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 (Supplementary material)

Hannes Eriksson[1, 2] **Debabrota Basu**[3, 4] **Mina Alibeigi**[1] **Christos Dimitrakakis**[2,5]

[1]Zenseact AB, Gothenburg, Sweden
[2]Chalmers University of Technology, Gothenburg, Sweden
[3]Scool, INRIA Lille-Nord Europe, Lille, France
[4]CRIStAL, CNRS, Lille, France
[5]University of Neuchatel, Switzerland and University of Oslo, Norway

Abstract

In this paper, we consider risk-sensitive sequential decision-making in Reinforcement Learning (RL). Our contributions are two-fold. First, we introduce a novel and coherent quantification of risk, namely composite risk, which quantifies the joint effect of aleatory and epistemic risk during the learning process. Existing works considered either aleatory or epistemic risk individually, or as an additive combination. We prove that the additive formulation is a particular case of the composite risk when the epistemic risk measure is replaced with expectation. Thus, the composite risk is more sensitive to both aleatory and epistemic uncertainty than the individual and additive formulations. We also propose an algorithm, SENTINEL-K, based on ensemble bootstrapping and distributional RL for representing epistemic and aleatory uncertainty respectively. The ensemble of K learners uses Follow The Regularised Leader (FTRL) to aggregate the return distributions and obtain the composite risk. We experimentally verify that SENTINEL-K estimates the return distribution better, and while used with composite risk estimates, demonstrates higher risk-sensitive performance than state-of-the-art risk-sensitive and distributional 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]. In addition, since in RL the environment is by definition unknown to the agent, exploring it so as to improve performance and eventually obtain the optimal policy entails risks. Although the 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, 2015, Prashanth and Fu, 2018]. In this paper, we employ a model-free 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 (MDPs) 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 recent literature, researchers have either quantified epistemic and aleatory risks separately [Mihatsch and Neuneier, 2002, Eriksson and Dimitrakakis, 2020] or considered an additive risk formulation where their weighted sum is minimised by an RL algorithm [Clements et al., 2019].

In this work, we propose a *composite risk* formulation in order to accurately capture the combined effect of aleatory and epistemic uncertainty for decision-making in RL (Section 4). Our composition of risks relies on *coherent* risk measures, for which we show that their composition remains coherent. Our choice of focusing on coherent risk measures is also motivated by its extensive use and corresponding benefits in control theory Majumdar et al. [2017], decision theory Pflug and Pichler [2016], and reinforcement learning theory [Tamar et al., 2016, Ruszczyński, 2010, and references therein].

We incorporate composite risk measures within the Distribu-

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

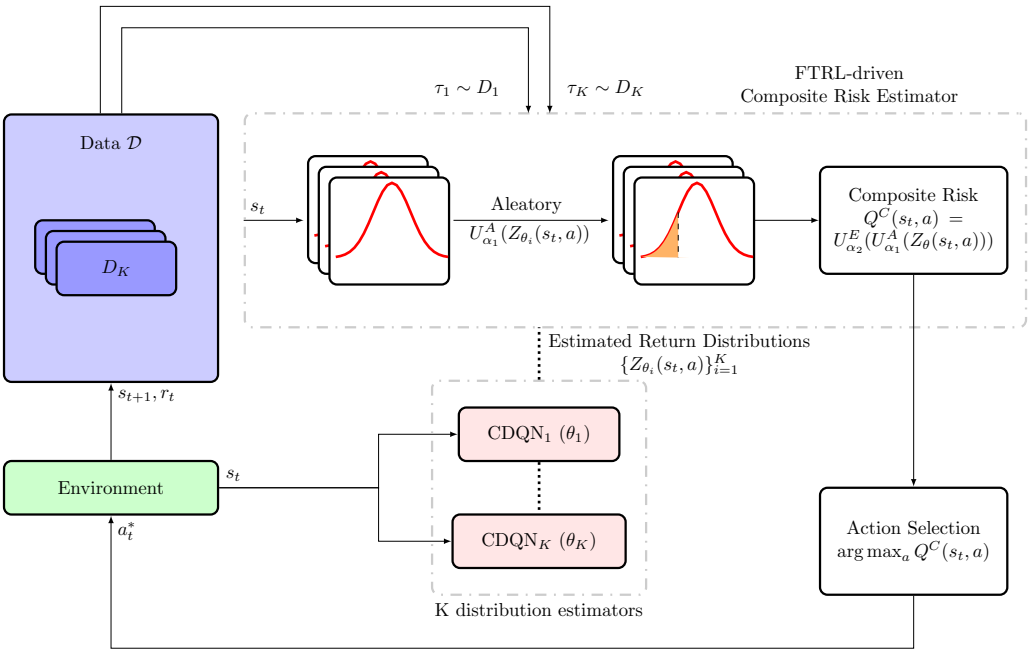


Figure 1: SENTINEL-K with FTRL-driven composite risk estimator and K CDQNs as return distribution estimators.

tional RL (DRL) framework [Bellemare et al., 2017, 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 3.2). 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. We us a bootstrapping [Efron and Tibshirani, 1985] framework to represent epistemic uncertainty. Our framework, which we call SENTINEL-K, is illustrated in Figure 1. At a high level, we use Categorical Deep Q Network (CDQN) [Bellemare et al., 2017] to model aleatory uncertainty and a bootstrapped ensemble for epistemic uncertainty. These can be used with any coherent measures and ensemble algorithm.

We discuss related work in Section 2. This is followed by some background on risk measures, Markov decision processes, and DRL in Section 3. SENTINEL-K is flexible enough to use any combination of coherent risk measures for aleatory and epistemic risks, as we explain in Section 4. The algorithm is described in detail in Section 5, with Section 5.1 and 5.2 showing how the ensemble is created and its members weighted respectively.

Section 6 examines the performance of SENTINEL-K with a composite CVaR metric on a highway environment with 10 cars. Our results show that our approach leads to fewer number of crashes than competing algorithms: Variational DQN (VDQN) [Tang and Agrawal, 2018], CDQN [Bellemare et al., 2017], total variance decomposition Uncertainty Aware-DQN (UA-DQN) [Clements et al., 2019], as well as SENTINEL-K with additive CVaR estimate, which we used as an ablation test to showcase the importance of the using a coherent composite risk. The supplementary material includes further experiments, showing that SENTINEL-K features significantly improved estimates of return distributions, and shows that using FTRL for weighing the ensemble members measurably improves performance.

2 RELATED WORK

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 that may lead to unsafe decisions in real-life. 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 risk-sensitive RL papers [Howard and Matheson, 1972, Coraluppi and Marcus, 1999, Marcus et al., 1997] focused on discrete state-space MDPs and either aleatory or epistemic risk. We are interested in designing a general risk-sensitive framework applicable to any type of state space and risk.

Both *aleatory* and *epistemic* uncertainties are important for risk-sensitive RL. The former expresses the *randomness* inherent to the problem and the latter a *lack of knowledge* about the problem. Aleatory risk-sensitivity in MDPs was first considered by [Howard and Matheson, 1972], who in-

troduced the idea of exponential utilities for the return.[1] Epistemic uncertainty in MDPs was investigated by [Satia, 1973], who provided game theoretic and Bayesian solution methods. Later works [Coraluppi and Marcus, 1999, Marcus et al., 1997, Mihatsch and Neuneier, 2002] extend risk-neutral methods to the risk-sensitive setting by using a non-linear utility [Garcıa and Fernández, 2015]. They consider aleatory risk-sensitive RL with exponential utility on the return [Mihatsch and Neuneier, 2002]. Follow-up works [Chow and Ghavamzadeh, 2014, C. et al., 2015] focus on scaling up these approaches. Other work on risk-sensitive RL focuses on CVaR [Chow and Ghavamzadeh, 2014, Tamar et al., 2015, Chow et al., 2015]. There have been recent works considering epistemic risk [Eriksson and Dimitrakakis, 2020], 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] considers the risk in the per-step rewards obtained in a MDP while Clements et al. [2019] proposes to use the additive formulation of epistemic and aleatory risks. Both of them use variance, which is not a coherent measure [Artzner et al., 1999]. Unlike previous work, our methodology of composite risk also allows us to apply any pair of coherent risk measures[2] to aleatory and epistemic uncertainty.

We instead define a generalised composite risk measure that takes into account both epistemic and aleatory uncertainty, and their entangled effect. Coherence is important, as we show that for any two coherent risk measures the composite risk retains coherence. This gives a principled approach for combining different application-appropriate risk measures for epistemic and aleatory uncertainties.

To express aleatory uncertainty, we rely on a distributional RL method called CDQN, which incorporates highly expressive approximators to model continuous and multimodal return distributions. In addition, we leverage ensemble methods to express epistemic uncertainty. Ensemble methods have first been used in risk-neutral RL by for representing epistemic uncertainty in order to improve exploration [Dimitrakakis, 2006, 2007]. This approach was later applied to MDPs by Osband et al. [2016]. On the other hand, Wiering and Van Hasselt [2008] used ensembles to combine policies instead. Ensembles have also been used to represent aleatory [Faußer and Schwenker, 2015, Pacchiano et al., 2020] uncertainty. Recently, [Depeweg et al., 2018, Clements et al., 2019] also use multiple Bayesian Neural Networks (BNNs) to estimate epistemic uncertainty. In the best of our knowledge, we are the first to use bootstrapped CDQNs for quantifying epistemic risk, which gives us freedom to model distributions on plausible MDPs without any structural assumptions, e.g. Gaussian distribution on parameters of Bayesian NNs or Gaussian distribution on state transitions [Clements et al., 2019]. An additional difference with prior work is that we use a follow the regularised leader (FTRL) algorithm to weigh the ensemble members in order to improve our uncertainty estimates.

