SENTINEL: Taming Uncertainty with Ensemble-based Distributional Reinforcement Learning

Hannes Eriksson, Debabrota Basu, Mina Alibeigi, Christos Dimitrakakis

SENTINEL: Taming Uncertainty with Ensemble-based Distributional Reinforcement Learning

Hannes Eriksson HANNES.ERIKSSON@ZENSEACT.COM
Zenseact AB, Gothenburg, Sweden
Chalmers University of Technology, Gothenburg, Sweden

Debabrota Basu
Scool, INRIA Lille- Nord Europe, Lille, France
CRIStAL, CNRS, Lille, France

Mina Alibeigi
Zenseact AB, Gothenburg, Sweden

Christos Dimitrakakis
University of Oslo, Oslo, Norway

Abstract

In this paper, we consider risk-sensitive sequential decision-making in model-based reinforcement learning (RL). We introduce a novel quantification of risk, namely *composite risk*, which takes into account both aleatory and epistemic risk during the learning process. Previous works have considered aleatory or epistemic risk individually, or, an additive combination of the two. We demonstrate that the additive formulation is a particular case of the composite risk, which underestimates the actual CVaR risk even while learning a mixture of Gaussians. In contrast, the composite risk provides a more accurate estimate. We propose to use a bootstrapping method, SENTINEL-K, for distributional RL. SENTINEL-K uses an ensemble of K learners to estimate the return distribution and additionally uses follow the regularized leader (FTRL) from bandit literature for providing a better estimate of the risk on the return distribution. Finally, we experimentally verify that SENTINEL-K estimates the return distribution better, and while used with composite risk estimate, demonstrates better risk-sensitive performance than competing RL algorithms.

1. Introduction

Reinforcement Learning (RL) algorithms with their recent success in games and simulated environments (Mnih et al., 2015) have drawn interest for real-world and industrial applications (Pan et al., 2017; Mahmood et al., 2018). Two aspects of RL algorithms constrain their applicability. Firstly, the large amount of data generally required by model-free RL algorithms. Secondly, since in reinforcement learning the environment is by definition is unknown to the agent, exploring it so as to improve performance and eventually obtain the optimal policy entails risks. Though risk is not an issue in simulation, it is important to consider risks when interacting in the real world (Pinto et al., 2017; Garcıa and Fernández,

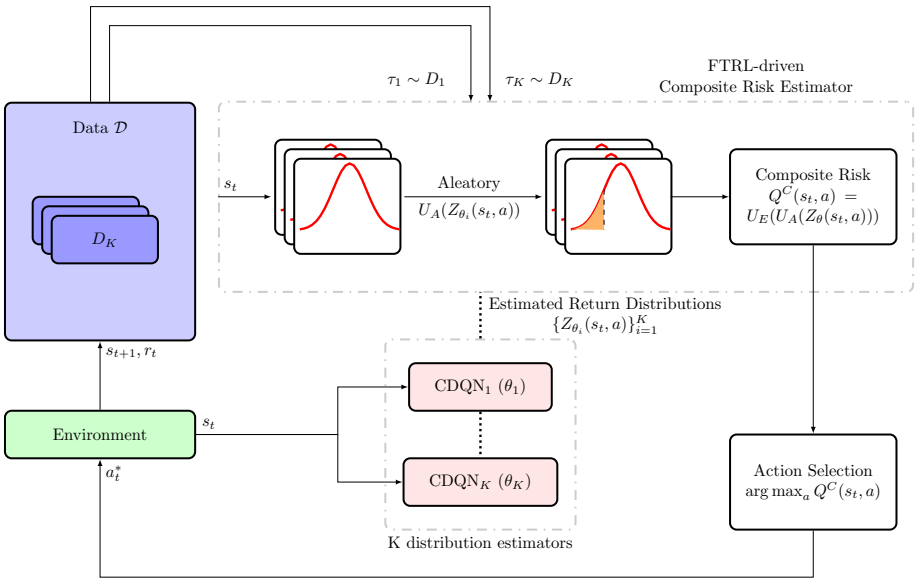


Figure 1: SENTINEL-K with FTRL-driven composite risk estimator and K CDQNs as distribution estimators.

2015; Prashanth and Fu, 2018). In this paper, we employ a model-based approach that enables us both to efficient in terms of the amount of data needed, and to be flexible with respect to the risk metric the agent should consider when making decisions.

Risk sensitivity in reinforcement learning and Markov decision processes has sometimes been considered under a minimax formulation over plausible MDPs (Satia, 1973; Heger, 1994; Tamar et al., 2014). Alternative approaches include maximising a risk-sensitive statistic instead of the expected return (Chow and Ghavamzadeh, 2014; Tamar et al., 2015; Clements et al., 2019). In this paper, we focus on the second approach due to its flexibility. Either approach requires estimating the uncertainty associated with the decision-making procedure. This uncertainty includes both the inherent randomness in the model and the uncertainty due to imperfect information about the true model. These two type of uncertainties are called *aleatory* and *epistemic* uncertainty respectively (Der Kiureghian and Ditlevsen, 2009).

In this work, we propose a *composite risk* formulation in order to capture the combined effect of aleatory and epistemic uncertainty for decision-making in RL (Section 4). In recent literature, researchers have either quantified epistemic and aleatory risks separately (Mihatsch and Neuneier, 2002; Eriksson and Dimitrakakis, 2019) or considered an additive risk formulation where their weighted sum is minimized by an RL algorithm (Clements et al., 2019). In a reductive experiment (Figure 2), we show that using an additive risk, which is the sum of separately computed epistemic and aleatory CVaR[1], strictly underestimates the total CVaR (Rockafellar et al., 2000), and the deviation is significant as CVaR focuses more on less probable events. In contrast, the composite risk takes into consideration the combined effect of two types of uncertainty, and better reflects the underlying risk. Finally, we show that additive risk is essentially a special case of composite risk.

