Policy Gradients for Cumulative Prospect Theory in Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

We derive a policy gradient theorem for Cumulative Prospect Theory (CPT) objectives in finite-horizon Reinforcement Learning (RL), generalizing the standard policy gradient theorem and encompassing distortion-based risk objectives as special cases. Motivated by behavioral economics, CPT combines an asymmetric utility transformation around a reference point with probability distortion. Building on our theorem, we design a first-order policy gradient algorithm for CPT-RL using a Monte Carlo gradient estimator based on order statistics. We establish statistical guarantees for the estimator and prove asymptotic convergence of the resulting algorithm to first-order stationary points of the (generally non-convex) CPT objective. Simulations illustrate qualitative behaviors induced by CPT and compare our first-order approach to existing zeroth-order methods.

1 Introduction

In classical reinforcement learning (RL), rational agents make decisions to maximize their expected cumulative rewards through interaction with their environment. This paradigm is primarily guided by expected utility theory, which has long been the dominant framework for decision-making under risk. Within this framework, agents are generally considered to be risk-neutral; however, risk-seeking and risk-averse behaviors can also be modeled by modifying the utility function, thereby adjusting the policy optimization objective (see e.g. Prashanth et al. (2022); Biswas & Borkar (2023); Bäuerle & Jaśkiewicz (2024) for recent surveys).

While risk-sensitive RL extends beyond risk-neutral settings to capture individual risk preferences (using e.g. variance or conditional value at risk metrics), many commonly used risk-sensitive objectives capture particular aspects of risk preferences (e.g., tail risk) but do not simultaneously model gain–loss asymmetry around a reference point together with probability distortion. In particular, many standard risk-sensitive criteria do not model the asymmetric perception of gains and losses, as well as the probability distortions inherent in human cognition, such as the tendency to overestimate rare events and underestimate frequent ones (see App. F for an illustration in a simulation). These behavioral phenomena, relevant in human-facing decision problems under uncertainty, are beyond the scope of traditional risk-sensitive RL frameworks, necessitating a more comprehensive approach.

Cumulative Prospect Theory (CPT), introduced by Kahneman and Tversky in their seminal works combining cognitive psychology and economics (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992), provides a descriptive behavioral model of decision-making under risk to explain several empirical observations that challenge the standard expected utility theory. CPT models how individuals perceive outcomes asymmetrically, being risk-averse in the domain of gains and risk-seeking in the domain of losses, and distort probabilities to reflect cognitive biases. These insights have led to widespread applications of CPT in stateless static settings in domains such as healthcare in psychiatry (Sip et al., 2018; George et al., 2019; Mkrtchian et al., 2023), chronic diseases treatment (Zhao et al., 2023) and emergency decision making (Sun et al., 2022) where modeling behavioral factors can be important, as well as other application domains such as energy (Ebrahimigharehbaghi et al., 2022; Dorahaki et al., 2022) and finance (Ladrón de Guevara Cortés et al., 2023; Luxenberg et al., 2024) to name a few. However, these applications often overlook the sequential decision-making nature of many real-world problems, where the outcome of one decision can affect future choices, a critical aspect of RL.

The integration of CPT into RL provides a promising avenue for behaviorally-aligned sequential decision-making, as it allows RL agents to consider both risk preferences and probability distortions in dynamic environments. While a few isolated recent works have explored this integration (L.A. et al., 2016; Borkar & Chandak, 2021; Ramasubramanian et al., 2021; Danis et al., 2023; Foo et al., 2023), the understanding and practical impact of CPT in RL remains limited. Specifically, the computational challenges and theoretical underpinnings of CPT-based RL models are still underexplored.

Contributions. In this work, we focus on the policy optimization problem where the objective is the CPT value of the cumulative sum of rewards, induced by a parametrized policy in a Markov Decision Process. Our main contributions are summarized as follows:

(i) Policy gradient theorem for CPT-RL. We establish a policy gradient theorem providing a closed form expectation expression for the gradient of our CPT-value objective w.r.t. the policy parameter under suitable regularity conditions on the utility and probability distortion functions. This result generalizes the standard policy gradient theorem in RL.

(ii) Policy gradient algorithm for CPT-RL. Building on our policy gradient theorem, we design a policy gradient (PG) algorithm to solve the CPT policy optimization problem. Our PG algorithm for CPT-RL uses first-order information and does not rely on zeroth-order policy gradient estimation.

(iii) Convergence and sample complexity. We show that our PG estimator is consistent and we analyze its sample complexity. Notably, our sample complexity scales logarithmically in the policy-parameter dimension (under our estimator and regularity assumptions), in contrast to existing zeroth-order PG estimators. We further show that the iterates of our PG algorithm converge asymptotically to the set of stationary points of the CPT-RL objective which is typically non-convex in the policy parameters. In particular, our analysis does not require increasing batch sizes to establish asymptotic convergence.

(iv) Simulations. We test our PG algorithm on several applications to illustrate the flexibility of CPT-RL compared to standard and risk-sensitive RL in leading to more nuanced behavior. We also compare the performance of our PG algorithm to the previously proposed zeroth order algorithm to compare against a zeroth-order baseline as the problem size increases in our simulation settings.

Closest related work. Closest to our work are policy gradient methods for distortion risk measures (DRMs) (Vijayan & L.A, 2024), which optimize distorted-expectation objectives via first-order gradients. CPT shares the probability distortion component but additionally incorporates a reference point and an asymmetric utility transformation (and potentially separate gain/loss distortions), enabling objectives beyond distortion-only criteria. Our policy gradient expression reduces to DRM policy gradients under the corresponding restrictions (e.g., identity utility and no reference point), while providing a first-order method for the more general CPT objective. We provide additional discussion after Theorem 3. A more comprehensive related work discussion can be found in section 6 and App. B.

2 From Classical RL To CPT-RL

2.1 MDPs, CPT and Notation

Markov Decision Process. A discrete-time Markov Decision Process (MDP) (Puterman, 2014) is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, H, \rho)$, where $\mathcal{S}, \mathcal{A}$ are respectively the state and action spaces, supposed to be finite for simplicity, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the state transition probability kernel, $r : \mathcal{S} \times \mathcal{A} \to [-r_{\max}, r_{\max}]$ is the reward function bounded by $r_{\max} > 0$, ρ is the initial state probability distribution and $H \geq 1$ is a finite horizon. A randomized stationary Markovian policy, which we will simply call a policy, is a mapping $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ which specifies for each $s \in \mathcal{S}$ a probability measure over the set of actions $\mathcal{A}$ by $\pi(\cdot|s) \in \Delta(\mathcal{A})$ where $\Delta(\mathcal{A})$ is the simplex over the finite action space $\mathcal{A}$. Each policy π induces a discrete-time Markov reward process $\{(s_t, r_t := r(s_t, a_t))\}_{t \in \mathbb{N}}$ where $s_t \in \mathcal{S}$ represents the state of the system at time t and r_t corresponds to the reward received when executing action $a_t \in \mathcal{A}$ in state $s_t \in \mathcal{S}$. We denote by $\mathbb{P}_{\rho,\pi}$ the probability distribution of the Markov chain $(s_t, a_t)_{t\in\mathbb{N}}$ generated by the MDP controlled by policy π with initial state distribution ρ. We use $\mathbb{E}_{\rho,\pi}$ (or often simply $\mathbb{E}$) to denote the expectation. In classical RL, the goal of the agent is to find a policy π maximizing the expected return $J(\pi) := \mathbb{E}_{\rho,\pi}[\sum_{t=0}^{H-1} r_t]$ where $s_0 \sim \rho$ and $H \geq 1$ is a finite horizon.

Policy classes. We now introduce different sets of policies which will be important for stating our results. Each policy class is defined according to the information history the policies have access to for selecting actions. Here, a history $h_t \in \mathcal{H} := \bigcup_{h\in[H]} \mathcal{H}_h$ (where $\mathcal{H}_h := \mathcal{S}^h \times \mathcal{A}^h \times [-r_{\max}, r_{\max}]^h$) is a finite sequence of successive states, actions and rewards: $(s_0, a_0, r_0, ..., s_{t-1}, a_{t-1}, r_{t-1})$.[1] More specifically, throughout this work, we will consider the following sets of policies: $\Pi_{NM} := \{\mathcal{H} \to \Delta(\mathcal{A})\}$ is the set of policies that are not necessarily Markovian, $\Pi_{\Sigma,NS} := \{\mathcal{S} \times \mathbb{R} \times \mathbb{N} \to \Delta(\mathcal{A})\}$ is the set of non-stationary policies that only depend on the current state, the sum of rewards accumulated so far and the timestep (i.e. $\pi(s, \sum_{k=0}^{t-1} r_k, t)$), $\Pi_{\Sigma,S} := \{\mathcal{S} \times \mathbb{R} \to \Delta(\mathcal{A})\}$ is the set of policies that only depend on the state and the sum of rewards (i.e. $\pi(s, \sum_{k=0}^{t-1} r_k)$), $\Pi_{M,NS} := \{\mathcal{S} \times \mathbb{N} \to \Delta(\mathcal{A})\}$ is the set of (non-stationary) Markovian policies (i.e. $\pi(s,t)$) and $\Pi_{M,S} := \{\mathcal{S} \to \Delta(\mathcal{A})\}$ is the set of stationary Markovian policies, i.e. Markovian policies which are time-independent. Deterministic policies assign a single action to each state. For each set of policies defined above, we define their corresponding subset of deterministic policies with a superscript D, e.g. Π^D_{NM}.

Remark 1. $\Pi_{M,S} \subseteq \Pi_{M,NS} \subseteq \Pi_{\Sigma,NS} \subseteq \Pi_{NM}$ *and* $\Pi_{M,S} \subseteq \Pi_{\Sigma,S} \subseteq \Pi_{\Sigma,NS} \subseteq \Pi_{NM}$ *, see Fig. 4 in App. A.*

Cumulative Prospect Theory Value. Instead of the expected return, CPT prescribes to consider the CPT value which relies on three distinct elements:

(a) A reference point x_0. Rewards larger than the reference are perceived as gains whereas lower values are viewed as losses.

(b) A utility function $\mathcal{U} : \mathbb{R} \to \mathbb{R}_+$. The agent's utility is a continuous and non-decreasing function which is not necessarily linear w.r.t. the total reward received by the agent. The function $u^+ : \mathbb{R} \to \mathbb{R}_+$ describing the gains is defined for every $x \in \mathbb{R}$ by $u^+(x) = \mathcal{U}(x)$ if $x \geq x_0$ and zero otherwise. Similarly, the function $u^- : \mathbb{R} \to \mathbb{R}_-$ which encodes the losses is defined by $u^-(x) = -\mathcal{U}(x)$ if $x \leq x_0$ and zero otherwise. Typically, the utility function is concave (resp. convex) for positive (resp. negative) rewards w.r.t. the reference point, i.e. u^+ is concave on $\mathbb{R}_+$ and $-u^-$ is convex on $\mathbb{R}_-$. For concreteness, we will use Kahneman & Tversky (1979)'s utility function as a running example: $\mathcal{U}(x) = (x - x_0)^\alpha$ if $x \geq x_0$ and $\mathcal{U}(x) = -\lambda(x - x_0)^\alpha$ if $x < x_0$, where $\lambda = 2.25, \alpha = 0.88$ are recommended hyperparameters. See Fig. 11 for an illustration with $x_0 = 0$.

(c) A probability distortion function $w : [0,1] \to [0,1]$. This is a continuous non-decreasing weight function that distorts the probability distributions of the gain and loss variables. Monotonicity here preserves ordering: it is natural to suppose that events with (truly) higher probability remain perceived as more probable than less truly probable ones after distortion. This distortion function w typically captures the human tendency to overestimate the probability of rare events and underestimate the probability of more certain ones. Similarly to the utility function, we denote by w_+ (resp. w_-) the function that warps the cumulative distribution function for gains (resp.for losses). Both functions are required to be defined on $[0,1]$, with values in $[0,1]$ and to be non-decreasing, continuous, with $w_+(0) = w_-(0) = 0$ and $w_+(1) = w_-(1) = 1$. Examples of such weights functions in the literature include $w : p \mapsto p^\eta(p^\eta + (1-p)^\eta)^{-\frac{1}{\eta}}$ (Kahneman & Tversky, 1979) and $w : p \mapsto \exp(-(-\ln p)^\eta)$ (Prelec, 1998) where $\eta \in (0,1)$ is a hyperparameter. We refer the reader to App. I.3 for examples and plots of utility and probability weight functions.

The CPT value of a real-valued random variable X is

$$\mathbb{C}(X) = \int_0^{+\infty} w_+(\mathbb{P}(u^+(X) > z))dz - \int_0^{+\infty} w_-(\mathbb{P}(u^-(X) > z))dz \,, \tag{1}$$

where appropriate integrability assumptions are assumed. While the CPT value $\mathbb{C}(X)$ accounts for the human agent's distortions in perception, it also recovers the expectation $\mathbb{E}(X)$ with weight functions w_+, w_- and utility functions u^+ (resp. $-u^-$) restricted to $\mathbb{R}_+$ (resp. $\mathbb{R}_-$) are set to be the identity functions. In addition, several risk measures such as variance, Conditional Value at Risk (CVaR) and distortion risk

[1] Rewards can be discarded from the history when they are deterministic functions of state-action pairs.

measures are also particular cases of CPT values with *discontinuous* probability weighting functions (see App. I and Table 1 therein).

2.2 CPT-RL Problem Formulation

In this work, we will focus on the CPT Policy Optimization (CPT-PO) problem where the objective is the CPT value of the random variable $X = \sum_{t=0}^{H-1} r_t$ recording the cumulative rewards induced by the MDP and the policy π for the finite horizon $H \geq 1$:

$$\max_{\pi \in \Pi_{NM}} \mathbb{C}\left[\sum_{t=0}^{H-1} r_t\right]. \tag{CPT-PO}$$

For an illustration of the CPT-RL problem formulation and its elements in a personal treatment for pain management application, see App. G.5. In the particular case of CPT-PO in which w_+, w_- are set to the identity, namely the Expected Utility Policy Optimization (EUT-PO) objective, only returns are distorted by the utility function:

$$\max_{\pi \in \Pi_{NM}} \mathbb{E}\left[\mathcal{U}\left(\sum_{t=0}^{H-1} r_t\right)\right]. \tag{EUT-PO}$$

Challenges. We outline the main difficulties in solving CPT-PO. First, the CPT value does not satisfy a Bellman equation: the nonlinear utility and weight functions violate the additivity and linearity of the standard expected return. While the special case EUT-PO has been studied (see e.g. Bäuerle & Rieder (2014); Fei et al. (2020); Wu & Xu (2023)), the CPT setting introduces fundamental differences. In particular, probability distortion breaks the dynamic programming structure, and optimal policies may be stochastic even under identity utility. Crucially, this aspect, central to CPT-RL and enabling richer behavioral modeling, has not been addressed in prior work on EUT-PO. Second, CPT-PO is highly nonconvex. The utility function is nonconvex in general (convex over gains, concave over losses), and the probabilities are also distorted by a nonconvex weight function. While the standard policy optimization problem is already nonconvex, CPT-PO compounds this difficulty with additional sources of nonconvexity.

Scope of this work. We assume that the utility and weighting functions are known, e.g., from domain experts, surveys, prior empirical studies on target user groups, or behavioral studies (Mkrtchian et al., 2023; George et al., 2019; Sip et al., 2018), see App. I.7 for more details on the choice of the utility function. Our goal is to align the agent's behavior with the given preferences (encoded in utility and weighting functions) by optimizing for the CPT value of returns. Typically, utility and weight functions are chosen as the KT model and hyperparameters of this model are estimated from data. We leave the question of discovering or inferring human preferences (e.g. RL from Human Feedback (RLHF)) to future work.[2] Note that the transition model $\mathcal{P}$ and the reward function r are still supposed to be unknown in our setting. In particular, we do not estimate them as in a model-based approach and our algorithm only uses sampled state-action-reward trajectories.

2.3 Peculiarities of Optimal Policies in CPT-RL

In this section, we highlight the properties of optimal policies to CPT-PO when they exist and contrast them with classical RL and EUT-PO (details in App. I.1):

The need for stochastic policies. In MDPs, optimal *deterministic* stationary policies always exist; in CPT-RL, optimal policies may need to be stochastic. In general, stochasticity of the policy is essential in solving CPT-PO. See App. C.1 for a proof.

Importance of probability weighting. Under EUT-PO, deterministic non-Markovian policies suffice (see e.g. Theorem 1 in Bäuerle & Rieder (2014)), but this is not the case for CPT-PO with probability weighting in general. Therefore, the need for stochasticity in the optimal policy is clearly due to the

[2] See section H for an extended discussion comparing CPT-RL to RLHF as preference learning paradigms, their pros and cons and opportunities for future work in combining them as they are not mutually exclusive. See also App. G for applications.

probability distortions in the CPT value.

The need for non-Markovian policies. Under EUT-PO, optimal *Markovian* policies exist for special cases of utility functions. Bäuerle & Rieder (2014) establish a characterization of such utility functions which turn out to be either affine or exponential (when $\mathcal{U}$ is continuous and increasing). This highlights the role of the (nonlinear) utility functions on the nature of optimal policies. However, this result cannot be extended to CPT-PO in general as we show next.

Proposition 2. *There exist instances of CPT-PO where $\mathcal{U}$ is of the form $x \mapsto A + B\exp(Cx)$ for positive constants A, B, C and CPT-PO does not admit an optimal policy in $\Pi_{M,NS}$.*

Even exponential utilities, which guarantee the existence of optimal Markovian policies in EUT-PO, may fail to admit Markovian optimal policies in CPT-PO.

3 Policy Gradient for CPT-value maximization

In this section, we design a policy gradient algorithm for solving CPT-PO. From this section on, we parametrize policies $\pi \in \Pi_{NM}$ by a vector $\theta \in \mathbb{R}^d$ and we denote by π_θ the parametrized policy. As a consequence, the CPT objective in CPT-PO becomes a function of the policy parameter θ and we use the notation $J(\theta)$ for the corresponding CPT objective value.

3.1 CPT Policy Gradient Theorem

Our key result enabling our algorithm design is a PG theorem for CPT value maximization.

Theorem 3. ***(Policy Gradient for CPT-RL).*** *Suppose that the utility functions u^-, u^+ are continuous and that the weight functions w_-, w_+ are Lipschitz and differentiable. Assume in addition that the policy parametrization $\theta \mapsto \pi_\theta(a|h)$ (for any $h, a \in \mathcal{H} \times \mathcal{A}$) are both differentiable. Then, for every $\theta \in \mathbb{R}^d$, the gradient of the CPT-PO objective J w.r.t. the policy parameter θ is given by:*

$$\nabla J(\theta) = \mathbb{E}\left[\varphi\left(R(\tau)\right) \sum_{t=0}^{H-1} \nabla_\theta \log \pi_\theta(a_t|h_t)\right],$$

where $R(\tau) := \sum_{t=0}^{H-1} r_t$ with $\tau := (s_t, a_t, r_t)_{0 \leq t \leq H-1}$ is a trajectory of length H generated from the MDP by following policy π_θ [3] *and*

$$\varphi(R(\tau)) := \int_{z=0}^{u^+(R(\tau))} w'_+(\mathbb{P}(u^+(R(\tau)) > z))dz - \int_{z=0}^{u^-(R(\tau))} w'_-(\mathbb{P}(u^-(R(\tau)) > z))dz\,, \tag{2}$$

where w'_+, w'_- denote the derivatives of the functions w_+, w_- respectively.