[1]Here, we use return to mean the total discounted reward

[2]For example, CVaR, Wang risk measure [Wang, 2002], Standard Deviation (SD).

3 BACKGROUND

3.1 RISK MEASURES: COHERENCE

The idea of quantifying risk in decision making is long-studied in decision theory and has found multiple applications in finance and actuarial science. A *risk measure* maps a real-valued distribution to a real number, and quantifies the probability of occurrence of an event away from the expectation [Szegö, 2002]. Some well-known risk measures are variance, Value at Risk (VaR) and Conditional Value at Risk (CVaR). *Coherent* risk measures obey a set of axioms Artzner et al. [1999]: normalisation, monotonicity, subadditivity, homogeneity, and translation invariance. Not all risk measures are coherent: CVaR is coherent, but variance and VaR do not satisfy respect homogeneity and subadditivity respectively [Artzner et al., 1999].

If a coherent risk measure also satisfies comonotonic subadditivity [Song and Yan, 2009, Axiom 4], it can be expressed as an expectation over a distorted distribution function, for a concave *distortion function* $U_\alpha : [0,1] \to [0,1]$. Specifically (see [Wang et al., 1997, Theorem 2]) a random variable Z with associated probability measure P and cumulative distribution function F_Z satisfies:

$$\begin{aligned}\mathrm{Risk}_{U_\alpha}(Z) &\triangleq \int_{\mathcal{Z}} Z \,\mathrm{d}(U_\alpha \circ P) \\ &= \int_{\mathcal{Z}} U_\alpha(1 - F_Z(z))\,\mathrm{d}z = \int_0^1 U_\alpha(t)\,\mathrm{d}q(1-t), \quad (1)\end{aligned}$$

where $(U_\alpha \circ P)(A) \triangleq U_\alpha[P(A)]$ for any $A \subseteq \mathcal{Z}$. The last line is obtained from substitution of variables [Wirch and Hardy, 2001]. Here, q is the quantile function, i.e. $q(1-t) = \inf\{z \geq 0 | F_Z(z) \geq 1-t\} = F_Z^{-1}(1-t)$, $U(0) = 0$, and $U(1) = 1$. Since in this paper we use the risk measures for decision making, we represent a coherent risk measure through its corresponding *distortion function* U_α.

In this paper we focus on the *CVaR* [Rockafellar et al., 2000] risk measure. It is extensively used in risk-sensitive RL as it is coherent, applies to general L_p spaces, and captures the heaviness of the tail of a distribution. It is the expectation of the worst α-quantile of a probability distribution, with $\alpha \in [0,1]$:

$$CVaR_\alpha(Z) \triangleq \mathbb{E}[Z \,|\, Z \leq \nu_\alpha \wedge \mathbb{P}(Z \geq \nu_\alpha) = 1 - \alpha]. \quad (2)$$

For CVaR, $U_\alpha(t) = \min\{\frac{t}{1-\alpha}, 1\}$, For $\alpha = 1$, CVaR reduces to the expected value, and thus risk-neutrality.

Due to generality of our methodology and the composite risk formulation, we are able to incorporate other coherent risk measures such as the Wang risk measure [Wang, 2002], and standard deviation [Cirillo, 2017] (Fig. 4).

3.2 RL: MDP AND DISTRIBUTIONAL RL

MDPs. We consider 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 space 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. In the risk-neutral setup, the goal of the agent is to find a policy $\pi : \mathcal{S} \to \mathcal{A}$ to maximise expected value of cumulative rewards given a time horizon T: $V^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R(s_t, a_t)\right]$. Here, $s_t \sim \mathcal{T}(.|s_{t-1}, a_{t-1})$, $a_t = \pi(s_t)$, $s_0 = s$, $a_0 = a$, and the discount factor $\gamma \in (0,1)$.
Distributional RL. 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.

In this work, we choose to extend CDQN by Bellemare et al. [2017], 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 multimodal 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 represented as 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 μ and policy π, and thus persists even after correctly estimating the model μ. Let us now define risk measures for aleatory and epistemic uncertainties, and then combine them into a composite risk measure.

Aleatory Risk. Given a coherent risk measure with distortion function U_α^A, the aleatory risk is quantified as the deviation of total risk of individual models from the risk of the average model.

$$A(U_\alpha^A, \beta) \triangleq \int_\Theta \int_\mathcal{Z} Z \,\mathrm{d}(U_\alpha^A \circ \mathbb{P})(Z|\theta)\,\mathrm{d}\beta(\theta) - \int_\Theta \int_\mathcal{Z} \hat{Z}\,\mathrm{d}(U_\alpha^A \circ \mathbb{P})(\hat{Z})$$

Here, $\mathbb{P}(\hat{Z}) = \int_\Theta \mathbb{P}(Z|\theta)\,\mathrm{d}\beta(\theta)$, i.e. the return distribution of the average model. The centered definition of aleatory risk is necessary to show that additive risk is a special case of composite risk.

Epistemic Risk. Given a coherent risk measure with distortion function U_α^E, the epistemic risk quantifies the uncertainty invoked by not knowing the true model. Thus, the risk can be computed over any statistics of the models, such as expectation.

$$E(U_\alpha^E, \beta) \triangleq \int_\Theta \int_\mathcal{Z} Z\,\mathrm{d}\mathbb{P}(Z|\theta)\,\mathrm{d}(U_\alpha^E \circ \beta)(\theta)$$

Composite Risk under Model and Inherent Uncertainty. In typical risk-sensitive RL settings, the true MDP model is both unknown and inherently stochastic. Thus, the overall uncertainty is a composition of aleatory and epistemic uncertainties. For that reason, quantify it using what we call the *composite risk*.

Definition 1 (Composite Risk)**.** *For two coherent risk measures with distortion functions $U_{\alpha_1}^A$ and $U_{\alpha_2}^E$, belief distribution β on model parameters $\theta \in \Theta$, and a random variable $Z \in \mathcal{Z}$, the composite risk of epistemic and aleatory uncertainties is defined as*

$$\begin{aligned} F^C(U_{\alpha_1}^A, U_{\alpha_2}^E, \beta) &\triangleq \mathrm{Risk}_{U_{\alpha_2}^E}(\mathrm{Risk}_{U_{\alpha_1}^A}(Z|\theta)|\beta) \\ &= \int_\Theta \int_\mathcal{Z} Z\,\mathrm{d}(U_{\alpha_1}^A \circ \mathbb{P})(Z|\theta)\,\mathrm{d}(U_{\alpha_2}^E \circ \beta)(\theta) \\ &= \int_0^1 \int_0^1 U_{\alpha_2}^E(v) U_{\alpha_1}^A(u)\,\mathrm{d}q_{Z|\theta}(1-u)\,\mathrm{d}q_\beta(1-v) \quad (3) \end{aligned}$$

Here, $q_{Z|\theta}$ and q_β are quantile functions of Z conditioned on θ and that of θ respectively. For brevity, we also denote $F^C(U_{\alpha_1}^A, U_{\alpha_2}^E, \beta)$ as $\mathrm{Risk}_{U_{\alpha_2}^E} \circ \mathrm{Risk}_{U_{\alpha_1}^A}$ (e.g. CVaR $\circ$ CVaR), whenever it is clear from the context.

Theorem 2 (Coherence)**.** *If $U_{\alpha_1}^A$ and $U_{\alpha_2}^E$ are distortion functions for two coherent risk measures, the composite risk measure $F^C(U_{\alpha_1}^A, U_{\alpha_2}^E, \beta)$ is also coherent.*

The proof of Theorem 2 is available in Appendix B. The generic nature of our composite risk definition allows us to use different risk measures compatible with epistemic and aleatory risks. This is demonstrated in experiments (Figure 4) using different combinations of CVaR, Wang risk, and standard deviation for quantifying epistemic and aleatory uncertainties. This flexibility was absent in previous risk-sensitive RL literature [Eriksson and Dimitrakakis, 2020, Depeweg et al., 2018, Clements et al., 2019].

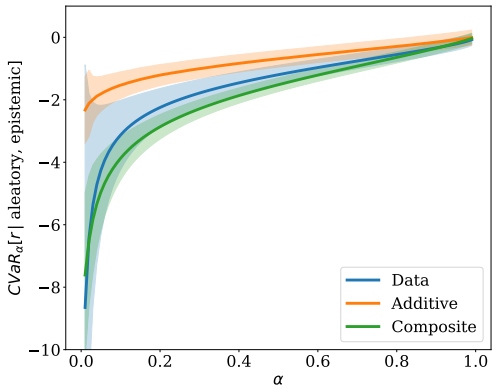


Figure 2: Estimation of total $CVaR_\alpha$ from a mixture of 100 Gaussians sampled from a posterior distribution. Total $CVaR_\alpha[Data]$ is based on the marginal distribution of r as in Example 1. We compare this with composite and additive estimates and illustrate results over 100 runs. Here, lower value of CVaR indicates higher mass on the left tail of the distribution and higher risk of obtaining low returns.

