Context-Aware Distributionally Robust Deep Q-Learning (CARDQN)

Giulio Golinelli
golinelli.giulio13@gmail.com

August 2026

Abstract

Robust trading agents are deliberately cautious: by hedging against the worst plausible market outcome they stay stable under model error, but they give up much of the available return. We make such an agent less cautious without making it reckless. Our method, CARDQN, makes the safety margin *context-aware*: it tightens the margin only in market conditions it can estimate reliably, and keeps it wide when uncertain. On the empirical S&P 500 (1995–2024) CARDQN reaches 3.40× terminal wealth at a 10-episode budget—above the 2.89× reported by the original authors—at even volatility, with higher Sharpe and Sortino and a comparable drawdown. We further describe, but do not validate here, a complementary *regime-conditional opportunism* rule that relaxes the worst-case criterion in clearly favorable regimes; by construction it is never optimistic.

1 Introduction

Distributionally Robust Deep Q-Learning (RDQN) learns a trading policy that hedges against the worst plausible evolution of the market rather than trusting a single fitted model. The price of that safety is conservatism: on a strongly drifting asset such as the S&P 500 the agent stays stable but forfeits much of the buy-and-hold return. This paper removes that conservatism at its source.

We trace the conservatism to its source: RDQN pools every historical transition into one reference measure, ignoring that bull and bear, calm and turbulent periods behave differently. The heterogeneity it averages away inflates uncertainty and forces an oversized safety margin (the radius ε, with Sinkhorn regularization δ). CARDQN makes this margin context-aware.

The context-aware system (§2) uses a low-dimensional, past-only regime tag τ to label the current market condition, and a confidence score $\phi(\tau)$—high only when the regime is well sampled and predicts its own future—to shrink the radius (via a knob β) and shift the reference measure toward the regime's own dynamics. The margin thus tightens where the agent is confident and stays wide where it is not. We additionally describe, as a proposed but not-yet-validated extension (§3), a regime-conditional opportunism rule: in clearly favorable, well-identified regimes the agent would blend the worst-case target toward the risk-neutral one by a confidence-gated, capped weight $\lambda(\tau)$ (capped at $\lambda_{\max} < 1$), defaulting to full robustness elsewhere and, by construction, never becoming optimistic.

We state the operators, prove their key properties, and show empirically that CARDQN surpasses the original authors' reported result—3.40× terminal wealth at a 10-episode budget versus their 2.89×, at even volatility with higher Sharpe and Sortino and a comparable drawdown—on the empirical S&P 500 (1995–2024). Data-driven radii that shrink with sample size are classical in distributionally robust optimization (Mohajerin Esfahani & Kuhn, 2018); our contribution is the contextual and criterion-level extension inside a Sinkhorn-dual deep RL agent.

2 Context-Conditional Ambiguity Sets

2.1 A past-only regime tag

We condition the ambiguity set on a low-dimensional *context descriptor* $\tau \in \mathcal{T}$, computed from the recent return history in the state (hence past-only, no look-ahead). For single-asset trading we use a composite tag $\tau = (\tau_{\text{trend}}, \tau_{\text{vol}}, \tau_{\text{pos}})$ with three components each in $\{-1, 0, +1\}$ (so $|\mathcal{T}| = 27$):

$$\tau_{\text{trend}} = \operatorname{sign}\Big(\tfrac{1}{k}\sum_{j=1}^{k}(x^{(1)}_{t-j} - x^{(1)}_{t-j-1})\Big), \tag{1}$$

$$\tau_{\text{vol}} = \operatorname{quantize}\Big(\sqrt{\tfrac{1}{k}\sum_{j=1}^{k}(x^{(1)}_{t-j})^2}\Big), \tag{2}$$

$$\tau_{\text{pos}} = \begin{cases} +1 & x^{(1)}_t \geq \max_{j\leq w} x^{(1)}_{t-j} - \sigma \\ -1 & x^{(1)}_t \leq \min_{j\leq w} x^{(1)}_{t-j} + \sigma \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $x^{(1)}$ is the most recent log-return. The volatility quantizer uses tercile thresholds computed *once* on a fixed reference history and *frozen*: recomputing them on a growing online sample would relabel the same day differently as training proceeds, making τ non-stationary. Freezing keeps the regime definition stable with no per-decision look-ahead.

