mathematics

MDPI

Article

Integrating Risk-Averse and Constrained Reinforcement Learning for Robust Decision-Making in High-Stakes Scenarios

Moiz Ahmad [1], Muhammad Babar Ramzan [2], Muhammad Omair [3] and Muhammad Salman Habib [4,*]

[1] Department of Industrial and Manufacturing Engineering, University of Engineering and Technology, Lahore 54700, Pakistan
[2] School of Engineering and Technology, National Textile University, Faisalabad 37610, Pakistan
[3] Department of Materials and Production, Aalborg University, 9220 Aalborg Øst, Denmark
[4] Institute of Knowledge Services, Center for Creative Convergence Education, Hanyang University ERICA Campus, Ansan-si 15588, Gyeonggi-do, Republic of Korea
* Correspondence: salmanhabib@hanyang.ac.kr

Abstract: This paper considers a risk-averse Markov decision process (MDP) with non-risk constraints as a dynamic optimization framework to ensure robustness against unfavorable outcomes in high-stakes sequential decision-making situations such as disaster response. In this regard, strong duality is proved while making no assumptions on the problem's convexity. This is necessary for some real-world issues, e.g., in the case of deprivation costs in the context of disaster relief, where convexity cannot be ensured. Our theoretical results imply that the problem can be exactly solved in a dual domain where it becomes convex. Based on our duality results, an augmented Lagrangian-based constraint handling mechanism is also developed for risk-averse reinforcement learning algorithms. The mechanism is proved to be theoretically convergent. Finally, we have also empirically established the convergence of the mechanism using a multi-stage disaster response relief allocation problem while using a fixed negative reward scheme as a benchmark.

Keywords: robust decision-making; dynamic decision-making; non-convexities; constrained reinforcement learning; augmented Lagrangian; Markov risk

MSC: 60J05; 60J10; 60J20; 68T05

Citation: Ahmad, M.; Ramzan, M.B.; Omair, M.; Habib, M.S. Integrating Risk-Averse and Constrained Reinforcement Learning for Robust Decision-Making in High-Stakes Scenarios. *Mathematics* **2024**, *12*, 1954. https://doi.org/10.3390/math12131954

Academic Editors: Ana M. Madureira, Joao Ferreira and André Santos

Received: 10 May 2024
Revised: 5 June 2024
Accepted: 17 June 2024
Published: 24 June 2024

1. Introduction

In certain sequential decision-making situations, e.g., disaster response or highly disruption-prone supply chains [1–4], decision outcomes are not only highly uncertain because of the gradual revelation of uncertainty, but their worst effects cannot be tolerated. To be robust against these worst-case outcomes, we need to use constraints to eliminate those decisions from the feasible solution space that may lead to them, while for additional robustness, we must use a dynamic risk-averse objective as a representation of decision makers' risk-averseness (a behavioral consideration). In the literature, constraints are specifically used for robust decision-making, especially in the safe reinforcement learning domain [5–7], while the use of a risk-averse objective function is also prevalent for decision robustness [8–12]. Recently, the simultaneous use of constraints and risk aversion is also regarded as an effective approach for robust decision-making [1,13,14]. This leads us towards multi-stage risk-averse constrained optimization models. Studies have used stochastic dual dynamic programming (SDDP) for the solution of these kinds of models [9,15]. However, SDDP assumes the problem's convexity and a finite number of random scenarios for its optimal convergence, which is restrictive as many real-world problems are non-convex, e.g., humanitarian relief allocation with deprivation costs [16], requiring multi-stage solution approaches which can handle non-convexities. Although [17] did

not presume the problem's convexity while using disciplined convex–concave programming for multi-stage risk-averse constrained optimization, optimal convergence was not established either theoretically or empirically. In this regard, as a model-free machine learning-based framework [18,19], reinforcement learning (RL) can be a better solution approach to non-convex problems. It has shown state-of-the-art performance in various real-world domains, e.g., cloud computing [20], finance [21,22], transportation [23,24], energy [25,26], inventory management [27,28], manufacturing or production scheduling [29], and routing [30]. However, in all of the above applications, robust and high-stack decision-making is not required, and thus expectation-based unconstrained RL is used. As we are considering high-stakes decision-making under uncertainty, risk-averse RL algorithms with constraint-handling abilities are required.