Comparison with Additive Risk Formulations. Clements et al. [2019], Depeweg et al. [2018] use a weighted sum of epistemic and aleatory variances as their risk measure. This formulation has mainly two problems. First, variance is not a coherent risk measure as it does not follow the homogeneity and subadditivity properties, as shown in [Cirillo, 2017]. Secondly, we show that even if we replace the variance with a coherent risk measure, the additive formulation is equivalent to considering U^E_α as an identity function. Thus, it is less sensitive to the effect of epistemic uncertainty than composite risk. More formally:

Theorem 3. *We are given two sources of aleatory and epistemic uncertainties ξ_1 and ξ_2. If $U^A_{\alpha_1}$ and $U^E_{\alpha_2}$ are distortion measures for two coherent risk measures quantifying aleatory and epistemic risks respectively, then, i) $F^A(U^A_{\alpha_1}, \beta) = F^C(U^A_{\alpha_1}, I, \beta)$, where I is the identity function, and ii) $F^C(U^A_{\alpha_1}, U^E_{\alpha_2}, \beta) \geq F^A(U^A_{\alpha_1}, \beta)$, if $\alpha_2 \neq 1$.*

Example 1 (A Reductive Empirical Evaluation of Composite and Additive Risks)**.** *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 $CVaR_\alpha[r]$ using the data generated from this mixture over 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 (specifically, $\alpha \leq 0.5$), 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.*

This means that for given sources of aleatory and epistemic uncertainties the additive risk which only considers expectation over epistemic uncertainty will always underestimate 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 and Figure 3).

5 ALGORITHM: SENTINEL-K

Now, we outline the algorithmic details of SENTINEL-K that estimates composite risk over returns using an ensemble of K distributional RL estimators, namely CDQN, in tandem with an adaptation of FTRL for estimator selection, and leverage the estimates for decision making.

Sketch of the Algorithm. Pseudocode of SENTINEL-K with composite risk is given in Algorithm 1. It has two main blocks: obtaining K estimates of return distribution with distributional RL framework (Lines 4- 13), and using them to compute composite risk for each action (Lines 15- 21). Finally, following the mechanism of Q-learning [Watkins and Dayan, 1992], it chooses the action with maximal composite risk in the decision making step (Line 23).

In the first block (Lines 4- 13), we specifically use an ensemble of K CDQNs. Each CDQN uses target and value networks for estimating the return distribution. We set a schedule for updating the target networks Γ_1 and a more frequent one ($\Gamma_1 \cup \Gamma_2$) for the value networks (Section 5.1).

The second block (Lines 15- 21) 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 $Q^A_i(s_t, a)$ of each of the estimators to provide a final estimate of the composite risk $Q^C(s_t, a)$ for each action, and then selecting the action with highest $Q^C(s_t, a)$.

5.1 ENSEMBLING AND BOOTSTRAPPING K-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 dataset $\mathcal{D}$ 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, on an average, each estimator i has access to $\frac{1}{3}$ of the dataset. Details about data masking are in Appendix D.1.

After preparing the datasets for the estimators, the target and value networks of the CDQN have to be updated and

optimised. 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 $Q_i^A(s_t, a) = \int_{\mathcal{Z}} Z \, \mathrm{d}(U_{\alpha_1}^A \circ \mathbb{P})(Z|\theta_i)$ 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 $Q^C(s_t, a) = \mathrm{Risk}_{U_{\alpha_2}^E}(\{Q_i^A(s_t, a)\}_{i=1}^K)$. Here, $\mathrm{Risk}_{U_{\alpha_2}^E}$ is the risk measure corresponding to the distortion $U_{\alpha_2}^E$. 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 WEIGHING ESTIMATES WITH FTRL

Now, the question is to adaptively and accurately aggregate the K estimated return distributions. Pacchiano et al. [2020] shows that adaptive model selection can boost performance in comparison to model averaging. The rationale for this can be given by seeing that some estimators might be overly optimistic or pessimistic. By weighing these less, you can effectively have a more robust ensemble. Further discussion of this issue is given in Appendix D.2.

We adapt the Follow The Regularised Leader (FTRL) algorithm [Cesa-Bianchi and Lugosi, 2006] studied in bandits and online learning for adaptively weighing the estimators. FTRL puts exponentially more weight on an estimator depending on its accuracy of estimating the return distribution. Since we do not 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 marginalised over $\beta(\theta)$, i.e. $l(\theta_i, \beta) \triangleq D_{\mathrm{KL}}\Big(\mathbb{P}(\hat{Z}) \,||\, \mathbb{P}(Z \mid \theta_i)\Big)$, as proxy of estimation loss of estimator i. FTRL selects estimator i with weight

$$w_i = \frac{e^{\lambda l(\theta_i, \beta)}}{\sum_j e^{\lambda l(\theta_j, \beta)}}, \quad \lambda \in [0, \infty). \tag{4}$$

Using FTRL weights for aggregating the K return distributions is analogous to using an exponentially weighted average forecaster [Cesa-Bianchi and Lugosi, 2006] on the K learners to create a final estimate of the return distribution and corresponding composite risk. This leads to a better aggregation of individual estimates than equally weighted average or a greedy selection of the best estimate [Cesa-Bianchi and Lugosi, 2006, Theorem 2.2]. 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, $Q_i^A(s_t, a) = \int_{\mathcal{Z}} Z \, \mathrm{d}(U_{\alpha_1}^A \circ \mathbb{P})(Z|\theta_i)$ (Line 18), and then the composite risk is computed by $Q^C(s_t, a) = \mathrm{Risk}_{U_{\alpha_2}^E}(\{w_i Q_i^A(s_t, a)\}_{i=1}^K)$ (Line 20). Here, $\lambda \in [0, \infty)$ 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. We experimentally show that performing FTRL with a reasonable λ value, namely 1, leads to better performance.

Action Selection. The algorithm always selects the action with the high composite risk Q^C. Its behaviour depends on the choice of risk measures or distortion utility functions $U_{\alpha_1}^A$ and $U_{\alpha_2}^E$. SENTINEL-K reduces to a risk-neutral algorithm if we choose both $U_{\alpha_1}^A, U_{\alpha_2}^E$ as identity functions, and to additive risk-sensitive algorithm if we choose $U_{\alpha_2}^E$ as identity. Designing it to accommodate composite risk provides us the flexibility to be risk-sensitive, risk-neutral, and treating epistemic and aleatory risk with different metrics.

6 EXPERIMENTAL EVALUATION

We test the risk-sensitive performance of SENTINEL-K with composite CVaR risk in two environments with continuous state spaces. We also display the flexibility of our composite risk formulation by evaluating heterogeneous risks with SENTINEL-K.[3] Settings for each of these experiments and results are elaborated in corresponding subsections. In all the experiments, we use 4 CDQNs in the ensemble and call it SENTINEL-4. We justify this choice of $K = 4$ in Appendix C.1. For each experiment, we report the mean and standard error of the mean over 20 runs for 10^5 steps.

Risk-sensitive Performance. In order to demonstrate performance in a larger domain, we opt to evaluate SENTINEL-4 in the *highway* [Leurent, 2018] environment. 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,

[3] Ablation studies for risk-neutral SENTINEL are in Appendix.

Table 1: Performance of risk-neutral (VDQN, CDQN, SENTINEL-K), aleatory risk-sensitive VDQN-CVaR, UA-DQN 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$_{RN}$ Tang and Agrawal [2018]	23.30 ± 0.36	14.29 ± 0.80	1252.33 ± 170.35
CDQN$_{RN}$ Bellemare et al. [2017]	25.96 ± 0.51	19.50 ± 1.44	839.53 ± 150.20
SENTINEL-4$_{RN}$	26.56 ± 0.32	20.88 ± 1.25	617.11 ± 100.15
VDQN-CVaR$_A$ Tang and Agrawal [2018]	24.39 ± 0.50	16.64 ± 1.25	871.33 ± 171.23
UA-DQN$_{E+A}$ Clements et al. [2019]	24.46 ± 0.29	16.9 ± 0.44	1060.65 ± 13.94
SENTINEL-4$_{E+A}$	26.82 ± 0.42	21.54 ± 1.40	645.55 ± 127.59
SENTINEL-4$_{E\circ A}$	$\mathbf{27.43 \pm 0.13}$	$\mathbf{24.16 \pm 0.54}$	$\mathbf{341.18 \pm 43.86}$