Remark 1 (State augmentation)**.** *Conditioning on τ is equivalent to augmenting the state to $\tilde{\mathcal{X}} = \mathcal{X} \times \mathcal{T}$, on which the Markov property holds. If τ is a deterministic function of the recent history already in x, it introduces no new information; it merely exposes structure the aggregated reference measure had averaged away.*

2.2 Fidelity

Not all regimes warrant the same confidence. We define a *fidelity* $\phi : \mathcal{T} \to [0, 1]$,

$$\phi(\tau) = \min\Big\{1, \tfrac{N_\tau}{N_{\min}}\Big\} \cdot \Big(1 - \tfrac{H_\tau}{\log B}\Big), \tag{4}$$

where N_τ counts transitions tagged τ, and H_τ is the Shannon (differential) entropy of the next-return distribution within regime τ,

$$H_\tau = -\int p_\tau(r)\, \log p_\tau(r)\, dr, \tag{5}$$

with p_τ the density of the next log-return conditional on regime τ. In practice p_τ is unknown; H_τ is *estimated* by discretizing the next-return into B bins. The value we use is the *held-out cross-entropy*: each regime's transitions are split into a fit half and a held-out half, $\hat{p}_\tau$ is estimated on the fit half, and the score is the cross-entropy of the held-out half under $\hat{p}_\tau$. Conditioning on τ alone (not the full (x, a, τ)) avoids the degeneracy $N(x, a, \tau) \approx 1$ of continuous states.

Remark 2 (Precision versus accuracy)**.** *An in-sample entropy rewards regimes whose data are tightly clustered—i.e.* precision*, not* accuracy*. A regime can be precisely estimated and still fail to generalize at a regime change; an in-sample ϕ would then shrink the ball exactly when the estimate is most likely wrong. The held-out cross-entropy makes ϕ a measure of* predictive *validity: if the regime does not generalize, $H_\tau > \log B$ and the $(1 - H_\tau / \log B)$ factor is clamped to zero, withholding confidence.*

2.3 Blended reference and adaptive radius

With ϕ in hand we form a context-aware reference by convex shrinkage toward the regime-specific estimate,

$$\tilde{P}(x, a, \tau) = (1 - \phi(\tau))\, \hat{P}(x, a) + \phi(\tau)\, \hat{P}(x, a, \tau), \tag{6}$$

and make the radius confidence-adaptive,

$$\tilde{\varepsilon}(\tau) = \varepsilon\, (1 - \beta\, \phi(\tau)). \tag{7}$$

High-fidelity regimes (many samples, low held-out entropy) receive a *tighter* ball around a *more local* center; low-fidelity regimes revert to the global, wider ball. Substituting (6)–(7) into the RDQN robust operator (Appendix A) gives the context-conditional robust Q-operator.

Proposition 1 (Asymptotic consistency)**.** *If regimes are identifiable, $P^*(\cdot|\tau_1) \neq P^*(\cdot|\tau_2)$ for $\tau_1 \neq \tau_2$, then as $N_\tau \to \infty$, $W_\delta(\tilde{P}(x, a, \tau), P^*(x, a, \tau)) \to 0$ in probability at a faster rate than for the context-agnostic estimator $\hat{P}(x, a)$.*

3 Regime-Conditional Opportunism

The context-aware system of §2 shrinks the ambiguity *set* but keeps a single decision *criterion*—the worst case—everywhere. The rule below, which we propose but do not validate empirically in this paper, would make the criterion itself regime-dependent.