1. $CVaR_\alpha$ captures the expected value of $\alpha\%$ of events in the left tail.

We then incorporate this composite risk measure within the Distributional RL (DRL) framework (Dabney et al., 2018b; Tang and Agrawal, 2018; Rowland et al., 2019). The DRL framework aims to model the distribution of returns of a policy for a given environment (Section 3.3). This highly expressive distributional representation allows us to both estimate appropriate risk measures and to incorporate them in final decision making. However, DRL approaches are typically limited to modelling aleatory uncertainty, with epistemic uncertainty due to partial information not being explicitly modelled in terms of the return distribution. In this paper, we propose a bootstrapping (Efron and Tibshirani, 1985) based framework to estimate the return distribution.

As we explain in Section 5, we use an ensemble of K distribution estimators, such as CDQNs (Dabney et al., 2018b), obtained through bootstrapping, to learn the return distribution. We use these return distributions to estimate the aleatory and composite risks for the corresponding RL method (Section 5.1). In order to perform the estimation accurately and efficiently, we adapt the Follow The Regularised Leader (FTRL) (Cesa-Bianchi and Lugosi, 2006) algorithm in order to weigh the estimators in our ensemble, as we describe in Section 5.2.

Our framework, which we call SENTINEL-K, is illsutrated in Figure 1. We instantiate SENTINEL-K to perform risk-sensitive model-based distributional RL by incorporating the composite CVaR estimate with FTRL-driven bootstrapped CDQN algorithm (Dabney et al., 2018b). We experimentally show in Section 6 that the FTRL-driven bootstrapping method of SENTINEL-K generates accurate estimates of true return distributions for even suboptimal actions and multimodal return distributions, where the vanilla distributional RL algorithm fails to do so. Estimation of SENTINEL-K even without risk-sensitive objective converges faster. We also show that our FTRL-based approach is more accurate than uniform or greedy aggregation of K approximations of the return distribution. Finally, we verify the risk-sensitive performance of SENTINEL-K with composite CVaR metric on the highway environment with 10 cars. Experimental results show that our approach leads to a higher estimate of underlying risk and thus, less number of crashes than competing distributional algorithms, which are VDQN (Tang and Agrawal, 2018), CDQN, and SENTINEL-K with additive CVaR estimate.

Before proceeding to the details of our contributions, we posit our work in the existing literature in Section 2. Following that, we provide a primer on risk measure, Markov decision processes, and DRL in Section 3 to elucidate our contributions.

2. Related Works

For RL applications in the real world, such as for autonomous driving and robotics, *risk-sensitive* RL approaches can avoid the negative consequences of excessive exploration. This has initiated a spate of research efforts (Howard and Matheson, 1972; Satia, 1973; Coraluppi and Marcus, 1999; Marcus et al., 1997; Mihatsch and Neuneier, 2002; Prashanth and Fu, 2018) spanning five decades. But the majority of these works focus only on discrete state-space MDPs. We are interested in designing a general framework applicable to both discrete and continuous state-spaces. Thus, we adopt the framework of distributional RL, specifically CDQN, that incorporates highly expressive approximators to model continuous and multimodal return distributions.

Both *aleatory* and *epistemic* are important for risk-sensitive RL (Der Kiureghian and Ditlevsen, 2009). The former expresses the randomness inherent to the problem and the latter uncertainty about the problem respectively. A common approach to make an algorithm risk-sensitive (Garcıa and Fernández, 2015) is to use a utility function that is nonlinear with respect to the return, or the expected return.[2] For example (Mihatsch and Neuneier, 2002) consider aleatory risk-sensitive RL by transforming the return. Follow-up works (Chow and Ghavamzadeh, 2014; C. et al., 2015) focus on scaling up these approaches. There have been recent works considering epistemic risk (Eriksson and Dimitrakakis, 2019), wherein problem uncertainty is expressed in a Bayesian framework as a distribution over MDPs. Depeweg et al. (2018); Clements et al. (2019) intuitively incorporates both of these risks in decision making. Depeweg et al. (2018) consider the risk in the individual costs in RL. (Clements et al., 2019) consider the additive formulation of epistemic and aleatory risks. They use variance as the risk measure which is not a coherent measure (Artzner et al., 1999). In order to rectify such varied choices, we define a composite risk that considers and quantifies the entangled effect of epistemic and aleatory uncertainties. We also show that for any coherent risk measure, such as CVaR, the composite risk retains coherence.

Ensemble-based RL has been done previously with great success (Wiering and Van Hasselt, 2008; Faußer and Schwenker, 2015; Osband et al., 2016; Pacchiano et al., 2020). This process typically involves creating an ensemble of well-known RL agents, such as Deep Q-Networks (DQN) (Mnih et al., 2015), where each estimator has its own dataset, and the final decision maker considers the joint prediction of the ensemble into account. Typically, the final estimate averages the individual estimators. In particular, adding additional estimators to form an ensemble of estimators not only improves performance for risk-neutral decision-making but also allows the consideration of the distribution of estimators. *This enables epistemic risk-sensitive decision-making.* We incorporate bootstrapping approach to ensemble K different estimations of the return distribution, and introduce the FTRL algorithm to estimate the return distribution accurately and efficiently.

3. Background

In this section, we introduce the notion of risk measures, the risk-sensitive Markov decision process formulation, and the distributional RL framework.

3.1 Risk measure

The idea of quantifying risk in decision making is long-studied in decision theory and has found multiple applications in finance and actuarial science. Researchers proposed multiple measures of risk, such as variance, Value at Risk (VaR), Conditional Value at Risk (CVaR), etc. to quantify the probability of occurrence of an event away from the expectation of corresponding distribution (Szegö, 2002). Artzner et al. (1999) have established a basic set of axioms to be satisfied for a *coherent risk measure*: normalization, monotonicity, sub-additivity, homogeneity, and translation invariance. For example, CVaR is a coherent risk measure whereas variance and VaR are not. Thus, in this work, we choose *CVaR* (Rockafellar et al., 2000) as the risk measure of interest.