Algorithm 1 SENTINEL-K with Composite Risk

Input: Initial state s_0, action set $\mathcal{A}$, distortion measures $U^A_{\alpha_1}, U^E_{\alpha_2}$, hyperparameter λ, target networks $[\theta_1^-, ..., \theta_K^-]$, value networks $[\theta_1, ..., \theta_K]$, update schedule Γ_1, Γ_2.
for $t = 1, 2, \ldots$ **do**
  //* Update K-value and target networks for estimating return distributions *//
  for $t' \in \Gamma_1 \cup \Gamma_2$ **do**
    Generate $\{D_1, ..., D_K\} \leftarrow \text{DataMask}(\mathcal{D}^{t'})$
    for $i = 1, \ldots, K$ **do**
      Sample mini batch $\tau \sim D_i$
      Estimate (3) $F^C(Z(s_t, a)|U^A_{\alpha_1}, U^E_{\alpha_2}, \beta)$ using τ and K-target networks $\{\theta_i^-\}_{i=1}^K$.
      Get $a^* = \arg\max_a F^C(Z(s_t, a)|U^A_{\alpha_1}, U^E_{\alpha_2}, \beta)$
      Update value network θ_i using τ, a^*
      Update target network θ_i^- using τ, a^* if $t' \in \Gamma_1$
    end for
  end for
  //* Estimate the composite risk of each action using the estimated return distributions *//
  for $a \in \mathcal{A}$ **do**
    Compute weights $\mathbf{w} = w_1, ..., w_K$ from Eq. 4.
    for i **in** K **do**
      Compute aleatory risks $Q_i^A(s_t, a)$ from $\int_{\mathcal{Z}} Z \, \mathrm{d}(U^A_{\alpha_1} \circ \mathbb{P})(Z|\theta_i)$
    end for
    Compute composite risk over weighted aleatory estimates $Q^C(s_t, a) = \text{Risk}_{U^E_{\alpha_2}}(\{w_i Q_i^A(s_t, a)\}_{i=1}^K)$
  end for
  //* Action selection *//
  Take action $a_t = \arg\max_a Q^C(s_t, a)$
  Observe s_t and update the dataset $\mathcal{D}^t \leftarrow \mathcal{D}^{t-1} \cup \{s_t, a_{t-1}, s_{t-1}, r_{t-1}\}$
end for

and the speed of the other vehicles.

We test the risk-neutral CDQN and VDQN algorithms, an aleatory risk-sensitive VDQN and the total variance decomposition algorithm UA-DQN 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 $U^A_{\alpha_1}[R \,|\, \pi, \mu]$, we use CVaR as $U^A_{\alpha_1}$ 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 3. 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 outperforms the competing algorithms in terms of all three metrics. The simultaneous improvement in both the value function and #crashes is due to the fact that *highway* is designed to have a reward function that penalises unsafe driving. 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 3 resonates with these observations in terms of the total number of crashes.

Heterogeneous Risk Measures. In order to demonstrate the flexibility of the composite risk framework estimated with SENTINEL, we investigate performance using heterogeneous coherent risk measures, that composes different coherent risk measures for aleatory and epistemic risk. The chosen risk measures are aleatory and epistemic CVaR, aleatory and epistemic Wang risk, aleatory CVaR with epistemic standard deviation, and aleatory standard deviation with epistemic CVaR. Note that any combination of coherent risk measures is possible. We evaluate SENTINEL-4 in the *CartPole-v0* environment [Brockman et al., 2016]. This environment is a popular test-bed 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 terminated. This means that the undiscounted return attained per episode is in $[0, 200]$. Thus, we

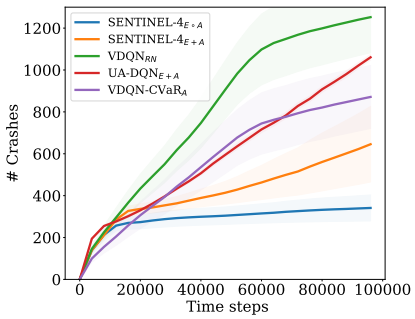


Figure 3: The total number of crashes in highway environment with 10 vehicles over 20 runs and horizon 10^6. Fewer #crashes indicate better risk-sensitive performance.

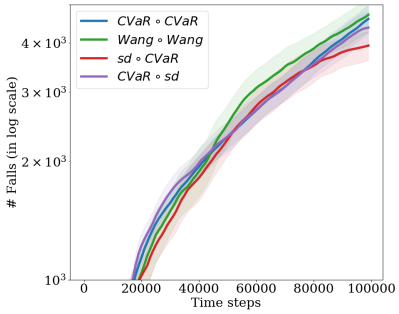


Figure 4: Performance and convergence of SENTINEL-4 using different risk measures. We show the number of falls in the *CartPole* environment over 20 runs with different initialisation.

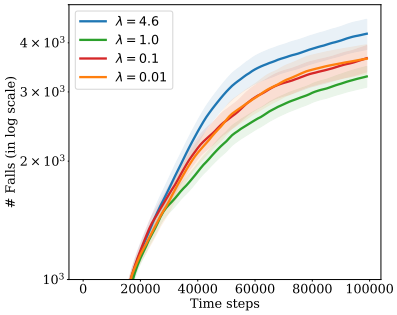


Figure 5: Performance and convergence of SENTINEL-4 (risk-neutral) for different values of λ. We show the number of falls in *CartPole* environment over 20 runs with different initialisation.

choose $V_{min} = 0, V_{max} = \frac{1-\gamma^{200}}{1-\gamma}$ as the histogram support of CDQN. The results are shown in Figure 4, which demonstrates than SENTINEL-4 performs flexibly and comparably for these composite risks.

FTRL vs. Average vs. Greedy. We choose $[0.01, 0.1, 1.0, \ln 100]$ as the different values of the regularising hyperparameter λ and test the performance of SENTINEL-4 for *CartPole-v0*. As $\lambda \to 0$, we perform standard model averaging which is sensitive to outliers. As $\lambda \to \infty$, model selection gets greedily biased towards the best average estimator while not providing other estimators a chance to improve. A sound value of λ would be one that excludes outlier estimators while still involves most of the other estimators. Figure 5 shows performance in terms of cumulative # Falls (lower is better) for the λ values with $CVaR_{0.25} \circ CVaR_{0.25}$. 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 #Falls is significantly less than that of other values and thus, stability of performance.

Summary of Results. Fig. 3 shows the risk-sensitive performance of VDQN, CDQN, aleatory CVaR, total variance decomposition UA-DQN and SENTINEL-4 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. Fig. 4 demonstrates the ability to chose any coherent risk measure for SENTINEL-K, including different risk measures for both epistemic and aleatory risk. Fig. 5 shows that selecting λ is important in bootstrapped RL, and tuning it yields better performance over model averaging ($\lambda \to 0$) and greedy selection ($\lambda \to \infty$). We defer the results on the choice of K in ensemble, convergence in return distribution, and improved efficiency in estimating multi-modal return distributions, to Appendix.

7 DISCUSSION

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 learning. With a reductive experiment, we show that composite risk estimates the total risk involved in a problem more accurately than existing additive formulations. The second one is *SENTINEL-K* which ensembles K distributional RL estimators, namely CDQNs, to provide an accurate estimate of the return distribution. We adopt FTRL from bandit literature as a means of model selection. FTRL weighs each estimator adaptively and leads to better experimental performance than greedy selection and model averaging. Experiments show that SENTINEL-K achieves superior risk-sensitive performance while used with composite CVaR estimate, and can operate on composition of different risks unlike existing works.

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

Acknowledgements

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 enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at C3SE partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

References

Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012.

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.

Pasquale Cirillo. About the Coherence of Variance and Standard Deviation as Measures of Risk. `https://courses.edx.org/c4x/DelftX/TW3421x/asset/coherence.pdf`, 2017. [Online; accessed 24-May-2021].

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.

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.

Christos Dimitrakakis. Nearly optimal exploration-exploitation decision thresholds. In *International Conference on Artificial Neural Networks*, pages 850–859. Springer, 2006.

Christos Dimitrakakis. *Ensembles for sequence learning*. PhD thesis, 2007.

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. In *ESANN*, pages 339–344, 2020.

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

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.

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.

Anirudha Majumdar, Sumeet Singh, Ajay Mandlekar, and Marco Pavone. Risk-sensitive inverse reinforcement learning via coherent risk models. In *Robotics: Science and Systems*, volume 16, page 117, 2017.

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.

Georg Ch Pflug and Alois Pichler. Time-consistent decisions and temporal decomposition of coherent risk functionals. *Mathematics of Operations Research*, 41(2):682–699, 2016.

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.

Andrzej Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical programming*, 125(2):235–261, 2010.

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

Yongsheng Song and Jia-An Yan. Risk measures with comonotonic subadditivity or convexity and respecting stochastic orders. *Insurance: Mathematics and Economics*, 45(3):459–465, 2009. ISSN 0167-6687. doi: https://doi.org/10.1016/j.insmatheco.2009.09.011. URL https://www.sciencedirect.com/science/article/pii/S0167668709001280.

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.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Sequential decision making with coherent risk. *IEEE Transactions on Automatic Control*, 62(7): 3323–3338, 2016.

Yunhao Tang and Shipra Agrawal. Exploration by distributional reinforcement learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2710–2716, 2018.