In this context, when it comes to constrained RL, besides other approaches such as mapping any action chosen by the policy to a constraint satisfying one which may need domain knowledge for its implementation [31], the most common and, so far, the best approach for solving constrained MDP is via Lagrangian dual (or augmented Lagrangian dual) optimization [32,33]. In this method, a constrained multi-stage optimization problem is converted into a min–max unconstrained problem, which is then solved by traditional RL algorithms. Besides its advantages, such as allowing non-convex constraints and not requiring any prior knowledge, it has theoretical guarantees for convergence as well. In this regard, it was proven that constrained MDPs have zero duality gap with no assumption on the problem's convexity [34]. This result explains the performance of these primal-dual methods, such as primal-dual policy optimization (PPO) [35] and Lyapunove-based policy learning [36]. However, in all these methods, cumulative constraints are considered, the satisfaction of which does not mean the satisfaction of instantaneous constraints [37]. This is why an augmented Lagrangian-based mechanism for constrained RL has recently been developed for instantaneously constrained MDPs by [37], which extended the theoretical results from [34]. However, all of the above-constrained RL developments are in the realm of expectation-based RL [12,38], while no research has been performed on handling constraints in risk-averse RL. The reason behind this paucity of research is that the risk-averse RL is in its embryonic stage itself [12], adding constraints to it is the next stage of investigation, which no one has entered yet.

Research Contributions

Therefore, we have initiated the new research domain of constrained risk-averse RL to ensure optimal convergence of multi-stage risk-averse constrained mathematical programs, which can be easily reformulated into their version of a stochastic decision-making process, namely risk-averse Markov decision processes with constraints [39,40]. In doing so, we will not make any assumption on the problem's convexity. However, some not-so-restrictive assumptions are made on the constraints and risk-averse objective. All the constraints used in this research are non-risk constraints. For the risk-averse MDPs with cumulative expectation-based constraints, we will assume those conditional spectral risk measures that satisfy conditions related to the derivative and integral of their risk-aversion function, given on page 7–9. Furthermore, for the incorporation of deterministic instantaneous constraints, the **translation invariance** property of coherent risk measures is made as one additional assumption on conditional spectral risk measures. This development will help the solution of real-world problems in many domains which are not yet considered solvable. For example, in humanitarian logistics, deprivation cost is an important metric for measuring human suffering due to deprivation from necessities. However, many researchers do not consider this cost to make the models tractable [41]. Despite the need for accurate and robust decision-making, no model in the disaster management literature has yet combined constrained optimization, risk aversion, and deprivation costs. This is due to the computational intractability issue which is solved by this research [42].

The main contributions of our research are summarized as follows:

- To prove the strong duality of spectral risk-averse MDPs with expectation-based cumulative and/or deterministic instantaneous constraints without making any assumption on the problem's convexity.
- To propose a constraint handling mechanism for risk-averse reinforcement learning based on the strong duality results.
- To theoretically and empirically establish the convergence of the proposed constraint handling mechanism.

For ease of understanding, the graphical representation of this research is shown in Figure 1. Subsequently, we will first specify the preliminaries and mathematical notations in Section 2, which will be crucial for understanding our theoretical results in Sections 3 and 4 for risk-averse MDPs with cumulative and instantaneous constraints, respectively. Based on the results, we will develop a constraint-handling mechanism in Section 5 where we will also prove the convergence of this newly developed approach. After this, in Section 6, we will empirically validate the convergence of the proposed mechanism using a multi-stage risk-averse constrained optimization problem. Findings will be discussed in Section 7, and finally, we will conclude in Section 8.

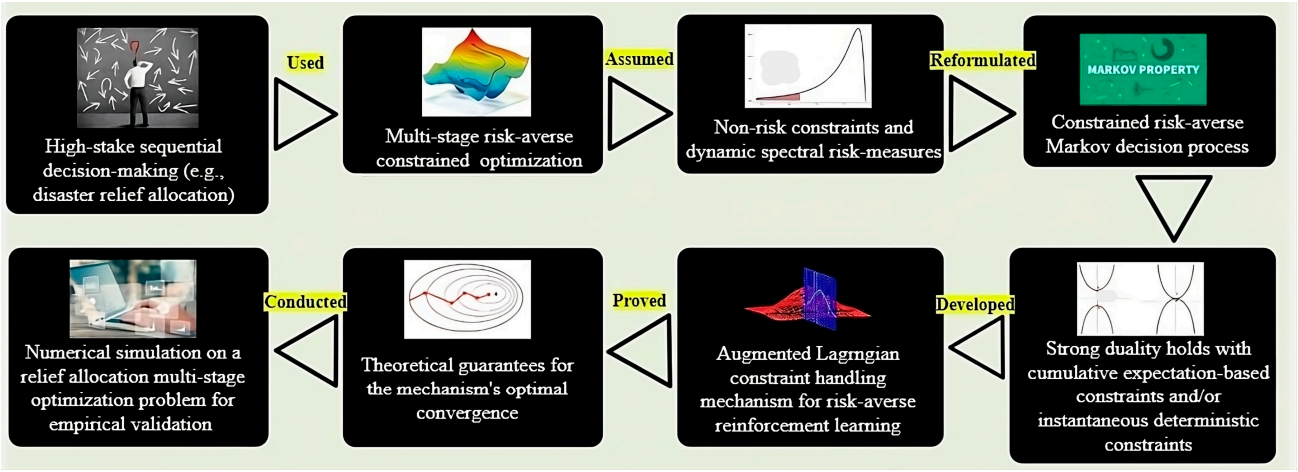


Figure 1. Graphical representation of this research.