2. Here, we use return to mean the total discounted reward

$CVaR_\alpha$ quantifies expectation of the worst $\alpha\%$ of a probability distribution. For a random variable Z and $\alpha \in [0, 1]$,

$$CVaR_\alpha(Z) \triangleq \mathbb{E}[Z \,|\, Z \geq \nu_\alpha \wedge \Pr(Z > \nu_\alpha) = \alpha] \tag{1}$$

CVaR is widely used in risk-sensitive RL (Chow and Ghavamzadeh, 2014; Tamar et al., 2015; Chow et al., 2015) as it is coherent, applies to general L_p spaces, and captures the heaviness of the tail of a distribution. For $\alpha = 0$, CVaR reduces to the expected value, and thus, the corresponding risk-sensitive RL algorithm behaves analogously to a risk-neutral one. Kolla et al. (2019) shows that CVaR of a distribution can be accurately estimated using i.i.d. samples.

3.2 Markov Decision Process

In this work, we are considering decision-making problems that can be modelled by a Markov Decision Process (MDP) (Sutton and Barto, 2018). An MDP is a tuple $\mu \triangleq (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma)$. $\mathcal{S} \in \mathbb{R}^d$ is a state representation of dimension d. $\mathcal{A}$ is the set of admissible actions. $\mathcal{T}$ is a transition kernel that determines the probability of successor states s' given the present state s and action a. The reward function $\mathcal{R}$ quantifies the goodness of taking action a in state s. The goal of the agent is to find a policy $\pi : \mathcal{S} \to \mathcal{A}$ to maximise expected value of a *utility function* U (Friedman and Savage, 1948) computed over a reward sequence given a time horizon T: $U^\pi(s, a) = \mathbb{E}\left[U(\sum_{t=0}^{T} \gamma^t R(s_t, a_t))\right]$. Here, $s_t \sim \mathcal{T}(.|s_{t-1}, a_{t-1})$, $a_t = \pi(s_t)$, $s_0 = s$, and $a_0 = a$.

When the utility function U is an identity function, $U^\pi(s, a)$ reduces to the Q-function which is the expected long-term discounted reward. If the utility function U is a coherent risk measure, such as CVaR, it leads to a risk-sensitive formulation of MDP (Mihatsch and Neuneier, 2002; Prashanth and Fu, 2018).

3.3 Distributional RL

Typically, the variable at the core of both risk-neutral and risk-sensitive RL is usually the accumulated discounted reward $Z^\pi(s, a) \triangleq \sum_{t=0}^{T} \gamma^t R(s_t, a_t)$. $Z^\pi(s, a)$ is called the return of a policy π. In distributional RL, the goal is to learn the return distribution $Z^\pi(s, a)$ obtained by following policy π from state x and action a under the given MDP.

Different methods are proposed to parametrize the return distribution. Bellemare et al. (2017) propose *CDQN*, a categorical distribution with N atoms and, with support in $[V_{MIN}, V_{MAX}]$. The mass of the atom z_i is then given by $\frac{e^{\theta_i(s,a)}}{\sum_j e^{\theta_j(s,a)}}$. Tang and Agrawal (2018), Dabney et al. (2018a), and Rowland et al. (2019) use unimodal Gaussians, quantiles, and expectiles to model the return distribution respectively. In this work, we choose to extend CDQN, as it permits richer representations of distributions, and flexibility to compute different statistics.

The intuition of using this distributional framework for risk-sensitive RL is its flexibility to model multi-modal and asymmetrical distributions, which is important for an accurate estimate of risk.

4. Quantifying Composite Risk

In risk-sensitive RL, we encounter two types of uncertainties: *aleatory* and *epistemic*. Aleatory uncertainty is engendered by the stochasticity of the MDP model μ and the policy π. Epistemic uncertainty exists due to the fact that the MDP model μ is unknown, In the Bayesian setting, this is seen as having a belief distribution β over a set of plausible MDPs Θ. Hence, risk measures can also be defined with respect to the MDP distribution. Consequently, as an agent learns more about the underlying MDP, the epistemic risk vanishes. The aleatory risk is inherent to the MDP model μ and policy π, and thus persists even after correctly estimating the model μ. Let us first define risk measures for aleatory and epistemic uncertainty separately. We then combine them into a composite risk measure.

Aleatory Risk. Given a coherent risk measure U_A, the aleatory risk is quantified as the deviation of total risk of individual models from the risk of the average model.

$$\begin{aligned} A(U_A, \beta) &\triangleq \mathbb{E}_\beta[\mathbb{E}_{\Pr(.|\theta)}[U_A(Z)] - U_A(\mathbb{E}_{\Pr(.|\theta)}[Z])] \\ &= \int_\Theta \int_{\mathcal{Z}} (U_A(z) - U_A(\mu_z)) \,\mathrm{d}\Pr(z \,|\, \theta) \,\mathrm{d}\beta(\theta), \end{aligned}$$

$U(\mu_z) \triangleq U\Big(\int_\Theta \mathbb{P}(z \,|\, \theta) \,\mathrm{d}\beta(\theta)\Big)$, the utility of the average model given a belief distribution β over the plausible set of models Θ. The centered definition of aleatory risk is necessary for the additive formulation to be a special case of the composite formulation.

Epistemic Risk. Given a coherent risk measure U_E, the epistemic risk quantifies the uncertainty invoked by not knowing the plausible models. Thus, the risk can be computed over any statistics of the models, such as expectation.

$$\begin{aligned} E(U_E, \beta) &\triangleq \mathbb{E}_\beta[U_E(\mathbb{E}_{\Pr(.|\theta)}[Z])] \\ &= \int_\Theta U_E \left(\int_{\mathcal{Z}} z \,\mathrm{d}\Pr(z \,|\, \theta) \right) \mathrm{d}\beta(\theta) \end{aligned}$$

Composite Risk under Model and Inherent Uncertainty. In typical risk-sensitive RL settings, the true MDP model is unknown, as well as the MDPs are inherently stochastic. Thus, the total uncertainty to be considered is a composition of aleatory and epistemic uncertainties. In order to quantify the total uncertainty under consideration, we propose the *composite risk*.