3.1 Trend quality and the opportunism weight

Define a *trend-quality* $q : \mathcal{T} \to [0, 1]$, read directly off τ, that is high only for unambiguously favorable, past-only regimes (positive drift, low volatility, proximity to local maxima) and driven to 0 otherwise ($q(\tau) = 0$ whenever $\tau_{\text{trend}} \leq 0$ or $\tau_{\text{vol}} = \text{high}$). The *opportunism weight* is

$$\lambda(\tau) = \lambda_{\max}\, \phi(\tau)\, q(\tau) \in [0, \lambda_{\max}]. \tag{8}$$

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

3.2 The blended Bellman target

Let Q^{R} denote the worst-case target of the context-aware system (§2) and Q^{N} the *nominal* (risk-neutral) target under the blended reference $\tilde{P}$:

$$Q^{\mathrm{R}}(x,a,\tau) = \inf_{P\in\mathcal{B}_{\tilde{\varepsilon}(\tau)}(\tilde{P}(\tau))} \mathbb{E}_P\big[r + \alpha \max_b Q^*(X',b,\tau')\big], \tag{9}$$

$$Q^{\mathrm{N}}(x,a,\tau) = \mathbb{E}_{\tilde{P}(\tau)}\big[r + \alpha \max_b Q^*(X',b,\tau')\big]. \tag{10}$$

The CARDQN operator blends them:

$$\boxed{Q^*(x,a,\tau) = \big(1-\lambda(\tau)\big)\, Q^{\mathrm{R}}(x,a,\tau) + \lambda(\tau)\, Q^{\mathrm{N}}(x,a,\tau).} \tag{11}$$

Theorem 1 (Never-optimistic bound)**.** *For all τ, $Q^{\mathrm{R}}(x,a,\tau) \le Q^*(x,a,\tau) \le Q^{\mathrm{N}}(x,a,\tau)$. In particular the CARDQN target never exceeds the nominal: the agent de-hedges toward risk-neutrality but never becomes risk-seeking.*

Proof. The worst case is an infimum, $Q^{\mathrm{R}} = \inf_{P\in\mathcal{B}_{\tilde{\varepsilon}}(\tilde{P})} \mathbb{E}_P[\cdot]$. The reference $\tilde{P}$ itself belongs to the ball $\mathcal{B}_{\tilde{\varepsilon}}(\tilde{P})$, so the infimum is bounded above by the value at $\tilde{P}$, which is exactly Q^{N}; hence $Q^{\mathrm{R}} \le Q^{\mathrm{N}}$. The blend (11) is a convex combination with $\lambda \in [0, \lambda_{\max}] \subseteq [0,1]$, so $Q^* \in [Q^{\mathrm{R}},\, (1-\lambda_{\max})Q^{\mathrm{R}} + \lambda_{\max} Q^{\mathrm{N}}] \subseteq [Q^{\mathrm{R}}, Q^{\mathrm{N}}]$. □ □

Thus in adverse or low-fidelity regimes $\lambda \to 0$ and (11) reduces to the pure robust operator; in favorable, high-fidelity regimes $\lambda \to \lambda_{\max}$ and it approaches, but never reaches, the nominal.

Remark 3 (A regime-dependent risk measure)**.** *The blend* (11) *is a convex combination of two coherent risk measures—the worst case and the expectation—and is therefore a coherent,* regime-dependent *spectral risk measure: the linear interpolation between the endpoints of a CVaR family $\rho_\eta = \mathrm{CVaR}_{\eta(\tau)}$ with η large (tail-sensitive) in adverse regimes and small (near the mean) in favorable ones.*

Remark 4 (Why smooth, not switched)**.** *A hard if/else switch of the objective at a regime boundary would make the Bellman target discontinuous in the transient context, destabilizing learning. The smooth ϕ-gated blend avoids this; conditioning Q on τ lets the network represent regime-conditional values consistently.*