2. Preliminaries

Let T be the index set of time periods(with indices starting from one) in a finite or infinite horizon MDP, then the key elements of the MDP are as follows:

2.1. Action Space

$a_t \in \mathcal{A} \subseteq \mathbb{R}^n \ \forall t \in T$ is the feasible action RL agent can take in the period $t \in T$ which belongs to the action space $\mathcal{A} \subseteq \mathbb{R}^n$ with n being the action dimension or the number of decision or control variables (continuous or discrete). It is assumed to be compact.

2.2. State Space

$s_t \in \mathcal{S} \subseteq \mathbb{R}^m \ \forall t \in T \cup \{|T|+1\}$ is the state belonging to the state space $\mathcal{S} \subseteq \mathbb{R}^m$ with m being the state dimension or the number of state variables (continuous or discrete). Similar to action space, it is also assumed to be compact. The state variables specify the state of the MDP in a particular period.

2.3. Policy Space

Here, $\pi \in \mathcal{P}$ is the feasible policy in the arbitrary policy space $\mathcal{P}$. $\mathcal{P}$ forms a measure space $(\mathcal{A} \subseteq \mathbb{R}^n, \mathcal{B}(\mathcal{A} \subseteq \mathbb{R}^n), \mathcal{P})$ with $\mathcal{B}(\mathcal{A} \subseteq \mathbb{R}^n)$ being a collection of Boral sets of the action space $A \in \mathbb{R}^n$.

2.4. Reward and Constraint Functions

There are cost functions $g_{it} = g_i(s_t \in \mathcal{S}, a_t \in \mathcal{A})$, $\forall i \in P \cup \{0\}, t \in T$ producing cost signals which are received by the RL agent in each period. These are the mappings from the current time's state $s_t \in \mathcal{S} \subseteq \mathbb{R}^m$ and the current time's action $a_t \in A \subseteq \mathbb{R}^n$ to the real number line $\mathbb{R}$. These costs are considered to be random variables either due to stochastic policy $\pi \in \mathcal{P}$ and/or uncertainty in the MDP environment dynamics. In this regard, P is the index set (with indices starting from one) of these cumulative or instantaneous constraint cost functions in a particular period $t \in T$, and the subscript "0" in (2) is regarded as the index of the reward function.

2.5. Discount Factor

$\gamma \in (0, 1)$ is the discount factor which is assumed to be in the open interval $(0, 1)$. This specifies the importance of the future rewards in the value of the current period's state. The greater its value, the more important are the future rewards for the current state's value.

2.6. Reference Equations

Before diving into the theory, we will first mention three Equations (1)–(3), which will be referred to further in this paper.

$$V_i(\pi \in \mathcal{P}) = \mathbb{E}\left[\sum_{t \in T}\left(\gamma^{t-1} g_{it}\right)\right], \quad \forall i \in P \tag{1}$$

$$M_0(\pi \in \mathcal{P}) = \mathbb{M}\Big[g_{01} + \gamma \mathbb{M}\Big[g_{02} + \ldots + \gamma \mathbb{M}\Big[g_{0|T|}\Big]\Big]\Big] = \mathbb{M}\Big[g_{0,1:|T|}\Big] \tag{2}$$

where $\mathbb{E}[.]$ and $\mathbb{M}[.]$ represents expectation and conditional (conditioned on the realized trajectory up to the current state in an MDP) spectral risk measure operators, respectively, and the corresponding two equations, Equations (1) and (2), are the cumulative expectation and dynamic risk measure functions, respectively. In both Functions (1) and (2), periodic costs $g_{it} = g_i(s_t \in \mathcal{S}, a_t \in \mathcal{A})$, $\forall i \in P, t \in T$ are used, which are discussed above in Section 2.4.

Another compact notation for the dual risk-averse objective we will use in Theorem 2 is as follows:

$$\mathbb{M}\Big[g_{0:|P|,1:|T|}\Big] = \mathbb{M}\left[g_{01} + \sum_{i \in P}(\lambda_{i1} g_{i1}) + \gamma \mathbb{M}\left[g_{02} + \sum_{i \in P}(\lambda_{i2} g_{i2}) + \ldots + \gamma \mathbb{M}\left[g_{0|T|} + \sum_{i \in P}\left(\lambda_{i|T|} g_{i|T|}\right)\right]\right]\right] \tag{3}$$

where, corresponding to every instantaneous constraint cost function g_{it}, $\forall i \in P, t \in T$ and $\lambda_{it} \in \mathbb{R} \backslash \mathbb{R}^+$, $\forall i \in P, t \in T$ are the components of the Lagrangian vector λ' for the dual of "Problem 2", referred to as "Dual 2" in Section "Strong Duality for "Problem 2"", on page 13.

