# BAYESIAN ADVERSARIAL DISTRIBUTIONAL REINFORCEMENT LEARNING FOR SAMPLE-EFFICIENT RISK-SENSITIVE CONTROL

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Distributional Reinforcement Learning (DRL) enables risk-sensitive decision-making by learning the distribution of cumulative returns rather than only their expectation. However, accurately estimating return distributions can require substantial interaction with the environment, making DRL difficult to apply when data collection is risky, costly, or limited. With limited experience, uncertainty in the learned return distribution can remain substantial, yet many widely used DRL methods do not explicitly quantify epistemic uncertainty in the learned return model. We propose *GAN-DRL*, a Bayesian Generative Adversarial Network (GAN) framework for DRL that combines Wasserstein-based return-distribution learning with Bayesian inference to quantify epistemic uncertainty in the learned return model. We use posterior information gain as an intrinsic signal to guide the agent toward informative experience, improving learning efficiency in stochastic risk-sensitive environments. We evaluate GAN-DRL on two continuous-control problems: European option hedging and F1TENTH autonomous driving. Across both domains, GAN-DRL reaches comparable or improved risk-sensitive performance using substantially fewer environment interactions than distributional and actor-critic baselines. In European option hedging, GAN-DRL improves tail-risk performance, as measured by Conditional Value-at-Risk (CVaR), while maintaining competitive average performance. In autonomous driving, GAN-DRL improves tail returns and reduces rollover rates on both the Offroad-Flat and Offroad-Bumpy tasks. Overall, these results show that uncertainty-guided distributional learning can substantially improve sample efficiency while maintaining or improving risk-sensitive performance.

## 1 INTRODUCTION

Many sequential decision-making problems are inherently risk-sensitive: two policies with similar expected returns may have substantially different probabilities of rare but costly outcomes. This distinction is particularly important in domains such as financial decision-making and autonomous control, where tail performance can be as important as average performance. Reinforcement Learning (RL) methods typically optimize expected cumulative return, which may obscure important variation in the distribution of possible outcomes. Distributional Reinforcement Learning (DRL) goes beyond this expectation-based objective by learning the distribution of cumulative returns Bellemare et al. (2017); Dabney et al. (2018). Access to the return distribution makes it possible to account explicitly for tail behavior and risk measures such as Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR), making DRL a natural framework for risk-sensitive control Clavier et al. (2022); Cao et al. (2023).

A central challenge, however, is *sample efficiency*. Deep RL methods can require large numbers of environment interactions before learning reliable policies Schwarzer et al. (2021); D'Oro et al. (2023); Bhatt et al. (2024). In DRL, limited interaction data must additionally support estimation of the return distribution itself. This requirement is particularly problematic in risk-sensitive settings, where collecting additional experience may be costly, slow, or unsafe. In financial applications, real market observations cannot be generated on demand, while exploratory behavior in physical control systems can lead to undesirable or unsafe outcomes. The relevant question is therefore not only



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.

whether a DRL method can ultimately learn an effective risk-sensitive policy, but whether it can do so with substantially fewer interactions *without sacrificing final risk-sensitive performance*.

Generative approaches provide a flexible way to represent return distributions. In particular, the distributional Bellman update can be formulated as a distribution-matching problem, allowing Generative Adversarial Networks (GANs) to implicitly represent complex return distributions Doan et al. (2018); Freirich et al. (2019). Wasserstein-based objectives are particularly natural in this setting because they provide a principled discrepancy between a learned return distribution and its Bellman target Bellemare et al. (2017); Freirich et al. (2019). However, adversarial distribution matching alone does not explicitly quantify the learner's epistemic uncertainty about its current return estimate. With limited interaction data, some regions of the state–action space may be well characterized while others remain poorly understood, motivating a mechanism that can identify where additional experience would be most informative.

Bayesian inference provides a principled framework for representing such epistemic uncertainty. Changes in a posterior distribution induced by new observations can be quantified through *information gain* and used to direct exploration toward informative experience Houthooft et al. (2016); Li & Faisal (2021). Motivated by this connection, we investigate whether Bayesian uncertainty can be coupled with Wasserstein adversarial return-distribution learning to substantially reduce interaction requirements while preserving, or even improving, risk-sensitive performance.

To this end, we propose *GAN-DRL*, a Bayesian adversarial framework for DRL that combines Wasserstein-based return-distribution learning with Bayesian inference over the learned return model. Posterior information gain is used as an intrinsic exploration signal to direct the agent toward informative experience. The adversarial component provides a flexible representation of the return distribution, while the Bayesian component quantifies uncertainty in that representation and guides data collection toward experience that is most informative for learning it. The resulting framework is designed to improve sample efficiency without giving up the risk-sensitive benefits of distributional learning.

We evaluate GAN-DRL on two continuous-control problems with distinct manifestations of risk: 30-day European option hedging and F1TENTH autonomous driving. In option hedging, risk appears through severe losses in the tail of the hedging-return distribution; in autonomous driving, it appears through tail returns and physical failure events such as vehicle rollovers. Across both domains, GAN-DRL attains comparable or improved risk-sensitive performance using substantially fewer environment interactions than distributional and actor–critic baselines.

Our main contributions are:

- We develop *GAN-DRL*, a risk-sensitive DRL framework that combines Bayesian adversarial posterior inference, Wasserstein-based return-distribution learning, and posterior information-gain exploration with a quantile-based distributional critic for direct risk-sensitive policy optimization.
- We demonstrate substantial gains in sample efficiency across option hedging and autonomous driving, with GAN-DRL requiring significantly fewer environment interactions to learn effective risk-sensitive policies than the considered baselines.
- We show that these gains are accompanied by strong and stable final risk-sensitive performance. GAN-DRL achieves improved tail CVaR performance across both domains, exhibits consistent performance across independent runs, and reduces rollover rates in autonomous driving.