We provide a few comments regarding this result:

- Theorem 3 recovers the celebrated policy gradient theorem for standard RL (Sutton et al., 1999) by setting w_+ (resp. w_-) to the identity function (on $\mathbb{R}_+$ (resp. $\mathbb{R}^-$) in which case w'_+ is the constant function equal to 1 and hence $\varphi(R(\tau)) = R(\tau)$.
- In the special case where (i) the CPT utilities are identity functions ($u^+(x) = x, u^-(x) = x$) and (ii) the gain/loss weighting functions derive from a single distortion function g (i.e. $w_+(t) = 1 - g(1-t), w_-(t) = g(t)$), the CPT value reduces to a DRM objective (see App. I.4) and Theorem 3 recovers the DRM policy gradient of Vijayan & L.A (2024). CPT objectives are strictly more expressive than distortion risk metrics: DRMs impose the constraints $u^+(x) = x, u^-(x) = x, w_+(t) + w_-(1-t) = 1$. This duality constraint implies that overweighting low probabilities of large losses implies a specific fixed way of transforming probabilities of large gains. As a consequence, behaviors like 'extreme pessimism about rare losses and

[3] The integral $\varphi(R(\tau))$ is finite under our continuity assumptions since the return $R(\tau)$ is bounded.

Algorithm 1 CPT-Policy Gradient (CPT-PG)

```
**Input:** $\theta_0 \in \mathbb{R}^d$, utility functions $u^+, u^-$, weight functions $w_+, w_-$, step sizes $(\alpha_k)$ .
**for** $k = 0, \cdots, K$, **do**
  /Policy gradient estimation
    Sample $m$ trajectories $\tau_l := (s_t^l, a_t^l, r_t^l)_{0 \leq t \leq H-1}$, $1 \leq l \leq m$ with $s_0^l \sim \rho$ following $\pi_{\theta_k}$
    // Quantile estimation
    Sample $n$ trajectories $\tau_j := (s_t^j, a_t^j, r_t^j)_{0 \leq t \leq H-1}$, $1 \leq j \leq n$ with $s_0^j \sim \rho$ following $\pi_{\theta_k}$
    Compute and order $R(\tau_j)$, label them as $R(\tau_{[1]}) < R(\tau_{[2]}) < \cdots < R(\tau_{[n]})$
    $\hat{\xi}^+_{\frac{i}{n}} = u^+(R(\tau_{[i]}))$; $\hat{\xi}^-_{\frac{i}{n}} = u^-(R(\tau_{[i]}))$
    //Approximation of $\varphi(R(\tau))$
    $\hat{\phi}^{\pm}_{k,n} = \sum_{i=0}^{j_n-1} w'_{\pm}\left(\frac{i}{n}\right)\left(\hat{\xi}^{\pm}_{\frac{n-i}{n}} - \hat{\xi}^{\pm}_{\frac{n-i-1}{n}}\right) + w'_{\pm}\left(\frac{j_n}{n}\right)\left(R(\tau) - \hat{\xi}^{\pm}_{\frac{n-j_n-1}{n}}\right)$
    $\psi_{k,n} = \frac{1}{m}\sum_{l=1}^{m}\sum_{t=0}^{H-1} \nabla_\theta \log \pi_{\theta_k}(a_t^l | h_t^l)$
    $\hat{\nabla}_{n,m} J(\theta_k) = (\hat{\phi}^+_n - \hat{\phi}^-_n) \cdot \psi_{k,n}$
  /Policy gradient update
    $\theta_{k+1} = \theta_k + \alpha_k \hat{\nabla}_{n,m} J(\theta_k)$
**end for**
```

mild optimism about rare gains' cannot be captured by DRMs. CPT is more expressive as it allows arbitrary u_+, u_-, w_+, w_- with no duality constraint. In particular, it captures reference dependence, loss aversion, unequal tail distortions and asymmetric risk attitudes. CPT uses different utility and probability weight functions for gains and loss regions (possibly asymmetric), allowing for more nuanced behaviors, e.g. risk seeking for gains and risk averse for losses at the same time.

- We stated the theorem in the general setting where the policy is non-Markovian. In practice, it is also possible to use a parametrization of a smaller policy set such as $\Pi_{\Sigma,NS}$ or even $\Pi_{M,S}$ in which the policy is a function of $(t, s_t, \sum_{k=0}^{t-1} r_k)$ or only s_t respectively.

3.2 Stochastic PG Algorithm for CPT-RL

In the light of Theorem 3, we will perform a policy gradient ascent on the objective J to solve CPT-PO. Our general PG algorithm is presented in Algorithm 1. As usual, since we only have access to sampled trajectories from the MDP, we need a stochastic policy gradient to estimate the true unknown gradient given by the theorem. In particular, we need an approximation of $\varphi(R(\tau))$ for any sampled trajectory τ from the MDP following policy π_θ. In the case of EUT-PO in which w is the identity, the unknown quantity $\varphi(R(\tau))$ reduces to $\mathcal{U}(R(\tau))$ which can be easily computed as $\mathcal{U}$ is known and $R(\tau)$ is the cumulative reward.

In the more general setting, the approximation task requires to compute the term $\int w'_+(\mathbb{P}(u^+(R(\tau)) > z)dz$ (and likewise for the second integral term). We address this challenge using the following result which is a variation of Proposition 6 in L.A. et al. (2016) in which the integrand is the derivative w'_+ (instead of w_+) and the integral is taken over a bounded interval. While L.A. et al. (2016) use this result to approximate the CPT value, we use it for approximating our special integral terms involving the derivatives of the weight functions as they appear in the policy gradient. We obtain a different approximation formula which is tailored to our setting. The approximation is essentially a Riemann sum using simple staircase functions.

Proposition 4. *Let X be a real-valued random variable. Suppose that the functions w'_+, w'_- are Lipschitz and that $u^+(X), u^-(X)$ have bounded first moments. Let $\xi^+_{\frac{i}{n}}$ and $\xi^-_{\frac{i}{n}}$ denote the $\frac{i}{n}$th quantile of $u^+(X)$ and $u^-(X)$, respectively. Then, we have for any $v \geq 0$,*

$$\int_0^v w'_+(\mathbb{P}(u^+(X) > z))dz = \lim_{n\to\infty} u_n \,,$$

$$u_n := \sum_{i=0}^{j_n-1} w'_+\left(\frac{i}{n}\right)\left(\hat{\xi}^+_{\frac{n-i}{n}} - \hat{\xi}^+_{\frac{n-i-1}{n}}\right) + w'_+\left(\frac{j_n}{n}\right)(v - \hat{\xi}^+_{\frac{n-j_n-1}{n}})\,,$$

where $j_n \in [0, n-1]$ is s.t. $v \in [\xi^+_{\frac{n-j_n-1}{n}}, \xi^+_{\frac{n-j_n}{n}}]$. The same identity holds when replacing $u^+(X), \xi^+_\alpha, w_+$ by $u^-(X), \xi^-_\alpha, w_-$ where ξ^-_α is the α^{th} quantile of $u^-(X)$.

Using Proposition 4, we approximate the integral using a finite sum with a given number of samples n. As for the quantiles $\xi^+_{\frac{i}{n}}$ we compute them using standard order statistics. Overall, compared to a vanilla PG algorithm, our additional required quantile estimation procedure requires a mild sorting step which can be executed in $\mathcal{O}(n \ln n)$ running time (without even invoking parallel implementations) where n is the length of the rewards to be sorted (see Algorithm 1).

Comparison to the CPT-SPSA-G algorithm. Our algorithm is designed for maximizing the CPT value of a sum of rewards generated by an MDP while the CPT Simultaneous Perturbation Stochastic Approximation Gradient (CPT-SPSA-G) algorithm in L.A. et al. (2016) can be used to maximize the CPT value of any real-valued random variable. However, we highlight that (a) this cumulative reward return structure is natural and ubiquitous in RL and economics applications and foremost (b) thanks to this problem structure, our PG algorithm leverages first-order information whereas CPT-SPSA-G only uses zeroth-order information, i.e. CPT value estimations. This difference is crucial as zeroth-order optimization algorithms are known to suffer from the curse of dimensionality. Our algorithm can scale better to higher dimensional problems as it is notoriously known for PG algorithms in classic RL. We provide empirical evidence of this fact in section 5 to further support the benefits of our algorithm.

4 Asymptotic Convergence and Sample Complexity

In this section, we establish asymptotic convergence and sample complexity guarantees for both our PG estimator and Algorithm 1.

First, we show that our PG estimator is consistent.

Proposition 5 (Consistency). *Suppose that the utility functions u^+ and u^- are continuous, strictly increasing and uniformly bounded by a positive constant M_u. Assume in addition that the functions w'_+, w'_- are L_w-Lipschitz. Then for any $\theta \in \mathbb{R}^d, \hat{\nabla}_{n,m} J(\theta) \to \nabla J(\theta)$ almost surely as the batch size parameters $n, m \to \infty$.*

To prove this result, we combine consistency of the Monte Carlo estimators of the expected score function ($\nabla \ln \pi_\theta$ sum term) and the expected distorted reward ($\phi(R(\tau))$ term) in Thm. 3, using independent random trajectories. Consistency of the first follows from the law of large numbers whereas the second follows from using a generalized dominated convergence theorem combined with the Glivenko-Cantelli theorem (to control the asymptotics of the empirical distribution function stemming from quantile estimation), similarly to the proof of Prop. 3 in L.A. et al. (2016).

Beyond consistency, we now quantify the number of trajectories ($n+m$) required to obtain an ε-approximate policy gradient for any policy parameter. Under the same assumptions as for the consistency result, we make an additional score boundedness assumption which is standard in the analysis of PG methods (see e.g. Papini et al. (2018); Yuan et al. (2022); Fatkhullin et al. (2023)).

Proposition 6 (Sample complexity). *Let the assumptions of Prop. 5 hold. Suppose in addition that the score function is bounded, i.e. there exists $M_\psi > 0$ s.t. $\|\nabla \ln \pi_\theta(a|s)\|_2 \leq M_\psi$ for all $\theta \in \mathbb{R}^d, (s,a) \in \mathcal{S} \times \mathcal{A}$. Then there exists $c > 0$ s.t. for any $\varepsilon > 0, \delta \in (0,1)$ and for all $\theta \in \mathbb{R}^d$, we have $\|\hat{\nabla}_{n,m} J(\theta) - \nabla J(\theta)\|_2 \leq \varepsilon$ with probability at least $1-\delta$, if $n \geq \frac{(cHM_\psi M_u L_w)^2 \ln(1/\delta)}{\varepsilon^2}$ and $m \geq \frac{(cHM_\psi M_u L_w)^2 \ln(2d/\delta)}{\varepsilon^2}$.*

The sample complexity result coincides with the classical $n + m = \tilde{\mathcal{O}}(\varepsilon^{-2})$ statistical rate of Monte Carlo estimation. The sample complexity increases with the curvature of the probability weight function, the magnitude of the utility function and the horizon length. The proof relies on using concentration inequality results, namely (i) a Hoeffding's style inequality for bounded vector-valued random variables (Jin et al., 2019) for the score function and (ii) Dvoretzky-Kiefer-Wolfowitz inequality to quantify concentration of the empirical distribution of a random variable. This second part of the proof follows similar lines as the proof of Prop. 3 in L.A. et al. (2016) where it rather used for CPT value estimation rather than our distorted reward (see Thm. 3 and Prop. 4). Note though that the direct dependence on the dimension d of the policy parameter is only logarithmic in contrast to the sample complexity of zeroth-order PG estimation which scales with the dimension d in L.A. et al. (2016) due to the need to estimate each policy gradient coordinate.

We close this section by discussing the convergence of the policy parameter sequence produced by Algorithm 1. The CPT-RL objective is non-convex in the policy parameter due to non-convexity of both utility and probability weighting functions. While this lack of structure makes global optimality out of reach, we can still obtain a standard asymptotic convergence result towards the set of stationary points using the machinery of stochastic approximation (see e.g. Borkar (2008); Benaïm (2006)).

Proposition 7 (Asymptotic convergence)**.** *Under the same assumptions as in Prop. 5, suppose in addition that the positive step sizes* (α_k) *satisfy the Robbins-Monro conditions* $\sum_k \alpha_k = +\infty$ *and* $\sum_k \alpha_k^2 < +\infty$. *Then the sequence of iterates* (θ_k) *generated by Algorithm 1 converges to the set of stationary points of* J *almost surely, i.e.* $\theta_k \to \{\theta \in \mathbb{R}^d : \nabla J(\theta) = 0\}$ *almost surely as* $k \to \infty$ *if the sequence* (θ_k) *is bounded.*

Compared to the asymptotic convergence result established for the zeroth-order CPT-SPSA-G algorithm proposed in L.A. et al. (2016), note that we do not use increasing batch sizes to reduce the bias due to zeroth order PG estimation as our PG estimator is unbiased. Our boundedness assumption can be simply relaxed by adding a projection in the gradient ascent step in Algorithm 1 and modifying the limit set to account for the projection similarly to the statement of Thm. 1 in L.A. et al. (2016). We prefer our simpler statement.

5 Numerical Simulations

While our main contributions are theoretical and methodological, we provide simulations to illustrate our findings. Our main goals in this section are: (a) to show how our CPT-PG algorithm produces policies illustrating more nuances in capturing human behavior compared to standard expected utility theory and risk sensitive RL; (b) to show that, as expected, our algorithm scales better to larger state spaces than existing zeroth-order methods in a grid MDP with increasing size; (c) to test our algorithm on a finance application to show the flexibility of CPT-RL.

(a) Nuances of CPT-RL. We test our CPT-PG algorithm and compare it to vanilla PG (vPG) and an exponential risk-sensitive PG (ERS-PG) algorithm on two simple 2-bandit action problem instances. In *Environment 1 (Gain Bandit):* A safe action guarantees a certain gain of 2 and a risky action gives a reward 5 or 0 each with probability 1/2. Clearly the risky action has a higher expected return of 2.5. In *Environment 2 (Loss Bandit):* We flip the signs of the rewards. The safe action guarantees a certain negative reward of -2 whereas the risky action yields either -5 or 0 with probability 1/2 each. In this case, the risky action has a smaller expected return of -2.5.

Setting: We train a softmax policy using CPT-PG, vPG and ERS-PG. Vanilla PG optimizes for the standard expected return (identity for u and w), CPT-PG using a KT model (as defined in section 2-2.1) with S-shaped utility using parameters ($\alpha = 0.6$, $\lambda = 2.5$) and an S-shaped probability function w under-weighting the probability 1/2. ERS-PG is run using an exponential (concave) risk sensitive utility $u(x) = \eta^{-1}(1 - \exp(-\eta x))$ with $\eta = 0.5$. The reference point is set to be $x_0 = 0$ in this experiment. We use the *exact same* parameters for both environments.

Results and interpretation: As shown in Fig. 1 (right), in the Gain Bandit vPG selects the risky arm (higher expected return), whereas both CPT-PG and ERS-PG choose the safe arm—reflecting human risk aversion over gains. In the Loss Bandit, only CPT-PG "flips" to the risky arm, capturing human risk-seeking over losses. Crucially, CPT-PG does this with the *same* S-shaped utility and probability-weighting parameters in *both* settings, whereas the concave ERS-PG objective remains risk-averse throughout. This exact "safe in gains, risky in losses" pattern—known as the reflection effect—is the hallmark of Prospect Theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992) from which our simple example is inspired. Neither expected-utility nor a single-parameter risk-sensitive criterion can reproduce both behaviors simultaneously, but CPT-RL can, using a fixed, psychologically-grounded model. While risk-sensitive RL can, with appropriate parameter tuning, replicate some aspects of human decision-making, it often lacks the asymmetric treatment of gains and losses and the probabilistic distortion that characterize human behavior. Overall our main message is that CPT-RL offers more flexibility in modeling. Note also that CPT captures several risk measures as particular cases using discontinuous weights (see App. I.3- I.4).

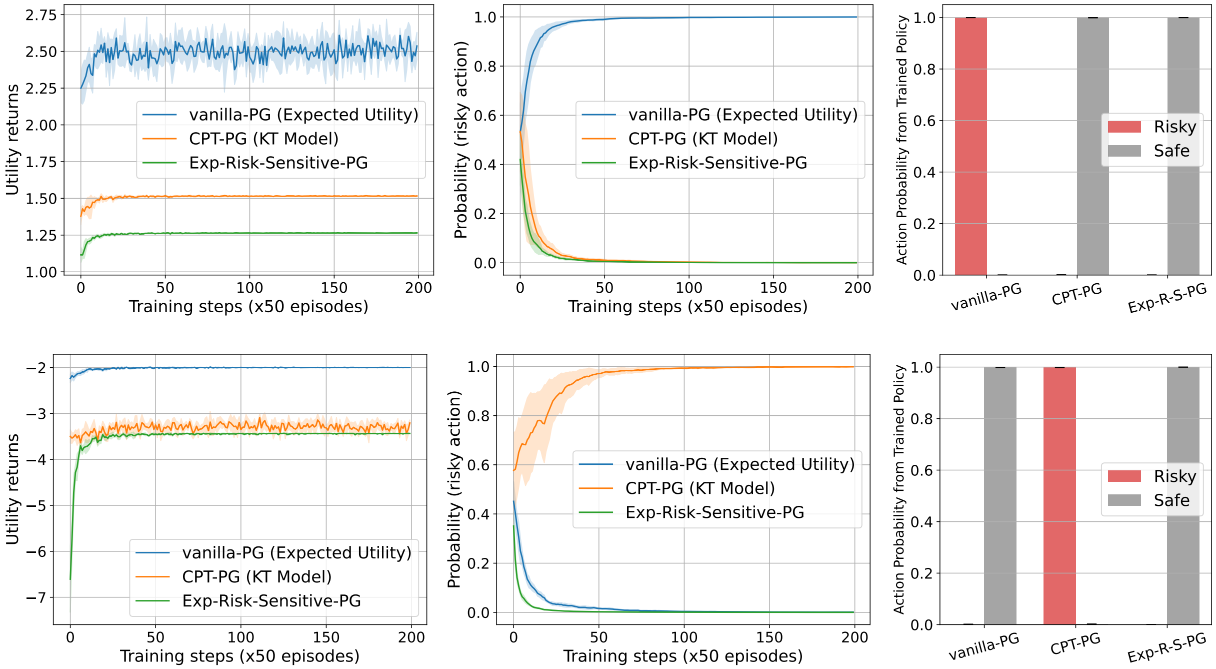


Figure 1: Comparison of our CPT-PG algorithm with vanilla PG (vPG) and exponential risk-sensitive (ERS-PG) on a simple 2-action bandit setting. (Upper fig.) Gain lottery setting: vPG trains a policy picking the risky action whereas CPT-PG and ERS-PG choose the safe one. (Lower fig.) Loss lottery: Only CPT picks the risky action. (Left) Recorded distorted returns, (center) evolution of probability of risky action along training steps, (right) Actions prescribed by trained policies. The shaded area is a range of ± one standard deviation with 5 independent runs with different seeds.

(b) Robustness to state space size. We compare our PG algorithm with the zeroth-order (CPT-SPSA-G) of L.A. et al. (2016) on MDPs with increasing size $n \times n$. Fig. 3 shows that our CPT-PG algorithm scales better to larger grid sizes than CPT-SPSA-G as expected due to its use of (first-order) gradient information. See App. J.5 for details.

(c) Application to finance. The goal is to train RL trading agents using our PG algorithm in the CPT-RL setting using a gym trading environment and data from the Bitcoin USD market. See App. J.8 for more details. We test several utility and probability weighting functions including a risk averse exponential of the form $x \mapsto \frac{1}{\beta}(1 - \exp(-\beta x))$ with different values of β as well as the KT (Kahneman and Tversky) function with different values of the reference point x_0 to illustrate its influence. In Figure 2, we make three observations. First, the reference point shifts the values of the achieved CPT returns: The smaller the reference point, the larger are the returns (Fig. 2, left). This is because only values larger than the reference point are perceived as positive returns. This illustrates how the subjective perception of the agent of the returns is taken into account by the model. Second, different values of β lead to different trajectories overall which can translate to different levels of risk aversion. In particular, the curves do not match the identity utility case in the first episodes and show more or less risk taken towards optimizing the CPT returns. Third, the exponent α in the utility distorts the function and shifts the returns significantly (Fig. 2, right). Lower values of α lead to higher returns in this setting. This parameter α provides a degree of freedom to model the behavior of the agent as per their perception of the returns. Different values of α modify the curvature of the utility function (w.r.t. the reference point $x_0 = 0$ here) which is concave for gains and convex for losses.