Lastly, we defined $\mathcal{M}(\mathcal{S}, \mathcal{A})$ as the set of all measures defined onset of ordered pairs of states and actions $(\mathcal{S} \times \mathcal{A})$, and $\mathcal{T}$ as occupation measures' set originated from feasible policy space $\mathcal{P}$ as

$$\mathcal{T} = \left\{\rho(s \in \mathcal{S}, a \in \mathcal{A}) \in \mathcal{M}(\mathcal{S}, \mathcal{A}) | \rho(s \in \mathcal{S}, a \in \mathcal{A}) = (1-\gamma)\sum_{t \in T}\left(\gamma^{t-1} p_\pi(s_t = s \in \mathcal{S}, a_t = a \in \mathcal{A})\right)\right\} \tag{4}$$

where $p_\pi(s_t = s \in \mathcal{S}, a_t = a \in \mathcal{A})$ is given as (A4) in Appendix A.

To initiate the field of constrained risk-averse RL, we will prove the following:

"Risk-averse MDPs with dynamic spectral risk objective and expectation-based cumulative constraints have zero duality gap in generic policy space $\mathcal{P}$ (as Theorem 1, in Section "Strong Duality for "Problem 1"", on page 6). Based on this Theorem 1, risk-averse MDPs with deterministic instantaneous constraints also have zero duality gap (as Theorem 2, in Section "Strong Duality for "Problem 2"", on page 12) without any assumption on the problem's convexity."

These results imply that these two classes of problems can be solved perfectly in the domain of dual variables where the space becomes convex (in terms of dual variables).

3. Risk-Averse Markov Decision Processes with Cumulative Constraints

In the literature [17], constrained risk-averse MDP refers to the MDP where both the objective function and constraints are risk-averse. However, we will use a varied version of it in which risk constraints [17] are substituted with cumulative expectation constraint(s) (9), which is the same as the cumulative constraints in constrained MDP [34]. The mathematical program for this varied version is as follows, which we will refer to as "Problem 1":

$$\underset{\pi \in \mathcal{P}}{max} M_0(\pi \in \mathcal{P}) \tag{5}$$

Subject to the following:

$$s_{t+1} \sim p_t(\mathcal{S} \subseteq \mathbb{R}^m | s_t \in \mathcal{S} \subseteq \mathbb{R}^m, a_t \in \mathcal{A} \subseteq \mathbb{R}^m), \quad \forall t \in T \tag{6}$$

$$a_t \sim \pi \in \mathcal{P}, \quad \forall t \in T \tag{7}$$

$$s_1 \sim p_0(\mathcal{S} \subseteq \mathbb{R}^m) \tag{8}$$

$$V_i(\pi \in \mathcal{P}) \leq 0, \quad \forall i \in P \tag{9}$$

where in "Problem 1" above, (5)–(8) are nested risk-averse objective, transition function set, policy inference, and initial state being sampled from its probability distribution, respectively.

Strong Duality for "Problem 1"

Let λ be the Lagrangian vector, with its components $\lambda_i \in \mathbb{R} \backslash \mathbb{R}^+, \forall i \in P$ as Lagrangian multipliers corresponding to all the cumulative constraints (9) in "Problem 1" above. Then, the following (10) is the Lagrangian function:

$$L\left(\pi \in \mathcal{P}, \left\{\lambda_i \in \mathbb{R} \backslash \mathbb{R}^+, \forall i \in P\right\}\right) = M_0(\pi \in P) + \sum_{i \in P} (\lambda_i (V_i(\pi \in \mathcal{P}))) \tag{10}$$

With fixed Lagrangian multipliers $\lambda_i \in \mathbb{R} \backslash \mathbb{R}^+, \forall i \in P$, if the Lagrangian function (10) is maximized over the policy space $\mathcal{P}$, then it will become the dual function, which forms a constrained optimization problem with (6)–(8). The solution of it provides an upper bound (for a maximization, or lower bound for a minimization problem) on the optimal value of "Problem 1" (primal problem).

In addition to optimizing (maximizing) policy space $\mathcal{P}$, if we also optimize (minimize) it over its Lagrangian multipliers $\lambda_i \in \mathbb{R} \backslash \mathbb{R}^+, \forall i \in P$ (dual variables) as in (11) below, the resulting min–max optimization problem will be the dual problem of "Problem 1", which we will refer to later as "Dual 1". Upon its solution, we will have the tightest (best) upper bound on the optimal solution of "Problem 1" [17]:

$$D^* = \underset{\lambda_i \in \mathbb{R} \backslash \mathbb{R}^+, \, \forall i \in P}{min} \; \underset{\pi \in \mathcal{P}}{max} L\left(\pi \in \mathcal{P}, \left\{\lambda_i \in \mathbb{R} \backslash \mathbb{R}^+, \;\; \forall i \in P\right\}\right) \tag{11}$$

Subject to the following: (6)–(8).