## 2 Related Work

**Risk-sensitive RL for financial decision-making.** RL has been extensively applied to portfolio optimization, trading, and option hedging Nevmyvaka et al. (2006); Buehler et al. (2019); Halperin (2020); Kolm & Ritter (2019); Cao et al. (2021); Vittori et al. (2020); Murray et al. (2022); Gao et al. (2023); Daluiso et al. (2023); Neagu et al. (2025). Recent work has increasingly incorporated tail-sensitive objectives and measures, including expectiles and CVaR, as well as uncertainty in the assumed market model Marzban et al. (2023); Wu & Jaimungal (2023); Zheng et al. (2023); Lütkebohmert et al. (2022). Within option hedging, distributional and quantile-based methods have been used to estimate and optimize tail-risk measures directly Cao et al. (2023); Sharma et al. (2024). Because extreme quantiles can be difficult to estimate from limited tail observations, related work has also combined DRL with extreme-value models to improve tail estimation Malekzadeh et al.

(2024). These approaches primarily improve the representation or estimation of tail risk rather than uncertainty about the learned return model itself.

**Generative distributional RL.** GAN-based approaches provide an implicit representation of return distributions and formulate Bellman learning as an adversarial distribution-matching problem Doan et al. (2018); Freirich et al. (2019). GAN Q-learning uses a generator to model the state–action return distribution, while Bellman-GAN develops a Wasserstein-based formulation in which the learned return distribution is matched to its Bellman target. Bellman-GAN further proposes exploration based on distributional prediction discrepancy Freirich et al. (2019). Our adversarial component builds on this general perspective, while augmenting the learned return model with Bayesian uncertainty estimation.

**Bayesian uncertainty and robust decision-making.** Bayesian exploration methods use information gain to identify experience that is most informative for reducing uncertainty, either in the environment dynamics or in the learned return distribution Houthooft et al. (2016); Li & Faisal (2021). Our approach combines Bayesian adversarial inference with Wasserstein-based return-distribution learning to quantify uncertainty in the learned return model and guide exploration. It further couples uncertainty-guided exploration with a quantile-based distributional critic for direct risk-sensitive policy optimization.

A complementary line of work addresses uncertainty in the environment model itself. In option hedging, robust methods seek strategies that remain effective under parameter misspecification or deviations from a reference market model Lütkebohmert et al. (2022); Limmer & Horvath (2024). For example, robust hedging GANs adversarially perturb the market model and penalize deviations from a reference distribution. In contrast, GAN-DRL uses adversarial inference within the learned return model rather than perturbing the environment dynamics.

## 3 Background

We review distributional value learning, adversarial latent inference, transport costs, and the risk measures used in our framework. Throughout the paper, uppercase letters denote random variables and lowercase letters denote their realizations.

### 3.1 Distributional Reinforcement Learning

We consider a Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $P(\cdot \mid s, a)$ is the transition kernel, $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the one-step reward function, and $\gamma \in (0, 1]$ is the discount factor. An agent follows a policy $\pi(\cdot \mid s)$ over actions.

For a horizon $T$, the discounted return from a state–action pair $(s, a)$ is the random variable

$$G^\pi(s, a) = \sum_{t=0}^{T-1} \gamma^t R_t, \qquad S_0 = s,\ A_0 = a, \tag{1}$$

where $R_t = r(S_t, A_t, S_{t+1})$, $S_{t+1} \sim P(\cdot \mid S_t, A_t)$, and $A_{t+1} \sim \pi(\cdot \mid S_{t+1})$. We denote the corresponding return distribution by $Z^\pi(s, a)$. Conventional value learning estimates its expectation, $Q^\pi(s, a) = \mathbb{E}[Z^\pi(s, a)]$ , whereas DRL models the return distribution itself.

Under a fixed policy $\pi$, the distributional Bellman equation is

$$Z^\pi(s, a) \overset{D}{=} \mathcal{T}^\pi Z^\pi(s, a) := r(s, a, S') + \gamma Z^\pi(S', A'), \tag{2}$$

where $S' \sim P(\cdot \mid s, a)$, $A' \sim \pi(\cdot \mid S')$, and $\overset{D}{=}$ denotes equality in distribution.

For $\gamma < 1$ and distributions with finite $p$th moments, the fixed-policy distributional Bellman operator is a $\gamma$-contraction under the maximal $p$-Wasserstein metric Bellemare et al. (2017),

$$\bar{d}_p(Z_1, Z_2) = \sup_{s,a} W_p(Z_1(s, a), Z_2(s, a)) , \tag{3}$$

such that

$$\bar{d}_p \left(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2\right) \leq \gamma\, \bar{d}_p(Z_1, Z_2). \tag{4}$$

This result motivates learning return distributions by matching the current estimate to its distributional Bellman target. The contraction result applies to policy evaluation under a fixed policy; the

distributional optimality operator is not, in general, a contraction under the same metric Bellemare et al. (2017).

**Quantile representation.** A common parametric representation of the return distribution is Quantile Regression (QR) Dabney et al. (2018). Using $N$ equally weighted quantile locations,

$$Z_\theta(s,a) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\beta_\theta^{\hat{\tau}_i}(s,a)}, \qquad \hat{\tau}_i = \frac{i-\frac{1}{2}}{N}, \tag{5}$$

where $\beta_\theta^{\hat{\tau}_i}(s,a)$ approximates the $\hat{\tau}_i$-quantile of the return distribution.