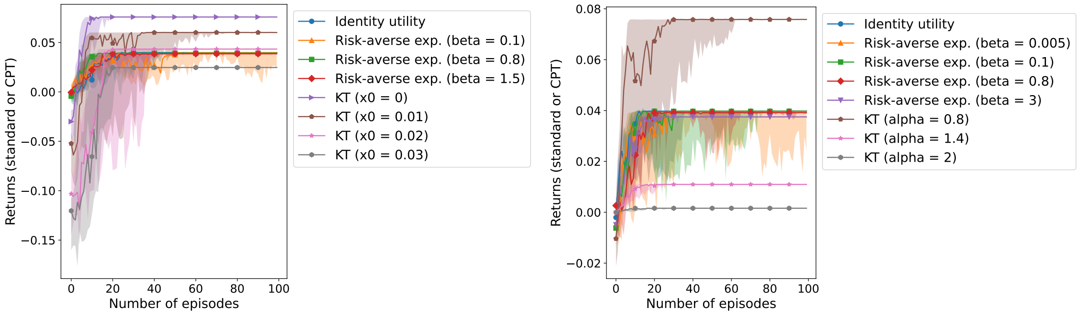


Figure 2: Performance of our PG algorithm on a financial trading application. KT refers to Kahneman and Tversky's utility function, x_0 is the reference point used in that utility, exp. refers to exponential and α is the parameter used in the definition of KT's utility. Shaded areas are interquantile (25-75%) margins and curves report the median values over 10 different runs.

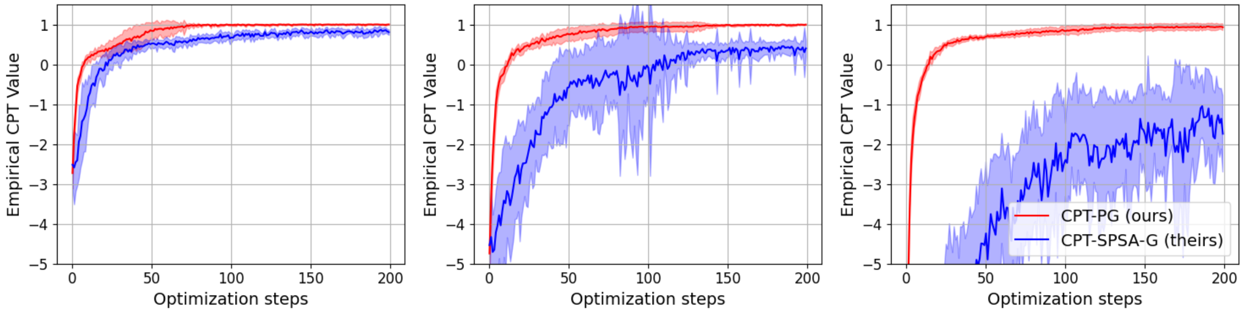


Figure 3: Compared performance of our algorithm and CPT-SPSA-G for $n = 3, 5, 9$. The shaded area is a range of ± one standard deviation over 10 runs.

Remark 8. ***(Reference point).*** *The reference point is typically learned from the data in specific human-related applications by personalized tuning. In our experiment (see Fig. 2), we vary this reference point to show how the performance is immediately influenced by this parameter of the model. If the agent has different perceptions of gains and losses with a different reference/tolerance point, then policy optimization takes this into account.*

Remark 9. ***(History dependent policies and non-Markovianity).*** *In our experiments, we incorporated partial history by expanding the input window of the networks to represent multiple past states. Typically, the input size was chosen to capture sufficient context without encoding the full trajectory length. This limited form of history dependence was sufficient in our settings, although incorporating more expressive temporal models could further enhance performance. The question of scaling to higher dimensional problems deserves further investigation. It is worth noting that the memory requirement is heavily correlated with the temporal structure of the data, especially in time series data like in finance. In our simple simulations, a small expanded input window was enough to obtain a descent performance. Our main goal was to show sensitivity of the KT model to different hyperparameters rather than optimizing performance.*

Remark 10. ***(Markovian vs non-Markovian policies).*** *In App. J.4 (see Fig. 22), we provide a simple example where we show how a non-Markovian policy performs better than a Markovian one when running CPT-PG.*

Additional simulations. We provide more simulations demonstrating the applicability of our CPT-PG algorithm in different settings, including illustrations of our theoretical results (App. J.3-J.4) and applications to a traffic control example over a grid (App. J.6), an electricity management example (App. J.7) and a control application (App. J.9).

6 Related Work

Prospect Theory and its sibling, CPT (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992; Barberis, 2013), were first integrated with RL by L.A. et al. (2016). Since then, only a few studies have explored the CPT-RL framework (Borkar & Chandak, 2021; Ramasubramanian et al., 2021; Ethayarajh et al., 2024). Notably, Borkar & Chandak (2021) proposed a Q-learning algorithm for CPT-based policy optimization, while Ramasubramanian et al. (2021) developed value-based methods estimating CPT values using dynamic programming. Their approach optimizes a sum of CPT-transformed period costs, making it amenable to dynamic programming (see remark 1 therein). In contrast, our CPT formulation is different: we maximize the CPT value of the return of a policy CPT-PO. This objective lacks an additive structure, hence does not satisfy a Bellman equation, rendering dynamic programming approaches inapplicable. Additionally, prior value-based methods are limited to finite state-action spaces, whereas our PG algorithm is also suitable for continuous state action settings, as shown in our experiments. More recently, Ethayarajh et al. (2024) incorporated CPT (without probability distortion) for fine-tuning large language models with human feedback. Our work complements prior CPT-RL studies (L.A. et al., 2016; Jie et al., 2018) that rely on zeroth-order SPSA methods (Spall, 1992). Instead, we introduce a PG algorithm that leverages first-order information, exploiting the structure of CPT values applied to cumulative rewards (see Section 3 for further comparison). Unlike existing PG approaches in risk-sensitive RL, our method explicitly accounts for probability distortion and S-shaped utility transformations, key aspects of CPT. For the special case of DRMs, as previously discussed in more details, Vijayan & L.A (2024) proposed a policy gradient method for maximizing DRM objectives and provided non-asymptotic first-order stationary guarantees. Recently, Pachal et al. (2025) proposed a policy Newton algorithm for maximizing DRM objectives and established a a non-asymptotic bound that establishes the convergence of the algorithm to approximate second-order stationary policies. We focus on first-order PG algorithms without resorting to higher-order Hessian information and we consider the more general class of CPT objectives.

For a broader discussion on CPT-RL, convex RL, and risk-sensitive RL, see App. B. A diagram summarizing these connections is provided in App. I.2.

7 Conclusion

We developed a policy-gradient framework for CPT-based policy optimization in finite-horizon MDPs, including a policy gradient theorem, a Monte Carlo gradient estimator, and convergence guarantees for a first-order CPT-PG algorithm. Our simulations illustrate qualitative behaviors induced by CPT objectives and compare first-order updates to existing zeroth-order approaches. A natural direction for future work is to relax the assumption that the utility and probability-distortion functions are specified, e.g., by learning or calibrating them from data and studying the resulting statistical and optimization trade-offs. Another direction is to extend CPT-based objectives to multi-agent or social settings, where reference points and probability distortions may interact with strategic incentives.

Broader Impact Statement

This work develops optimization tools for sequential decision-making objectives inspired by behavioral economics, specifically Cumulative Prospect Theory (CPT), which models probability distortion and asymmetric valuation around a reference point. Such objectives are relevant in settings where decision-makers exhibit systematic deviations from expected-utility assumptions, including human-in-the-loop or preference-driven applications. Our primary contribution is theoretical and methodological: we provide a policy-gradient framework for CPT-RL and demonstrate it in simulation environments drawn from several application themes (e.g., finance, traffic, and energy) as illustrative case studies.

Potential risks include mis-specification of CPT components (utility and distortion functions), reinforcement of undesirable biases if these components are learned from data, and misuse of behavioral models to manipulate users. Any deployment in real-world, human-facing systems would require careful calibration and validation, transparency about modeling assumptions, and appropriate safeguards to ensure responsible use.

References

Qinbo Bai, Mridul Agarwal, and Vaneet Aggarwal. Joint optimization of concave scalarized multi-objective reinforcement learning with policy gradient based algorithm. *Journal of Artificial Intelligence Research*, 74, September 2022. 19

Alejandro Balbás, José Garrido, and Silvia Mayoral. Properties of distortion risk measures. *Methodology and Computing in Applied Probability*, 11(3):385–399, 2009. 33

Anas Barakat, Ilyas Fatkhullin, and Niao He. Reinforcement learning with general utilities: Simpler variance reduction and large state-action space. In *International Conference on Machine Learning*, 2023. 19

Anas Barakat, Souradip Chakraborty, Peihong Yu, Pratap Tokekar, and Amrit Singh Bedi. On the global optimality of policy gradient methods in general utility reinforcement learning. In *Annual Conference on Neural Information Processing Systems*, 2025. 19

Nicholas C Barberis. Thirty years of prospect theory in economics: A review and assessment. *Journal of economic perspectives*, 27(1):173–196, 2013. 11, 19, 35

Nicole Bäuerle and Anna Jaśkiewicz. Markov decision processes with risk-sensitive criteria: an overview. *Mathematical Methods of Operations Research*, 99(1):141–178, 2024. 1

Nicole Bäuerle and Ulrich Rieder. More risk-sensitive markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014. 4, 5, 32

Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional reinforcement learning.* MIT Press, 2023. 18

Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pp. 1–68. Springer, 2006. 8, 26

S. Bhatnagar, H. Prasad, and L. Prashanth. *Gradient Schemes with Simultaneous Perturbation Stochastic Approximation*, pp. 41–76. Springer London, London, 2013. 18

Anup Biswas and Vivek S Borkar. Ergodic risk-sensitive control—a survey. *Annual Reviews in Control*, 55: 118–141, 2023. 1

Vivek S Borkar. Q-learning for risk-sensitive control. *Mathematics of operations research*, 27(2):294–311, 2002. 18

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 9. Springer, 2008. 8, 26

Vivek S Borkar and Siddharth Chandak. Prospect-theoretic q-learning. *Systems & Control Letters*, 156: 105009, 2021. 2, 11, 19

Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for cvar optimization in mdps. *Advances in neural information processing systems*, 27, 2014. 18

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167):1–51, 2018. 18

Dominic Danis, Parker Parmacek, David Dunajsky, and Bhaskar Ramasubramanian. Multi-agent reinforcement learning with prospect theory. *2023 Proceedings of the Conference on Control and its Applications (CT)*, pp. 9–16, 2023. 2, 19

Riccardo De Santi, Manish Prajapat, and Andreas Krause. Global reinforcement learning : Beyond linear and convex rewards via submodular semi-gradient methods. In *International Conference on Machine Learning*, 2024. 19

Sobhan Dorahaki, Masoud Rashidinejad, Seyed Farshad Fatemi Ardestani, Amir Abdollahi, and Mohammad Reza Salehizadeh. A home energy management model considering energy storage and smart flexible appliances: A modified time-driven prospect theory approach. *Journal of Energy Storage*, 48:104049, 2022. 1, 29, 36

Shima Ebrahimigharehbaghi, Queena K Qian, Gerdien de Vries, and Henk J Visscher. Application of cumulative prospect theory in understanding energy retrofit decision: A study of homeowners in the netherlands. *Energy and Buildings*, 261:111958, 2022. 1, 29, 36

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *International Conference on Machine Learning*, 2024. 11, 19, 32

Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies. In *International Conference on Machine Learning*, 2023. 7

Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 2020. 4

Marcus Jun Rong Foo, Nixie S Lesmana, and Chi Seng Pun. Drl trading with cpt actor and truncated quantile critics. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 574–582, 2023. 2

Dongli Gao, Wei Xie, Ruifeng Cao, Jingwen Weng, and Eric Wai Ming Lee. The performance of cumulative prospect theory's functional forms in decision-making behavior during building evacuation. *International Journal of Disaster Risk Reduction*, pp. 104132, 2023. 29, 36

Song Gao, Emma Frejinger, and Moshe Ben-Akiva. Adaptive route choices in risky traffic networks: A prospect theory approach. *Transportation research part C: emerging technologies*, 18(5):727–740, 2010. 29

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015. 18

Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Oliver Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: The mean-field game viewpoint. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, pp. 489–497, 2022. 19

Sophie A George, Jony Sheynin, Richard Gonzalez, Israel Liberzon, and James L Abelson. Diminished value discrimination in obsessive-compulsive disorder: A prospect theory model of decision-making under risk. *Frontiers in Psychiatry*, 10:469, 2019. 1, 4

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, 2019. 19

Cheng Jie, LA Prashanth, Michael Fu, Steve Marcus, and Csaba Szepesvári. Stochastic optimization in a cumulative prospect theory framework. *IEEE Transactions on Automatic Control*, 63(9):2867–2882, 2018. 11, 19

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019. 7, 25, 26

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. 1, 3, 8, 11, 19, 35

Prashanth L.A., Cheng Jie, Michael Fu, Steve Marcus, and Csaba Szepesvari. Cumulative prospect theory meets reinforcement learning: Prediction and control. In *International Conference on Machine Learning*, 2016. 2, 6, 7, 8, 9, 11, 19, 25, 26, 35

Rogelio Ladrón de Guevara Cortés, Leticia Eva Tolosa, and María Paula Rojo. Prospect theory in the financial decision-making process: An empirical study of two argentine universities. *Journal of Economics, Finance and Administrative Science*, 28(55):116–133, 2023. 1, 29

Eric Luxenberg, Philipp Schiele, and Stephen Boyd. Portfolio optimization with cumulative prospect theory utility via convex optimization. *Computational Economics*, pp. 1–21, 2024. 1, 30

Anahit Mkrtchian, Vincent Valton, and Jonathan P Roiser. Reliability of decision-making and reinforcement learning computational parameters. *Computational Psychiatry*, 7(1):30, 2023. 1, 4

Mehrdad Moharrami, Yashaswini Murthy, Arghyadip Roy, and Rayadurgam Srikant. A policy gradient algorithm for the risk-sensitive exponential cost mdp. *Mathematics of Operations Research*, 2024. 18

Mirco Mutti, Riccardo De Santi, and Marcello Restelli. The importance of non-markovianity in maximum state entropy exploration. In *International Conference on Machine Learning*, 2022a. 19

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Challenging common assumptions in convex reinforcement learning. *Advances in Neural Information Processing Systems*, 2022b. 19

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Convex reinforcement learning in finite trials. *Journal of Machine Learning Research*, 24(250):1–42, 2023. 19, 35, 36

Erfaun Noorani, Christos Mavridis, and John Baras. Risk-sensitive reinforcement learning with exponential criteria. *arXiv preprint arXiv:2212.09010*, 2022. 18, 33

Soumen Pachal, Mizhaan Prajit Maniyar, and Prashanth L. A. Policy newton methods for distortion riskmetrics, 2025. URL https://arxiv.org/abs/2508.07249. 11

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *International conference on machine learning*, 2018. 7

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703. 37

LA Prashanth, Michael C Fu, et al. Risk-sensitive reinforcement learning via policy gradient search. *Foundations and Trends® in Machine Learning*, 15(5):537–693, 2022. 1, 18

Drazen Prelec. The probability weighting function. *Econometrica*, 66(3):497–527, 1998. 3

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014. 2, 32

Bhaskar Ramasubramanian, Luyao Niu, Andrew Clark, and Radha Poovendran. Reinforcement learning beyond expectation. In *IEEE Conference on Decision and Control*, 2021. 2, 11, 19, 35

Marc Oliver Rieger, Mei Wang, and Thorsten Hens. Estimating cumulative prospect theory parameters from an international survey. *Theory and Decision*, 82:567–596, 2017. 29, 31, 36

Ulrich Schmidt and Horst Zank. What is loss aversion? *Journal of risk and uncertainty*, 30:157–167, 2005. 30

Ekaterina N Sereda, Efim M Bronshtein, Svetozar T Rachev, Frank J Fabozzi, Wei Sun, and Stoyan V Stoyanov. Distortion risk measures in portfolio optimization. *Handbook of portfolio construction*, pp. 649–673, 2010. 33

Kamila E Sip, Richard Gonzalez, Stephan F Taylor, and Emily R Stern. Increased loss aversion in unmedicated patients with obsessive–compulsive disorder. *Frontiers in Psychiatry*, 8:309, 2018. 1, 4

James C Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 37(3):332–341, 1992. 11, 19

Jiayi Sun, Xiang Zhou, Juan Zhang, Kemei Xiang, Xiaoxiong Zhang, and Ling Li. A cumulative prospect theory-based method for group medical emergency decision-making with interval uncertainty. *BMC Medical Informatics and Decision Making*, 22(1):124, 2022. 1

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12, 1999. 5

Aviv Tamar, Dotan Di Castro, and Shie Mannor. Policy gradients with variance related risk criteria. In *International conference on machine learning*, 2012. 18

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012. 47

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5:297–323, 1992. 1, 8, 11, 19, 35

Nithia Vijayan and Prashanth L.A. A policy gradient approach for optimization of smooth risk measures. In *Conference on Uncertainty in Artificial Intelligence*, 2023. 18

Nithia Vijayan and Prashanth L.A. Policy gradient methods for distortion risk measures, 2024. URL https://arxiv.org/abs/2107.04422. 2, 5, 11

Julia L Wirch and Mary R Hardy. Distortion risk measures: Coherence and stochastic dominance. In *International congress on insurance: Mathematics and economics*, pp. 15–17, 2001. 34

Zhengqi Wu and Renyuan Xu. Risk-sensitive markov decision process and learning under general utility functions. *arXiv preprint arXiv:2311.13589*, 2023. 4

Qianqian Yan, Tao Feng, and Harry Timmermans. Investigating private parking space owners' propensity to engage in shared parking schemes under conditions of uncertainty using a hybrid random-parameter logit-cumulative prospect theoretic model. *Transportation Research Part C: Emerging Technologies*, 120: 102776, 2020. 29, 36

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021. 29

Rui Yuan, Robert M. Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, 2022. 7

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. *Advances in Neural Information Processing Systems*, 2021. 19

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020. 19

Junyu Zhang, Chengzhuo Ni, Csaba Szepesvari, and Mengdi Wang. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 2021. 19

Meng Zhao, Yajun Wang, Xinyu Meng, and Huchang Liao. A three-way decision method based on cumulative prospect theory for the hierarchical diagnosis and treatment system of chronic diseases. *Applied Soft Computing*, 149:110960, 2023. 1

Contents

A Notation for Policy Classes

Throughout this work, we will consider the following sets of policies:

- $\Pi_{NM} := \{\mathcal{H} \to \Delta(\mathcal{A})\}$ is the set of non-Markovian policies,[4]
- $\Pi_{\Sigma,NS} := \{\mathcal{S} \times \mathbb{R} \times \mathbb{N} \to \Delta(\mathcal{A})\}$ is the set of policies that only depend on the current state, the timestep and the sum of discounted rewards accumulated so far: The RL agent in state s at timestep t following policy $\pi \in \Pi_{\Sigma,NS}$ samples its next action from the distribution $\pi(s, \sum_{k=0}^{t-1} \gamma^k r_k, t)$,

[4] By 'non-Markovian', we mean '*non necessarily* Markovian' policies including Markovian ones. Elements of $\Pi_{NM} - \Pi_{M,NS}$ can be designated as 'stricly non-Markovian' policies. Likewise, by 'non stationary', we mean 'non necessarily stationary', and by 'stochastic' we mean 'non necessarily deterministic'.

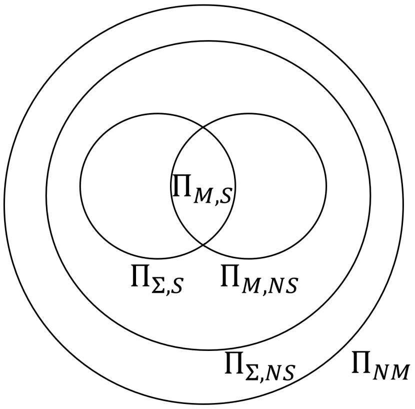


Figure 4: Policy classes (see Rem. 1).