In general, solving "Dual 1" does not result in the solution of "Problem 1" as there may not be any relation between the optimal values of both problems. However, if strong duality for "Problem 1" holds, i.e., the duality gap Δ between "Problem 1" (primal) and "Dual 1" (dual) is zero, then the optimal value of "Problem 1" P^* will be equal to that of the "Dual 1" D^*, i.e., $P^* = D^*$. In Theorem 1 below, we will prove just that (with certain assumptions on the risk-aversion function in (5) given in Theorem 1's proof)for both finite and infinite planning horizons with a discount factor $\gamma \in (0, 1)$ using perturbation theory.

We define the perturbation function associated with "Problem 1" for any value of perturbation components $\xi_i, \forall i \in P$ of the perturbation vector ξ as follows, which we will refer to as "The Perturbation function":

$$P(\xi) \triangleq \underset{\pi \in \mathcal{P}}{max} M_0(\pi \in \mathcal{P}) \tag{12}$$

Subject to the following: (6)–(8)

$$V_i(\pi \in \mathcal{P}) \leq \xi_i, \quad \forall i \in P \tag{13}$$

Proposition 1. *If we assume the Slater's condition for "Problem 1", and concavity of "The Perturbation function", then strong duality will hold for "Problem 1".*

Theorem 1. *If the costs $g_{it}, \forall i \in P \cup \{0\}, t \in T$ are bounded with the satisfaction of Slater's condition for "Problem 1", then strong duality holds for "Problem 1".*

The above theorem will be proved by connecting the strong duality of "Problem 1" with the convexity of "The Perturbation function" as follows in "Proposition 1" and further.

Proof. "Proposition 1" has been proven by [43] in Corollary 30.2.2, and as Slater's condition holds by the assumption of Theorem 1, the only thing to prove here is the concavity of "The Perturbation function", i.e., for any two perturbation vectors $\xi^1, \xi^2 \in \mathbb{R}^{|P|}$ and $\mu \in (0,1)$, the following holds

$$P\left(\mu\xi^1 + (1-\mu)\xi^2\right) \geq \mu P\left(\xi^1\right) + (1-\mu)P\left(\xi^2\right), \quad \forall \mu \in (0,1) \tag{14}$$

In the case when either perturbation $\xi^1, \xi^2 \in \mathbb{R}^{|P|}$ makes "The Perturbation function" infeasible, then either $P(\xi^1) = -\infty$ or $P(\xi^2) = -\infty$ and, consequently, "The Perturbation function" is trivially concave. However, if the problem remains feasible for these perturbations $\xi^1, \xi^2 \in \mathbb{R}^{|P|}$, then these perturbations have their respective optimal policies, $\pi_1 \in \mathcal{P}$ and $\pi_2 \in \mathcal{P}$, respectively, which can achieve values of respective perturbation functions such that $P(\xi^1) = \underset{\pi \in \mathcal{P}}{max} M_0(\pi \in \mathcal{P}) = M_0(\pi_1 \in \mathcal{P})$ is subject to $V_i(\pi_1 \in \mathcal{P}) \leq \xi_i^1, \forall i = P$ and (6)–(8) and $P(\xi^2) = \underset{\pi \in \mathcal{P}}{max} M_0(\pi \in \mathcal{P}) = M_0(\pi_2 \in \mathcal{P})$ is subject to $V_i(\pi_2 \in \mathcal{P}) \leq \xi_i^2, \forall i = P$ and (6)–(8).

Now, for proving the concavity of "The Perturbation function", it is sufficient to show that a policy $\pi_\mu \in \mathcal{P}, \forall \mu \in (0,1)$ will exist which will satisfy $V_i(\pi_\mu \in \mathcal{P}) \leq \mu\xi_i^1 + (1-\mu)\xi_i^2, \forall i \in P, \mu \in (0,1)$, and $M_0(\pi_\mu \in \mathcal{P}) = \mu M_0(\pi_1 \in \mathcal{P}) + (1-\mu)M_0(\pi_2 \in \mathcal{P})$, $\forall \mu \in (0,1)$ holds, so the policy $\pi_\mu \in \mathcal{P}, \forall \mu \in (0,1)$ is not only feasible but (15) naturally follows (which will result in the implication of (14)):

$$P\left(\mu\xi^1 + (1-\mu)\xi^2\right) \geq M_0(\pi_\mu \in \mathcal{P}) = \mu M_0(\pi_1 \in \mathcal{P}) + (1-\mu)M_0(\pi_2 \in \mathcal{P}) = \mu P\left(\xi^1\right) + (1-\mu)P\left(\xi^2\right), \forall \mu \in (0,1) \tag{15}$$

So, we will prove the existence of such a policy $\pi_\mu \in \mathcal{P}, \forall \mu \in (0,1)$ for "Problem 1" to prove its strong duality.