S. Wang. A risk measure that goes beyond coherence. 2002.

Shaun S. Wang, Virginia R. Young, and Harry H. Panjer. Axiomatic characterization of insurance prices. *Insurance: Mathematics and Economics*, 21(2):173–183, 1997. ISSN 0167-6687. doi: https://doi.org/10.1016/S0167-6687(97)00031-0. URL https://www.sciencedirect.com/science/article/pii/S0167668797000310. in Honor of Prof. J.A. Beekman.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

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.

Julia L Wirch and Mary R Hardy. Distortion risk measures: Coherence and stochastic dominance. In *International congress on insurance: Mathematics and economics*, pages 15–17, 2001.

A COHERENT RISK MEASURES

In the following section we expand on details that did not make it into the main paper.

A.1 FORMAL DEFINITIONS

Definition 4 (Coherent Risk Measure). *A coherent risk measure is a mapping $U : \Delta(\mathcal{X}) \to \mathbb{R}$ from a set of distributions on $\mathcal{X}$ to the real numbers, satisfying four axioms:*

Axiom 5 (Monotonicity). *If $X \leq Y$ almost surely, $U(X) \leq U(Y)$.*

Axiom 6 (Positive homogeneity). *For any $c \geq 0$, $U(cX) = cU(X)$.*

Axiom 7 (Translation invariance). *For any constant $a \in \mathbb{R}$, $U(X + a) = U(X) + a$.*

Axiom 8 (Subadditivity). *For $X, Y \in \mathcal{X}$, $U(X + Y) \leq U(X) + U(Y)$.*

Definition 9 (Conditional Value-at-Risk). *For a random variable Z quantifying risk and $\alpha \in [0, 1]$,*

$$CVaR_\alpha(Z) \triangleq \mathbb{E}[Z \mid Z \leq \nu_\alpha \wedge \mathbb{P}(Z \geq \nu_\alpha) = 1 - \alpha] \tag{5}$$

Definition 10 (Entropic Value-at-Risk). *For a random variable Z quantifying risk and $\alpha \in [0, 1]$,*

$$EVaR_\alpha(Z) = \inf_{\lambda > 0} \Big\{ \lambda^{-1} \ln \Big(\frac{M_Z(\lambda)}{1 - \alpha} \Big) \Big\} \tag{6}$$

Here, $M_Z(\lambda) \triangleq \mathbb{E}[\exp(\lambda Z)]$ is the moment generating function (MGF) of Z for any $\lambda \in \mathbb{R}$. For $\alpha = 0$, EVaR reduces to entropic risk measure or exponential utility based risk measure.

Definition 11 (Wang risk measure). *For a random variable Z quantifying risk and $\alpha \in [0, 1]$,*

$$WR_\alpha(Z) = \Phi[\Phi^{-1}(F(Z)) - \Phi^{-1}(\alpha)]. \tag{7}$$

Here, $F(Z)$ is the Cumulative Distribution Function (CDF) of Z. Φ and Φ^{-1} are the standard normal CDF and inverse normal CDF respectively. For $\alpha \leq 0.5$, this induces risk aversion and for $\alpha \geq 0.5$, it causes risk attraction.

A.2 OUR APPROACH OF COMPUTING RISK OVER RETURN DISTRIBUTIONS

The aforementioned coherent risk measures can also be written as an expectation over a distorted cumulative distribution function for a given distortion function g. Here, $g : [0, 1] \to [0, 1]$, $g(0) = 0$, and $g(1) = 1$. When combined with a probability measure P the distortion function defines a new function on events:

$$(g \circ P)(A) \triangleq g[P(A)].$$

The distortion function allows us to treat different samples with different risk-sensitive weights unlike standard expectation where $g(t) = t$.
Thus, given a distortion function g, we can compute the corresponding risk measure as

$$\begin{aligned}
\text{Risk}_g(Z|\alpha) &= \int_{\mathcal{Z}} z \, \mathrm{d}(g_\alpha \circ \mathbb{P})(z) \\
&= \int_{\mathcal{Z}} g_\alpha(1 - F_Z(z)) \, \mathrm{d}z \\
&= \int_0^1 g_\alpha(t) \, \mathrm{d}q(1 - t) \\
&= \int_0^1 q(1 - t) \, \mathrm{d}g_\alpha(t) = \int_0^1 q(1 - t) g'_\alpha(t) \, \mathrm{d}t.
\end{aligned} \tag{8}$$

Here, q is the quantile function, i.e. $q(1 - \alpha) = \inf\{x \geq 0 | F_Z(x) \geq 1 - \alpha\} = F_Z^{-1}(1 - \alpha)$. This is called the Wang transform, leads us to following observations:

1. For Categorical estimate of return distribution, computing risk measures will require defining a quantile function and then multiplying it with corresponding $g(\alpha_i)$ and adding over multiple $\alpha_i \in [0, 1]$.
2. For CVaR, $g_\alpha(t) = \min\{\frac{t}{1-\alpha}, 1\}$.
3. For EVaR, $g_\alpha(t) = \min_{\lambda>0}\Big\{\lambda^{-1} \ln\Big(\exp(\lambda t - \ln(1-\alpha))\Big)\Big\}$.
4. For the Wang risk measure, $g_\alpha(t) = \Phi[\Phi^{-1}(t) - \Phi^{-1}(\alpha)]$.

These observations allow us to compute the corresponding risk measures using quantile functions of the return distributions estimated using CDQNs. Note that, every coherent risk measure can be written using the Wang transformation or distorted expectation using quantile functions if and only if there exists a monotonic concave distortion function $g_\alpha(t)$ corresponding to it.

B DETAILED PROOFS

Proof of Theorem 2. Now, let us denote the aleatory and epistemic uncertainties of a random variable X as ξ_1 and ξ_2. Let us represent two coherent risk measures $U_A : \mathcal{X}|_{\xi_1} \to \mathcal{Z} \subseteq \mathbb{R}$ and $U_E : \mathcal{Z}|_{\xi_2} \to \mathbb{R}$ corresponding to distorted utility functions $U^E_{\alpha_2}$ and $U^A_{\alpha_1}$. Now, if U_A and U_E are coherent risk measures, we obtain

1. Monotonicity: If $U_A(X_1|_{\xi_1}) = Z_1$, $U_A(X_2|_{\xi_1}) = Z_2$, and $X_1|_{\xi_1} \leq X_2|_{\xi_1}$ almost surely,

$$U_E(U_A(X_1|_{\xi_1})|_{\xi_2}) = U_E(Z_1|_{\xi_2}) \leq U_E(Z_2|_{\xi_2}) = U_E(U_A(X_2|_{\xi_1})|_{\xi_2}).$$

   The inner inequality is true because if U_A is coherent risk measure, then $Z_1 = U_A(X_1|_{\xi_1}) \leq U_A(X_2|_{\xi_1}) = Z_2$.
2. Positive Homogeneity: For any $c \geq 0$,

$$U_E(U_A(c\,X|_{\xi_1})|_{\xi_2}) = U_E(c\,U_A(X|_{\xi_1})|_{\xi_2}) = c\,U_E(U_A(X|_{\xi_1})|_{\xi_2}).$$

3. Translation invariance: For any constant $a \in \mathbb{R}$,

$$U_E(U_A(X|_{\xi_1} + a)|_{\xi_2}) = U_E(U_A(X|_{\xi_1})|_{\xi_2} + a) = U_E(U_A(X|_{\xi_1})|_{\xi_2}) + a.$$

4. Subadditivity: For $X_1, X_2 \in \mathcal{X}$,

$$\begin{aligned} U_E(U_A((X_1 + X_2)|_{\xi_1})|_{\xi_2}) &\leq U_E((U_A(X_1|_{\xi_1}) + U_A(X_2|_{\xi_1}))|_{\xi_2}) \\ &\leq U_E(U_A(X_1|_{\xi_1})|_{\xi_2}) + U_E(U_A(X_2|_{\xi_1})|_{\xi_2}). \end{aligned}$$

Thus, composition of two coherent risk measures U_A and U_E quantifying the aleatory and epistemic uncertainties ξ_1 and ξ_2 is also a coherent risk measure. □

We observe that the existence of distorted utility functions $U^E_{\alpha_2}$ and $U^A_{\alpha_1}$ are not necessary to prove Theorem 2. We state the theorem statement with distorted utilities to maintain the flow of the text in the main paper.

Rather, we can leverage the observation that if a coherent risk measure U also satisfies comonotonic subadditivity [Song and Yan, 2009], we always get a concave distortion function g_α corresponding to it such that $g_\alpha(0) = 0$ and $g_\alpha(1) = 1$. Here,

1. **Comonotonic Sub-additivity:** If X_1 and $X_2 \in \mathcal{X}$ are comonotonic, then

$$U(X_1 + X_2) \leq U(X_1) + U(X_2).$$

2. **Comonotonicity:** Two random variables $X_1, X_2 \in \mathcal{X}$ are comonotonic, if and only if

$$[X_1(\omega_1) - X_1(\omega_2)][Y_1(\omega_1) - Y_2(\omega_2)] \geq 0$$

   almost surely for all ω_1 and ω_2 in the event space Ω.

In that case, the aforementioned proof of coherence of composite risk $U_E \circ U_A$ naturally extends for the composition of corresponding distorted utility functions $U^E_{\alpha_2}$ and $U^A_{\alpha_1}$.

