Maintain the original state space $\mathcal{X}$ but allow ambiguity sets to depend on context: $\mathcal{B}_{\varepsilon,\delta}(\hat{P}(x, a, \tau))$.*

The first interpretation is more theoretically rigorous as it explicitly maintains the Markovian framework. However, if τ can be computed deterministically from the recent history encoded in state x (e.g., if x contains a window of past returns), then τ is already implicitly available and no additional information is introduced.

3 Adaptive Ambiguity Radius via Context Fidelity

3.1 Context-Dependent Confidence Measure

Not all contexts provide equal statistical confidence. We introduce a **fidelity function** $\phi : \mathcal{T} \to (0, 1]$ measuring the reliability of context-conditional estimates:

$$\phi(\tau) = \min\left\{1, \frac{N_\tau}{N_{\min}}\right\} \cdot \left(1 - \frac{H_\tau}{\log B}\right) \tag{8}$$

where:

- N_τ is the number of historical transitions in regime τ (conditioned on τ *alone*, not on the full (x, a, τ), for continuous-state tractability; see Remark 2)
- $N_{\min}$ is a threshold for adequate sample size (default: $N_{\min} = 100$)
- H_τ is an *out-of-sample* entropy of the next-return distribution within regime τ: we split each regime's transitions into a fit half and a held-out half, estimate $\hat{P}(\tau)$ on the fit half, and take H_τ as the cross-entropy of the held-out half under $\hat{P}(\tau)$ (1-D histogram with B bins, Laplace-smoothed). This estimates *predictive* validity rather than in-sample fit — see Remark 3.
- The normalized entropy term $\frac{H_\tau}{\log B} \in [0, 1]$ measures distributional uncertainty

Remark 2 (Continuous-state operationalization)**.** *In the continuous 63-dimensional state space, conditioning on the full (x, a, τ) is degenerate (no two transitions share the same continuous (x, a)), so $N(x, a, \tau) \approx 1$. We therefore condition ϕ on τ* alone*: N_τ counts all transitions whose context tag is τ. Similarly, H_τ is computed over a* scalar projection *(the next return $x'^{(1)}$) discretized into B bins, rather than the full next-state vector, yielding a reliable, well-normalized entropy estimate. This is the standard operationalization of discrete-form formulas for continuous state spaces.*

Remark 3 (Precision versus accuracy: why H_τ is held-out)**.** *An in-sample entropy rewards regimes whose observed transitions are tightly clustered — that is* precision *(self-agreement of the data), not* accuracy *(agreement with the data-generating process). A regime may contain many tightly clustered samples yet still fail to predict a regime change; an in-sample ϕ would then shrink the ambiguity ball precisely when the estimate is most likely wrong — rewarding the absence of past variation and punishing the agent the moment variation arrives. Estimating H_τ on held-out transitions converts ϕ into a measure of predictive validity: a regime contributes confidence only if its estimated transition law generalizes to unseen transitions; otherwise its cross-entropy exceeds $\log B$ and the $(1 - H_\tau / \log B)$ term is clamped to zero. This does not weaken the asymptotic consistency of Proposition 1 (held-out cross-entropy is a stricter, not weaker, consistency diagnostic); it removes the most dangerous failure mode of the in-sample estimator.*

Intuition:

- High $\phi(\tau)$: Many samples and low conditional entropy $\Rightarrow$ high confidence $\Rightarrow$ use smaller ambiguity ball
- Low $\phi(\tau)$: Few samples or high conditional entropy $\Rightarrow$ low confidence $\Rightarrow$ use larger ambiguity ball

3.2 Adaptive Reference Measure and Radius

We define the context-aware reference measure via shrinkage:

$$\tilde{P}(x, a, \tau) = (1 - \phi(\tau))\hat{P}(x, a) + \phi(\tau)\hat{P}(x, a, \tau) \tag{9}$$

where $\phi(\tau)$ controls the blend between the global (context-agnostic) estimate and the local (context-specific) estimate. Operationally, this is realized as a ϕ-weighted mixture of next-state samples: each inner-loop candidate is drawn from the global reference cloud with probability $1 - \phi(\tau)$ and from the regime-τ empirical next-return cloud with probability $\phi(\tau)$.

The context-dependent ambiguity radius is:

$$\tilde{\varepsilon}(\phi(\tau)) = \varepsilon_{\text{base}} \cdot (1 - \beta\phi(\tau)) \tag{10}$$

where:

- $\varepsilon_{\text{base}}$ is the baseline radius from standard RDQN
- $\beta \in [0, 1)$ controls how aggressively we shrink the ambiguity radius based on context confidence
- Higher $\phi(\tau)$ leads to smaller $\tilde{\varepsilon}$, reflecting increased confidence

Proposition 1 (Asymptotic Consistency)**.** *As $N_\tau \to \infty$, if the true conditional distributions differ across contexts, i.e., $P^*(\cdot|\tau_1) \neq P^*(\cdot|\tau_2)$ for $\tau_1 \neq \tau_2$, then:*

$$W_\delta(\tilde{P}(x,a,\tau), P^*(x,a,\tau)) \xrightarrow{P} 0 \tag{11}$$

at a faster rate than for the context-agnostic estimator $\hat{P}(x,a)$.