Expectation-Based Constraints

The manipulations of expectation-based constraints (9) we followed in this theorem are from [34]. However, due to notational differences, we have included all of these manipulations in terms of our notation in Appendix A for the readers' ease of understanding.

Risk-Averse Objective

Now, for the spectral risk-averse objective, we can rewrite $M_0(\pi \in P)$ in (5) as follows:

$$M_0(\pi \in \mathcal{P}) = \int\limits_{(\mathcal{S}\times\mathcal{A})^{|T|}} \left(\varphi\Big(F\Big(R^{ds}\Big(d^{trj} \in D^{trj}\Big)\Big)\Big) \sum_{t\in T} \Big(\gamma^{t-1} g_{0t}\Big) p_\pi\Big(d^{trj} \in D^{trj}\Big) \right) ds_1 da_1 ds_2 da_2 \dots ds_{|T|} da_{|T|} \tag{16}$$

In (16) above, $R^{ds}\big(d^{trj} \in D^{trj}\big) = \sum\limits_{t\in T}\big(\gamma^{t-1} g_{0t}\big)$, which is the mapping from a trajectory $d^{trj} \in D^{trj}$ to its discounted sum of rewards. $F(.)$ is the cumulative probability distribution function that maps the discounted sum of rewards $\sum\limits_{t\in T}\big(\gamma^{t-1} g_{0t}\big)$ (associated with a certain trajectory $d^{trj} \in D^{trj}$) to the closed interval $[0, 1]$. $\varphi(.)$ is the risk aversion function (in a spectral risk measure, which depends on the decision maker's risk-aversion) mapping the output of $F(.)$ to the interval $[0, 1]$. $\varphi(.)$ and, its single variable and cross derivatives of all orders (with respect to any state and decision variables) is assumed to exist and be piecewise C^∞ function in terms of all state and decision variables. In the end, $p_\pi\big(d^{trj} \in D^{trj}\big)$ is the probability of a particular trajectory $d^{trj} \in D^{trj}$, given a policy $\pi \in \mathcal{P}$ and $|T|$ being the cardinality of the set T (number of periods in the planning horizon).

Since the cost functions g_{0t} are bounded (by the assumption of Theorem 1), the dominance convergence theorem (DCT) will hold, allowing us to reverse the order of integration and summation in (16). With the use of conditional probabilities and Markov's property of the MDPs, we can rewrite (16) above as follows:

$$M_0(\pi \in \mathcal{P}) = \sum_{t\in T} \left(\gamma^{t-1} \int\limits_{(\mathcal{S}\times\mathcal{A})^{|T|}} \left(\varphi\Big(F\Big(R^{ds}\Big(d^{trj} \in D^{trj}\Big)\Big)\Big) g_{0t} \prod_{u\in T\setminus\{1\}} (p(s_u|s_{u-1}, a_{u-1}) p_\pi(a_u|s_u)) p_0(s_1) p_\pi(a_1|s_1) \right) ds_1 da_1 ds_2 da_2 \dots ds_{|T|} da_{|T|} \right) \tag{17}$$

In (17), $p(s_u|s_{u-1}, a_{u-1})$ represents the probability of state s_u given the previous state s_{u-1} and action a_{u-1}. $p_\pi(a_u|s_u)$ is the probability of the action a_u given s_u and $\pi \in \mathcal{P}$ are current state and policy, respectively, and $p_0(s_1)$ is the probability of the initial state s_1.

Now let

$$f_1 = \varphi\Big(F\Big(R^{ds}\Big(d^{trj} \in D^{trj}\Big)\Big)\Big) \tag{18}$$

$$f_2 = g_{0t} \prod_{u\in T\setminus\{1\}} (p(s_u|s_{u-1}, a_{u-1}) p_\pi(a_u|s_u)) p_0(s_1) p_\pi(a_1|s_1) \tag{19}$$

Then, (17) above can be written as follows:

$$M_0(\pi \in \mathcal{P}) = \sum_{t\in T} \left(\gamma^{t-1} \int\limits_{(\mathcal{S}\times\mathcal{A})^{|T|}} (f_1 f_2) ds_1 da_1 ds_2 da_2 \dots ds_{|T|} da_{|T|} \right), \ \forall i \in P \tag{20}$$

Now, using integration by part formula, we can write (20) as follows:

$$M_0(\pi \in \mathcal{P}) = \sum_{t\in T} \left(\gamma^{t-1} \int\limits_{\mathcal{S}^{|T|-1}} \int\limits_{\mathcal{A}^{|T|}} \left(f_1 \int\limits_{\mathcal{S}} f_2 ds_1 - \int\limits_{\mathcal{S}} \left(f_1' \int\limits_{\mathcal{S}} f_2 ds_1 \right) ds_1 \right) da_1 ds_2 da_2 \dots ds_{|T|} da_{|T|} \right) \tag{21}$$