Remark 12. *If the random variable Z follows a distribution P, any coherent risk measure $U : \Delta(\mathcal{Z}) \to \mathbb{R}$ can be written in a dual form:* $\mathrm{Risk}_{1-\alpha}(Z|P) = \sup_{Q \in Q_\alpha} \mathbb{E}[Z]$. *Here, Q_α is a set of distributions defined around P constrained by α and on support of P. For example, in case of CVaR, $Q_\alpha = \{Q \ll P : \frac{\mathrm{d}Q}{\mathrm{d}P} \leq \frac{1}{\alpha}$ almost surely$\}$, and in case of Entropic VaR [Ahmadi-Javid, 2012], $Q_\alpha = \{Q \ll P : D_{KL}(Q||P) \leq -\ln \alpha\}$. For $\alpha = 1$, the risk measures reduce to expectation and $Q_{\alpha=1} = \{P\}$.*

Proof of Theorem 3. Let $F^C(U_A, U_E, \beta) \triangleq U_E(U_A(X|_{\xi_A})|_{\xi_E})$ be the primal form of composite risk, and the dual form is: $F^C(U_A, U_E, \beta) \triangleq \sup_{\beta' \in \mathcal{B}eta_{\alpha_2}} \mathbb{E}_{\theta \sim \beta'} \left[\sup_{Q \in Q^\theta_{\alpha_1}} \mathbb{E}_{Z \sim Q(.|\theta)}[Z] \right]$.

Part a: By replacing the variance with a dual of a coherent risk measure in [Clements et al., 2019], we obtain:

$$\begin{aligned} F^A(U^A, \beta) &= A(U^A, \beta) + \sup_{Q \in Q^\theta_\alpha} \int_\Theta \int_{\mathcal{Z}} z \, \mathrm{d}Q(z|\theta) \, \mathrm{d}\beta(\theta) \\ &= \mathbb{E}_{\theta \sim \beta} \left[\sup_{Q \in Q^\theta_\alpha} \mathbb{E}_{Z \sim Q(.|\theta)}[Z] \right] \\ &= F^C(U_A, I, \beta). \end{aligned}$$

The penultimate equality is obtained by the centered definition of aleatory risk. The last inequality is a direct consequence of the definition of the composite risk.

Part 2: The second claim follows from Remark 12. First, we observe that $\beta \in \mathcal{B}eta_\alpha$ as $\mathcal{B}eta_\alpha = \{\beta' \ll \beta \,|\, f_1\left(\frac{\mathrm{d}\beta'}{\mathrm{d}\beta}\right) \leq f_2(\alpha)\}$ and $f_1\left(\frac{\mathrm{d}P}{\mathrm{d}P}\right) = 0 \leq f_2(\alpha)$ for any $\alpha \in [0, 1]$. Now, let us denote

$$\beta^* \triangleq \arg\sup_{\beta' \in \mathcal{B}eta_\alpha} \mathbb{E}_{\beta'}[-Z|_{\xi_E}].$$

Thus, if $\beta^* \neq \beta$, then $F^C(U_A, U_E, \beta) \geq F^C(U_A, I, \beta) = F^A(U_A, \beta)$. We conclude the proof by observing that for $\alpha \neq 1$, $\beta^* \neq \beta$.

□

The aforementioned proof is independent of the existence of distortion function. If it exists for the epistemic risk measure, the proof is even straightforward. If we assume that there exists a distortion function $U^E_{\alpha_2}$ for epistemic risk, we get $U^E_{\alpha_2}(t) \geq t$ for all $t \in [0, 1]$. Because $U^E_{\alpha_2}$ is *concave*, and $U^E_{\alpha_2}(0) = 0$ and $U^E_{\alpha_2}(1) = 1$. Thus, it will be almost always above t by definition of concave function.

C ADDITIONAL EXPERIMENTAL RESULTS

In this section we provide additional experiment results that did not make it into the main paper. These involve testing the distributional fit of the return distribution using two different distributional RL frameworks, an empirical experiment demonstrating why composite risk is preferable over additive risk and

C.1 EFFECT OF ENSEMBLE SIZE ON PERFORMANCE AND COMPUTATION TIME

In the following experiment we investigate how the number of estimators in the ensemble affect the results and running time of the algorithm.

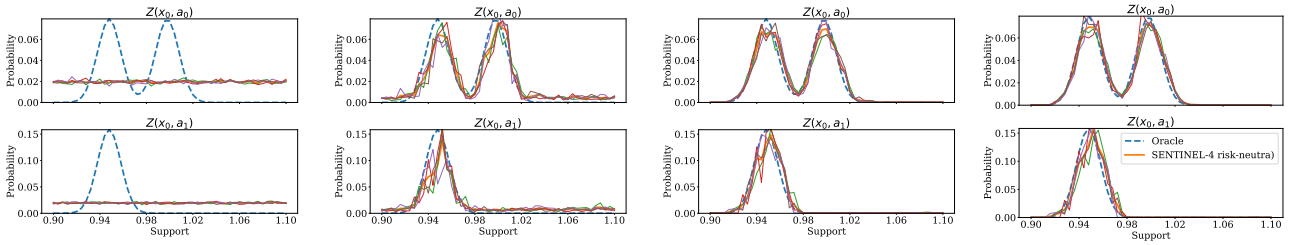
Table 2: Performance of risk-neutral SENTINEL-K in the *CartPole-v0* environment. Shown is number of *falls* (lower is better) and the time elapsed per experiment in seconds. The results were taken over 20 independent runs for each ensemble size. σ is the standard error of the mean.

Ensemble size	# Falls $\pm\sigma$	Time elapsed (s) per experiment $\pm\sigma$
$K = 1$	5332.4 ± 404.86	469.8 ± 14.2
$K = 2$	4627.8 ± 386.2	1886.3 ± 130.6
$K = 4$	4357.9 ± 334.4	4285.6 ± 204.5
$K = 8$	3532.8 ± 207.5	16528.2 ± 1479.9

In Table 2 we can see a monotonic increase in performance with the size of the ensemble. However, with each added estimator we can also observe a sizeable increase in computation time per experiment. Thus, there is a trade-off between computation time and performance and while more estimators would be preferable to use we chose to use $K = 4$ for most of the experiments.

C.2 RETURN DISTRIBUTION ESTIMATION

In these experiments, we verify the goodness of fit of the used DRL framework (CDQN) and compare the results with another DRL framework (VDQN).



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

Figure 6: 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 $\mathbb{P}(\hat{Z})$ with SENTINEL-4. The thin lines are the posteriors of the individual estimators.

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 $Z(s_0, a_1) \sim \mathcal{N}(\mu_1, \sigma_1)$. Here, $\Phi = [0.5, 0.5]$, $\mu = [1.0, 0.95]$, $\sigma = [0.1, 0.1]$. Figure 6 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 7, 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.

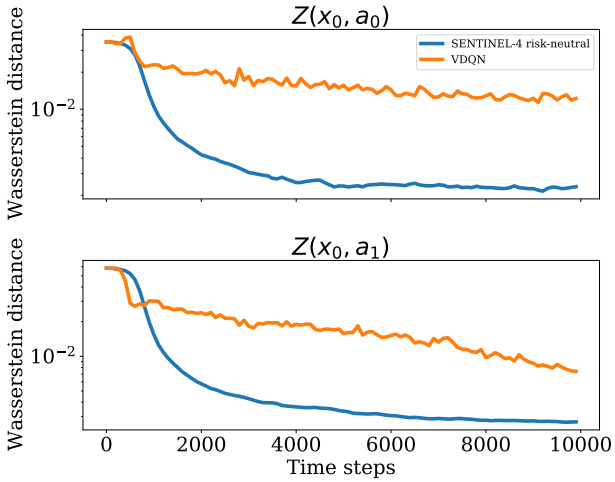


Figure 7: Shows convergence in distribution of SENTINEL-4 (risk-neutral) and VDQN by measuring the Wasserstein distance between the categorical approximation of $Z(s_0, a_0), Z(s_0, a_1)$ and the estimated distributions by the two agents, for each action.

C.3 COMPOSITE RISK VS. ADDITIVE RISK

In order to compare the risk estimation using additive and composite formulations, we consider an example of estimating CVaR over a Gaussian mixture.

Example 2. *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 8 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 *(specifically,* $\alpha \leq 0.5$*), 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.*

In the following example the sensitivity w.r.t the parameter uncertainty of the composite risk formulation is shown. As the belief concentrates, both the composite and additive risk formulations ends up with the optimal behaviour for this problem, as seen in Figure 8. The main difference in behaviour arises in situations with high parameter uncertainty.

Example 3 (Composite vs. Additive risk for Gaussian estimators)**.** *Let* $Z(z; s_1, a_1) = \mathcal{N}(z; \theta_1, \theta_2), Z(z; s_1, a_2) = \omega\mathcal{N}(z; \theta_3, \theta_4) + (1 - \omega)\mathcal{N}(z; \theta_5, \theta_6)$ *for* $\omega \in [0, 1]$*. Let* $\omega \sim Beta(\theta_7, \theta_8)$*. Then, let* $\theta = [0, 1, 1, 1, -1, 1, 1, 1]^\top$*. In Figure 8 the example is illustrated.*

D ADDITIONAL DETAILS

In this section we further describe the experimental details and the parameters chosen for said experiment.