Definition 1 (Composite Risk) *For two coherent risk measures U_A and U_E, belief distribution β on model parameters θ, and a random variable Z, the composite risk of epistemic and aleatory uncertainties is defined as*

$$F^C(U_A, U_E, \beta) \triangleq \int_\Theta U_E \left(\int_{\mathcal{Z}} U_A(z) \,\mathrm{d}\Pr(z \,|\, \theta) \right) \mathrm{d}\beta(\theta).$$

The composite risk is flexible to use two different risk measures for quantifying epistemic and aleatory uncertainties.

Lemma 2 (Coherence) *If U_A and U_E are two coherent risk measures, the composite risk measure $F(U_A, U_E, \beta)$ is also coherent.*

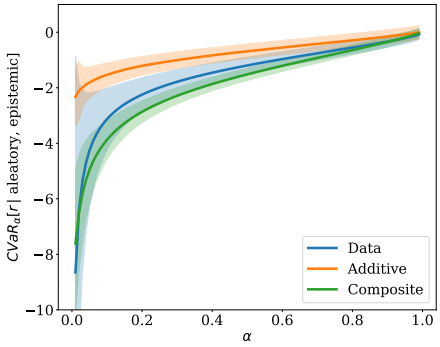


Figure 2: Estimation of total $CVaR_\alpha$ from a mixture of 100 Gaussians sampled from a posterior distribution. Total $CVaR_\alpha[Data]$ is based on the marginal distribution of the r as given in Example 1. We compare this with composite and additive estimates and illustrate results over 100 runs.

Additive Risk Measure. If U_E is the identity function, the composite risk is reduced to an additive risk measure.

$$F^A(U_A, \beta) \triangleq \int_\Theta \int_\mathcal{Z} U_A(z) \, \mathrm{d} \Pr(z \mid \theta) \, \mathrm{d}\beta(\theta)$$
$$= A(U_A, \beta) + E(U_A, \beta)$$

Often the additive risk measure or weighted sum of the epistemic and aleatory uncertainty is used in risk-sensitive RL literature (Clements et al., 2019). However, the additive risk formulation strictly underestimates the composite effect of epistemic risk. Thus, we observe that additive risk leads to worse risk-sensitive performance than composite risk in RL problems (Table 1). In order to compare the risk estimation using additive and composite formulations, we consider an example of estimating CVaR over a Gaussian mixture.

Example 1 *We consider a mixture of* 100 *Gaussians:* $p(r) = \sum_{i=1}^{100} \phi_i \mathcal{N}(\mu_i, \sigma_i^2)$, *where* $\Phi \sim Dir([0.5]^{100}), \mu \sim \mathcal{N}(0, 1)$, *and* $\sigma^2 \sim \Gamma^{-1}(2, 0, 1)$. *We compute* $CVaR_\alpha[r]$ *from the data generated from such mixture for 100 runs. We further estimate composite risk with* $U_E, U_A = CVaR_\alpha$ *and additive risk with* $U_A = CVaR_\alpha$. *The results illustrated in Figure 2 show that the additive CVaR risk strictly underestimates the total CVaR risk computed from the data, whereas the composite risk is closer to the one computed from data. Specifically, for lower values of* α, *i.e. towards the extreme end of the left tail where events occur with low probability, the additive CVaR risk deviates significantly from data whereas the composite measure yields closer estimation. Such values of* α*'s are typically interesting for risk-sensitive applications.*

5. Algorithm: SENTINEL-K

In this section, we outline the algorithmic details of SENTINEL-K as an ensemble of K distributional RL estimators, such as CDQN (Bellemare et al., 2017), along with an adaptation of FTRL for estimator selection. We further evaluate the composite risk using return distribution estimated by SENTINEL-K for decision making.

Algorithm 1 SENTINEL-K with Composite Risk

1: **Input:** Initial state s_0, action set $\mathcal{A}$, risk measures U_A, U_E, hyperparameter λ, target networks $[\theta_1^-, ..., \theta_K^-]$, value networks $[\theta_1, ..., \theta_K]$, update schedule Γ_1, Γ_2.
2: **for** $t = 1, 2, \ldots$ **do**
3: //* Update K-value and target networks for estimating return distributions*//
4: **for** $t' \in \Gamma_1 \cup \Gamma_2$ **do**
5: Generate $\{D_1, ..., D_K\} \leftarrow \text{DataMask}(\mathcal{D}^{t'})$
6: **for** $i = 1, \ldots, K$ **do**
7: Sample mini batch $\tau \sim D_i$
8: Estimate $F^C(Z(s_t, a)|U_A, U_B, \beta)$ using τ and K-target networks $\{\theta_i^-\}_{i=1}^K$.
9: Get $a^* = \arg\max_a F^C(Z(s_t, a)|U_A, U_B, \beta)$
10: Update value network θ_i using τ, a^*
11: Update target network θ_i^- using τ, a^* if $t' \in \Gamma_1$
12: **end for**
13: **end for**
14: //* Estimate the composite risk of each action using the estimated return distributions*//
15: **for** $a \in \mathcal{A}$ **do**
16: Compute weights $\mathbf{w} = w_1, ..., w_K$ from Eq. 2.
17: **for** i **in** K **do**
18: Compute aleatory risks $Q_i^A(s_t, a)$ from $\int_{\mathcal{Z}} U_A(z)\, \mathrm{d}\mathbb{P}(z \,|\, \theta_i)$.
19: **end for**
20: Compute composite risk over weighted aleatory estimates $Q^C(s_t, a) = U_E\Big(\mathbf{w} \cdot \mathbf{Q}^A(s_t, a)\Big)$
21: **end for**
22: //*Action selection*//
23: Take action $a_t = \arg\max_a Q^C(s_t, a)$
24: Observe s_t and update the dataset $\mathcal{D}^t \leftarrow \mathcal{D}^{t-1} \cup \{s_t, a_{t-1}, s_{t-1}, r_{t-1}\}$
25: **end for**