Given a sampled transition $(s,a,r,s')$, target quantiles can be constructed from the distributional Bellman update as

$$y_j = r + \gamma \beta_{\bar{\theta}}^{\hat{\tau}_j}(s',a'), \qquad a' \sim \pi(\cdot \mid s'), \tag{6}$$

where $\bar{\theta}$ denotes target-network parameters when a target network is used. The predicted quantiles are then fitted to these target samples, typically using the quantile Huber loss Dabney et al. (2018). QR provides a tractable finite representation of the return distribution, but its estimation quality ultimately depends on the interaction data available for constructing Bellman targets.

### 3.2 Wasserstein Distribution Matching

Generative models provide a flexible implicit representation of continuous return distributions. In generative DRL, the distributional Bellman update can be viewed as a distribution-matching problem between the learned return distribution and its Bellman target Doan et al. (2018); Freirich et al. (2019).

For probability distributions $P$ and $Q$, the $p$-Wasserstein distance is defined as

$$W_p^p(P,Q) = \inf_{\Gamma \in \Pi(P,Q)} \mathbb{E}_{(X,Y)\sim\Gamma}\left[d_p^p(X,Y)\right], \tag{7}$$

where $\Pi(P,Q)$ denotes the set of couplings with marginals $P$ and $Q$, and $d_p$ is the corresponding ground metric. This transport-cost formulation provides the basis for the return-reconstruction objective used in our generative model.

### 3.3 Bayesian Posterior Inference and Information Gain

A standard generative model maps latent variables to observations but does not by itself provide posterior inference from an observation back to its latent representation. Adversarial inference methods, such as adversarially learned inference Dumoulin et al. (2017), introduce an encoder $q_\psi(z \mid x)$ and match the joint distribution induced by the encoder with that induced by the generative model. In a conditional setting, these joint distributions can be written as

$$q_\psi(x,z \mid s,a) = P_X(x \mid s,a)\, q_\psi(z \mid x,s,a), \tag{8}$$

and

$$p_\theta(x,z \mid s,a) = P_Z(z \mid s,a)\, p_\theta(x \mid z,s,a). \tag{9}$$

Matching the two joint distributions provides an amortized approximation to the latent posterior associated with the generative model.

Posterior changes induced by newly observed information can be quantified through information gain. Given a posterior distribution $q_{\text{post}}$ and a reference distribution $q_{\text{prior}}$, we define

$$\text{IG} = D_{\text{KL}}\left(q_{\text{post}} \,\|\, q_{\text{prior}}\right), \tag{10}$$

where $D_{\text{KL}}$ denotes the Kullback–Leibler (KL) divergence. Information-gain-based exploration uses this posterior change as a signal for identifying experience that is informative about the learner's current model Houthooft et al. (2016); Li & Faisal (2021).

### 3.4 Risk Measures and Return Distributions

Throughout this work, larger cumulative returns correspond to more favorable outcomes, and adverse outcomes therefore lie in the lower tail of the return distribution. This convention is independent of the particular application; the physical interpretation of a low return is specified separately for each experimental environment.

Let $F_{s,a}$ denote the cumulative distribution function of $Z^\pi(s,a)$, and define its quantile function as

$$q_u(s,a) = F_{s,a}^{-1}(u) = \inf\{z : F_{s,a}(z) \geq u\}, \qquad u \in (0,1). \tag{11}$$

For a confidence level $\alpha \in (0,1)$, we define the tail VaR as $\mathrm{VaR}_\alpha(Z^\pi(s,a)) = q_{1-\alpha}(s,a)$, and the corresponding CVaR as

$$\mathrm{CVaR}_\alpha(Z^\pi(s,a)) = \frac{1}{1-\alpha}\int_0^{1-\alpha} q_u(s,a)\,du. \tag{12}$$

For continuous return distributions, CVaR is equivalently the expected return within the worst $100(1-\alpha)\%$ of outcomes. Under this return-based convention, larger VaR and CVaR values indicate better tail performance. For QR-based return models, VaR can be approximated by the estimated quantile nearest to level $1-\alpha$, while CVaR can be approximated from the estimated quantiles below this level. For generative return models, the same quantities can be estimated empirically from samples drawn from the learned return distribution.

At high confidence levels, only a small fraction of observations contributes to these tail quantities. For example, $\mathrm{CVaR}_{0.95}$ depends on the worst $5\%$ of the return distribution. Accurate estimation of such quantities can therefore be particularly data demanding under limited interaction budgets Bai et al. (2021); Beranger et al. (2021). This motivates learning reliable tail behavior while using interaction data efficiently.

## 4 Proposed Methodology

We propose *GAN-DRL*, a Bayesian adversarial framework for DRL that combines Bayesian adversarial posterior inference, Wasserstein-based return-distribution learning, posterior information-gain exploration, and risk-sensitive policy optimization. Figure 1 illustrates the overall framework.

Given an observed transition $(s_t, a_t, r_t, s_{t+1})$, the generative return model produces a current return sample at $(s_t, a_t)$ and a next-state return sample used to construct its distributional Bellman target. An inference network maps both the current return sample and its Bellman target to latent posterior distributions, whose discrepancy defines an information-gain signal for exploration. Separately, a quantile-based distributional critic estimates the return distribution used for direct risk-sensitive policy optimization.