D.1 DATA MASKING

Similar to [Osband et al., 2016], we use data masking to ensure the estimators have access to different parts of the data. The authors in that paper mention a few ways of doing this, namely using a Bernoulli mask, an exponential mask and a Poisson mask. In this work we chose to use a Bernoulli mask with parameter $p = \frac{1}{3}$. This means that in expectation, each estimator has access to a third of the full data set $\mathcal{D}$.

Upon observing a transition s, a, r, s', we sample K parameters from $Ber(\frac{1}{3})$, and assign those parameters to each estimator,

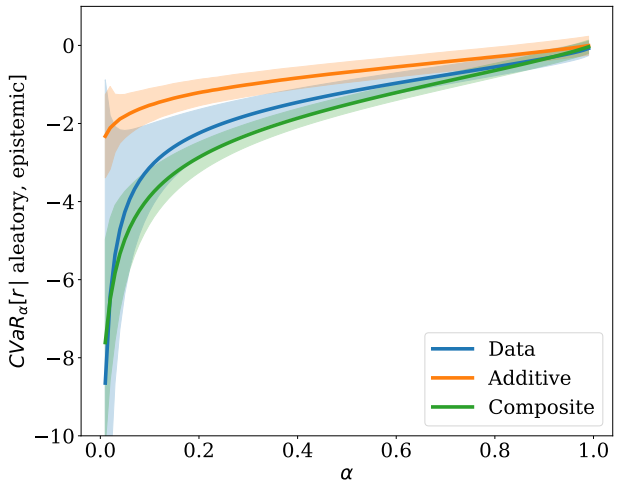


Figure 8: Estimation of total $CVaR_\alpha$ from a mixture of 100 Gaussians sampled from a posterior distribution. Total $CVaR_\alpha[Data]$ is based on the marginal distribution of the r as given in Example 2. We compare this with composite and additive estimates and illustrate results over 100 runs. Here, lower value of CVaR indicates higher mass on the left tail of the distribution and thus, higher risk of obtaining low returns.

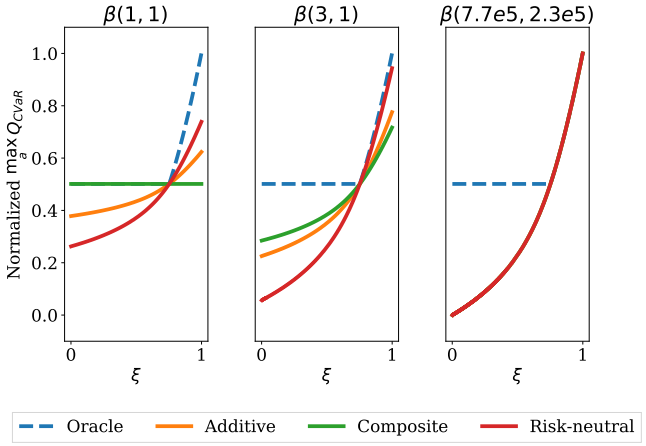


Figure 9: Illustrates the behaviour of a risk-neutral decision-maker and two risk-sensitive decision-makers, with additive and composite risk respectively. Shown in the figure is the normalized $\max_a Q_{CVaR}$ for $\alpha = 0.25$.

respectively, for that particular transition. As an example, consider the following table in Table 3, where the columns of the data $\mathcal{D}$ has been augmented with the Bernoulli mask.

In this example, the first estimator will have the first and the last transition available to it, since $m_{1,t} = m_{1,T} = 1$, while the second and the K'th estimator will not have access to the first transition, since $m_{2,t} = m_{K,t} = 0$.

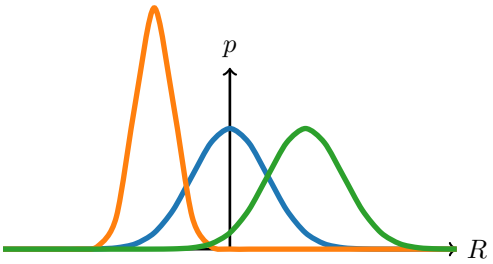
Table 3: An example of data masking on transitions $[\tau_t, \tau_T]$, where s_t denotes the state at time t, a_t the action taken at time t, r_t the reward received at time t and s_{t+1} the successor state of s_t at time t. $m_{1,t}$ denotes the availability of transition t to the first estimator.

s	a	r	s'	m_1	m_2	$\cdots$	m_K
s_t	a_t	r_t	s_{t+1}	1	0	$\cdots$	0
s_{t+1}	a_{t+1}	r_{t+1}	s_{t+2}	0	0	$\cdots$	1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
s_{T-1}	a_{T-1}	r_{T-1}	s_T	1	1	$\cdots$	1

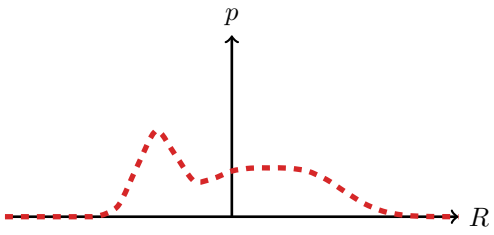
D.2 ADDENDUM ON *FOLLOW THE REGULARISED LEADER*

Follow the regularised leader in our setting can be seen as a mapping $f_\lambda : \mathcal{X}^K \to \mathcal{X}$, where the mapping is taken over a convex combination over the K densities. Let $f_k(x)$ be the k'th probability density function over x, then,

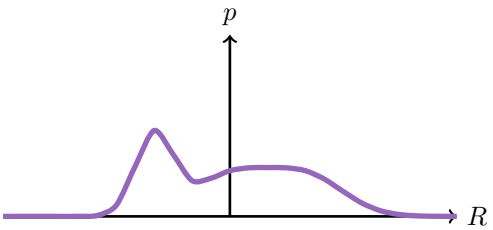
$$f_\lambda(x) = \sum_{k=1}^{K} w_k f_k(x)$$
$$\text{where, } ||\mathbf{w}||_1 = 1 \wedge \forall_k w_k \geq 0.$$



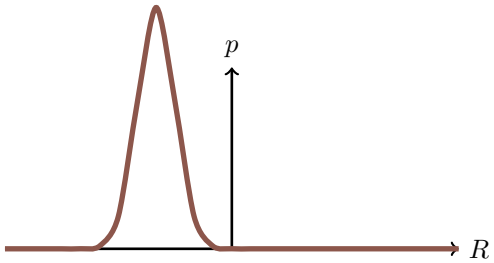
(a) Estimator posteriors $f_k(x)$



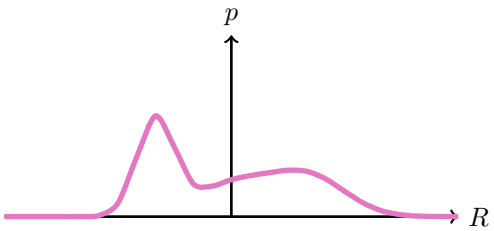
(b) Marginal over estimators $\frac{1}{K}\sum_k f_k(x)$



(c) FTRL($\lambda = 0$)



(d) FTRL($\lambda \to \infty$)



(e) FTRL($\lambda = 1$)

Figure 10: Shows how FTRL transforms a set of probability distributions into a weighted average, with the weights depending on the regularising parameter λ. In (a), three Gaussian distributions are given and in (b), the marginal distribution over returns can be seen, having marginalised out the estimators. In (c, d, e) the weighted distributions can be seen when varying λ. If $\lambda = 0$, as in (c), then the weighted distribution is the same as in (b). If $\lambda \to \infty$ as in (d), then the estimator that is closest to the marginal (using Kullback-Leibler divergence) will be assigned all weight. Finally, for $0 \leq \lambda < \infty$ the weighting is somewhere in between *Average* and *Greedy* selection.

Consider the following example with Gaussian estimators.

Example 4 (FTRL with Gaussian estimators)**.** *Let $f_k(x) = \mathcal{N}(\mu_k, \sigma_k^2), \mu = [0, -2, 2], \sigma = [1, 0.5, 1]$, respectively. We can compute the Kullback-Leibler divergence from each of the estimators to the marginal experimentally, and get approximately the following, $D_{KL}\Big(\frac{1}{K}\sum_k f_k(x) \,||\, f_i(x)\Big) = [0.48, 0.92, 0.74]$. Now, using the exponentiated FTRL approach as defined in Section 5, we get that $w = \exp\Big([0.48\lambda, 0.92\lambda, 0.74\lambda]\Big)/(w_1 + w_2 + w_3)$. This leads to $w = [0.26, 0.40, 0.34]$ for $\lambda = 1$ as used in most of the experiments. If $\lambda = 0$, then $w = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$. Finally, if $\lambda \to \infty$, then $w = [0, 1, 0]$, which assigns all probability to the estimator that is the most similar to the marginal distribution. The distributions given from FTRL, the original estimators and the marginal distribution can be seen in Figure 10.*

D.3 COMPUTE SPECIFICATIONS AND TOTAL COMPUTE

In this section we explain the specifications of the computers the experiments were ran in, and compare the compute time for the different algorithms. In Table 4 the compute time is shown for CDQN [Bellemare et al., 2017], UA-DQN [Clements et al., 2019], VDQN [Tang and Agrawal, 2018] and the proposed algorithm in this paper.