Sketch of the Algorithm. Pseudocode of SENTINEL-K with composite risk is described in Algorithm 1. Algorithm 1 has mainly two functional blocks: yielding K estimations of return distribution with distributional RL framework (Lines 4- 13), and using such K estimates to compute composite risk of each of the actions (Lines 15- 21). Finally, following the mechanism of Q-learning, it chooses the action with maximal composite risk in the decision making step (Line 23).

In the first functional block, we specifically use an ensemble of K CDQNs. Each CDQN uses target and value networks for estimating the return distribution. We set a schedule of updating the target networks Γ_1 and a more frequent schedule $\Gamma_1 \cup \Gamma_2$ to update the value networks. The details of this procedure is elaborated in Section 5.1.

The second functional block is used for decision-making and iterated at every time step. It adapts the FTRL algorithm (Section 5.2) for aggregating the K estimated return

distributions and to compose aleatory risk $Q_i^A(s_t, a)$ of each of the estimators to provide a final estimate of the composite risk $Q^C(s_t, a)$ for each action.

5.1 Ensembling & bootstrapping Estimators

The ensemble of SENTINEL-K consists of K distribution estimators. Each estimator gets its own dataset $\{D_i\}_{i=1}^K \subseteq \mathcal{D}$, value network $\{\theta_i\}_{i=1}^K$ and target network $\{\theta_i^-\}_{i=1}^K$. The K datasets are created from the original data set $\mathcal{D}$ by *data masking* (Line 5). For each transition s_t, a_t, r_t, s_{t+1}, a fixed weight vector $\mathbf{u}_t \in [0,1]^K$ is generated such that $u_t^j \sim Ber(\frac{1}{3})$. Thus, each estimator i has access to on an average $\frac{1}{3}$ of the whole dataset.

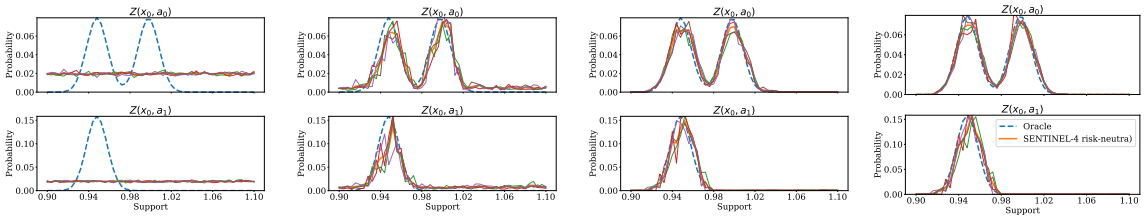
After preparing the datasets for the estimators, the target and value networks of the CDQN have to be updated and optimized. For i-th estimator, it begins with sampling mini batches of data τ from the respective dataset D_i (Line 7). Then, this dataset is used to compute the composite risk for all actions $a \in \mathcal{A}$ and to obtain a^* (Lines 8- 9). Obtaining the composite risk first involves estimating the aleatory risk with $Q_i^A(s_t, a) = \int_{\mathcal{Z}} U_A(z)\, \mathrm{d}\mathbb{P}(z \mid \theta_i)$ for a particular estimator i. This quantity can be attained by considering each of the estimators separately, however, as we turn to compute the epistemic risk the estimators jointly contribute to this risk. Then, we compose the aleatory risk of all the estimators to compute $Q^C(s_t, a) = \sum_i U_E(Q_i^A(s_t, a))$. Finally, the optimal action $a^* = \arg\max_a Q^C(s_t, a)$, and the risk estimates $Q^C(s_t, a)$ are used to update the value and network parameters $\{\theta_i\}_{i=1}^K$ and $\{\theta_i^-\}_{i=1}^K$ (Lines 10- 11) by minimising the cross-entropy loss of the current parameters and the projected Bellman update as described in (Bellemare et al., 2017).

Ensembling estimators have been shown to outperform individual estimators as seen in (Wiering and Van Hasselt, 2008; Faußer and Schwenker, 2015; Osband et al., 2016; Pacchiano et al., 2020). Further, incorporating multiple estimators introduces uncertainty over the estimators. Because of having separate data sets, each of the estimators learn different parts of the MDP. Thus, uncertainty over estimators acts as a quantifier of the model uncertainty. In Section 6, we show that this ensemble-based approach leads SENTINEL-K to achieving superior performance.

5.2 Follow the regularised leader

Now, the question is to aggregate the K estimated return distributions in one such that the final estimation is as accurate as possible, where each of the estimators may vary in terms of learning and accuracy. Pacchiano et al. (2020) shows that model selection can boost performance than model averaging. The rationale for this can be given by seeing that some estimators might be overly optimistic or pessimistic. By considering these outliers less, you can effectively have a more robust ensemble.