- $\Pi_{\Sigma,S} := \{\mathcal{S} \times \mathbb{R} \to \Delta(\mathcal{A})\}$ is the set of policies that only depend on the state and the sum of discounted rewards: The RL agent in state s at timestep t following policy $\pi \in \Pi_{\Sigma,S}$ samples its next action from the distribution $\pi(s, \sum_{k=0}^{t-1} \gamma^k r_k)$,

- $\Pi_{M,NS} := \{\mathcal{S} \times \mathbb{N} \to \Delta(\mathcal{A})\}$ is the set of Markovian policies: An agent in state s at timestep t following policy $\pi \in \Pi_{M,NS}$ samples its next action from the distribution $\pi(s, t)$.

- $\Pi_{M,S} := \{\mathcal{S} \to \Delta(\mathcal{A})\}$ is the set of stationary Markovian policies, i.e. Markovian policies which are time-independent.

B Extended Related Work Discussion

B.1 Risk-sensitive RL

There is a rich literature around risk sensitive control and RL that we do not hope to give justice to here. We refer the reader to recent comprehensive surveys on the topic (Garcıa & Fernández, 2015; Prashanth et al., 2022) and the references therein. Let us briefly mention that there exist several approaches to risk sensitive RL. These include formulations such as constrained stochastic optimization to control the tolerance to perturbations and stochastic minmax optimization to model robustness with respect to worst case perturbations for instance. Another approach which is more relevant to our paper discussion consists in regularizing or modifying objective functions. Such modifications are based on considering different statistics of the return deviating from the standard expectation such as the variance or the conditional value at risk (e.g. Tamar et al. (2012); Chow & Ghavamzadeh (2014); Chow et al. (2018)) or even considering the entire distribution of the returns like distributional RL (Bellemare et al., 2023). Another popular objective modification consists in maximizing an exponential criterion (e.g. Borkar (2002); Noorani et al. (2022)) to obtain robust policies w.r.t noise and perturbations of system parameters or variations in the environment. Noorani et al. (2022) designed a model-free REINFORCE algorithm and an actor-critic variant of the algorithm leveraging an (approximate) multiplicative Bellman equation induced by the exponential objective criterion. Moharrami et al. (2024) recently proposed and analyzed similar PG algorithms for the same exponential objective. Vijayan & L.A. (2023) introduced a PG algorithm for solving risk-sensitive RL for a class of smooth risk measures including some distortion risk measures and a mean-variance risk measure. Their approach is based on simultaneous perturbation stochastic approximation (SPSA) (Bhatnagar et al., 2013) using zeroth-order information to estimate gradients. Our CPT-PO problem covers several of the aforementioned objectives including smooth distortion risk measures and exponential utility as particular cases (see App. I for more details).

B.2 Convex RL/RL with General Utilities

In the last few years, convex RL (a.k.a. RL with general utilities) (Hazan et al., 2019; Zhang et al., 2020; Zahavy et al., 2021; Geist et al., 2022) has emerged as a framework to unify several problems of interest such as pure exploration, imitation learning or experiment design. More precisely, this line of research is concerned with maximizing a given functional of the state(-action) occupancy measure w.r.t. a policy. To solve this problem, several policy gradient algorithms have been proposed in the literature (Zhang et al., 2021; Bai et al., 2022; Barakat et al., 2023; 2025). Mutti et al. (2022b;a; 2023) challenged the initial problem formulation and proposed a finite trial version of the problem which is closer to practical concerns as it consists in maximizing a functional of the empirical state(-action) distribution rather than its true asymptotic counterpart. The particular case of our CPT policy optimization problem without probability distortion (see EUT-PO below) coincides with a particular case of the single trial convex RL problem (Mutti et al., 2023) in which the function of the empirical visitation measure is a linear functional of the reward function (see App. I.6 for details). However, our general problem is not a particular case of convex RL which does not account for probability distortions. Furthermore, our utility function is in general nonconvex in our setting (see example in Fig 11) and our policy gradient algorithm is not model-based in the sense that we do not estimate the state transition model. More recently, De Santi et al. (2024) introduced a *global* RL problem formulation where rewards are globally defined over trajectories instead of locally over states and used submodular optimization tools to solve the resulting non-additive policy optimization problem. While global RL allows to account for trajectory-level global rewards, it does not take into consideration probability distortions. In addition, their investigation is restricted to the setting where the transition model is known whereas our PG algorithm does not require to know or estimate the transition model.

B.3 Cumulative Prospect Theoretic RL

Motivated by Prospect Theory and its sibling CPT (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992; Barberis, 2013), L.A. et al. (2016) first proposed to combine CPT with RL to obtain a better model for human decision making. Following this first research effort, only few isolated works (Borkar & Chandak, 2021; Ramasubramanian et al., 2021; Ethayarajh et al., 2024) considered a similar CPT-RL setting. In particular, Borkar & Chandak (2021) proposed and analyzed a Q-learning algorithm for CPT policy optimization. Ramasubramanian et al. (2021) further developed value-based algorithms for CPT-RL by estimating the CPT value of an action in a given state via dynamic programming. More precisely, they were concerned with maximizing a sum of CPT value period costs which is amenable to dynamic programming. In contrast to their accumulated CPT-based cost (see their remark 1), our CPT policy optimization problem formulation is different: we maximize the CPT value of the return of a policy (see CPT-PO). In particular, this objective does not enjoy an additive structure and hence does not satisfy a Bellman equation. Moreover, their work relying on value-based methods is restricted to finite discrete state action spaces. Our PG algorithm is also suitable for continuous state action settings as we demonstrate in our experiments. More recently, Ethayarajh et al. (2024) incorporated CPT (without probability distortion) into RL from human feedback for fine-tuning large language models. CPT has also been recently exploited for multi-agent RL (Danis et al., 2023). Our work is complementary to this line of research, especially to L.A. et al. (2016) and its extended version (Jie et al., 2018) which are the most closely related work to ours. While their algorithm design makes use of simultaneous perturbation stochastic approximation (SPSA) (Spall, 1992) using only zeroth order information, we rather propose a PG algorithm exploiting first-order information thanks to our special problem structure involving the CPT value of a cumulative sum of rewards. See section 3 for further details regarding this comparison.

We refer the reader to App. I.2 for a summarizing diagram illustrating the relationships between CPT-RL, convex RL and risk-sensitive RL.

C Proofs for Section 2.3

C.1 The need for stochastic policies

To prove the result (i.e. the need for stochastic policies), we consider a simple MDP with only two states (an initial state and a terminal one) and two actions (A and B). See Fig. 14a below. We choose the identity as utility. Action A yields reward 1 with probability 1 and action B yields either 0 or $\frac{3}{2}$ with probability $\frac{1}{2}$ each. We further consider the following probability distortion function $w_+ : [0,1] \to [0,1]$ defined for every $x \in [0,1]$ as follows:

$$w_+(x) = \begin{cases} 5x & \text{if } x \leq 0.1, \\ \frac{1}{2} + \frac{5}{9}(x - 0.1) & \text{otherwise}, \end{cases} \tag{3}$$

and we set $w_- = 0$. All the policies can be described with a single scalar $p \in [0,1]$, the probability of choosing B instead of A.

The CPT value of the reward X is:

$$\mathbb{C}(X) = w_+\left(1 - \frac{p}{2}\right) + \frac{1}{2} w_+\left(\frac{p}{2}\right) . \tag{4}$$

There are only two possible deterministic policies:

- For the policy corresponding to $p = 0$, $\mathbb{C}(X) = 1$.
- For the policy corresponding to $p = 1$, $\mathbb{C}(X) = \frac{3}{2} w_+(\frac{1}{2}) = \frac{13}{12} \approx 1.08$.

However, with the non-deterministic policy $p = 0.2$, we get:

$$\mathbb{C}(X) = w_+(0.9) + \frac{1}{2} w_+(0.1) = \frac{17}{18} + \frac{1}{4} = \frac{43}{36} \approx 1.19$$

which is larger than the CPT values of both deterministic policies. We conclude that there are no deterministic policies solving the CPT problem in this case.

Remark 11. *We provided a counterexample with random rewards, but there also exist counterexamples with deterministic rewards. One way to build such a counterexample is to start from the MDP we just studied and 'transfer' the randomness from the reward functions to the probability transition, by constructing a larger -but equivalent- MDP, with intermediate states like in Fig. 6.*

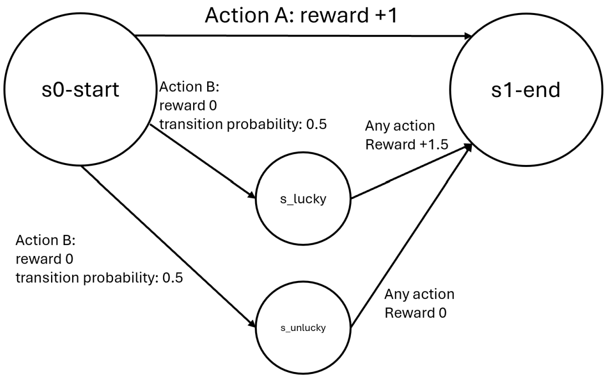
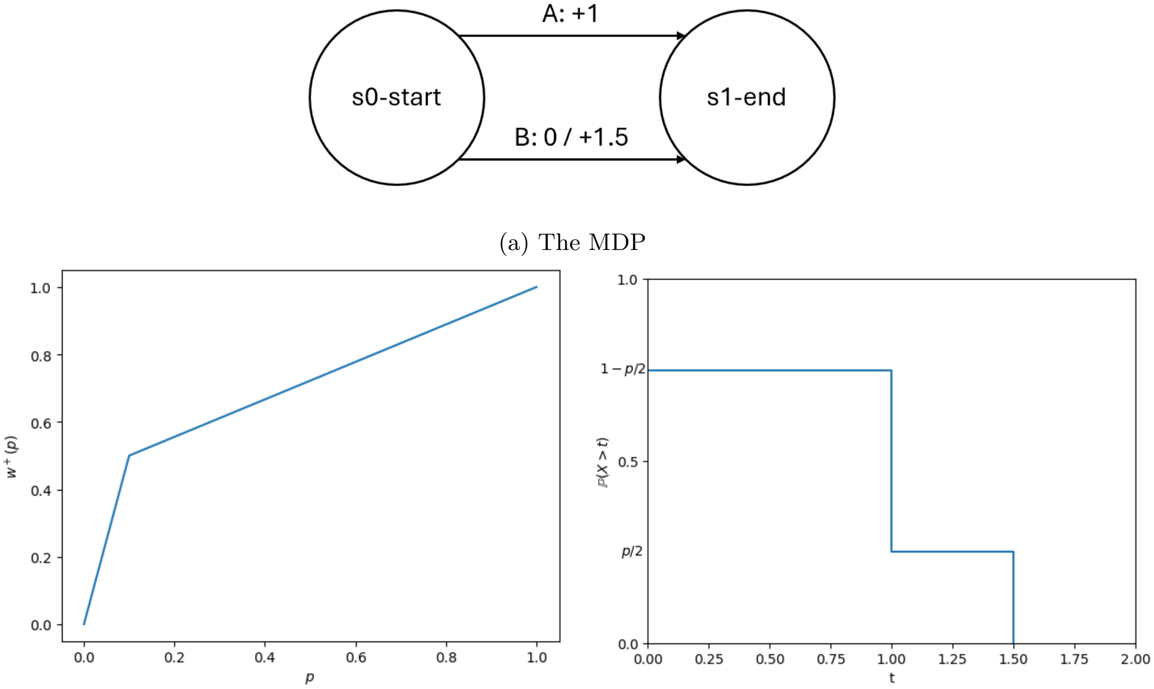


Figure 6: An equivalent example with deterministic rewards



(a) The MDP

(b) The w_+ function

(c) The $t \mapsto \mathbb{P}(X > t)$ function

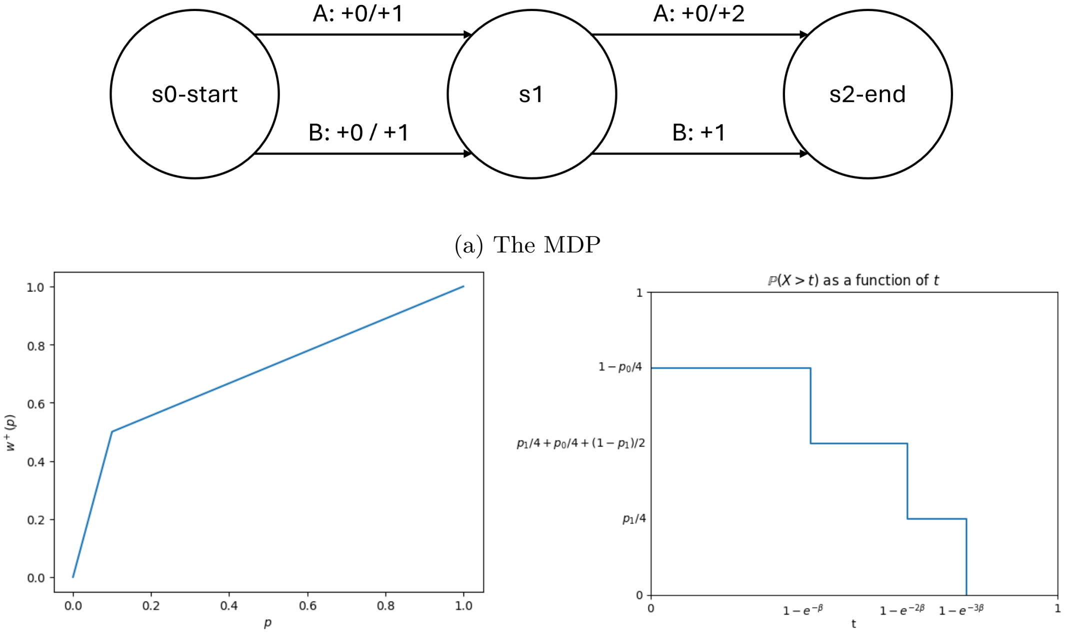
Figure 5: Problem instance for the proof.

C.2 Proof of Proposition 2

We proceed by providing a counterexample. We consider the utility function $\mathcal{U} : x \mapsto 1 - \exp(-\beta x)$ with $\beta = \frac{1}{2}$, and the weight function:

$$w_+(x) = \begin{cases} 5x & \text{if } x \leq 0.1, \\ \frac{1}{2} + \frac{5}{9}(x - 0.1) & \text{otherwise.} \end{cases}$$

We also set $w_- = 0$. However, we consider another MDP. Our MDP has three states: an initial state s_0, an intermediate state s_1, and a terminal state s_2. There are two actions: A and B. All trajectories start in s_0. Any action from s_0 leads to s_1 with probability 1 and yields reward +1 with probability $\frac{1}{2}$ and 0 otherwise. The action taken when in s_0 is completely irrelevant. Any action taken in s_1 leads to s_2 with probability 1 and the episode stops as soon as s_2 is reached. When taking action A in s_1, the reward is either 0 or +2, with probability $\frac{1}{2}$ each. When taking action B in s_1, the reward is +1 with probability 1. All policies in Π_{NM} can be described by $(p_{\text{start}}, p_0, p_1)$, where p_{start} is the probability of choosing action A when in s_0, p_0 is the probability of choosing action A in s_1 if the transition from s_0 to s_1 yielded reward 0 and p_1 is the probability of choosing action A in s_1 if the transition from s_0 to s_1 yielded reward 1. p_{start} is irrelevant to the performance of the policy so we can ignore it. The set of Markovian policies here is the set of policies such as $p_0 = p_1$. $\mathbb{C}(\pi)$ is a piecewise affine function of p_0 and p_1 and it can therefore be directly maximized. We omit the calculations here: one can check that the best achievable CPT value for Markovian policies is ≈ 0.616 for $p_0 = p_1 = 0.4$ but that a CPT value of ≈ 0.625 is achievable for $p_0 = 0$ and $p_1 = 0.4$, proving the lemma.



(a) The MDP

(b) The w_+ function

(c) The $t \mapsto \mathbb{P}(X > t)$ function

Figure 7: Figures for the proof of Proposition 2

D Proofs and Additional Details for Section 3

D.1 Proof of Theorem 3

The CPT value is a difference between two integrals (see definition in (1)). In what follows, we compute the derivative of the first integral assuming that the second one is zero in the CPT value. A similar treatment can be applied to the second integral. We skip these redundant details for conciseness.

Remark 12. *As we consider a finite horizon setting with finite state and action spaces, the integral on trajectories τ are in fact finite sums, allowing us to differentiate freely. We leave the interesting and technical question of the extension to continuous state-action spaces to future work.*

Using the shorthand notation $X = \sum_{t=0}^{H-1} r_t$, we first observe that:

$$\mathbb{C}(X) = \int_{z=0}^{+\infty} w(\mathbb{P}(\mathcal{U}(X) > z)dz = \int_{z=0}^{+\infty} w\left(\int_{\tau \text{ s.t.} \mathcal{U}(R(\tau))>z} \rho_\theta(\tau)d\tau\right) dz\,, \tag{5}$$

where ρ_θ is the trajectory probability distribution induced by the policy π_θ defined for any H-length trajectory $\tau = (s_0, a_0, \cdots, s_{H-1}, a_{H-1})$ as follows:

$$\rho_\theta(\tau) = p(s_0) \prod_{t=0}^{H-1} \pi_\theta(a_t|h_t)p(s_{t+1}|h_t, a_t)\,. \tag{6}$$

Remark 13. *Recall that we have ignored the second integral in the CPT value definition for conciseness.*

Starting from the above expression (5), it follows from using the chain rule that:

$$\begin{aligned}
\nabla_\theta \mathbb{C}(X) &= \int_{z=0}^{+\infty} w'(\mathbb{P}(\mathcal{U}(X > z)) \nabla_\theta \left(\int_{\tau \,\text{s.t.}\, \mathcal{U}(R(\tau)) > z} \rho_\theta(\tau) d\tau \right) dz \\
&= \int_{z=0}^{+\infty} w'(\mathbb{P}(\mathcal{U}(X > z)) \int_{\tau \,\text{s.t.}\, \mathcal{U}(R(\tau)) > z} \nabla_\theta \rho_\theta(\tau) d\tau dz \\
&= \int_\tau \int_{z=0}^{\mathcal{U}(R(\tau))} w'(\mathbb{P}(\mathcal{U}(X) > z)) \nabla_\theta \rho_\theta(\tau) dz d\tau \\
&= \int_\tau \phi(\mathcal{U}(R(\tau))) \nabla_\theta \rho_\theta(\tau) d\tau \,, \qquad (7)
\end{aligned}$$

where $\phi(t) := \int_{z=0}^t w'(\mathbb{P}(\mathcal{U}(X) > z)) dz$ for any real t.

We now use the standard log trick to rewrite our integral as an expectation:

$$\nabla_\theta \mathbb{C}(X) = \int_\tau \phi(\mathcal{U}(R(\tau))) \rho(\tau) \nabla_\theta \log \rho(\tau) d\tau = \mathbb{E}_{\tau \sim \rho}[\phi(\mathcal{U}(R(\tau))) \nabla_\theta \log \rho(\tau)] \,.$$

Furthermore, we can expand the gradient of the score function using (6) as follows:

$$\log \rho_\theta(\tau) = \log p(s_0) + \sum_{t=0}^{H-1} \log \pi_\theta(a_t | h_t) + \sum_{t=0}^{H-1} \log p(s_{t+1} | h_t, a_t) \,, \qquad (8)$$

$$\nabla_\theta \log \rho_\theta(\tau) = \sum_{t=0}^{H-1} \nabla_\theta \log \pi_\theta(a_t | h_t) \,, \qquad (9)$$

where the last step follows from observing that only the policy terms involve a dependence on the parameter θ. Combining (7) and (9) leads to our final policy gradient expression:

$$\nabla_\theta \mathbb{C}(X) = \mathbb{E}\left[\phi\left(\sum_{t=0}^{H-1} r_t \right) \sum_{t=0}^{H-1} \nabla_\theta \log \pi_\theta(a_t | h_t) \right] \,. \qquad (10)$$

Note that we have used the notation ϕ above instead of φ used in Theorem 3 to avoid the confusion with the full definition of φ which involves both integrals.