Remark 5 (Asymmetric by design)**.** *The default is the worst case; relaxation requires* both *a favorable regime (q high) and confidence (ϕ high), capped below nominal. This bounds the cost of the irreducible failure mode—a favorable label assigned just before an adverse move—at the finite gap $Q^{\mathrm{N}} - Q^{\mathrm{R}}$, never at an unbounded optimistic bet.*

4 Theoretical Properties of CARDQN

CARDQN inherits RDQN's robustness where it matters and relaxes it where it is safe:

- **Robustness where uncertain.** In any regime with $\phi(\tau) = 0$ or $q(\tau) = 0$, $\lambda = 0$ and (11) reduces to the worst-case operator on $\mathcal{B}_{\tilde{\varepsilon}}(\tilde{P})$—full RDQN robustness, with the ball tightened by the context-aware radius where ϕ is high.
- **Bounded opportunism where confident.** Where both ϕ and q are high, the target lifts by at most $\lambda_{\max}(Q^{\mathrm{N}} - Q^{\mathrm{R}})$, a finite, data-dependent margin (Theorem 1).
- **Consistency.** As data accumulates, $\tilde{P} \to P^*(\cdot|\tau)$ faster than $\hat{P} \to P^*$ (Proposition 1), so the nominal anchor Q^{N} itself becomes a reliable estimate of the regime-conditional value.

5 Results: Original RDQN versus CARDQN

5.1 Setup and compute

We compare the original RDQN against CARDQN—here the context-aware system of §2; the opportunism rule of §3 is a proposed extension and is not part of the evaluated system. Both agents train on a signature-MMD market simulator—synthetic paths fitted to the S&P 500—and are evaluated out-of-sample on the empirical S&P 500, 1995–2024, with 0.05% proportional transaction cost; figures are means over 5 seeds of a 10-episode budget. Runs were executed on the UNIBO DISI *giano* cluster under SLURM, one NVIDIA RTX 2080 Ti per run, with the market environment on the CPU and the Q-network plus the inner Sinkhorn-dual solve on the GPU; the dual variable λ is vectorized as a single batched tensor rather than one parameter per sample, which is the main throughput consideration. An episode takes roughly 10–14 minutes on this hardware.

Reading. CARDQN preserves the defining RDQN property—much lower volatility and drawdown than buy-and-hold—while surpassing the original authors' reported result on the headline metrics: more terminal wealth (3.40× vs 2.89×), higher Sharpe and Sortino, at even volatility and a comparable drawdown. As a same-codebase control we also reproduce the authors' RDQN through the CARDQN code path (context-aware layer off) at

Agent	Wealth	Volatility	Sharpe	Sortino	Max drawdown
S&P 500 buy-and-hold	9.524	0.191	0.410	0.569	−0.568
Lu et al. (2025) RDQN, reported	2.89	0.121	0.265	0.373	−0.371
CARDQN (ours)	**3.401**	0.132	**0.297**	**0.413**	−0.387
RDQN, our reproduction (same budget)	1.588	0.108	0.131	0.182	−0.377

Table 1: Empirical S&P 500, 1995–2024. CARDQN (5-seed mean, 10-episode budget) versus the original authors' reported RDQN: CARDQN delivers more terminal wealth at even volatility, with higher Sharpe and Sortino and a comparable drawdown. The final row is our independent reproduction of the authors' RDQN through the CARDQN code path at the same budget (a same-codebase control).

the same 5-seed, 10-episode budget; its 1.59× wealth confirms the gain is attributable to the context-aware machinery, not to implementation differences.

6 Conclusion

CARDQN makes the RDQN ambiguity set context-aware, replacing an aggregated, oversized ball with a regime-aware one whose center and radius adapt to local, held-out-validated confidence. The operator is coherent, asymptotically consistent, and empirically surpasses the original RDQN's reported result—more terminal wealth at even volatility, with higher Sharpe and Sortino and a comparable drawdown—while preserving its low volatility and drawdown. The regime-conditional opportunism rule of §3 is proposed as a complementary extension and is left to future validation.