We adapt the Follow The Regularised Leader (FTRL) algorithm (Cesa-Bianchi and Lugosi, 2006) studied in bandits and online learning for selecting the estimators. FTRL puts exponentially more weight on an estimator depending on its accuracy of estimating the return distribution. Since we don't know the 'true' return distribution, we use the KL-divergence from the posterior of a single estimator i, $\mathbb{P}(z \mid \theta_i)$, to the posterior marginalized over $\beta(\theta)$, i.e. $l(\theta_i, \beta) \triangleq D_{\mathrm{KL}}\Big(\int_\Theta \int_{\mathcal{Z}} z\, \mathrm{d}\mathbb{P}(z \mid \theta)\, \mathrm{d}\beta(\theta) \,||\, \int_{\mathcal{Z}} z\, \mathrm{d}\mathbb{P}(z \mid \theta_i) \Big)$, as the proxy of



(a) $n = 0$ (b) $n = 1000$ (c) $n = 5000$ (d) $n = 10000$

Figure 3: Return distributions of a_0 and a_1 for $0, 1000, 5000$ and 10000 data points (n) respectively. The blue dashed line is the categorical approximation of $Z(s_0, a_0)$ and $Z(s_0, a_1)$ respectively. The thick orange line is the marginal posterior $\int_\Theta \mathbb{P}(z \mid \theta)\, \mathrm{d}\beta(\theta)$ with SENTINEL-4. The thin lines are the posteriors of the individual estimators.

estimation loss of estimator i. FTRL selects estimator i with weight

$$w_i = \frac{\exp\Big(\lambda l(\theta_i, \beta)\Big)}{\sum_{j=1}^{K} w_j}, \tag{2}$$

$\lambda \in [0, \infty)$ is a regularising parameter that determines to what extent estimators far away from the marginal estimator should be penalised. If $\lambda \to 0$, we obtain standard model averaging. If $\lambda \to \infty$, it reduces to greedy selection.

Having computed the weights $\mathbf{w}$ (Line 16), we compute the weighted composite risk measure by first computing the aleatory risk of each of the estimators, $Q_i^A(s_t, a) = \int_{\mathcal{Z}} U_A(z)\, \mathrm{d}\mathbb{P}(z \mid \theta_i)$ (Line 18), and then the composite risk is computed by $Q^C(s_t, a) = U_E(\mathbf{w} \cdot \mathbf{Q}^A(s_t, a))$ (Line 20). Here, $\cdot : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}^K$ is the pointwise product. We experimentally show that performing FTRL with a reasonable λ value, namely 1, leads to better performance.

SENTINEL-K reduces to a risk-neutral algorithm if we choose both U_A, U_E as identity functions, and to additive risk-sensitive algorithm if we choose U_E as identity. Designing it to accommodate composite risk provides us this flexibility. We use risk-neutral SENTINEL-K to validate its efficiency to estimate return distributions, and the one with composite CVaR risk to perform risk-sensitive RL tasks.

6. Experimental Evaluation

In this section, we experimentally validate the performance of risk-neutral SENTINEL-K in terms of estimating the return distribution of different actions and improvement of FTRL over model averaging or greedy model selection. We also test the risk-sensitive performance of SENTINEL-K with composite CVaR risk in a large enough environment with continuous state space. Settings for each of these three experiments and results are elaborated in corresponding subsections. In all the experiments, we use 4 CDQNs in the ensemble and call it SENTINEL-4.

Return Distribution Estimation. In order to demonstrate uncertainty estimation and convergence in distribution of SENTINEL-K framework, we test SENTINEL-4 on an MDP environment with known multimodal return distribution. The MDP contains three states and two actions such that the return distribution of a_0 from state s_0 is a mixture of Gaussians $Z(s_0, a_0) \sim \sum_{i=0}^{N} \Phi_i \mathcal{N}(\mu_i, \sigma_i)$ and the return distribution of action a_1 is $Z(s_0, a_1) \sim$

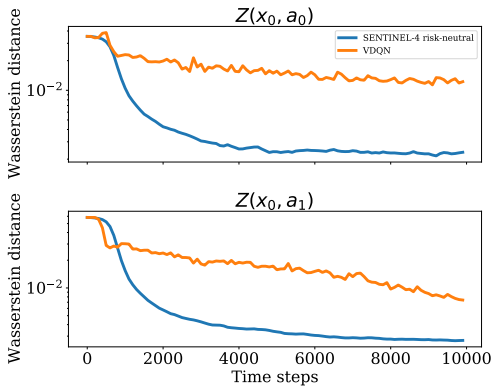


Figure 4: Shows convergence in distribution of SENTINEL-4 (risk-neutral) and VDQN by measuring the Wasserstein distance between the categorical approximation of $Z(s_0, a_0), Z(s_0, a_1)$ and the estimated distributions by the two agents, for each action.

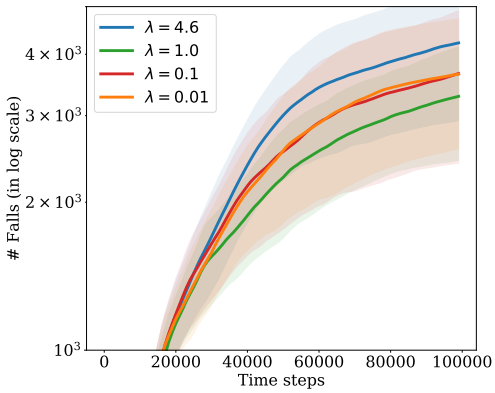


Figure 5: Performance and convergence of SENTINEL-4 (risk-neutral) for different parameter values of λ. Shown is the number of falls in the *CartPole* environment. Experimental results are computed over 20 runs with different initialisation and the shaded region represents $\mu_t \pm \sigma_t$.

$\mathcal{N}(\mu_1, \sigma_1)$. Here, $\Phi = [0.5, 0.5]$, $\mu = [1.0, 0.95], \sigma = [0.1, 0.1]$. Figure 3 shows convergence in distribution of SENTINEL-4. We observe that SENTINEL-4 estimates the return distributions of both the actions considerably well after using 5000 data points.

In Figure 4, we further illustrate the Wasserstein distance of the distributions estimated by risk-neutral SENTINEL-4 and VDQN algorithms from the true return distribution. We show that the VDQN fails to converge to the true return distribution whereas SENTINEL-4 converges to the true return distribution in significantly less number of steps.