### 4.1 Bellman Target Construction and Return Generation

Let $G_\theta(s,a,z)$ denote the conditional return generator, where $z \sim P_Z$ is a latent variable and $\theta$ denotes the generator parameters. For the current state–action pair $(s_t, a_t)$, we draw $z_t \sim P_Z$ and generate $g_t = G_\theta(s_t, a_t, z_t)$, which is a sample from the current estimate of the return distribution $Z_\theta(s_t, a_t)$.

We sample $z'_t \sim P_Z$ and construct a Bellman-target return sample as

$$x_t = r_t + \gamma G_{\bar{\theta}}(s_{t+1}, a_{t+1}, z'_t), \tag{13}$$

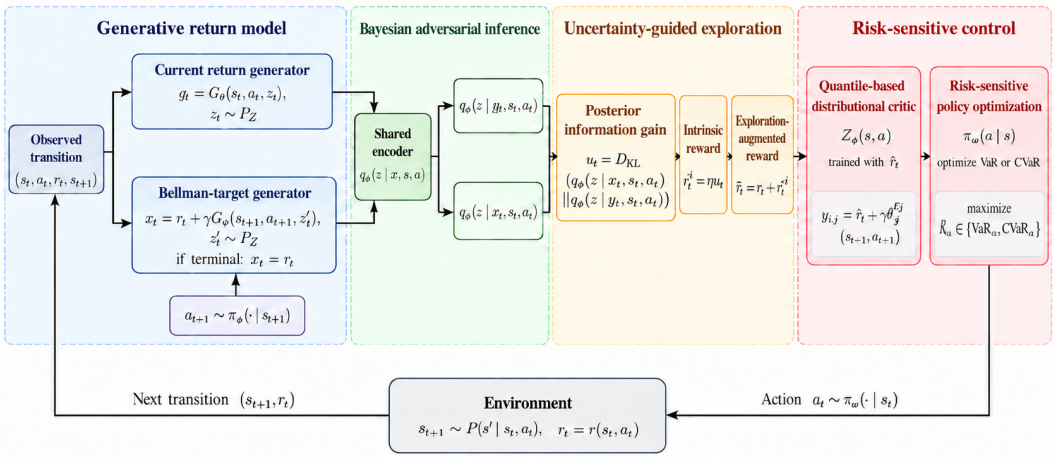



Figure 1: Overview of the proposed GAN-DRL framework.

where $\bar{\theta}$ denotes the target-generator parameters and $a_{t+1}$ is selected according to the policy used for the Bellman backup. If $s_{t+1}$ is terminal, we set $x_t = r_t$. Thus, $g_t$ represents the current return prediction at $(s_t, a_t)$, whereas $x_t$ represents its one-step distributional Bellman target.

### 4.2 Bayesian Adversarial Inference

To infer uncertainty in the learned return representation, we introduce an encoder $q_\psi(z \mid x, s, a)$, parameterized by $\psi$, that maps a return observation and its state–action condition to a distribution over latent variables.

For each state–action pair $(s, a)$, the target and generative joint distributions are factorized as

$$q_\psi(x, z \mid s, a) = P_{\text{target}}(x \mid s, a) q_\psi(z \mid x, s, a), \tag{14}$$

and

$$p_{\bar{\theta}}(x, z \mid s, a) = P_Z(z)\, p_{\bar{\theta}}(x \mid z, s, a), \tag{15}$$

respectively. For a deterministic generator, $p_{\bar{\theta}}(x \mid z, s, a)$ is concentrated at $G_{\bar{\theta}}(s, a, z)$.

Following adversarial posterior inference, the generator is held fixed while the discriminator and encoder are trained by matching the two joint distributions:

$$\begin{aligned} \min_{\psi} \max_{\phi} \; & \mathbb{E}_{z \sim P_Z} \left[ \log D_\phi\big(s, a, G_{\bar{\theta}}(s, a, z), z\big) \right] \\ & + \mathbb{E}_{x \sim P_{\text{target}}(\cdot \mid s, a)} \mathbb{E}_{z \sim q_\psi(\cdot \mid x, s, a)} \left[ \log\left(1 - D_\phi(s, a, x, z)\right) \right]. \end{aligned} \tag{16}$$

No gradients are propagated through $G_{\bar{\theta}}$ during this step. The adversarial matching encourages $q_\psi(z \mid x, s, a)$ to approximate the latent posterior induced by the current generative return model.

### 4.3 Wasserstein Return Reconstruction

With the posterior encoder fixed, the generator is updated using a Wasserstein-motivated reconstruction objective:

$$\min_{\theta} \; \mathbb{E}_{x \sim P_{\text{target}}(\cdot \mid s, a)} \mathbb{E}_{z \sim q_\psi(\cdot \mid x, s, a)} \left[ d_p^p\left(x, G_\theta(s, a, z)\right) \right], \tag{17}$$

where gradients are not propagated through $q_\psi$ during this update and $d_p$ denotes the ground metric associated with the $p$-Wasserstein distance. With a Euclidean ground metric and $p = 2$, this reduces to the squared reconstruction loss $\|x - G_\theta(s, a, z)\|_2^2$.

Training therefore alternates between adversarial posterior inference in equation 16 and Wasserstein return reconstruction in equation 17.