Remark 4 (Honest evaluation protocol)**.** *Two precautions are required for a trustworthy empirical comparison against plain RDQN. First, the aggressiveness β (and any hyperparameter of ϕ) must be selected on a temporal* validation *window disjoint from the reported* test *backtest, lest β be fit to the test curve. Second, terminal wealth is highly non-monotone across training checkpoints, so each variant is evaluated over several seeds and the mean±standard deviation of risk-adjusted metrics (Sharpe, Sortino, maximum drawdown) is reported rather than any single checkpoint; the checkpoint with the best validation wealth (or a seed average) should be used for A/B comparisons.*

4 Modified Robust Bellman Operator

The robust value iteration becomes:

$$V_\delta(x,\tau) = \max_{a\in\mathcal{A}} \min_{P\in\mathcal{B}_{\tilde{\varepsilon}(x,a,\tau)}(\tilde{P}(x,a,\tau))} \mathbb{E}_P[r(x,a,X') + \alpha V_\delta(X',\tau')] \tag{12}$$

where τ' is computed from the local features at next state X'.
Similarly, the robust Q-function satisfies:

$$Q^*_\delta(x,a,\tau) = \min_{P\in\mathcal{B}_{\tilde{\varepsilon}(x,a,\tau)}(\tilde{P}(x,a,\tau))} \mathbb{E}_P[r(x,a,X') + \alpha \max_{b\in\mathcal{A}} Q^*_\delta(X',b,\tau')] \tag{13}$$

5 Regime-Conditional Opportunism: Blending the Robust and Nominal Objectives

5.1 Motivation: a regime-dependent criterion

The constructions of Sections 3–4 reduce conservatism by shrinking the *ambiguity set* (the radius $\tilde{\varepsilon}$) in high-fidelity contexts, but they retain the *same decision criterion everywhere* — the agent always hedges against the worst-case distribution inside the ball. In benign regimes this is wasteful: when the local context is favorable and reliably estimated, hedging against an adversarial distribution forfeits expected return for protection that is unlikely to be needed. As a proposed extension (not empirically validated here), this section makes the *decision criterion itself* regime-dependent, interpolating between the worst-case (robust) objective and the nominal (risk-neutral) objective as a function of the context. The agent would hedge in adverse or uncertain regimes and relax toward the nominal expectation in favorable, well-estimated ones.

5.2 Trend quality and the opportunism weight

Define a *trend-quality* score $q : \mathcal{T} \to [0, 1]$ that scores how favorable a regime is, read directly off the context descriptor $\tau = (\tau_{\text{trend}}, \tau_{\text{vol}}, \tau_{\text{pos}})$. A regime is favorable when it exhibits positive drift, low volatility, and proximity to local maxima; we set $q(\tau)$ high for such combinations and drive it toward 0 for adverse or neutral ones (in particular $q(\tau) = 0$ whenever $\tau_{\text{trend}} \leq 0$ or $\tau_{\text{vol}} = \text{high}$). The exact map is a design choice; what matters is that q is high only for unambiguously favorable, past-only regimes.

The *opportunism weight* is then

$$\lambda(\tau) \;=\; \lambda_{\max} \cdot \phi(\tau) \cdot q(\tau) \;\in\; [0, \lambda_{\max}], \tag{14}$$

with $\lambda_{\max} < 1$ a cap (default $\lambda_{\max} = 0.5$). Two safeguards are built in: (i) the product with $\phi(\tau)$ *gates* relaxation on confidence — the agent relaxes only when the favorable regime is also reliably estimated; (ii) the cap $\lambda_{\max} < 1$ ensures the worst-case objective is never fully abandoned.

5.3 The blended Bellman target

Let $Q^{\mathrm{R}}_\delta(x, a, \tau)$ denote the worst-case (robust) target of Section 4,

$$Q^{\mathrm{R}}_\delta(x, a, \tau) \;=\; \min_{P \in \mathcal{B}_{\tilde{\varepsilon}(\tau)}(\tilde{P}(\tau))} \mathbb{E}_P \Big[r(x, a, X') + \alpha \max_{b \in \mathcal{A}} Q^*_\delta(X', b, \tau') \Big],$$

and let $Q^{\mathrm{N}}(x, a, \tau)$ be the *nominal* (risk-neutral) target evaluated under the context-aware reference measure $\tilde{P}(\tau)$,

$$Q^{\mathrm{N}}(x, a, \tau) \;=\; \mathbb{E}_{\tilde{P}(\tau)} \Big[r(x, a, X') + \alpha \max_{b \in \mathcal{A}} Q^*_\delta(X', b, \tau') \Big].$$