FTRL vs. Average vs. Greedy. In order to demonstrate the performance of the model selection algorithm, we evaluate SENTINEL-4 in the *CartPole-v0* environment (Brockman et al., 2016). This environment is a common testbed for continuous state-space RL tasks. In the environment, a reward of 1 is attained for every time step the pole is kept upright. If the pole falls to either of the sides or if the number of time steps reaches 200, the episode is

Table 1: Performance of risk-neutral (VDQN, CDQN, SENTINEL-4), and risk-sensitive (SENTINEL-4 with additive and composite CVaRs) for highway-v1 with 10 vehicles. Results are reported over 20 runs. SENTINEL-4 with composite CVaR performs better.

Agent	Value $\pm\sigma$	Aleatory metric $\pm\sigma$	# crashes $\pm\sigma$
VDQN risk-neutral	23.30 ± 1.59	14.29 ± 3.60	1252.33 ± 761.85
CDQN risk-neutral	25.96 ± 2.27	19.50 ± 6.43	839.53 ± 671.70
SENTINEL-4 risk-neutral	26.56 ± 1.45	20.88 ± 5.58	617.11 ± 447.89
SENTINEL-4 additive	26.82 ± 1.87	21.54 ± 6.24	645.55 ± 570.58
SENTINEL-4 composite	$\mathbf{27.43 \pm 0.60}$	$\mathbf{24.16 \pm 2.40}$	$\mathbf{341.18 \pm 196.13}$

terminated. This means that the undiscounted return attained per episode is within $[0, 200]$ and so we chose $V_{min} = 0, V_{max} = \frac{1-\gamma^{200}}{1-\gamma}$ as the histogram support of CDQN.

We choose $[0.01, 0.1, 1.0, 4.6]$ as the different values of the regularising hyperparameter λ. As $\lambda \to 0$, we are essentially doing standard model averaging. We expect this to have average performance since all estimators are weighted equally. This means that it might be overly sensitive to estimator outliers. As $\lambda \to \infty$, model selection gets greedily biased towards the best average estimator. In fact, we expect performance to be poor when λ is too high since it is putting almost all weight on one single estimator while not providing other estimators a chance to improve. A sound value of λ would be one that excludes outlier estimators while still involving most of the other estimators. We run each of the experiments for 10^5 steps and average the results over 20 runs. Figure 5 shows performance in terms of cumulative # Falls (lower is better) for the λ values with $\alpha = 0.25$. We observe that FTRL with reasonable $\lambda = 1.0$ shows better performance, i.e. less number of falls, than the ones with large $\lambda = 4.6$ and small λ's 0.01 and 0.1. We also observe that for $\lambda = 1$ the variance of the total number of falls is significantly less than that of other values. This indicates stability of performance.

Risk-sensitive Performance. In order to demonstrate performance in a larger domain, we opt to evaluate SENTINEL-4 in the *highway* (Leurent, 2018). Highway is an environment developed to test RL for autonomous driving. We use a version of the *highway-v1* domain with five lanes, and ten vehicles in addition to the ego vehicle. In this environment, the episode is terminated if any of the vehicles crash or if the time elapsed is greater than 40 time steps. The reward function is a combination of multiple factors, including staying in the right lane, the ego vehicle speed, and the speed of the other vehicles.

We test the risk-neutral CDQN and VDQN algorithms along with SENTINEL-4 with both additive and composite CVaRs. The typical performance metric for this scenario is the expected discounted return $\mathbb{E}^{\pi}_{\mu}[R]$. In order to test the risk-sensitive performance, we use two metrics. In order to measure aleatory risk $U_A[R \,|\, \pi, \mu]$, we use CVaR as U_A with threshold $\alpha = 0.25$. The CVaR metric is a statistic of the left-tail of the return distribution and higher values would mean better performance in the 25% worst-cases of performance. Finally, as a proxy for the epistemic risk, we use the number of crashes (lower is better).

Experimental results are illustrated in Table 1 and Figure 6. From Table 1, we observe that our algorithm with composite risk achieves a higher value, higher estimate of aleatory risk, and less number of crashes. Thus, SENTINEL-4 with composite CVaR aces the competing algorithms in all the three metrics of risk-sensitive and risk-neutral performances.

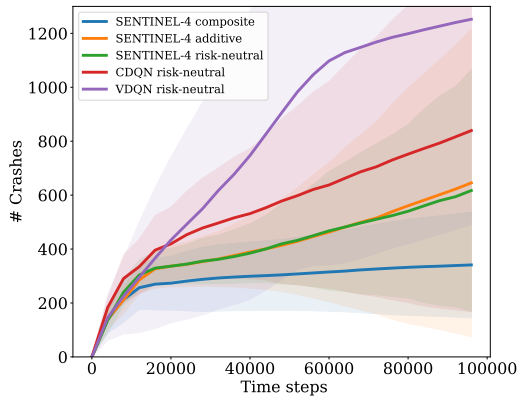


Figure 6: The total number of crashes in highway environment with 10 vehicles over 20 runs and horizon 10^6. Less #crashes indicate better risk-sensitive performance and the shaded region represents $\mu_t \pm \sigma_t$.

Additionally, we observe that the variance of performance metrics over 20 runs is the least for our algorithm with composite CVaR measure. This shows the stability of our algorithm which is another demonstration of good risk-sensitive performance. Figure 6 resonates with these observations in terms of the total number of crashes.

Summary of Results. Figure 4 shows that SENTINEL-K framework estimates even multimodal return distributions more efficiently than the classical distributional RL algorithms, such as VDQN. Figure 5 demonstrates that selecting λ is important in bootstrapped RL. We observe that it yields better performance over model averaging ($\lambda \to 0$) and greedy selection ($\lambda \to \infty$). Figure 6 shows the risk-sensitive performance of VDQN, CDQN, and SENTINEL-4 with risk-neutral, additive and composite CVaR risks on a large continuous state environment. SENTINEL-4 with composite risk outperforms competing algorithms in terms of the achieved value function and estimated aleatory risk. It causes the least number of crashes than competing algorithms.