### 4.4 Information-Gain Exploration

The inferred posterior enables us to quantify how informative an observed transition is about the current return model. Following the posterior information-gain construction, we compare the latent distribution inferred from the Bellman target with that inferred from the current return prediction:

$$u_t = D_{\text{KL}}\left(q_\psi\left(z \mid x_t, s_t, a_t\right) \,\|\, q_\psi\left(z \mid g_t, s_t, a_t\right)\right). \tag{18}$$

The distribution $q_\psi(z \mid g_t, s_t, a_t)$ represents the model's return belief before incorporating the new Bellman target, while $q_\psi(z \mid x_t, s_t, a_t)$ reflects the information supplied by the observed transition. Consequently, a larger $u_t$ indicates a larger change in the model's latent belief.

The information gain is converted into an intrinsic exploration reward, $r_t^{\text{IG}} = \eta_t u_t$, where $\eta_t = \eta_0 \sqrt{(\log t)/t}$ decays the influence of information-gain exploration over training, and $\eta_0 > 0$ denotes its initial scale. The exploration-augmented reward used to train the distributional critic is

$$\widetilde{r}_t = r_t + r_t^{\text{IG}}. \tag{19}$$

The original environment reward $r_t$, rather than $\widetilde{r}_t$, is used in equation 13 when training the generative return model. Thus, information gain guides exploration without changing the environment-return distribution learned by the Bayesian generative model.

### 4.5 Risk-Sensitive Policy Optimization

In addition to the generative return model used for posterior inference, GAN-DRL maintains a quantile-based distributional critic $Z_\varphi(s, a)$ for risk-sensitive policy optimization. The critic is trained using the exploration-augmented reward $\widetilde{r}_t$ from equation 19. For quantile locations $\{\tau_j\}_{j=1}^N$,

the Bellman targets are

$$y_{t,j} = \widetilde{r}_t + \gamma \beta_{\bar{\varphi}}^{\tau_j}(s_{t+1}, a_{t+1}), \tag{20}$$

where $a_{t+1}$ is selected according to the policy used for the Bellman backup, with the future-return term omitted for terminal transitions. The critic parameters $\varphi$ are updated using the quantile Huber loss.

The policy $\pi_\omega$ is optimized with respect to a selected risk measure, either VaR or CVaR. Let

$$j_\alpha = \arg\min_j \left|\tau_j - (1-\alpha)\right|, \qquad \mathcal{I}_\alpha = \{j : \tau_j \le 1-\alpha\}. \tag{21}$$

The corresponding estimates are

$$\widehat{\mathrm{VaR}}_\alpha = \beta_{\varphi}^{\tau_{j_\alpha}}(s,a), \qquad \widehat{\mathrm{CVaR}}_\alpha = \frac{1}{|\mathcal{I}_\alpha|} \sum_{j \in \mathcal{I}_\alpha} \beta_{\varphi}^{\tau_j}(s,a). \tag{22}$$

Let $\widehat{\mathcal{R}}_\alpha \in \left\{\widehat{\mathrm{VaR}}_\alpha, \widehat{\mathrm{CVaR}}_\alpha\right\}$ denote the selected risk objective. Since larger returns are preferred, the policy maximizes

$$J_{\mathrm{risk}}(\omega) = \mathbb{E}\left[\widehat{\mathcal{R}}_\alpha\left(Z_\varphi(S,A)\right)\right], \qquad A \sim \pi_\omega(\cdot \mid S), \tag{23}$$

where the expectation is taken over the state–action samples used by the underlying policy-learning algorithm. This formulation can be incorporated into either on-policy or off-policy actor–critic methods. The complete GAN-DRL training procedure is provided in Algorithm 1 in Appendix A.

## 5 Experiments

We evaluate GAN-DRL on two risk-sensitive continuous-control environments: European option hedging Cao et al. (2023) and F1TENTH autonomous driving O'Kelly et al. (2020). The former evaluates downside-risk control in a financial setting, while the latter evaluates risk-sensitive control in a safety-critical domain. Our experiments address three questions: *1)* How quickly does GAN-DRL learn an effective risk-sensitive policy relative to the baselines? *2)* Does improved learning efficiency preserve or improve final risk-sensitive performance? *3)* Does GAN-DRL learn consistently across independent runs?

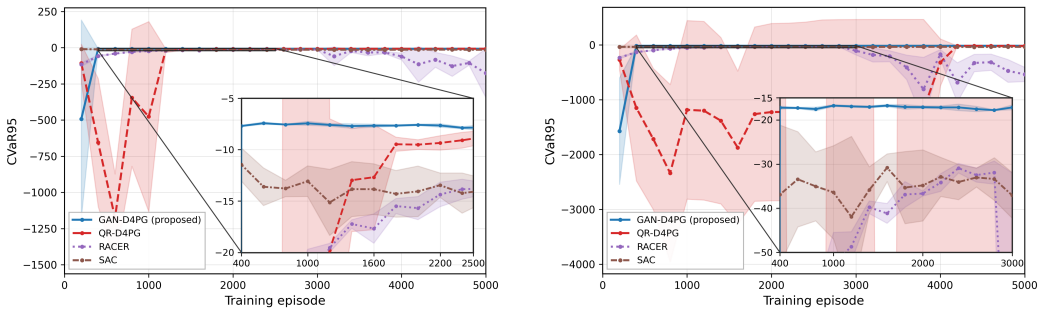
**Metrics.** Sample efficiency is evaluated using risk-sensitive learning curves as a function of training progress. For option hedging, we report $\mathrm{CVaR}_{0.95}$, following the evaluation setting used in the corresponding hedging literature Cao et al. (2023). For autonomous driving, we report $\mathrm{CVaR}_{0.90}$, consistent with the risk-sensitive evaluation protocol used in the reference driving implementation Stachowicz & Levine (2024). Using the task-specific risk levels from the corresponding reference settings allows comparison under established evaluation conventions. Final evaluation reports mean return and the task-specific CVaR measure, with rollover rate additionally reported for autonomous driving as a direct measure of safety-critical failures. Stability is assessed using variability in the learning curves and the standard deviation of final evaluation performance across independent runs.