A The Original RDQN System

This appendix records the legacy system CARDQN builds on: the trading MDP, the robust operator, the worst-case prior, and the Sinkhorn-dual inner solve.

The MDP. The trading task is a Markov decision process with state space $\mathcal{X}$ (a window of past log-returns augmented with log-wealth, position and a time feature), action space $\mathcal{A}$ (discrete portfolio weights), discount $\alpha \in (0, 1)$, and reward $r(x, a, x') = \log G$ where G is the portfolio's simple growth factor over the step (so cumulative reward telescopes to log terminal wealth).

The robust operator. Given a reference transition law $\hat{P}(x,a)$ and a Sinkhorn (entropic optimal-transport) ball

$$\mathcal{B}_{\varepsilon,\delta}(\hat{P}) = \big\{P \in \mathcal{M}_1(\mathcal{X}) \,:\, W_\delta(\hat{P},P) \le \varepsilon\big\}, \tag{12}$$

with W_δ the entropic-transport cost (regularization δ) and radius ε, the RDQN robust Q-operator is

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

The infimum is over an adversarially chosen transition law: the agent learns against the *worst* distribution within transport distance ε of $\hat{P}$, which is what makes the policy robust (and, by lowering the target relative to the nominal, conservative). The radius ε controls how much model error the agent tolerates: large ε is safe but conservative; small ε is aggressive but fragile.

The worst-case prior ν. The reference measure is represented operationally by a *cloud* of candidate next-states drawn from a sampling distribution ν (Student-t in our setting). ν is more than a numerical device: it is a *prior over the worst case*, because the inner transport problem aligns the adversarial distribution with the support and tail structure of ν. Gaussian ν is inadequate for returns—realised financial returns exhibit fat tails driven by behavioral mechanisms (herding, panic, overshooting) and exogenous shocks (defaults, geopolitical events). A heavy-tailed Student-t ν places mass on the tails the adversary may exploit, yielding a worst case that genuinely stress-tests drawdown behavior. The choice of ν is therefore a modeling decision: heavier-tailed ν for tail-risk-sensitive mandates; a bounded/uniform ν for zero-drift, high-volatility environments where no outcome should be privileged a priori.

The Sinkhorn-dual inner solve ("HQ"). Let $f(x') = r(x,a,x') + \alpha \max_b Q(x',b)$ and let the reference cloud be $\{x'_j\}_{j=1}^n$ with transport cost $c_{j\ell} = \|x'_j - x'_\ell\|$. The worst-case value admits the dual

$$\inf_{P:\, W_\delta(P,\hat{P})\le\varepsilon} \mathbb{E}_P[f] \;=\; \sup_{\lambda>0}\Big[-\lambda\Big(\varepsilon + \delta\,\mathbb{E}_{\text{outer}}\big[\log \mathbb{E}_{\text{inner}}[\exp\big((-f/\delta\lambda) - c/\delta\big)]\big]\Big)\Big], \tag{14}$$

a one-dimensional concave maximization in $\lambda > 0$: the primal *infimum* over distributions (the worst case) becomes a dual *supremum* over λ (the

inner expectation is over cloud points, the outer over the reference). For each minibatch transition the agent ascends λ by gradient ascent—we use Nesterov-accelerated Adam, which reaches the dual supremum faster than plain Adam without altering the optimum—to recover the worst-case target, fed into the standard squared Bellman error.

Remark 6 (Numerical invariants)**.** *The dual is invariant to additive shifts of f (handled by subtracting a running max before exponentiation, the log-sum-exp stabilization), and per-sample feasibility is respected: when the effective radius is non-positive the transition carries no usable hedge and is masked out of the loss. These are numerical necessities, not modeling choices.*