D.2 Alternative Practical Procedure for Computing Stochastic Policy Gradients

In this section, we discuss an alternative approximation procedure to the one proposed in section 3 for computing stochastic policy gradients. More precisely, we seek to approximate $\varphi(R(\tau))$ without the need for estimating quantiles and using order statistics for this. This alternatively procedure will be especially useful in practice when the probability distortion w is not necessarily differentiable or smooth. As discussed in the main part, one of the key challenges to compute stochastic policy gradients is to compute the integral terms appearing in the policy gradient expression. Our idea here is to approximate the probability distortion function w by a piecewise (linear or quadratic) function, leveraging the following useful lemma which shows that the integral is simple to compute when w is quadratic for instance.

Lemma 14. *Let X be a real-valued random variable and suppose that the weight function w is quadratic on an interval $[a, b]$ for some positive constants a, b, hence there exist $\alpha, \beta \in \mathbb{R}$ s.t. for all $x \in [a, b], w'(x) = \alpha x + \beta$. Let $Y_{a,b} := \min(\max(\mathcal{U}(X) - a, b - a), 0)$. Then, we have that $\int_a^b w'(\mathbb{P}(\mathcal{U}(X) > z) dz = \alpha \mathbb{E}[Y_{a,b}] + \beta(b - a)$.*

Proof. For any $a, b \in \mathbb{R}$ s.t. $a \leq b$, we have

$$\begin{aligned}
\int_a^b w'(\mathbb{P}(\mathcal{U}(X) > z)dz &= \int_0^{b-a} w'(\mathbb{P}(\mathcal{U}(X) - a > v))dv \\
&= \int_0^{b-a} (\alpha(\mathbb{P}(\mathcal{U}(X) - a > v)) + \beta)dv \\
&= \alpha \int_0^{b-a} \mathbb{P}(\mathcal{U}(X) - a > v)dv + \beta(b-a) \\
&= \alpha \int_0^{b-a} \mathbb{P}(Y_{a,b} > v)dv + \beta(b-a) \\
&= \alpha \int_0^{+\infty} \mathbb{P}(Y_{a,b} > v)dv + \beta(b-a) \\
&= \alpha \mathbb{E}[Y_{a,b}] + \beta(b-a) \,.
\end{aligned}$$

□

This result is convenient: Instead of estimating an entire probability distribution, we just have to estimate an expectation, which is much easier. However, we cannot reasonably approximate an arbitrary weight function by a quadratic function. Therefore, we consider the larger class of piecewise quadratic functions for which Lemma 14 extends naturally.

Proposition 15. *Let w be piecewise quadratic: there exists $q_1 < q_2 < .... < q_k$, with $q_1 = 0$ and $q_k = 1$, as well as reals $\alpha_1, ..., \alpha_k$, $\beta_1, ..., \beta_k$ and $\delta_1, ..., \delta_k$ such as $w(x) = \sum_{i=1}^{k-1} \mathbb{1}_{[q_i, q_{i+1}[}(t)(\frac{1}{2}\alpha_i t^2 + \beta_i t + \delta_i)$. For all $1 \leq i \leq k-1$, define the i-th quantile of $\mathcal{U}(X)$ as $\tilde{q}_i := \sup\{t \in \mathbb{R} \cup \{+\infty, -\infty\}, \mathbb{P}(\mathcal{U}(X) > t) \geq q_i\}$. Then, for any given $t \in [\tilde{q}_{j+1}, \tilde{q}_j[$:*

$$\int_0^t w'(\mathbb{P}(U(X) > z)dz = \sum_{i=j+1}^{k-1} (\alpha_i \mathbb{E}(Y_{\tilde{q}_{i+1}, \tilde{q}_i}) + \beta_i(\tilde{q}_i - \tilde{q}_{i+1})) + \alpha_j \mathbb{E}(Y_{\tilde{q}_j, t}) + \beta_j(t - \tilde{q}_{j+1}) \,.$$

Proof. We simply apply Lemma 14 to each segment:

$$\begin{aligned}
\int_0^t w'(\mathbb{P}(U(X) > z)dz &= \sum_{i=j+1}^{k-1} \int_{\tilde{q}_{i+1}}^{\tilde{q}_i} w'(\mathbb{P}(U(X) > z))dz + \int_{\tilde{q}_{j+1}}^{t} w'(\mathbb{P}(U(X) > z))dz \\
&= \sum_{i=j+1}^{k-1} (\alpha_i \mathbb{E}(Y_{\tilde{q}_{i+1}, \tilde{q}_i}) + \beta_i(\tilde{q}_i - \tilde{q}_{i+1})) + \alpha_j \mathbb{E}(Y_{\tilde{q}_j, t}) + \beta_j(t - \tilde{q}_{j+1}) \,.
\end{aligned}$$

□

The above lemma shows that we would have to estimate several quantiles and expectations to use this result. In particular, the expectation $\mathbb{E}(Y_{\tilde{q}_j, t})$ introduces some undesired computational complexity as the term differs for every t. However, if we rather consider a simpler piecewise affine approximation of w which can be computed once before any computation (independently from the rest) if the probability distortion function w is priorly known (which we implicitly suppose throughout this work), the expression is greatly simplified, yielding Lemma 16.

Lemma 16. *Suppose that the weight function $w : [0, 1] \mapsto [0, 1]$ is piecewise affine, i.e. there exists $q_1 < q_2 < .... < q_k$, with $q_1 = 0$ and $q_k = 1$, as well as reals $\beta_1, ..., \beta_k$ and $\delta_1, ..., \delta_k$ s.t. $w(x) = \sum_{i=1}^{k-1} \mathbb{1}_{[q_i, q_{i+1}[}(x)(\beta_i x + \delta_i)$ for any $x \in [0, 1]$. Let $\tilde{q}_i := \sup\{t \in \mathbb{R} \cup \{+\infty, -\infty\}, \mathbb{P}(\mathcal{U}(X) > t) \geq q_i\}$ for any $i = 1, \cdots, k$. Then for any $1 \leq j \leq k-1$ and any $t \in [\tilde{q}_{j+1}, \tilde{q}_j[$,*

$$\int_0^t w'(\mathbb{P}(\mathcal{U}(R(\tau)) > z)dz = \sum_{i=j+1}^{k-1} (\beta_i(\tilde{q}_i - \tilde{q}_{i+1})) + \beta_j(t - \tilde{q}_{j+1}) \,.$$

E Proofs for Section 4

The proofs in this section use similar techniques as the proofs developed in L.A. et al. (2016). Nevertheless, our policy gradient estimator is different and based on first-order information (using our policy gradient theorem) and this entails a few technical differences in the analysis.

E.1 Proof of Proposition 5: Consistency of the PG estimator

The proof of this result follows similar lines to the proof of Proposition 3 in L.A. et al. (2016) with the difference that we have a difference policy gradient estimator. The proof is based on applying a dominated convergence theorem (Theorem 17 below).

Observe first using the definition of $\hat{\phi}_n^+$ (see Algorithm 1) that we can write:

$$\begin{aligned} \hat{\phi}_n^+ &= \sum_{i=0}^{j_n-1} w'_+\left(\frac{i}{n}\right)\left(\hat{\xi}^+_{\frac{n-i}{n}} - \hat{\xi}^+_{\frac{n-i-1}{n}}\right) + w'_+\left(\frac{j_n}{n}\right)\left(R(\tau) - \hat{\xi}^+_{\frac{n-j_n-1}{n}}\right) \\ &= \sum_{i=0}^{j_n-1} \xi^+_{\frac{i}{n}}\left(w'_+\left(\frac{n-i}{n}\right) - w'_+\left(\frac{n-i-1}{n}\right)\right) + w'_+\left(\frac{j_n}{n}\right)\left(R(\tau) - \hat{\xi}^+_{\frac{n-j_n-1}{n}}\right) \\ &= \int_0^{R(\tau)} w'_+(1-\hat{F}_n^+(x))dx\,, \end{aligned} \tag{11}$$

where $\hat{F}_n^+(x)$ is the empirical distribution of $u^+(R(\tau))$ (where τ is a random trajectory).

Using this form, it remains to prove that $\int_0^{R(\tau)} w'_+(1-\hat{F}_n^+(x))dx \to \int_0^{R(\tau)} w'_+(\mathbb{P}(u^+(R(\tau) > x))dx$ using Theorem 17 with $f_n = w'_+(1-\hat{F}_n^+(x))$ and $g_n = L_w(1-\hat{F}_n^+(x))$ and using the fact that g_n converges almost surely to $\mathbb{P}(u^+(R(\tau) > x)$ uniformly in x by the Glivenko-Cantelli theorem. See proof of Proposition 3 in L.A. et al. (2016) p. 20 for details.

Theorem 17. ***(Generalized Dominated Convergence theorem)*** *Let (f_n) be a sequence of measurable functions on E that converge pointwise a.e. on a measurable space E to f. Suppose there exists a sequence (g_n) of integrable functions on E that converge pointwise almost everywhere on E to g such that $|f_n| \leq g_n$ for all $n \in \mathbb{N}$. If $\lim_{n\to\infty} \int_E g_n = \int_E g$, then $\lim_{n\to\infty} \int_E f_n = \int_E f$.*

E.2 Proof of Proposition 6: Sample complexity of the PG estimator

Observe first that our policy gradient estimator can be written as follow:

$$\hat{\nabla}_{n,m} J(\theta) = Y_n \cdot \left(\frac{1}{m}\sum_{l=1}^{m} X_l\right)\,, \quad \text{where} \quad Y_n := \hat{\phi}_n^+ - \hat{\phi}_n^-\,, \quad X_l := \sum_{t=0}^{H-1} \nabla_\theta \log \pi_{\theta_k}(a_t^l|h_t^l)\,, \tag{12}$$

see Algorithm 1. The idea of the proof is to show that both estimators in the policy gradient estimator concentrate around their mean. On the one hand, for the random variable Y_n we will use the following Dvoretzky-Kiefer-Wolfowitz inequality (see Lemma 18). On the other hand, for the variable X_l which is vector-valued, we will use a concentration inequality for bounded vector-valued random variables from the literature (Jin et al., 2019).

We start with the random variable Y_n and show how to control $\hat{\phi}_n^+$, $\hat{\phi}_n^-$ can be bounded in the same way. Recalling the rewritten form of $\hat{\phi}_n^+$ in (11), it appears that $\hat{\phi}_n^+$ is clearly comparable to $\varphi^+(v) := \int_{z=0}^{\max(v,0)} w'_+(\mathbb{P}(u^+(R(\tau)) > z))dz$ as defined in Theorem 3.

The rest of the proof for controlling this quantity follows similar lines to the proof of Proposition 3 in L.A. et al. (2016). Since $u^+(\tau)$ is supposed to be bounded by M_u and w'_+ (note here this is the derivative of w_+

unlike in Proposition 3 in L.A. et al. (2016)) is L_w-Lipschitz, we have

$$\begin{aligned}|\hat{\phi}_n^+ - \varphi^+(R(\tau))| &= \left|\int_0^{R(\tau)} w'_+(\mathbb{P}(u^+(R(\tau))) > z)dz - \int_0^{R(\tau)} w'_+(1-\hat{F}_n^+(z))dz\right|\\ &\leq \int_0^{M_u} \left|w'_+(\mathbb{P}(u^+(R(\tau))) > z) - w'_+(1-\hat{F}_n^+(z))\right| dz\\ &\leq \int_0^{M_u} L_w \left|\mathbb{P}(u^+(R(\tau))) < z) - \hat{F}_n^+(z)\right| dz\\ &\leq L_w M_u \sup_{z\in\mathbb{R}} \left|\mathbb{P}(u^+(R(\tau)) < z) - \hat{F}_n^+(z)\right| .\end{aligned}$$

Using the DKW inequality (Lemma 18 below) yields:

$$\mathbb{P}\left(|\hat{\phi}_n^+ - \varphi^+(R(\tau))| > \epsilon/2\right) \leq \mathbb{P}\left(L_w M_u \sup_{z\in\mathbb{R}} \left|\mathbb{P}(u^+(R(\tau)) < z) - \hat{F}_n^+(z)\right| > \epsilon/2\right) \leq 2e^{-n\frac{\epsilon^2}{2L_w^2 M_u^2}} . \tag{13}$$

A similar inequality holds for $|\hat{\phi}_n^- - \varphi^-(R(\tau))|$.

As for the random variable X_l, we observe that it is bounded by HM_g under our score function boundedness assumption. Applying Corollary 7 in Jin et al. (2019), there exists an absolute constant $c > 0$ such that with probability at least $1-\delta$ (for any $\delta \in (0,1)$), we have:

$$\left\|\frac{1}{m}\sum_{l=1}^{m} X_l - \mathbb{E}[X_1]\right\| \leq c\sqrt{\frac{(HM_\psi)^2 \log\left(\frac{2d}{\delta}\right)}{m}} . \tag{14}$$

Combining the above inequality with (13), we obtain with probability at least $1-\delta$,

$$\|\hat{\nabla}_{n,m} J(\theta) - \nabla J(\theta)\| \leq cHM_\psi \sqrt{\frac{4L_w^2 M_u^2 \log\left(\frac{1}{\delta}\right)}{n}} + cM_u L_w \sqrt{\frac{(HM_\psi)^2 \log\left(\frac{2d}{\delta}\right)}{m}} . \tag{15}$$

It remains to pick $n \geq \frac{(cHM_\psi M_u L_w)^2 \ln(1/\delta)}{\varepsilon^2}$ and $m \geq \frac{(cHM_\psi M_u L_w)^2 \ln(2d/\delta)}{\varepsilon^2}$ to conclude the proof.

Lemma 18. ***(Dvoretzky-Kiefer-Wolfowitz (DKW) inequality)*** *Let $\hat{F}_n(u) = \frac{1}{n}\sum_{i=1}^{n} 1_{((u(X_i))\leq u)}$ denote the empirical distribution of a random variable U, with $u(X_1), \ldots, u(X_n)$ being sampled from the random variable $u(X)$. Then, for any n and any $\epsilon > 0$, we have*

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}. \tag{16}$$

E.3 Proof of Proposition 7: Asymptotic convergence

This result follows from a standard application of the stochastic approximation framework (see e.g. Borkar (2008); Benaïm (2006)) similarly to L.A. et al. (2016). Under our regularity assumptions on the utility and weight functions, we can show asymptotic convergence of the iterates to the set of stationary points of our CPT objective. The idea is to show that the iterates track a gradient flow defined by the gradient of the CPT-PO objective with vanishing step sizes. The proof follows the same steps as the proof of Theorem 1 in L.A. et al. (2016) (see App. D, p. 25 therein), so we do not reproduce it here, up to the difference that our policy gradients are not estimated using zeroth-order information but rather using our PG Thm. 3 and Prop. 4.

To apply existing stochastic approximation results, it suffices to write:

$$\theta_{k+1} = \theta_k + \alpha_k(\nabla J(\theta_k) + \beta_t + \eta_t) , \tag{17}$$

where η_t is a martingale difference sequence, β_t is a possible non-zero bias (in our case we use unbiased estimates) and α_k satisfies the Robbins-Monro conditions.

F CPT-RL vs Risk-sensitive RL: An Illustration

In this section, we illustrate key features of CPT-RL compared to existing risk-sensitive RL approaches using a simple environment. This serves to provide intuition through an easy-to-grasp example. Specifically, we highlight how CPT-RL inflates low-probability (high-risk) events, distinguishing it from other methods. For comparison, we focus on an exponential risk-sensitive RL policy gradient algorithm, though similar results can also be shown for other risk-sensitive measures.

RiskyGridworld environment and reward structure. We consider a 5×5 custom gridworld environment The agent starts at $(0, 0)$ and must navigate to the goal state $(4, 4)$, choosing between safe and risky paths.

- States: each cell represents a state. There are 3 risky states, two of which ((1,0), (2,2)) correspond to penalty states and one is low probability high reward state (see reward description below), starting and goal states and the rest of the states are considered safe.
- Actions: The agent can move up, down, left, or right.
- Rewards: Safe steps incur a small penalty (-10). The risky state (2,4) offers a high reward (10^6) with low probability (0.01), otherwise a large penalty (-10^3). The remaining risky states give the same penalty. Reaching the goal provides a reward of $50,000$.
- Transitions: Movement is deterministic given the actions which are sampled according to the trained policy (which is not deterministic in our setting).

Algorithms for policy optimization. We compare our CPT-PG algorithm with an exponential risk-sensitive PG algorithm (using an exponential utility ($\mathcal{U}(x) = \frac{1}{\beta}\exp(-\beta x)), \beta = 0.1$, without probability weighting). For our CPT-PG algorithm, we use Kahneman and Tversky's utility function with parameters ($x_0 = 1, \lambda = 3, \alpha = 0.6$) (see section 2) and a probability weight function which is piecewise affine given by the following coefficients $w = [4, 0, 0.8, 0.2, 2.7, -1.7, 0.1, 0.9]$ where $[a_1, a_2, a_3, b_1, b_2, b_3, c_1, c_2]$ stands for a piecewise affine w function with $w(x) = a_i x + b_i$ for $c_{i-1} < x < c_i$. Note that this weight function inflates low probability events. This is one of the key features of CPT that we illustrate. We use a step size $\alpha = 0.01$ for both algorithms. For the policy network, we use a simple two-layer feed-forward neural network with a Leaky ReLU activation in the hidden layer and a softmax output layer for action selection.

Policy visualization. We provide heat maps visualizations of trained policy networks in the RiskyGridworld environment, representing the probability of selecting a risky action at each state. For each state we define risky actions as the actions leading to risky states. Therefore, the only states possibly leading to risky states are the states adjacent to risky ones. For each one of these states we encode the risky action as the one leading to the risky state (choose the riskier one if there are multiple ones). Then the heatmap assigns to each state the probability of selecting the risky action associated to it (probability as provided by the trained policy under consideration in that state).

Results. See figures 8 and 9 and their captions in the next page.

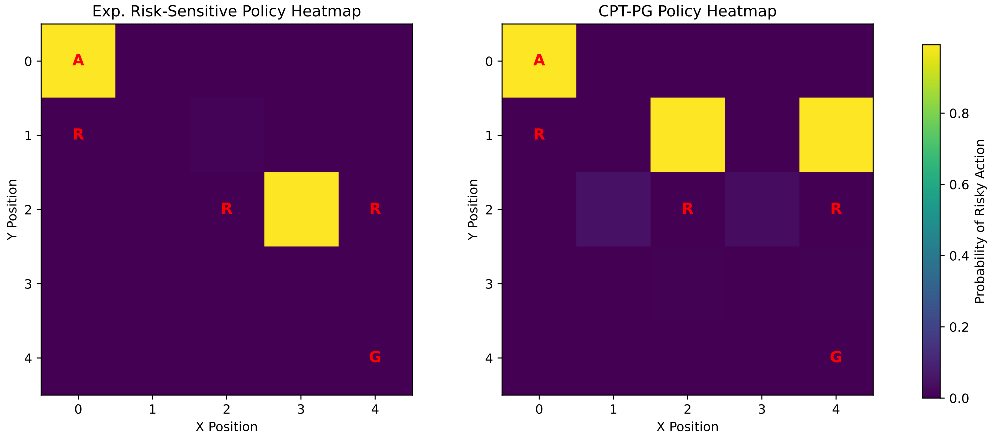


Figure 8: Heat maps representing the policies trained using our algorithm CPT-PG and an exponential risk-sensitive PG algorithm. Each cell in the 5 × 5 grid corresponds to a state, risky states are denoted by the letter 'R' in the cell, 'A' stands for the initial state and 'G' for the goal state, all the other states are considered safe (with a zero probability assigned). The color represents the probability of selecting a risky action at each state. The risky state (2,4) is risky in the sense that with low probability 0.01 it leads to high reward of 10^6 and a penalty of -10^3 otherwise. The main observation here is that CPT takes more risk in trying to end up in this risky state (low probability high reward) as you can see with the yellow cell above the risky state. Overall the risk profile is different for both methods. This is partly explained by the fact that CPT inflates low probability events thanks to the probability weight function.