**Baselines.** We instantiate GAN-DRL on D4PG, denoted *GAN-D4PG*, and compare it with QR-D4PG Cao et al. (2023), SAC Haarnoja et al. (2019), and RACER Stachowicz & Levine (2024), representing distributional, risk-neutral, and risk-sensitive RL baselines, respectively. Full implementation details are provided in Appendix C.

### 5.1 Option Hedging

For the gamma-hedging environment, we follow the experimental setting of Cao et al. (2023); Malekzadeh et al. (2024). Client orders arrive according to a Poisson process with intensity one option per day, each corresponding to a 60-day contract on $100$ units of the underlying asset and equally likely to be a long or short position. A 30-day at-the-money (ATM) option is used for hedging, transaction costs are fixed at $1\%$ of the option price, and the initial stock price is $\$10$. We consider volatility levels $\sigma \in \{0.3, 0.7\}$. Additional environment details are provided in Appendix B.

**1) Sample efficiency.** Figure 2 shows the evolution of $\mathrm{CVaR}_{0.95}$ during training. Higher values correspond to better tail returns; therefore, reaching a given strong CVaR level with fewer interactions




(a) $\sigma = 0.3$ (b) $\sigma = 0.7$

Figure 2: $\mathrm{CVaR}_{0.95}$ learning curves for the two volatility settings. Shading denotes $\pm$ one standard deviation across six runs.

indicates greater sample efficiency. SAC does not explicitly optimize CVaR; its CVaR is computed from realized evaluation returns.

At $\sigma = 0.3$, GAN-D4PG reaches a strong $\mathrm{CVaR}_{0.95}$ level early in training. QR-D4PG improves more gradually before eventually approaching GAN-D4PG, while RACER remains below GAN-D4PG for most of the training horizon. SAC also exhibits weaker tail performance, consistent with its risk-neutral objective.

At $\sigma = 0.7$, the difference is more pronounced. GAN-D4PG rapidly reaches strong $\mathrm{CVaR}_{0.95}$ performance and maintains its advantage through most of training. In contrast, QR-D4PG exhibits large early-training fluctuations before recovering, while RACER and SAC remain at substantially lower CVaR levels. Overall, the learning curves provide evidence that GAN-D4PG reaches strong risk-sensitive performance earlier than the considered baselines, particularly under higher market volatility.

**2) Final evaluation performance and stability.** Table 1(a) reports final evaluation performance in terms of mean return and $\mathrm{CVaR}_{0.95}$. GAN-D4PG achieves the strongest tail performance at both volatility levels, outperforming the considered baselines in final CVaR. These results show that the sample-efficiency gains observed during training are preserved in the final risk-sensitive evaluation.

The standard deviations in Table 1(a), together with the shaded regions in Fig. 2, characterize variability across independent runs. GAN-D4PG exhibits low variability in final CVaR at both volatility levels, particularly relative to RACER, and maintains stable learning behavior throughout training. Overall, GAN-D4PG combines strong final tail performance with consistent performance across independent runs.

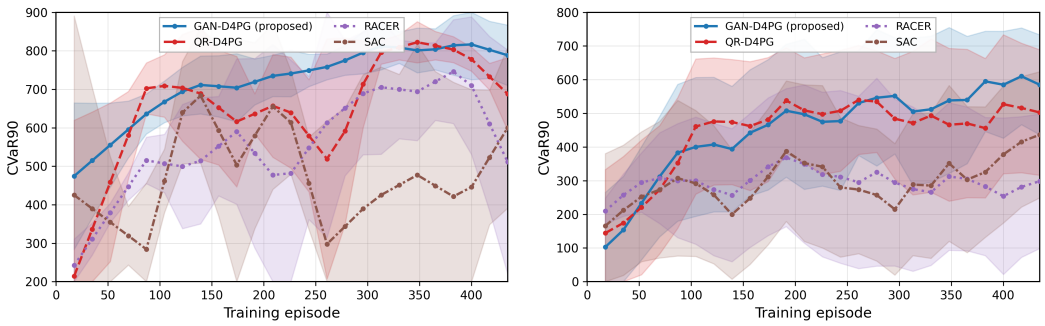
### 5.2 Autonomous Driving

For autonomous driving, we follow the F1TENTH experimental setting of O'Kelly et al. (2020) and consider the Offroad-Flat and Offroad-Bumpy tasks. These environments evaluate risk-sensitive continuous control, where the agent must achieve high return while limiting adverse tail outcomes and safety-critical failures such as vehicle rollovers. Additional training and environment details are provided in Appendix C.2.

**1) Sample efficiency.** Figure 3 shows the evolution of $\mathrm{CVaR}_{0.90}$ during training on Offroad-Flat and Offroad-Bumpy. GAN-D4PG improves rapidly in both environments and reaches strong tail performance early in training. The difference is particularly visible on Offroad-Bumpy, where adverse tail outcomes are more pronounced. Across both tasks, GAN-D4PG reaches competitive or improved risk-sensitive performance earlier than the considered baselines, providing evidence of improved sample efficiency.

**2) Final evaluation performance and stability.** Table 1(b) reports final performance in terms of mean return, $\mathrm{CVaR}_{0.90}$, and rollover rate. GAN-D4PG achieves the strongest tail performance and the lowest rollover rate in both driving environments. Thus, the improved sample efficiency observed during training is accompanied by stronger final tail performance and fewer safety-critical failures.