The regime-conditional blended Q-operator is

$$Q^*_\delta(x, a, \tau) \;=\; \big(1 - \lambda(\tau)\big)\, Q^{\mathrm{R}}_\delta(x, a, \tau) \;+\; \lambda(\tau)\, Q^{\mathrm{N}}(x, a, \tau). \tag{15}$$

Because the worst-case target never exceeds the nominal one, $Q^{\mathrm{R}}_\delta \leq Q^{\mathrm{N}}$, the blend satisfies for every τ

$$Q^{\mathrm{R}}_\delta(x, a, \tau) \;\leq\; Q^*_\delta(x, a, \tau) \;\leq\; (1 - \lambda_{\max}) Q^{\mathrm{R}}_\delta(x, a, \tau) + \lambda_{\max} Q^{\mathrm{N}}(x, a, \tau) \;\leq\; Q^{\mathrm{N}}(x, a, \tau). \tag{16}$$

Hence the blended target is *never optimistic*: it de-hedges toward risk-neutrality but never becomes risk-seeking. In adverse or low-fidelity contexts $\lambda(\tau) \to 0$ and (15) reduces to pure locality-aware RDQN; in favorable, high-fidelity contexts $\lambda(\tau) \to \lambda_{\max}$ and the operator approaches (but stays short of) the nominal objective.

5.4 Interpretation as a regime-dependent risk measure

Equation (15) is a convex combination of two coherent risk measures — the worst-case ($\sup_P$ over the ball, the most risk-averse) and the expectation (risk-neutral) — and is therefore itself a coherent risk measure that depends on context. It is a special case of a *regime-dependent spectral risk measure* ρ whose risk-aversion is a function of τ. An equivalent, common parameterization is a regime-dependent Conditional Value-at-Risk $\rho_{\eta(\tau)} = \mathrm{CVaR}_{\eta(\tau)}$, with $\eta(\tau)$ large (tail-sensitive, robust) in adverse regimes and small (near the mean, risk-neutral) in favorable ones; the linear blend (15) is the first-order interpolation between the worst-case and expectation endpoints of this family.

5.5 Design remarks

Remark 5 (Smooth blend over hard switching)**.** *A hard if/else switch of the objective at a regime boundary would make the Bellman target a discontinuous function of the transient context, destabilizing learning — the same state could be scored worst-case in one episode and nominally in the next. The smooth, ϕ-gated blend in* (15) *avoids this: $\lambda(\tau)$ varies continuously with the context, and conditioning Q on τ (Remark on state augmentation) lets the network represent regime-conditional values consistently.*

Remark 6 (Radius and criterion, used together or apart)**.** *The radius mechanism $\tilde{\varepsilon}(\tau)$ (Sections 3–4) and the criterion rule $\lambda(\tau)$ (this section) are complementary: the former controls the size of the ambiguity set, the latter the optimism of the objective within it. They may be used jointly, or the criterion rule may be taken as the more direct de-conservatism mechanism. The two should be ablated separately, since stacking them risks redundant over-relaxation.*

Remark 7 (The asymmetry is intentional)**.** *The design is deliberately asymmetric: the default is the worst case, and relaxation requires* both *a favorable regime (q high)* and *confidence it is real (ϕ high), capped below the nominal. This bounds the cost of the irreducible failure mode — a favorable label assigned just before an adverse move — at the finite gap $Q^N - Q^R$ between the nominal and worst-case targets, rather than at an unbounded optimistic bet.*

6 Conclusion and Expected Improvements

The proposed locality-aware extension of RDQN refines transition estimation by conditioning ambiguity sets on temporal context, yielding more confident and narrowly concentrated probability distributions. By incorporating local features such as trend, volatility, and proximity to extrema, the agent captures distinct market regimes and reduces epistemic uncertainty within stable contexts.

The fidelity function $\phi(\tau)$ enables the Sinkhorn ambiguity ball to adapt its radius to the reliability of each context. High-fidelity conditions—characterized by abundant data and low entropy—produce tighter ambiguity regions and less

conservative policies, while low-fidelity contexts revert to the global baseline $\hat{P}(x, a)$ to preserve robustness.

A complementary rule, proposed but not yet empirically validated, then converts the fixed worst-case criterion into a regime-dependent risk measure: a smooth blend of the worst-case and nominal Bellman targets, weighted by an opportunism factor $\lambda(\tau) = \lambda_{\max} \phi(\tau) q(\tau)$ that is large only in favorable *and* high-fidelity regimes and capped below one. This would concentrate the robustness budget on adverse or uncertain regimes while permitting risk-neutral behavior where the context is benign and reliably estimated, yet the target provably never exceeds the nominal (the agent de-hedges but never becomes risk-seeking). The asymmetry is deliberate and bounds the cost of any misclassified regime.

Overall, this formulation balances adaptivity and protection: it contracts epistemic uncertainty where confidence is high and safeguards against model misspecification where it is low. In practice, the resulting agent embodies a more aggressive yet robustly profitable systematic trading strategy—capable of exploiting favorable market regimes without compromising stability under uncertainty.