The majority of the computations were ran on NVIDIA Tesla T4 GPUs with 16 GB RAM and 16 core Intel(R) Xeon(R) Gold 6226R CPUs @ 2.90GHz with 768GB DDR4 RAM and the hyperparameters were selected such that the shared hyperparameters are the same and with the algorithm specific parameters chosen such that the overall compute time is similar in most cases.

Table 4: Compute time for the different algorithms in the *Highway-v1* environment. Shown is the time elapsed per experiment in seconds.

Algorithm	Time elapsed (s) per experiment
CDQN	≈ 10614
UA-DQN	≈ 13748
VDQN	≈ 13069
SENTINEL-4	≈ 36384

E HYPERPARAMETERS

In this section we show the hyperparameters used for the different experiments, including problem parameters, algorithm parameters and network structure. The choice of hyperparameters is firstly done to match the shared parameters, (such as minibatch size, update schedules and ensemble size), then secondly to match the number of learnable parameters for the model. Finally, we attempt to match the computation time.

E.1 FTRL VS. AVERAGE. VS GREEDY.

The experimental results for this experiment can be seen in Figure 5, where the *follow the regularised leader* parameter λ is varied across experiments.

Table 5: Hyperparameters for the FTRL vs. Average vs. Greedy experiment.

	Hyperparameter	Value
Problem parameters		
	Environment	*CartPole-v0*
	State dimensions $\lvert\mathcal{S}\rvert$	4
	Action dimensions $\lvert\mathcal{A}\rvert$	2
	Maximum episode length	200
Algorithm parameters		
	Discount γ	0.99
	Number of atoms	51
	Maximum steps in env	$1e5$
	Initial ϵ	1.0
	Final ϵ	0.05
	Samples from replay buffer	100
	Replay buffer size	$1e5$
	Minibatch size	32
	Regularising parameter λ	$[0.01, 0.1, 1.0, 4.6]$
	Return distribution range $[V_{min}, V_{max}]$	$[0, \frac{1-\gamma^{200}}{1-\gamma}]$
	Update ensembles at steps	$[100, 200, \cdots]$
	Update target ensembles at steps	$[1000, 2000, \cdots]$
	Ensemble size K	4
	Learnable parameters	$(32\,\lvert\mathcal{S}\rvert + 11891)K\lvert\mathcal{A}\rvert$
	Optimiser	*Adam*
	Learning rate	0.00025
	Network structure	$\underset{\text{input}}{4} \rightarrow \underset{\text{dense}}{32} \rightarrow \underset{\text{dense}}{32} \rightarrow \underset{\text{dense}}{128} \rightarrow \underset{\text{output}}{(51, 2)}$

E.2 EFFECT OF ENSEMBLE SIZE ON PERFORMANCE AND COMPUTATION TIME

In Table 2 the results are shown for varying the ensemble size. In this section, we hyperparameters used for that experiment are displayed.

Table 6: Hyperparameters for the Effect of Ensemble Size on Performance and Computation Time experiment.

	Hyperparameter	Value
Problem parameters		
	Environment	*CartPole-v0*
	State dimensions $\|\mathcal{S}\|$	4
	Action dimensions $\|\mathcal{A}\|$	2
	Maximum episode length	200
Algorithm parameters		
	Discount γ	0.99
	Number of atoms	51
	Maximum steps in env	$1e5$
	Initial ϵ	1.0
	Final ϵ	0.05
	Samples from replay buffer	100
	Replay buffer size	$1e5$
	Minibatch size	32
	Regularising parameter λ	1.0
	Return distribution range $[V_{min}, V_{max}]$	$[0, \frac{1-\gamma^{200}}{1-\gamma}]$
	Update ensembles at steps	$[100, 200, \cdots]$
	Update target ensembles at steps	$[1000, 2000, \cdots]$
	Ensemble size K	$[1, 2, 4, 8]$
	Learnable parameters	$(32\,\|\mathcal{S}\| + 11891)K\|\mathcal{A}\|$
	Optimiser	*Adam*
	Learning rate	0.00025
	Network structure	$\underset{\text{input}}{4} \rightarrow \underset{\text{dense}}{32} \rightarrow \underset{\text{dense}}{32} \rightarrow \underset{\text{dense}}{128} \rightarrow \underset{\text{output}}{(51, 2)}$

E.3 EXPERIMENTS WITH HETEROGENOUS RISK MEASURES

In Figure 5 the results are shown for experiment with different coherent risk measures. In this section the hyperparameters used for that experiment is shown.

Table 7: Hyperparameters for the Experiments With Different Risk Measures.

	Hyperparameter	Value
Problem parameters		
	Environment	*CartPole-v0*
	State dimensions $\|\mathcal{S}\|$	4
	Action dimensions $\|\mathcal{A}\|$	2
	Maximum episode length	200
Algorithm parameters		
	Discount γ	0.99
	Number of atoms	51
	Maximum steps in env	$1e5$
	Initial ϵ	1.0
	Final ϵ	0.05
	Samples from replay buffer	100
	Replay buffer size	$1e5$
	Minibatch size	32
	Regularising parameter λ	1.0
	Return distribution range $[V_{min}, V_{max}]$	$[0, \frac{1-\gamma^{200}}{1-\gamma}]$
	Update ensembles at steps	$[100, 200, \cdots]$
	Update target ensembles at steps	$[1000, 2000, \cdots]$
	Ensemble size K	4
	$CVaR \circ CVaR\,(\alpha_E, \alpha_A)$	(Epistemic 0.25, Aleatory 0.25)
	$Wang \circ Wang\,(\alpha_E, \alpha_A)$	(Epistemic 0.10, Aleatory 0.10)
	$CVaR \circ sd\,(\alpha_E, \alpha_A)$	(Epistemic 0.25, Aleatory 1.0)
	$sd \circ CVaR\,(\alpha_E, \alpha_A)$	(Epistemic 1.0, Aleatory 0.25)
	Learnable parameters	$(32\,\|\mathcal{S}\| + 11891)K\|\mathcal{A}\|$
	Optimiser	*Adam*
	Learning rate	0.00025
	Network structure	$\underset{\text{input}}{4} \rightarrow \underset{\text{dense}}{32} \rightarrow \underset{\text{dense}}{32} \rightarrow \underset{\text{dense}}{128} \rightarrow \underset{\text{output}}{(51, 2)}$

E.4 HYPERPARAMETERS FOR THE HIGHWAY EXPERIMENT

In Figure 3 and Table 1 the results are shown for the highway environment. In this section the hyperparameters are shown used for that experiment.

Table 8: Hyperparameters for the Large State Space Risk-Sensitive Experiments.

	Hyperparameter	Value
Problem parameters		
	Environment	*Highway-v1*
	State dimensions $\lvert\mathcal{S}\rvert$	25
	Action dimensions $\lvert\mathcal{A}\rvert$	5
	Maximum episode length	40
	Vehicles count	10
Algorithm parameters		
–Shared	Discount γ	0.99
	Maximum steps in env	$1e5$
	Initial ϵ	1.0
	Final ϵ	0.01
	Samples from replay buffer	1000
	Replay buffer size	$1e5$
	Update ensembles at steps	$[1, 3, 6, 10, \cdots]$
	Update target ensembles at steps	$[15, 55, 120, \cdots]$
	Ensemble size K	4
	Optimiser	*Adam*
	Learning rate	0.00025
	Minibatch size	32
–SENTINEL-K	Number of atoms	51
	Regularising parameter λ	1.0
	Return distribution range $[V_{min}, V_{max}]$	$[0, 40]$
	Risk-sensitive parameters (α_E, α_A)	$(0.25, 0.25)$
	Learnable parameters	$(32\,\lvert\mathcal{S}\rvert + 11891)K\lvert\mathcal{A}\rvert$
	Network structure	$\underset{\text{input}}{25} \rightarrow \underset{\text{dense}}{32} \rightarrow \underset{\text{dense}}{32} \rightarrow \underset{\text{dense}}{128} \rightarrow \underset{\text{output}}{(51, 5)}$
–UA-DQN	Risk-sensitive parameters (β, λ)	$(0.2, 0.1)$
	Learnable parameters	$(256\,\lvert\mathcal{S}\rvert + 1541)K\lvert\mathcal{A}\rvert$
	Network structure	$\underset{\text{input}}{25} \rightarrow \underset{\text{dense}}{256} \rightarrow \underset{\text{output}}{5}$
–VDQN-CVaR	Risk-sensitive parameter α	0.25
	Learnable parameters	$(400\lvert\mathcal{S}\rvert + 400)\lvert\mathcal{A}\rvert$
	Network structure	$\underset{\text{input}}{25} \rightarrow \underset{\text{dense}}{100} \rightarrow \underset{\text{output}}{5}$