The standard deviations in Table 1(b), together with the shaded regions in Fig. 3, characterize variability across independent runs. GAN-D4PG maintains strong tail performance across both envi-




(a) Offroad-Flat (b) Offroad-Bumpy

Figure 3: $\mathrm{CVaR}_{0.90}$ learning curves for F1TENTH Offroad-Flat and Offroad-Bumpy. Shading denotes $\pm$ one standard deviation across six runs.

Table 1: Evaluation results for option hedging and F1TENTH driving. Values are mean $\pm$ standard deviation over six runs; higher is better for return and CVaR, while lower is better for rollover.

**(a) Option Hedging,** $\mathrm{CVaR}_{0.95}$

| Vol. | Method | Mean | $\mathrm{CVaR}_{0.95}$ |
|---|---|---|---|
| 0.3 | GAN-D4PG | $-3.20\pm0.18$ | $\mathbf{-7.34\pm0.11}$ |
| | QR-D4PG | $-3.29\pm0.08$ | $-7.57\pm0.13$ |
| | RACER | $-2.97\pm1.38$ | $-175.39\pm163.54$ |
| | SAC | $\mathbf{-2.14\pm0.10}$ | $-14.09\pm1.40$ |
| 0.7 | GAN-D4PG | $-7.20\pm0.12$ | $\mathbf{-16.44\pm0.16}$ |
| | QR-D4PG | $-8.17\pm1.16$ | $-17.99\pm2.27$ |
| | RACER | $-12.72\pm1.96$ | $-565.77\pm156.94$ |
| | SAC | $\mathbf{-4.95\pm0.19}$ | $-36.71\pm2.78$ |

**(b) F1TENTH Driving,** $\mathrm{CVaR}_{0.90}$

| Env. | Method | Mean | $\mathrm{CVaR}_{0.90}$ | Roll. (%) |
|---|---|---|---|---|
| Flat | GAN-D4PG | $1033.16\pm36.53$ | $\mathbf{787.73\pm71.75}$ | $\mathbf{0.40\pm0.80}$ |
| | QR-D4PG | $992.26\pm21.03$ | $681.29\pm93.12$ | $2.80\pm2.40$ |
| | RACER | $1082.66\pm48.92$ | $511.31\pm260.13$ | $7.20\pm5.31$ |
| | SAC | $\mathbf{1089.41\pm17.46}$ | $599.64\pm186.78$ | $6.40\pm5.85$ |
| Bumpy | GAN-D4PG | $879.03\pm47.67$ | $\mathbf{639.11\pm177.19}$ | $\mathbf{1.60\pm2.33}$ |
| | QR-D4PG | $866.02\pm27.18$ | $571.40\pm150.57$ | $2.80\pm2.04$ |
| | RACER | $831.49\pm99.00$ | $353.85\pm258.01$ | $14.67\pm13.70$ |
| | SAC | $\mathbf{898.02\pm36.89}$ | $231.13\pm161.19$ | $14.00\pm5.48$ |

ronments with comparatively consistent final evaluation performance, while several baselines exhibit substantially larger variability. Additional information-gain ablation results are reported in Appendix C.3.

## 6 Conclusion

We proposed *GAN-DRL*, a Bayesian adversarial distributional reinforcement learning framework for sample-efficient risk-sensitive control. The method combines Bayesian adversarial posterior inference, Wasserstein-based return-distribution learning, posterior information-gain exploration, and a quantile-based distributional critic for direct risk-sensitive policy optimization.

Across European option hedging and F1TENTH autonomous driving, GAN-DRL achieves strong tail performance with improved sample efficiency relative to the considered baselines. In option hedging, GAN-DRL reaches strong risk-sensitive performance with fewer interactions and improves CVaR across the evaluated market conditions. In autonomous driving, it improves tail CVaR and substantially reduces rollover rates. Across independent runs, GAN-DRL also exhibits consistent learning and final evaluation performance. The ablation results further show that posterior information-gain exploration contributes to both learning efficiency and final risk-sensitive performance.

Overall, these results indicate that combining Bayesian uncertainty-guided exploration with distributional return learning provides an effective approach to sample-efficient risk-sensitive control. Future work will consider additional safety-critical domains, alternative risk objectives, and the computational trade-offs of the proposed framework.

### AI use statement

ChatGPT was used to assist with language editing and manuscript organization, including improving the clarity and conciseness of the text. GitHub Copilot in Visual Studio Code was used to assist with code completion and debugging. The underlying methodology, experimental design, figures, results, and scientific conclusions were developed by the authors. All AI-assisted text edits and code suggestions were reviewed and verified by the authors.

## REFERENCES


Yu Bai, Song Mei, Huan Wang, and Caiming Xiong. Understanding the under-coverage bias in uncertainty estimation. *Advances in Neural Information Processing Systems*, 34:18307–18319, 2021.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pp. 449–458. PMLR, 2017.

Boris Beranger, Simone A Padoan, and Scott A Sisson. Estimation and uncertainty quantification for extreme quantile regions. *Extremes*, 24(2):349–375, 2021.

Aditya Bhatt, Daniel Palenicek, Boris Belousov, Max Argus, Artemij Amiranashvili, Thomas Brox, and Jan Peters. CrossQ: Batch normalization in deep reinforcement learning for greater sample efficiency and simplicity. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PczQtTsTIX.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of political economy*, 81(3):637–654, 1973.