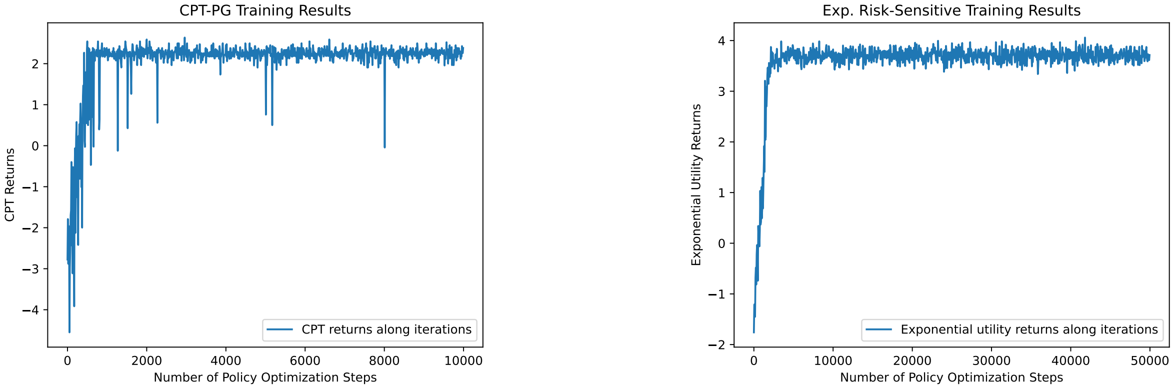


Figure 9: Policy training curves: (**Left**) CPT returns for CPT-PG. (**Right**) Exponential utility returns for Exponential Risk-Sensitive PG. Quantities are not comparable as one shows the CPT returns optimized in CPT policy optimization whereas the exponential risk-sensitive PG optimizes the exponential utility of the returns. Hence we represent them in separate graphs. This figure mainly serves the purpose of showing that the policies have been trained to maximize their respective objectives.

G Applications of CPT: From Prior Work in Stateless Settings to CPT-RL

In this section, we provide a discussion regarding the applications where CPT has already been successfully used (mainly in the static stateless setting) and potential applications in the dynamic (RL) setting with state transitions.

We highlight that CPT has been tested and effectively used in a large number of compelling behavioral studies that we cannot hope to give justice to here. Besides the initial findings of Tversky and Kahneman for which the latter won the Nobel Prize in economics in 2002, please see a few recent references below for a broad spectrum of real-world applications ranging from economics to transport, security and energy, mostly in the stateless (static) setting.

- Risk preferences across 53 countries worldwide in an international survey (Rieger et al., 2017). Estimates of CPT parameters from data illustrate economic and cultural differences whereas probability weighting also reflects gender differences as well as economic and cultural impacts. Note here the explainability feature of CPT.

G.1 Energy: Renovation and Home Energy Management

- A study of homeowners in the Netherlands to investigate energy retrofit decision using CPT (Ebrahimigharehbaghi et al., 2022). CPT is shown to predict the number of homeowners decisions to renovate their homes more accurately than Expected Utility Theory (EUT).
- Home energy management (Dorahaki et al., 2022). This work proposes a behavioral home energy management model to increase the user's satisfaction.

G.2 Security: Building Evacuation

- Application of CPT to building evacuation (Gao et al., 2023). CPT allows to take into account individual psychology and irrational behavior in modeling evacuations via pedestrian movement modeling. This is particularly important for designing and optimizing emergency and safety management strategies.

G.3 Urban Planning and Mobility

- **Parking:** Understanding private parking space owners' propensity to share their parking spaces by considering their psychological concerns as well as their socio-demographic and revenue characteristics for instance (Yan et al., 2020). This might be useful to help developing shared parking services.
- **Traffic routing.** Gao et al. (2010) model the travelers' strategic behavior for route choice in a stochastic network when adapting to traffic conditions which are revealed en route.

G.4 Finance

- Empirical study about financial decision making in two universities in Argentina (Ladrón de Guevara Cortés et al., 2023). In particular, it is shown that the financial decisions of the participants under uncertainty are more consistent with Prospect Theory than expected utility theory.

G.5 Example: Personalized Treatment for Pain Management

We illustrate our CPT-RL problem formulation with a concrete example in healthcare to give the reader more intuition about the different features of CPT-RL, its importance in applications when human perception and behavior matter and its differences compared to risk-sensitive RL. The goal is to help a physician manage a patient's chronic pain by suggesting a personalized treatment plan over time. The challenge here is to balance pain relief and the risk of opioid dependency or other side effects that might be due to the treatment, i.e. short-term relief and longer term risks. The idea is to train a CPT-RL agent to help the physician.

1. ***Why RL?*** (a) The physician needs to adjust treatment at each time step depending on the patient's reported pain level as well as the observed side effects. This is relevant to dynamic treatment regimes in general (such as for chronic diseases, see e.g. Yu et al. (2021) for a survey) in which considering delayed effects of treatments is also important and RL does account for such effects. (b) Decisions clearly impact the patient's immediate pain relief, dependency risks in the future and their overall health condition. A state is described by e.g. current pain level, dependency risk and side effect severity. Actions are treatments, e.g. no treatment, opioid or alternative treatment.

2. ***Why CPT-RL?*** Patients and clinicians make decisions influenced by psychological biases. We illustrate the importance of each one of the three features of CPT in section 2 (reference point, utility and probability distortion weight functions) via this example:

- *Reference points:* Patients assess and report pain levels according to their subjective (psychologically biased) baseline. Incorporating reference point dependence leads to a more realistic model of human decision-making taking into account *perceived* gains and losses. In our example, reducing pain from a level of 7 to 5 is not perceived the same way if the reference point of the patient is 3 or 5. In contrast, risk-sensitive RL treats every pain reduction as a uniform gain, regardless of the patient's starting reference pain level.

- *Utility transformation:* Patients might often show a loss averse behavior, i.e., they might perceive pain increase or withdrawal symptoms as worse than equivalent gains in pain relief. Note here that loss aversion should not be confused with risk aversion (Schmidt & Zank, 2005). In short, loss aversion can be defined as a *cognitive bias* in which the emotional impact of a loss is more intense than the satisfaction derived from an equivalent gain. For instance, in our example, a 2-point increase in pain might be seen as much worse than a 2-point reduction even if the change is the same in absolute value. This loss aversion concept is a cornerstone of Kahneman and Tversky's theory. In contrast, risk aversion rather refers to the *rational* behavior of undervaluing an uncertain outcome compared to its expected value. Risk sensitive approaches might be less adaptive to a patient's subjective preferences if they deviate from objective risk assessments.

- *Probability weighting:* Low probability events such as severe side effects (e.g., opioid overdose or dependency) might be overweighted or underweighted based on the patient's psychology.

3. *CPT-RL vs Risk-averse RL.* In terms of policies, risk-averse RL would favor non-opioid treatments unless extreme pain levels make opioids justifiable. In contrast, CPT-RL policies would prescribe opioids if pain significantly exceeds the patient's reference point. As dependency risk increases, CPT-RL policies would transition to non-opioid treatments as a consequence of overweighting the probability of rare catastrophic outcomes. Notably, CPT-RL policies can oscillate between risk-seeking (to address high pain) and risk-averse (to avoid severe side effects). In contrast, a risk-sensitive agent focuses on minimizing variability in health states and dependency risks and would likely avoid opioids in most cases unless pain levels become extreme. Such risk-sensitive policies favor stable strategies (e.g., consistent non-opioid use), prioritizing low variance in patient outcomes.

G.6 Further Applications of CPT-RL

Our CPT-RL problem formulation finds applications in a number of diverse areas. A nonexhaustive list includes:

- **Traffic control.** We refer the reader to our toy example in the main part. simulations for specific CPT-RL applications in simple settings for traffic control, electricity management and financial trading that we will not discuss again here.

- **Electricity management.** Please see simulations in the main part (section 5 and App. J.7) in a simple example setting to illustrate our methodology.

- **Finance:** portfolio optimization, risk management, behavioral asset pricing (e.g. influence of investor sentiment on price dynamics via e.g. over-weighting of low-probability events, including their preferences). For recent applications of CPT to finance, we refer the reader to a recent paper (Luxenberg et al., 2024) using CPT for portfolio optimization (in a stateless static setting). We also applied our methodology to financial trading (see App. J.8).

- **Health:** personalized treatment plans, (e.g. health insurance design for specific groups modeling risk and factoring perceived fairness).

On a more high-level note, we would like to mention that CPT-RL is of practical relevance for finance and healthcare for several reasons: in short, CPT allows for (a) **modeling human biases**, (b) **factoring risk**, and (c) **capturing individual preferences for personalization.** All these three points are essential in the above applications.

Many other meaningful human-centric applications are yet to be explored, including legal and ethical decision making, cybersecurity and human-robot interaction.

H CPT-RL and Trajectory-Based Reward RL as Preference Learning Paradigms

In this section, we compare the CPT-RL and trajectory-based reward RL (using a single reward for the entire trajectory, such as Reinforcement Learning from Human Feedback) seen as preference learning paradigms. In particular, we also discuss the pros and cons of each one of them.

Regarding the structure of the final reward and the metric learning you mention, this is a fair point and we agree that Our present work requires so far access to utility and weight functions whereas trajectory-based reward RL learns the metric to be optimized using human preference data. However, let us mention a few points:

(a) These can be readily available in specific applications (for risk modeling or even chosen at will by the users themselves);

(b) CPT relies on a predefined model, this can be beneficial in applications such as portfolio optimization or medical treatment where trade-offs have to be made and models might be readily available;

(c) Furthermore, we argue that having such a model allows it to be more explainable compared to a model entirely relying on human feedback and fine tuning, let alone the discussion about the cost of collecting human feedback. We also note that some of the most widely used algorithms in RLHF (e.g. DPO) do rely on the fact that the reward follows a Bradley-Terry model for instance (either for learning the reward or at least to derive the algorithm to bypass reward learning);

(d) Let us mention that one can also learn the utility and weight functions. We mentioned this promising possibility in our conclusion although we did not pursue this direction in this work. One can for instance represent the utility and weight functions by neural networks and train models to learn them using available data with relevant losses, jointly with the policy optimization task. One can also simply fit the predefined functions (say e.g. Tversky and Kahneman's function) to the data by estimating the parameters of these functions (see η with our notations and exponents of the utility function in Table 1 for the CPT row). This last approach is already commonly used in practice, see e.g. Rieger et al. (2017).

CPT vs RLHF: General comparison. CPT has been particularly useful when modeling specific biases in decision making under risk to account for biased probability perceptions. It allows to *explicitly* model cognitive biases. In contrast, RLHF has been successful in training LLMs which are aligned with human preferences where these are complex and potentially evolving and where biases cannot be explicitly and reasonably modeled. RLHF has been rather focused on learning *implicit* human preferences through interaction (e.g. using rankings and/or pairwise comparisons). Overall, CPT can be useful for tasks where risk modeling is essential and critical whereas RLHF can be useful for general preference alignment although RLHF can also be adapted to model risk if human preferences are observable and abundantly available at a reasonable cost. This might not be the case in healthcare applications for instance, where one can be satisfied with a tunable risk model. On the other hand, so far CPT does not have this ability to adapt to evolving preferences over time unlike RLHF which can do so via feedback.

CPT and RLHF: Pros and cons. To summarize the pros and cons of both approaches, we provide the following elements. As for the pros, CPT directly models psychological human biases in decision making via a structured framework which is particularly effective for risk preferences. RLHF can generalize to different scenarios with sufficient feedback and handle complex preferences via learning from diverse human interactions, it is particularly useful in settings where preferences are not explicitly defined such as for LLMs for aligning the systems with human preferences and values. As for the cons, CPT is a static framework since the utility and probability weight functions are fixed, it is hence less adaptive to changing preferences. It uses a predefined model of human behavior which is not directly using feedback. It also requires to estimate model parameters precisely, often for specific domains. As for RLHF on the other hand, the quality and the quantity of the human feedback is essential and this dependence on the feedback clearly impacts performance. This dependence can also cause undesirable bias amplification which is present in the human feedback. We also note that training such models is computationally expensive in large scale applications.

CPT and RLHF are not mutually exclusive. While CPT and trajectory-based RL (say e.g. RLHF) both offer frameworks for incorporating human preferences into decision making, we would like to highlight that CPT and RLHF are not mutually exclusive. We can for instance use CPT to design an initial reward structure reflecting human biases, then refine it with RLHF. We can also consider to further relax the requirement of sum of rewards (which already has several applications on its own) and think about incorporating CPT features to RLHF. Some recent efforts in the literature in this direction that we mentioned in our paper include the work of Ethayarajh et al. (2024) which combines prospect theory with RLHF (without probability weight distortion though, which limits its power). Note that the ideas of utility transformation and probability weighting are not crucially dependent on the sum of rewards structure and can also be applied to trajectory-based rewards or trajectory frequencies for instance. We believe this direction deserves further research, one interesting point would be how to incorporate risk awareness from human behavior to such RLHF models using ideas from CPT.

I More About CPT Values and CPT Policy Optimization

I.1 More about Optimal Policies in CPT-RL

The need for stochastic policies. We start our discussion by pointing out a stark difference between optimal policies in standard MDPs and CPT-PO. While there exists an optimal *deterministic* stationary policy for MDPs (see e.g. Thm. 6.2.10 in Puterman (2014)), this is not the case in general for CPT-PO. Indeed, there does not always exist an optimal policy for CPT-PO in Π^D_{NM} . In general, stochasticity of the policy is essential in solving our CPT-RL problem. To see this, consider an example built around a function w_+ that puts special emphasis on the 10% of the best outcomes. In this case, the optimal policy needs to be randomized to take advantage of this and obtain the highest returns with some probability without suffering from bad outcomes by deterministically committing to this riskier strategy (see App. C.1 for a proof).

Importance of probability weighting. When setting the probability weight distortion function w to the identity, i.e. when considering the particular case EUT-PO of CPT-PO, it appears that an optimal policy is not necessarily stochastic. Indeed, it has been shown that there exists an optimal policy for EUT-PO in $\Pi^D_{\Sigma,NS}$ (see e.g. Theorem 1 in Bäuerle & Rieder (2014)). Therefore, the need for stochasticity in the optimal policy is clearly due to the probability distortions in the CPT value. The aforementioned result allows to safely restrict the policy search to $\Pi_{\Sigma,NS}$ which is a much smaller policy space than the set of non-Markovian policies Π_{NM}. The fact that an optimal deterministic policy exists is a fundamental difference with the general CPT-PO setting. Whether there are specific weight functions (apart from the identity) for which there always exist a deterministic optimal policy remains an open question left for future work.

The need for non-Markovian policies. We now ask the next natural question: Can we further restrict our policy search to a smaller policy class compared to $\Pi_{\Sigma,NS}$? In particular, are there specific utility functions for which the resulting EUT-PO problem has optimal *Markovian* policies? Bäuerle & Rieder (2014) provide a positive answer by establishing a precise characterization of such utility functions which turn out to be either affine or exponential (when $\mathcal{U}$ is continuous and increasing). This highlights the role of the (nonlinear) utility functions on the nature of optimal policies. However, these results cannot be extended to CPT-PO in general.

I.2 Positioning CPT-RL in the literature

Remark 19. *For the infinite horizon discounted setting, the objective becomes the CPT value of the random variable $X = \sum_{t=0}^{+\infty} \gamma^t r_t$ recording the cumulative discounted rewards induced by the MDP and the policy π . The policy can further be parameterized by a vector parameter $\theta \in \mathbb{R}^d$.*

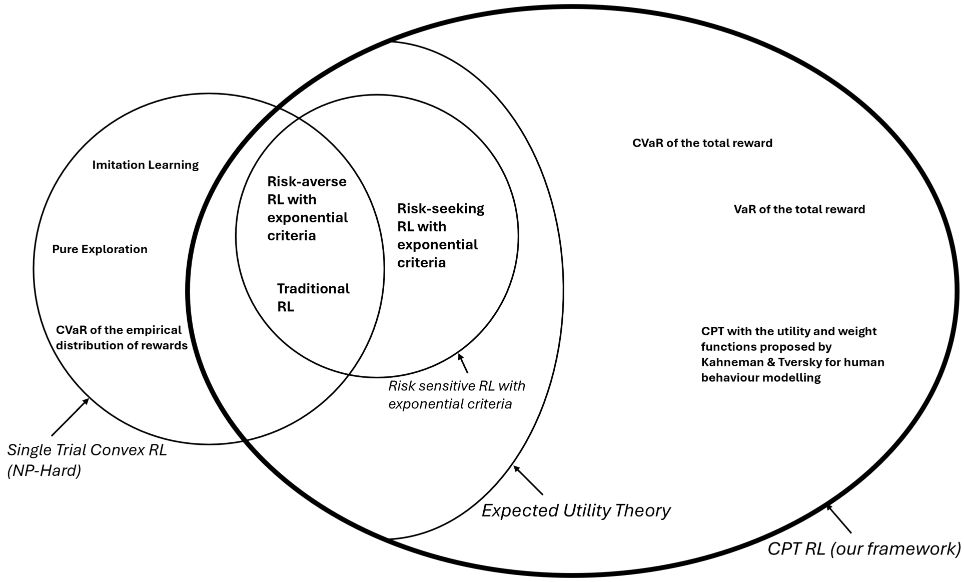

Imitation Learning
Pure Exploration
CVaR of the empirical distribution of rewards
Single Trial Convex RL (NP-Hard)
Risk-averse RL with exponential criteria
Traditional RL
Risk-seeking RL with exponential criteria
Risk sensitive RL with exponential criteria
Expected Utility Theory
CVaR of the total reward
VaR of the total reward
CPT with the utility and weight functions proposed by Kahneman & Tversky for human behaviour modelling
CPT RL (our framework)

Figure 10: A Venn Diagram representing our framework and some other frameworks in the literature

I.3 CPT value examples

Setting	**Utility function**	$\boldsymbol{w_+}$	$\boldsymbol{w_-}$
CPT	Any	Any	Any
CPT (Functions proposed by Kahneman and Tversky)	$\begin{cases}(x-x_0)^\alpha & \text{if } x \geq 0,\\ -\lambda(x-x_0)^\alpha & \text{if } x < 0\end{cases}$	$\frac{p^\eta}{(p^\eta+(1-p)^\eta)^{\frac{1}{\eta}}}$	$\frac{p^\delta}{(p^\delta+(1-p)^\delta)^{\frac{1}{\delta}}}$
EUT	Any	Identity function	Identity function
Distortion risk measure	Identity function	Any	$1-w_+(1-t)$
CVaR* ((Balbás et al., 2009))	Identity function	$1-w_-(1-t)$	$\begin{cases}\frac{x}{1-\alpha} & \text{if } 0 \leq x < 1-\alpha,\\ 1 & \text{if } 1-\alpha \leq x \leq 1\end{cases}$
VaR* (Balbás et al., 2009)	Identity function	$1-w_-(1-t)$	$\begin{cases}0 & \text{if } 0 \leq x < 1-\alpha,\\ 1 & \text{if } 1-\alpha \leq x \leq 1\end{cases}$
Risk-sensitive RL with exponential criteria (Noorani et al., 2022)	$\frac{1}{\beta}\exp(\beta x), \beta > 0$	Identity function	Identity function

Table 1: CPT value examples. *: Note that w_+ and w_- are discontinuous for VaR and CVaR.