7. Discussions

In this paper, we study the problem of risk-sensitive RL. We propose two main contributions. The first is the *composite risk* formulation that quantifies the holistic effect of aleatory and epistemic risk involved in the learning process. With a reductive experiment, we show that composite risk estimates the total risk involved in a problem more accurately than the additive formulation. The other one is *SENTINEL-K* which ensembles K distributional RL estimators, namely CDQNs, to provide an accurate estimate of the return distribution. We also reintroduce FTRL from bandit literature as a means of model selection. FTRL weighs each estimator differently depending on how far away they are from the average estimator. This leads to a better estimate of the composite risk over return. FTRL leads to better experimental performance than greedy selection and model averaging. Experiments also show that SENTINEL-K even in a risk-neutral setting estimates the return distribution of

all the actions better, and also achieves superior risk-sensitive performance while used with composite CVaR estimate.

Motivated by the experimental performances of SENTINEL-K, we aim to investigate the theoretical properties of FTRL-driven bootstrapped distributional RL with and without composite risk estimates.

Acknowledgments

We would like to thank Dapeng Liu for fruitful discussions in the beginning of the project, further, this work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation and the computations were performed on resources at Chalmers Centre for Computational Science and Engineering (C3SE) provided by the Swedish National Infrastructure for Computing (SNIC).

References

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 449–458. JMLR. org, 2017.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

Yinlam C., Mohammad G., Lucas J., and Marco P. Risk-constrained reinforcement learning with percentile risk criteria, 2015.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Y. Chow and M. Ghavamzadeh. Algorithms for cvar optimization in mdps. In *Advances in neural information processing systems*, pages 3509–3517, 2014.

Y. Chow, A. Tamar, S. Mannor, and M. Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. In *Advances in Neural Information Processing Systems*, pages 1522–1530, 2015.

William R Clements, Benoît-Marie Robaglia, Bastien Van Delft, Reda Bahi Slaoui, and Sébastien Toth. Estimating risk and uncertainty in deep reinforcement learning. *arXiv preprint arXiv:1905.09638*, 2019.

Stefano P Coraluppi and Steven I Marcus. Risk-sensitive and minimax control of discrete-time, finite-state markov decision processes. *Automatica*, 35(2):301–309, 1999.

W. Dabney, M. Rowland, Marc G. Bellemare, and R. Munos. Distributional reinforcement learning with quantile regression. In *AAAI*, 2018a.

Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018b.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1192–1201, 2018.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

Bradley Efron and Robert Tibshirani. The bootstrap method for assessing statistical accuracy. *Behaviormetrika*, 12(17):1–35, 1985.

Hannes Eriksson and Christos Dimitrakakis. Epistemic risk-sensitive reinforcement learning. *arXiv preprint arXiv:1906.06273*, 2019.

Stefan Faußer and Friedhelm Schwenker. Neural network ensembles in reinforcement learning. *Neural Processing Letters*, 41(1):55–69, 2015.

M. Friedman and L. J. Savage. The Utility Analysis of Choices Involving Risk. *The Journal of Political Economy*, 56(4):279, 1948.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Matthias Heger. Consideration of risk in reinforcement learning. In William W. Cohen and Haym Hirsh, editors, *Machine Learning Proceedings 1994*, pages 105–111. Morgan Kaufmann, San Francisco (CA), 1994.

Ronald A Howard and James E Matheson. Risk-sensitive markov decision processes. *Management science*, 18(7):356–369, 1972.

Ravi Kumar Kolla, Prashanth L. A., Sanjay P. Bhat, and Krishna P. Jagannathan. Concentration bounds for empirical conditional value-at-risk: The unbounded case. *Operations Research Letters*, 47(1):16–20, 2019.

Edouard Leurent. An environment for autonomous driving decision-making. `https://github.com/eleurent/highway-env`, 2018.

A Rupam Mahmood, Dmytro Korenkevych, Gautham Vasan, William Ma, and James Bergstra. Benchmarking reinforcement learning algorithms on real-world robots. In *Conference on robot learning*, pages 561–591. PMLR, 2018.

Steven I Marcus, Emmanual Fernández-Gaucherand, Daniel Hernández-Hernandez, Stefano Coraluppi, and Pedram Fard. Risk sensitive markov decision processes. In *Systems and control in the twenty-first century*, pages 263–279. Springer, 1997.

O. Mihatsch and R. Neuneier. Risk-sensitive reinforcement learning. *Machine learning*, 49 (2-3):267–290, 2002.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In *Advances in neural information processing systems*, pages 4026–4034, 2016.

Aldo Pacchiano, Philip Ball, Jack Parker-Holder, Krzysztof Choromanski, and Stephen Roberts. On optimism in model-based reinforcement learning. *arXiv preprint arXiv:2006.11911*, 2020.

Xinlei Pan, Yurong You, Ziyan Wang, and Cewu Lu. Virtual to real reinforcement learning for autonomous driving. *arXiv preprint arXiv:1704.03952*, 2017.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2817–2826, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

L. A. Prashanth and Michael C. Fu. Risk-sensitive reinforcement learning: A constrained optimization viewpoint. *arXiv*, 2018.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. *arXiv preprint arXiv:1902.08102*, 2019.

Roy E. Lave Jay K. Satia. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Giorgio Szegö. Measures of risk. *Journal of Banking & finance*, 26(7):1253–1272, 2002.

A. Tamar, S. Mannor, and H. Xu. Scaling up robust mdps using function approximation. In *International Conference on Machine Learning*, pages 181–189, 2014.

Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the cvar via sampling. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Yunhao Tang and Shipra Agrawal. Exploration by distributional reinforcement learning. *arXiv preprint arXiv:1805.01907*, 2018.

Marco A Wiering and Hado Van Hasselt. Ensemble algorithms in reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(4):930–936, 2008.