Hans Buehler, Lukas Gonon, Josef Teichmann, and Ben Wood. Deep hedging. *Quantitative Finance*, 19(8):1271–1291, 2019.

Jay Cao, Jacky Chen, John Hull, and Zissis Poulos. Deep hedging of derivatives using reinforcement learning. *The Journal of Financial Data Science*, 3(1):10–27, 2021.

Jay Cao, Jacky Chen, Soroush Farghadani, John Hull, Zissis Poulos, Zeyu Wang, and Jun Yuan. Gamma and vega hedging using deep distributional reinforcement learning. *Frontiers in Artificial Intelligence*, 6:1129370, 2023.

Pierre Clavier, Stéphanie Allassonière, and Erwan Le Pennec. Robust reinforcement learning with distributional risk-averse formulation, 2022. URL https://arxiv.org/abs/2206.06841.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Roberto Daluiso, Marco Pinciroli, Michele Trapletti, and Edoardo Vittori. Cva hedging with reinforcement learning. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 261–269, 2023.

Thang Doan, Bogdan Mazoure, and Clare Lyle. Gan q-learning. *arXiv preprint arXiv:1805.04874*, 2018.

Pierluca D'Oro, Max Schwarzer, Evgenii Nikishin, Pierre-Luc Bacon, Marc G. Bellemare, and Aaron C. Courville. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In *International Conference on Learning Representations*, 2023.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference, 2017. URL https://arxiv.org/abs/1606.00704.

Dror Freirich, Tzahi Shimkin, Ron Meir, and Aviv Tamar. Distributional multivariate policy evaluation and exploration with the bellman gan. In *International Conference on Machine Learning*, pp. 1983–1992. PMLR, 2019.

Kang Gao, Stephen Weston, Perukrishnen Vytelingum, Namid Stillman, Wayne Luk, and Ce Guo. Deeper hedging: A new agent-based model for effective deep hedging. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 270–278, 2023.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications, 2019.

Igor Halperin. Qlbs: Q-learner in the black-scholes (-merton) worlds. *The Journal of Derivatives*, 28(1):99–122, 2020.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Petter N Kolm and Gordon Ritter. Dynamic replication and hedging: A reinforcement learning approach. *The Journal of Financial Data Science*, 1(1):159–171, 2019.

Luchen Li and A. Aldo Faisal. Bayesian distributional policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8429–8437, 2021. doi: 10.1609/aaai.v35i10.17024.

Yannick Limmer and Blanka Horvath. Robust hedging gans: Towards automated robustification of hedging strategies. *Applied Mathematical Finance*, 31(3):164–201, 2024.

Eva Lütkebohmert, Thorsten Schmidt, and Julian Sester. Robust deep hedging. *Quantitative Finance*, 22(8):1465–1480, 2022. doi: 10.1080/14697688.2022.2056073.

Parvin Malekzadeh, Zissis Poulos, Jacky Chen, Zeyu Wang, and Konstantinos N Plataniotis. Ex-drl: Hedging against heavy losses with extreme distributional reinforcement learning. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, pp. 370–378, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400710810. doi: 10.1145/3677052.3698668.

Saeed Marzban, Erick Delage, and Jonathan Yu-Meng Li. Deep reinforcement learning for option pricing and hedging under dynamic expectile risk measures. *Quantitative Finance*, 23(10):1411–1430, 2023. doi: 10.1080/14697688.2023.2244531.

Robert C Merton. Theory of rational option pricing. *The Bell Journal of economics and management science*, pp. 141–183, 1973.

Phillip Murray, Ben Wood, Hans Buehler, Magnus Wiese, and Mikko Pakkanen. Deep hedging: Continuous reinforcement learning for hedging of general portfolios across multiple risk aversions. In *Proceedings of the Third ACM International Conference on AI in Finance*, pp. 361–368, 2022.

Andrei Neagu, Frédéric Godin, and Leila Kosseim. Deep reinforcement learning algorithms for option hedging, 2025. URL `https://arxiv.org/abs/2504.05521`.

Yuriy Nevmyvaka, Yi Feng, and Michael Kearns. Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 673–680, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143929.

Matthew O'Kelly, Hongrui Zheng, Dhruv Karthik, and Rahul Mangharam. F1tenth: An open-source evaluation environment for continuous control and reinforcement learning. In Hugo Jair Escalante and Raia Hadsell (eds.), *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, volume 123 of *Proceedings of Machine Learning Research*, pp. 77–89. PMLR, 08–14 Dec 2020.

Max Schwarzer, Ankesh Anand, Rishab Goel, R. Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. In *International Conference on Learning Representations*, 2021.

Anil Sharma, Freeman Chen, Jaesun Noh, Julio DeJesus, and Mario Schlener. Hedging beyond the mean: A distributional reinforcement learning perspective for hedging portfolios with structured products. *arXiv preprint arXiv:2407.10903*, 2024.

Kyle Stachowicz and Sergey Levine. Racer: Epistemic risk-sensitive rl enables fast driving with fewer crashes, 2024.

Edoardo Vittori, Michele Trapletti, and Marcello Restelli. Option hedging with risk averse reinforcement learning. In *Proceedings of the first ACM international conference on AI in finance*, pp. 1–8, 2020.

David Wu and Sebastian Jaimungal. Robust risk-aware option hedging. *Applied Mathematical Finance*, 30(3):153–174, 2023. doi: 10.1080/1350486X.2023.2301354.

Cong Zheng, Jiafa He, and Can Yang. Option dynamic hedging using reinforcement learning, 2023. URL https://arxiv.org/abs/2306.10743.