I.4 Proof: CVaR, Var and distortion risk measures are CPT values

For a random variable X and a non-decreasing function $g : [0,1] \to [0,1]$ with $g(0) = 0$ and $g(1) = 1$, the **distortion risk measure** (Sereda et al., 2010) is defined as:

$$\rho_g(X) := \int_{-\infty}^{0} \tilde{g}(F_{-X}(x))dx - \int_{0}^{+\infty} g(1 - F_{-X}(x))dx\,,$$

where $F_{-X} : t \mapsto \mathbb{P}(-X \leq t)$ and $\tilde{g} : t \mapsto 1 - g(1-t)\,.$

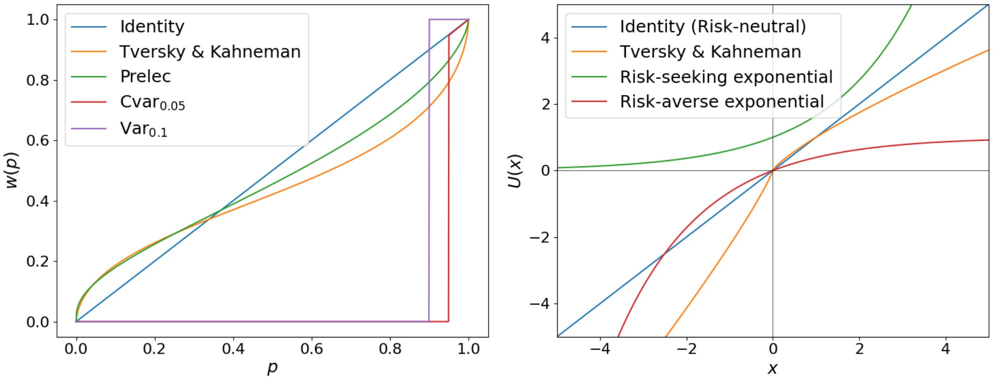


Figure 11: Various examples of probability weight functions (left) and utility functions (right).

Proposition 20. *Any distortion risk measure of a given random variable X can be written as a CPT value with $u^+ = id^+$, $u^- = -id^-$, $w_+ = \tilde{g}$ and $w_- = g$.*

Proof. It follows from the definition of the distortion risk measure together with a simple change of variable $x \mapsto -x$ that:

$$\begin{aligned}
\rho_g(X) &= \int_{-\infty}^{0} \tilde{g}(F_{-X}(x))dx - \int_{0}^{+\infty} g(1 - F_{-X}(x))dx \\
&= -\int_{+\infty}^{0} \tilde{g}(F_{-X}(-x))dx - \int_{0}^{+\infty} g(1 - F_{-X}(x))dx \\
&= \int_{0}^{+\infty} \tilde{g}(F_{-X}(-x))dx - \int_{0}^{+\infty} g(1 - F_{-X}(x))dx \\
&= \int_{0}^{+\infty} \tilde{g}(\mathbb{P}(-X \leq -x))dx - \int_{0}^{+\infty} g(1 - \mathbb{P}(-X \leq x))dx \\
&= \int_{0}^{+\infty} \tilde{g}(\mathbb{P}(X \geq x))dx - \int_{0}^{+\infty} g(\mathbb{P}(-X > x))dx \,.
\end{aligned}$$

Since $g(\mathbb{P}(-X > x)) = g(\mathbb{P}(-X \geq x))$ almost everywhere (in a measure theoretic sense) on $[0, +\infty($, and g is bounded, we obtain:

$$\rho_g(X) = \int_{0}^{+\infty} \tilde{g}(\mathbb{P}(X \geq x))dx - \int_{0}^{+\infty} g(\mathbb{P}(-X \geq x))dx \,.$$

We recognize the CPT-value of X with $u^+ = \mathrm{id}^+$, $u^- = -\mathrm{id}^-$, $w_+ = \tilde{g}$ and $w_- = g$. □

Remark 21. *When X admits a density function, Value at Risk (VaR) and Conditional Value at Risk (CVaR) (Wirch & Hardy (2001)) have been shown to be special cases of distortion risk measures and are therefore also instances of CPT-values.*

I.5 Simple examples and further insights about CPT

Human decision makers might not act rationally due to psychological biases and personal preferences, their decisions might not necessarily be dictated by expected utility theory. Consider this simple example as a first illustration: A player must choose between (A) receiving a payoff of 80 and (B) participating in a lottery and receive either 0 or 200 with equal probability. The player's preference depends on their attitude towards risk. While a risk-neutral agent will be satisfied with the immediate and safe payoff of 80, another individual might want to try to obtain the much higher 200 payoff. In particular, different agents might perceive the same utility and the same random outcome differently. Furthermore, they can exhibit both risk-seeking and risk-averse behaviors depending on the context.

Therefore, due to its failure to capture such settings as a descriptive model, the standard expected utility theory has been called into question by the pioneering behavioral psychologist Daniel Kahneman together with his colleague Amos Tversky (Kahneman & Tversky, 1979). In particular, Daniel Kahneman has been awarded the Nobel Prize in Economic Sciences in 2002 "for having integrated insights from psychological research into economic science, especially concerning human judgment and decision-making under uncertainty". In their seminal works combining cognitive psychology and economics, they laid the foundations of the so-called prospect theory and its cumulative version later on (Tversky & Kahneman, 1992) to explain several empirical observations that challenge the standard expected utility theory.

Let us illustrate this in a simple example borrowed from Ramasubramanian et al. (2021) (example 2 in section IV therein) for the purpose of our exposition. Consider a game where one can either earn \$100 with probability (w.p.) 1 or earn 10000 w.p. 0.01 and nothing otherwise. A human might rather lean towards the first option which gives a certain gain. In contrast, if the situation is flipped, i.e., a loss of 100 w.p. 1 versus a loss of \$10000 w.p. 0.01, then humans might rather choose the latter option. In both settings, the expected gain or loss has the same value (100). The CPT paradigm allows to model the tendency of humans to perceive gains and losses differently. Moreover, the humans tend to deflate high probabilities and inflate low probabilities (Tversky & Kahneman, 1992; Barberis, 2013). For instance, as exposed in L.A. et al. (2016), humans might rather choose a large reward, say 1 million dollars w.p. 10^{-6} over a reward of 1 w.p. 1 and the opposite when rewards are replaced by losses.

I.6 Connection to General Utility RL and Convex RL in finite trials

In this section, we elaborate in more details on one of the connections we noticed (and mentioned in related works) between our CPT-PO problem of interest and the literature of generality utility RL.

First, we recall a few notations complementing the preliminaries. Any fixed policy π and any initial state distribution ρ induce together a state occupancy measure d_ρ^π recording the visitation frequency of each state, it is defined at each state $s \in \mathcal{S}$ by $d_\rho^\pi(s) := \sum_{t=0}^{H-1} \mathbb{P}_{\rho,\pi}(s_t = s)$. The corresponding state-action occupancy measure is defined for every state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ by $\mu_\rho^\pi(s,a) := d_\rho^\pi(s)\pi(a|s)$. Recall that $J(\pi) = \langle \mu_\rho^\pi, r\rangle := \sum_{s\in\mathcal{S},a\in\mathcal{A}} \mu_\rho^\pi(s,a)r(s,a)$ for any policy π and any initial state distribution ρ.

The general utility RL problem consists in maximizing a (non-linear in general) functional of the occupancy measure induced by a policy. More formally, the general utility RL can be written as follows:

$$\max_\pi F(d_\rho^\pi)\,, \tag{18}$$

where F is the real valued utility function defined on the set of probability measures over the state or state-action space, ρ is the initial state distribution and d_ρ^π is the state (or sometimes state-action) occupancy measure induced by the policy π. This problem captures the standard RL problem as a particular case by considering a linear functional F defined using a fixed given reward function. Recently, motivated by practical concerns, Mutti et al. (2023) argued for the relevance of a variation of the problem under the qualification of *convex RL in finite trials*. They introduce for this the empirical state distributions $d_n \in \Delta(\mathcal{S})$ defined for every state $s \in \mathcal{S}$ by:

$$d_n^\pi(s) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=0}^{T-1} \mathbb{1}(s_{t,i} = s)\,, \tag{19}$$

where $s_{t,i}$ is the state at time t resulting from the interaction with the MDP (with policy π) in the i-th episode, among n independent trials. Their policy optimization problem is then as follows:

$$\max_\pi \xi_n(\pi) := \mathbb{E}[F(d_n^\pi)]\,. \tag{20}$$

Note that d_n^π is a random variable as it is an empirical state distribution. Observe also that $\lim_{n\to\infty} \xi_n(\pi) = F(d_\rho^\pi)$ under mild technical conditions (e.g. continuity and boundedness of F). This shows the connection between the above final trial convex RL objective and the general utility RL problem (18). The interesting differences between both problem formulations arise for small values of n. Of particular interest, both in this paper and in Mutti et al. (2023), is the *single trial RL* setting where $n = 1$.

Setting the probability distortion function w to be the identity, our CPT-PO problem becomes EUT-PO, i.e.:

$$\max_{\pi} \mathbb{E}\left[\mathcal{U}\left(\sum_{t=0}^{H-1} r_t\right)\right], \tag{21}$$

which is of the form $\xi_1(\pi)$, the single-trial RL objective as defined in Mutti et al. (2023). Indeed, it suffices to write the following to observe it:

$$\mathcal{U}\left(\sum_{t=0}^{H-1} r_t\right) = \mathcal{U}(\langle d_1^{\pi}, r\rangle), \tag{22}$$

where r is the reward function seen as a vector in $\mathbb{R}^{|\mathcal{S}|}$, $\langle\cdot,\cdot\rangle$ is the standard Euclidean product in $\mathbb{R}^{|\mathcal{S}|}$. Therefore, it appears that the above objective is indeed a functional of the empirical distribution d_1^{π}. Single trial general utility RL is more general than EUT-PO since it does not necessarily consider an additive reward inside the non-linear utility and can accommodate any (convex) functional of the occupancy measure. However, CPT-PO does not appear to be a particular case of single trial convex RL because of the probability distortion function introduced.

I.7 Choice of the utility function

We provide here additional comments regarding the choice of the utility function to complement our brief discussion in the main part of the paper. The problems themselves might dictate to the user or decision maker the utility function to be used. The user might also design their own according to their own beliefs, behaviors and objectives, based on the goal to be achieved (e.g. risk-seeking, risk-neutral, risk-averse). Specific applications might also suggest specific utility functions such as specific risk measures like in risk sensitive RL for instance. We have provided in table 1 a list of different examples one might consider. Learning the utility function is also an interesting direction to investigate as we mention in the conclusion. In practice, it is rather common to use the example we provide in table 1 (CPT row) with exponent parameters which are estimated using data. We provide a few concrete examples in the following. For instance, Rieger et al. (2017) adopt such an approach (see sections 3.1, 3.2 and 3.3 therein for a detailed discussion about parameter estimation). Ebrahimigharehbaghi et al. (2022) choose some similar variation of this utility (see eqs. 2-3 therein) while still using KT's probability weighting functions. Gao et al. (2023) compare different functions for different similar power utility functions with fitted parameters (see Tables 1, 2 and 3 therein p. 3, 4, 6 for extensive comparisons with the existing literature). Similar investigations were conducted in Yan et al. (2020). Dorahaki et al. (2022) consider psychological time discounted utility functions (variations of the same power functions) in their model with additional relevant hyperparameters, motivated by (domain-specific) psychological studies (see eq. (4) therein). It is worth noting that all these examples are only in the static stateless setting.

J More Details about Section 5 and Additional Experiments

Batch size and quantile estimation. Quantile estimation is generally more sensitive to batch size than expectation-based methods. However, in practice, we observed that relatively small batches (5–32 samples) performed robustly across all environments. As shown in Fig. 15, larger batches can improve performance, but small batches already yield competitive CPT values.

History dependence of policy networks. In most experiments, we used feed-forward networks without explicit history dependence. We found that Markovian policies performed well in our simple simulations for applications like financial markets and Mujoco control. In other domains (e.g., electricity management), we incorporated partial history by expanding the input window to represent multiple past states. Typically, the input size was chosen to capture sufficient context without encoding the full trajectory length (H). This limited form of history dependence was sufficient in our settings, though incorporating more expressive temporal models could enhance performance further.

Hardware and execution time. We have run part of the experiments on a laptop with a 13th Gen Intel Core i7-1360P2.20 GHz CPU and 32 GB of RAM, and the remaining simulations (section 5, App. J.8, J.9) on a MacBook Pro M4 with 48 GB of RAM. Experiments can take a few seconds for the simplest ones, to a few minutes for the ones in the main part. More complex ones take a couple of hours, e.g. for electricity management (App. J.7), Finance (App. J.8) and MuJoCo (App. J.9).

Software. Our experiments are coded in Python and mainly use Pytorch (Paszke et al., 2019).

Weight functions. The risk-neutral w_+ function is simply the identity function. As for the definition of other probability distortion functions w_+ we use for experiments, we define:

$$w_{ra}(x) := \begin{cases} 0.5x & \text{if } x \leq 0.9, \\ 5.5x - 4.5 & \text{otherwise.} \end{cases} \quad w_{rs}(x) := \begin{cases} 5x & \text{if } x \leq 0.1, \\ \frac{1}{2} + \frac{5}{9}(x - 0.1) & \text{otherwise.} \end{cases}$$

$$w_{sra}(x) := \begin{cases} 0.1x & \text{if } x \leq 0.9, \\ 9.1x - 8.1 & \text{otherwise.} \end{cases} \quad w_{srs}(x) := \begin{cases} 9x & \text{if } x \leq 0.1, \\ \frac{1}{9}x + \frac{8}{9} & \text{otherwise.} \end{cases}$$

Optimizer. Instead of vanilla stochastic gradient descent, we use the Adam optimizer to speed up convergence. In our Python implementation, we use the same batch of trajectories for estimating the function φ and for performing the stochastic gradient ascent step.

Neural net activation function. We use the tanh activation function before the last softmax layer to encourage exploration and reduce the risk of converging to local optima which may occasionally occur for some runs.

J.1 More details about bandit simulations in Section 5

In addition to the parameters specified in the main part, we used the following set of parameters to conduct the bandit experiments in Section 5:

Table 2: Hyperparameters for the bandit experiments in Section 5

Hyperparameter	**Value**
Optimizer	Adam
Learning rate	0.01
Number of episodes	1000
Batch size	10

We used a piecewise linear probability weight function given by the coefficients: $w = [1.0, 0.0, 1.0, 0.0, 0.5]$ where the notation $[a_1, a_2, \ldots, a_i, b_1, b_2, \ldots, b_i, c_1, \ldots, c_{i-1}])$ refers to a piecewise affine w function with $f(x) = a_i x + b_i$ for $c_{i-1} < x < c_i$.

In Fig. 12, we show a case where CPT-PG learns a stochastic policy.

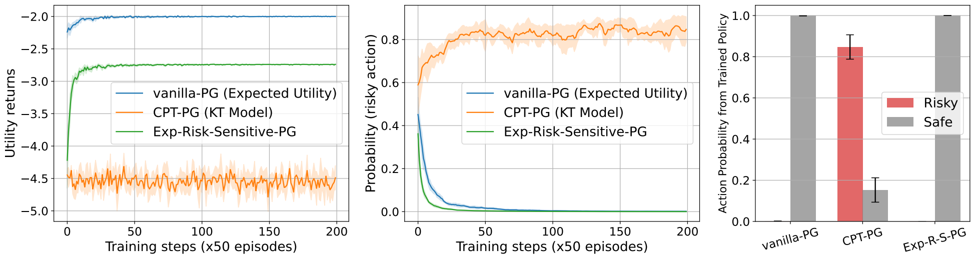


Figure 12: Comparison of our CPT-PG algorithm with exponential risk-sensitive PG and vanilla PG on a simple 2-action bandit setting. (Lower figure) Loss lottery setting: Only CPT picks the risky action with high probability. Note here that the policy trained by CPT-PG is stochastic.

J.2 CPT compared to Distortion Risk Measures

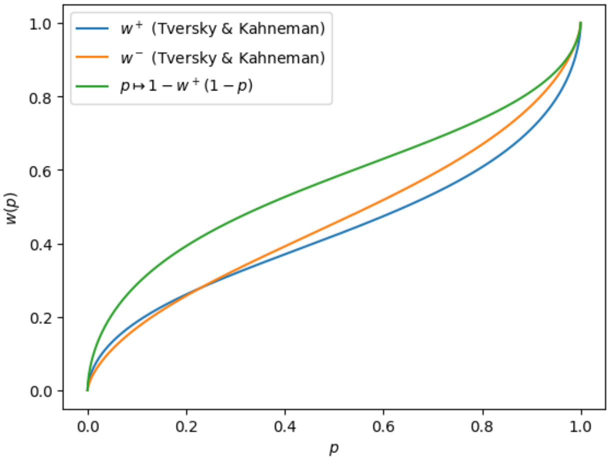
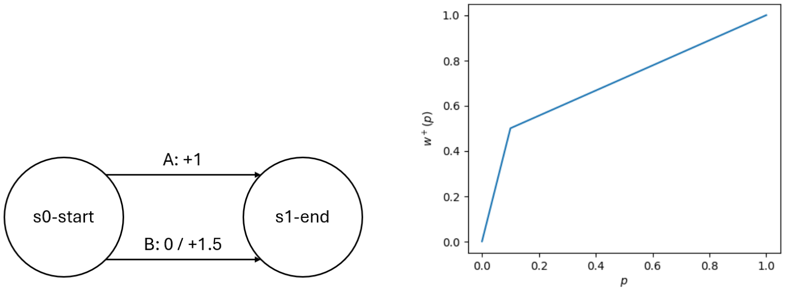


Figure 13: Illustration of the flexibility of CPT compared to the Distortion Risk Measure. Notice how w_- is distinct from both w_+ and $p \mapsto 1 - w_+(1-p)$.

J.3 Illustration of the need for stochastic policies in CPT-RL



(a) The MDP (b) The w_+ function

Figure 14: Setting of the experiment on non-deterministic policies and batch size influence

We study experimentally the behavior of our algorithm with regards to small batch sizes.

Setting. We use the barebones setting (Figure 14) with a w function that aggressively focuses on the 10% of favorable outcomes. Denoting by p the probability of choosing A for a given policy, we look at the value of p at convergence (1000 optimization steps) for various batch sizes. The optimal policy corresponds to $p = 0.8$.

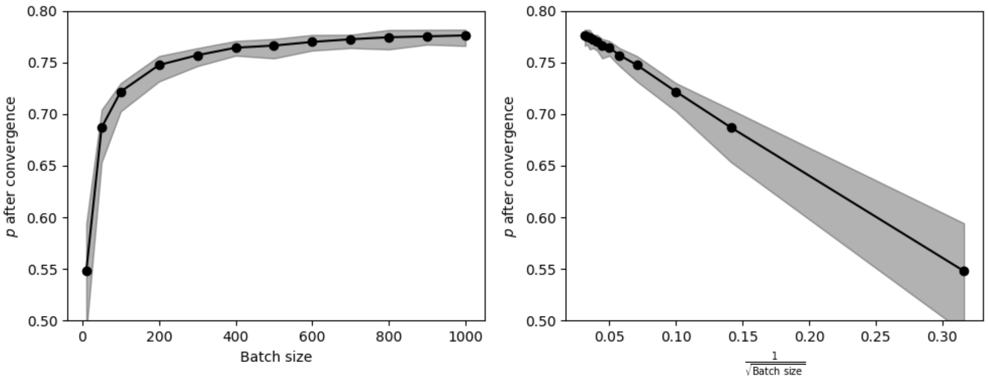


Figure 15: Results of the experiment on non-deterministic policies and batch size influence, over 100 runs. The black dots are the medians and the shaded area represents the interquartile range.

Insights. For each batch size we test, we run and plot a hundred training rounds (Figure 15). We fist observe that the policy we obtain with our algorithm indeed approaches the optimal $p = 0.8$ policy. The estimation error (w.r.t. the optimal theoretical value of $p = 0.8$) appears to be of order $\frac{1}{\sqrt{\text{batch size}}}$. It was to be expected that a small batch size would lead to a bias in CPT value and CPT gradient estimation, and, finally, in policy, as a small batch size renders impossible an accurate estimation of the probability distribution of the total return function. The fact that this bias appears to be proportional to the inverse of the square root of the batch size is in line with the standard statistical intuition (as e.g. per the central limit theorem). In our particular example, the estimated p is below (and not above) the theoretical p. This is likely because our w function places a strong weight on the top 10% of outcomes. Hence there is an imbalance between the impact of overestimating and underestimating the proportion of good outcomes in a given run: if we underestimate the probability of getting +1.5 with a given policy due to sampling, the effect will be stronger than the opposite effect we would get by overestimating the probability of the same error. As the batch size grows, the estimation error is reduced and the effect vanishes.