By applying integration by part formula recursively on the second term each time it is being generated upon the application of integration by part formula, we can write the above expression as follows:

$$M_0(\pi \in \mathcal{P}) = \sum_{t \in T} \left(\gamma^{t-1} \int_{\mathcal{S}^{|T|-1}} \int_{\mathcal{A}^{|T|}} \left(\begin{array}{l} (-1)^{1-1} {f_1}^{(1-1)} \int_{\mathcal{S}} f_2 ds_1 + (-1)^{2-1} {f_1}^{(2-1)} \int_{\mathcal{S}^2} f_2 ds_1 ds_1 \\ +(-1)^{3-1} {f_1}^{(3-1)} \int_{\mathcal{S}^3} f_2 ds_1 ds_1 ds_1 + (-1)^{4-1} {f_1}^{(4-1)} \int_{\mathcal{S}^4} f_2 ds_1 ds_1 ds_1 ds_1 \\ +(-1)^{5-1} {f_1}^{(5-1)} \int_{\mathcal{S}^5} f_2 ds_1 ds_1 ds_1 ds_1 ds_1 + \ldots \\ +(-1)^{(n-1)} {f_1}^{(n-1)} \int_{\mathcal{S}^n} f_2 (ds_1)^n + (-1)^n \int_{\mathcal{S}} \left({f_1}^n \int_{(s_1 \in \mathcal{S})^n} f_2 (ds_1)^n \right) ds_1 \end{array} \right) da_1 ds_2 da_2 \ldots ds_{|T|} da_{|T|} \right) \quad (22)$$

If we continue the application of integration by part formula recursively infinite times, we will have the following expression if we assume that an infinite derivative exists for the risk aversion function f_1 (concerning s_1) and $\int_{\mathcal{S}} \left({f_1}^n \int_{\mathcal{S}^n} f_2 (ds_1)^n \right) ds_1 \to 0$ as $n \to \infty$ (which holds in the case of CVaR and mean-CVaR given their piecewise constant risk aversion functions f_1 (which will have zero infinite derivatives for each piecewise component after the splitting of the integration interval), while for the rest of spectral risk measures, it may or may not hold depending on the risk aversion function itself along with the cumulative probability distribution function [44]):

$$M_0(\pi \in \mathcal{P}) = \sum_{t \in T} \left(\gamma^{t-1} \int_{\mathcal{S}^{|T|-1}} \int_{\mathcal{A}^{|T|}} \left(\begin{array}{l} (-1)^{1-1} {f_1}^{(1-1)} \int_{\mathcal{S}} f_2 ds_1 + (-1)^{2-1} {f_1}^{(2-1)} \int_{\mathcal{S}^2} f_2 ds_1 ds_1 \\ +(-1)^{3-1} {f_1}^{(3-1)} \int_{\mathcal{S}^3} f_2 ds_1 ds_1 ds_1 + (-1)^{4-1} {f_1}^{(4-1)} \int_{\mathcal{S}^4} f_2 ds_1 ds_1 ds_1 ds_1 \\ +(-1)^{5-1} {f_1}^{(5-1)} \int_{\mathcal{S}^5} f_2 ds_1 ds_1 ds_1 ds_1 ds_1 + \ldots + (-1)^{(n-1)} {f_1}^{(n-1)} \int_{\mathcal{S}^n} f_2 (ds_1)^n \\ +(-1)^n {f_1}^n \int_{\mathcal{S}^{n+1}} f_2 (ds_1)^{n+1} + \ldots + (-1)^\infty {f_1}^\infty \int_{\mathcal{S}^\infty} f_2 (ds_1)^\infty \end{array} \right) da_1 ds_2 da_2 \ldots ds_{|T|} da_{|T|} \right) \quad (23)$$

As a side note, it is important to search and develop risk-aversion functions for which we can write (23) for (22), which will be the direction of our future research.

Let $c = -1$ and $h_1 = \int_{\mathcal{S}} f_2 ds_1$, $h_2 = \int_{\mathcal{S}^2} f_2 ds_1 ds_1$, $h_3 = \int_{\mathcal{S}^3} f_2 ds_1 ds_1 ds_1$, ……, $h_n = \int_{\mathcal{S}^n} f_2 (ds_1)^n$, ……, $h_\infty = \int_{\mathcal{S}^\infty} f_2 (ds_1)^\infty$.