J.4 Markovian vs Non-Markovian Policies for CPT-RL

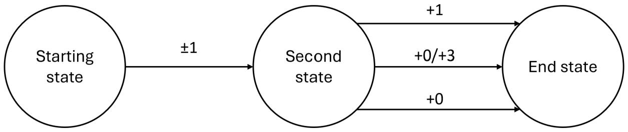


Figure 16: The environment for the experiment on non-Markovian policy

To illustrate the fundamental difference between memoryless utility functions and others we conduct a small experiment on a simple setting (Figure 16), similar to the one introduced in the proof of the theorem. We consider three states and three actions. From the starting state, any action leads to the second state with probability 1 and yields a reward of $+1$ with probability $\frac{1}{2}$ and of -1 with probability $\frac{1}{2}$. Once in the second state, the first action yields reward $+1$ with probability 1, the second action yields 0 or 3 with probability $\frac{1}{2}$ each, and the third action always yields 0. We compare the performance of a policy parametrized in $\Pi_{\Sigma,NS}$ and one in $\Pi_{M,NS}$.

Insights. The results (Figure 22) illustrate indeed the performance advantage of the non-Markovian policy compared to the Markovian one in the case of a non-affine, non-exponential utility function, and the absence thereof in the exponential setting.

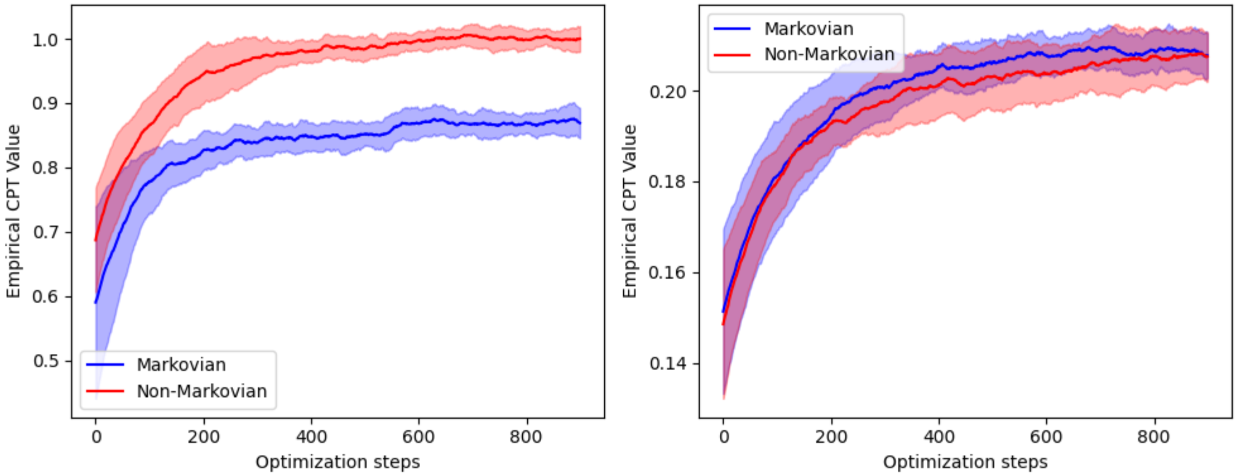


Figure 17: Comparison of Markovian and Non-Markovian policy performances for non-exponential (left) and exponential (right) utility functions. Shaded areas represent a range of ± one standard deviation over 20 runs.

J.5 Grid Environment

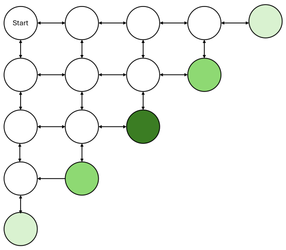


Figure 18: Scaling grid example.

↓	↓	↓	↓
→	↓	↓	↓
→	→	→	↓
+5	→	→	+6

(a) A risk-neutral optimal policy obtained with our algorithm

↓	↓	↓	↓
↓	↓	↓	↓
↓	→	→	↓
+5	→	→	+6

(b) An optimal policy obtained by training with the risk-averse utility $\mathcal{U} : x \mapsto \sqrt{x}$

Figure 19: Comparison of optimal policies under risk-neutral and risk-averse scenarios

Exploration. To avoid our gradient ascent algorithm getting stuck in a local optimum, we have to ensure enough exploration is going on. Therefore, we tweak the last layer of the neural network to prevent every action's probability from vanishing too soon. We choose a parameter α, choose our last layer as $x \mapsto \text{softmax}(\alpha \tanh(x/\alpha))$, and we let α slowly grow with the iterations. A small α forces exploration, larger α allows for more exploitation: this is similar to an ϵ-*greedy* scheme (with ϵ decaying as α grows), as it forces every action to be chosen with at least a small probability.

Scalability to larger state spaces. We consider a family of MDPs where the state space is a $n \times n$ grid for a given integer parameter n. The agent starts in the top right corner and has always four possible actions (up,down,left,right). Taking a step yields a reward of $\frac{-1}{n}$, attempting to leave the grid yields $\frac{-2}{n}$, and reaching the anti-diagonal ends the episode with a positive reward. All cells on the anti-diagonal yield the same expected reward, but with different levels of risk; the least risky reward is the deterministic one, in the center of the grid. We consider tabular policies and the initial policy is a random policy assigning the probability 1/4 to each action. We test the sensitivity of the performance of both algorithms to the size of the state space. The steps sizes of both algorithms have been tuned to approach their possible performance; we wish to draw attention not to the absolute performance of either algorithm on any particular example, but rather to the evolution of the performance of both as the size of the problem increases. We observe that the performance of CPT-SPSA-G suffers for larger state space size whereas our PG algorithm is robust to state space scaling. While both algorithms are gradient ascent based algorithms in principle, our stochastic policy gradients are different.

Influence of the utility function. We consider a 4x4 grid for our illustration purpose. Our agent starts on a random square on one of the three upper rows of the grid, and can move in all four directions. Any move to an empty square will award it a random reward of -1 with probability $\frac{1}{2}$ and of $+0.8$ with probability $\frac{1}{2}$. Therefore, longer trajectories are slightly costly in expectation, and generate significant variance. In two corners of the grid, we add cells that yields rewards of $+5$ for one or $+6$ for the other, and conclude the episode. Illegal moves (attempting to leave the grid) are punished by a negative reward. Our parameterized policy is a neural network whose last layer is activated with softmax and has 4 coordinates corresponding to the 4 different possible moving actions. We consider solving CPT-PO with different utility functions: risk-neutral identity utility, risk-averse KT utility, as well as exponential utility function. The obtained policies differ depending on the utility function. For examples of risk-neutral/averse policies obtained, see Fig. 19b.

J.6 Traffic Control over a Grid

We simulate a car agent navigating a city grid, where central roads are faster but risk higher delays (see fig. 20). The agent must balance speed against risk by avoiding the city center. In risk-neutral settings, the agent favors faster routes, while risk-averse policies avoid the risky central roads. Fig. 20 (center) demonstrates that our algorithm successfully adapts to different risk-weighted objectives. We also consider solving CPT-PO with different utility functions: risk-neutral identity utility, risk-averse KT utility, as well as exponential utility function. Fig. 23 show the corresponding different CPT returns observed. The obtained policies differ depending on the utility function. For examples of risk-neutral/averse policies obtained, see Fig. 19b in App. J.5.

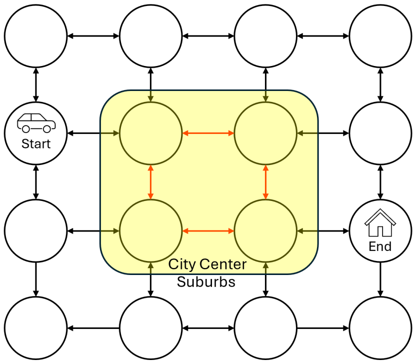


Figure 20: Traffic control environment: red roads in the city center are prone to congestion.

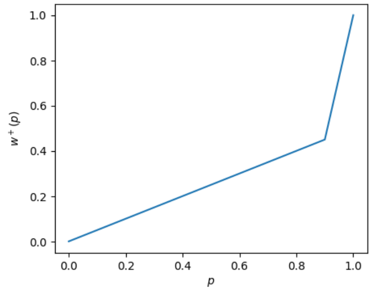


Figure 21: The probability distortion function w_+ used for the traffic control experiment.

	0	1	2
0	→	→	↓
1	↑	?	⌂
2	↑	?	↑

(a) Training with our w function for traffic control (3 × 3)

	0	1	2
0	→	→	↓
1	→	→	⌂
2	→	→	↑

(b) Risk-neutral reference

	0	1	2	3
0	→	→	?	↓
1	↑	?	→	↓
2	↑	↑	→	⌂
3	↑	↑	?	→

(c) Training with our w function for traffic control (4 × 4)

	0	1	2	3
0	→	→	→	↓
1	→	→	→	↓
2	→	→	→	⌂
3	→	→	→	↑

(d) Risk-neutral reference

Figure 22: Examples of policies obtained with our algorithm. Question marks indicate a non-deterministic action selection in a given state.

Implementation details. In both cases, the risk-neutral optimal solution (going around the city center) is also a local optimum for the risk-averse objective, and, because it is a shorter path, is easier to stumble upon by chance when exploring the MDP. This means we have to implement special measures to force exploration. The algorithm used *as is* is prone to get stuck from time to time in local minima on this example. It would seem that our w function, which is aggressively risk-averse, hinders exploration. To mitigate this, we introduce an entropy regularization term that we add to the score function with a decaying regularization

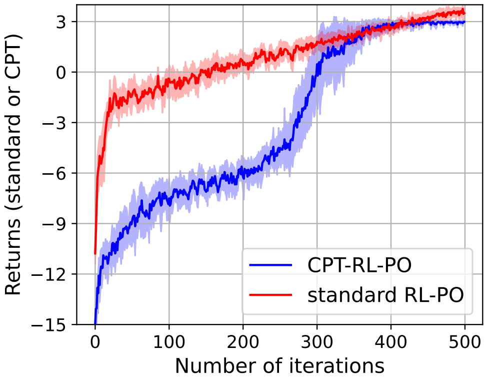


Figure 23: Returns along the iterations of our PG algorithm for CPT-PO for traffic control. Shaded areas indicate a range of ± one standard deviation over 20 runs. See App. J for details.

weight in the policy gradient found in Theorem 3, see App. J for further details. We incorporate entropy regularization in the policy gradient as follows:

$$\mathbb{E}\left[\varphi\left(\sum_{t=0}^{H-1} r_t\right)\sum_{t=0}^{H-1}\nabla_\theta\Big(\log\pi_\theta(a_t|s_t)+\underbrace{\alpha_n H(\pi_\theta(a_t|s_t))}_{\text{Entropy regularization term}}\Big)\right], \tag{23}$$

where α_n is the weight of the regularization. We found that a decaying α_n yielded the best results.

On the 4 × 4 grid, we also start by pretraining our model with a risk-neutral method for a few steps, to accelerate training and avoid some bad local optima we can stumble upon due to unlucky policy initialization, before carrying on with our risk-aware method.

J.7 Electricity Management

In this application involving *continuous* state and action spaces, we consider an individual home which has solar panels for producing electricity and a battery for storing energy (see Fig. 24, left). We consider a 24-hour time frame where the agent must decide the quantity of electricity to buy/sell, based on solar panel production, market prices, and battery levels. We use public data for selling prices recorded on a national electricity network. Public data is available online.[5] We experiment with risk-neutral, risk-averse, and risk-seeking objectives with 3 weight functions w and a Gaussian neural network policy. Our algorithm performs well across these scenarios, with the risk-averse policy minimizing downside risk, and the risk-seeking policy maximizing potential gains. Fig. (24) (right) shows the distribution of total returns for different objectives. The most rewarding time to sell electricity is around 4pm (see prices in Fig. 24, center). However, selling too much too soon exposes us to the risk of falling short of battery during the night and risking to buy it later for a higher price. The risk-averse policy avoids selling a lot of electricity and tends to keep it stored until the end of the day. Conversely, the risk-seeking policy aggressively sells energy when the markets are high at the cost of possibly having to buy it again later in the day. The risk-averse policy has the distribution with the best left tail (worst cases are not too bad), the risk-seeking distribution has the best right tail (best cases are particularly good). The risk-neutral policy has the best mean value.

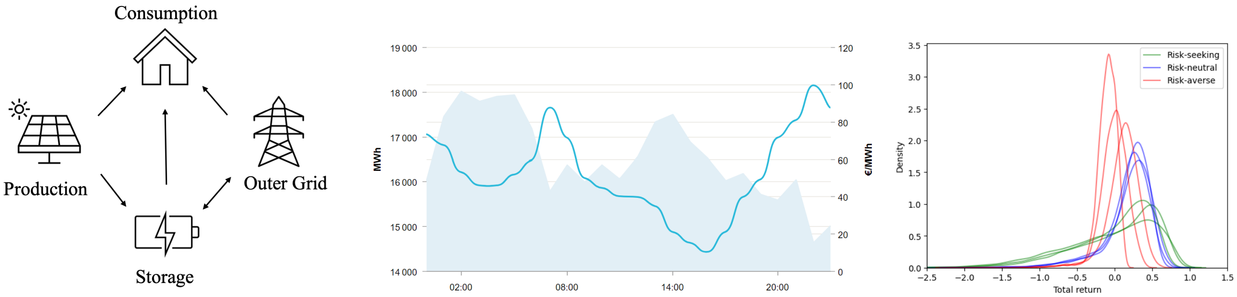


Figure 24: (**Left**) Electricity management environment: arrows refer to electricity flow. (**Center**) Electricity prices in a typical day, the blue line (right-hand side scale) records the electricity price on the European market, the shaded area (left-hand side scale) represents the total electricity production. (**Right**) Density of the empirical returns obtained by deploying different trained PG policies from different initializations, density computed using 10,000 runs for each curve.

[5] www.services-rte.com/en/view-data-published-by-rte/france-spot-electricity-exchange.html

J.8 Trading in Financial Markets

We discuss here an application of our methodology to financial trading. The goal is to train RL trading agents using our general PG algorithm in the setting of our CPT-RL framework.

Environment: general description. We consider a gym trading environment available online, all the details about this environment are available here: https://gym-trading-env.readthedocs.io/en/latest/. This environment simulates stocks and allows to train RL trading agents. For the interest of the reader, we provide a brief summary explaining how the environment works. The environment is build from a given dataframe and a list of possible positions. The dataframe contains market data throughout a given period. The list of possible positions will represent the set of possible actions the agent can take, We provide more details about our specific environment in the following paragraph.

Our trading environment. We use data from the Bitcoin USD (BTC-USD) market between May 15th 2018 and March 1st 2022 available in the aforementioned website. We note that the data used follows the same pattern as publicly available data after a few preprocessing steps, the reader can find such data examples at https://finance.yahoo.com/quote/BTC-USD/history including the date, a few extracted features ('open', 'high', 'low', 'close') which respectively represent the open price, i.e. the price at which the first trade occurred for the asset at the beginning of the time period, the highest, lowest and last such prices, and the volume in USD which is the total value of all trades executed in a given time period. In particular, we will consider static features (computed once at the beginning of the data frame preprocessing) and dynamic features (computed at each time step) such as the last position taken as introduced by the Gym Trading Environment.

- State space: We consider a seven dimensional continuous state space. Features are constructed from the raw stock market data as previously explained. State transitions are described using the provided time series. See the publicly available code of the environment for more details.

- Action space: We consider three classical types of positions the trader can take in a financial market: SHORT, OUT and LONG. These positions constitute the set of actions. These actions refer to whether the trader expects the price of an asset to rise or fall and how they are positioned to profit from that fluctuation. Extending this setting to a setting with a larger set of positions is straightforward as the environment implementation also supports more complex positions.

- Rewards: The rewards we consider are given by the log values of the ratio of the portfolio valuations at times t and $t-1$. Borrowing interest rates and trading fees are also considered in the computation. The reward function can also be easily modified in the environment thanks to the implementation of the Gym Trading Environment which builds on the standard Gym environments.

Remark 22. *One can easily build their own environment by downloading their own dataframe for any historical stock market data and performing their desired preprocessing as for the features they would like to consider to build their states.*

Experimental setting. We have tested several utility and probability weighting functions including a risk averse exponential of the form $x \mapsto \frac{1}{\beta}(1-\exp(-\beta x))$ with different values of β as well as the KT (Kahneman and Tversky) function as defined in the main part with different values of the reference point x_0 to illustrate its influence.

Hyperparameters. We used the following set of parameters to conduct the experiments:

Additional hyperparameters used are directly reported in the legends of the figures below.

Results. We refer the reader to Fig. 2. We make a few observations:

Table 3: Hyperparameters

Hyperparameter	**Value**
Optimizer	Adam
Learning rate	0.05
Number of episodes	100
Batch size	5
Number of steps per episode	25

- Influence of the reference point: It can be seen that the reference point shifts the values of the achieved CPT returns: The smaller the reference point, the larger are the returns. This is because only values larger than the reference point are perceived as positive returns given the definition of the KT utility. This illustrates how the subjective perception of the agent of the returns is taken into account by the model.
- Different return trajectories for different risk averse functions: Different values of β lead to different trajectories overall which can translate to different levels of risk aversion. In particular, the curves do not match the identity utility case in the first episodes and show more or less risk taken towards optimizing the CPT returns.
- Influence of the parameter α in KT's utility: Observe that the exponent α in the utility distorts the function and shifts the returns significantly. Lower values of α lead to higher returns in this setting where the returns (as per the ratio definition of the reward) are smaller than 1. This parameter α provides a degree of freedom to model the behavior of the agent as per their perception of the returns. Different values of α modify the curvature of the utility function (w.r.t. the reference point which is $x_0 = 0$ here) which is concave for gains and convex for losses.

J.9 Control on MuJoCo Environments

In this section we test our algorithm on the InvertedPendulum-v5 environment (Todorov et al., 2012) to demonstrate that our PG algorithm is also applicable to other control benchmarks with continuous state and action spaces.

Hyperparameters. We used the following set of parameters to obtain our results:

Table 4: Hyperparameters

Hyperparameter	**Value**
Optimizer	Adam
Learning rate	1e-4
Number of episodes	2000
Batch size	32
Maximum number of steps per episode	200

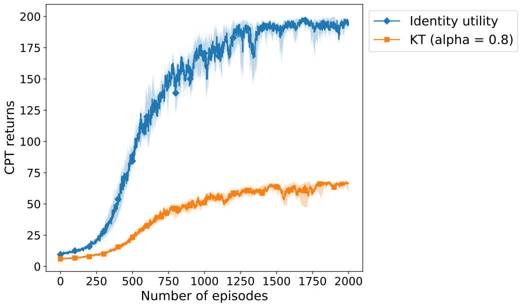


Figure 25: Performance of our PG algorithm on the InvertedPendulum-v5 environment (Todorov et al., 2012). KT refers to Kahneman and Tversky's utility function, alpha is the parameter used in the definition of KT's utility, exp. refers to exponential. Shaded areas are interquantile (25-75%) margins and curves report the median values over 10 different runs. All the CPT return curves are obtained with the same probability weighting function w which is piecewise affine with three segments (hence different from the standard RL identity setting).

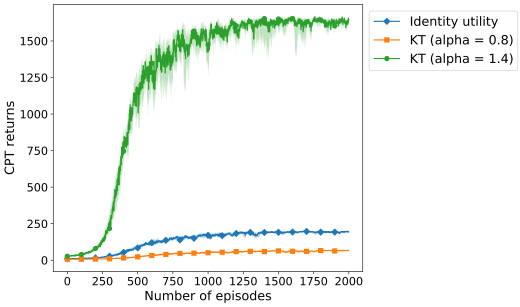


Figure 26: Performance of our PG algorithm on the InvertedPendulum-v5 environment (Todorov et al., 2012). This figure complements Fig 25 with the CPT returns using a KT utility with $\alpha = 1.4$. Notice that a much higher CPT return is achieved in that case. We also provide Fig. 25 for scaling purposes, the CPT returns being much higher for KT ($\alpha = 1.4$).