Now, to integrate all the individual terms with respect to the first action a_1, the expression (23) can be written as follows:

$$M_0(\pi \in \mathcal{P}) = \sum_{t \in T} \left(\gamma^{t-1} \int_{(\mathcal{S} \times \mathcal{A})^{|T|-1}} \left(\begin{array}{l} c^{1-1} \int_{\mathcal{A}} \left({f_1}^{(1-1)} h_1 \right) da_1 + c^{2-1} \int_{\mathcal{A}} \left({f_1}^{(2-1)} h_2 \right) da_1 da_1 + c^{3-1} \int_{\mathcal{A}} \left({f_1}^{(3-1)} h_3 \right) da_1 \\ +c^{4-1} \int_{\mathcal{A}} \left({f_1}^{(4-1)} h_4 \right) da_1 + c^{5-1} \int_{\mathcal{A}} \left({f_1}^{(5-1)} h_5 \right) da_1 + \ldots \\ +c^{(n-1)} \int_{\mathcal{A}} \left({f_1}^{(n-1)} h_n \right) da_1 + c^n \int_{\mathcal{A}} \left({f_1}^{(n)} h_{n+1} \right) da_1 + \ldots \\ +c^\infty \int_{\mathcal{A}} \left({f_1}^{(\infty)} h_\infty \right) da_1 \end{array} \right) ds_2 da_2 \ldots ds_{|T|} da_{|T|} \right) \quad (24)$$

In the above expression, the terms $\int_{\mathcal{A}} \left({f_1}^{(1-1)} h_1 \right) da_1$, $\int_{\mathcal{A}} \left({f_1}^{(2-1)} h_2 \right) da_1$, $\int_{\mathcal{A}} \left({f_1}^{(3-1)} h_3 \right) da_1$, …, $\int_{\mathcal{A}} \left({f_1}^{(\infty)} h_\infty \right) da_1$ can be expanded individually (similar to the expansion of (20)) by recursive application of integration by part formula (infinite times) given the derivatives of all orders of function f_1 are well-defined (with respect to a_1) and $\int_{\mathcal{A}} \left({f_1}^{(p-1)(m)} \int_{\mathcal{A}^m} h_p (da_1)^m \right) da_1 \to 0, \forall p \in \{1, 2, \ldots, \infty\}$ as $m \to \infty$, where, in ${f_1}^{(p-1)(m)}$, $(p-1)$ is the order of f_1 derivative with respect to s_1, and m is the order of f_1 derivative with respect to a_1.

After the expansion of these terms, we will have similar terms (as the individual terms present in (23)). The integrals (with respect to other states and actions, e.g., $s_2, a_2, s_3, a_3 \ldots$)

of these resulting terms can be expanded similarly by using the mechanism of recursive by-part integration as used above with s_1 and a_1.

Let

$$f_3 = \int\limits_{(\mathcal{S}\times\mathcal{A})^{|T|}} \left(g_{0t} \prod_{u\in T\setminus\{1\}} (p(s_u|s_{u-1},a_{u-1})p_\pi(a_u|s_u)p_0(s_1)p_\pi(a_1|s_1)) \right) ds_1 da_1 ds_2 da_2 \dots ds_{|T|} da_{|T|} \tag{25}$$

After recursive integration (with the application of by-part integration formula recursively infinite times) with respect to states and actions in all the time periods $s_1, a_1, s_2, a_2, s_3, a_3, \dots, s_{|T|}, a_{|T|}$ and given the fact that all the resulting multiple integral terms will always have an integral with respect to each state and action $s_1, a_1, s_2, a_2, s_3, a_3, \dots, s_{|T|}, a_{|T|}$, we will have the expression for $M_0(\pi \in \mathcal{P})$ in the following form:

$$M_0(\pi\in\mathcal{P}) = \sum_{t\in T}\left(\gamma^{t-1}\left(\begin{array}{l} c_1 \frac{dd\dots ddd\dots df_1}{ds_1 ds_2\dots ds_{|T|} da_1 da_2\dots da_{|T|}} \left(\int\limits_{\mathcal{A}}\cdots\int\limits_{\mathcal{A}}\int\limits_{\mathcal{S}}\cdots\int\limits_{\mathcal{S}} f_3 ds_1\dots ds_1 da_1\dots da_1 ds_2\dots ds_2 da_2\dots da_2\dots ds_{|T|}\dots ds_{|T|} da_{|T|}\dots da_{|T|}\right) \\ +c_2 \frac{dd\dots ddd\dots df_1}{ds_1 ds_2\dots ds_{|T|} da_1 da_2\dots da_{|T|}} \left(\int\limits_{\mathcal{A}}\cdots\int\limits_{\mathcal{A}}\int\limits_{\mathcal{S}}\cdots\int\limits_{\mathcal{S}} f_3 ds_1\dots ds_1 da_1\dots da_1 ds_2\dots ds_2 da_2\dots da_2\dots ds_{|T|}\dots ds_{|T|} da_{|T|}\dots da_{|T|}\right)+\dots \\ +c_n \frac{dd\dots ddd\dots df_1}{ds_1 ds_2\dots ds_{|T|} da_1 da_2\dots da_{|T|}} \left(\int\limits_{\mathcal{A}}\cdots\int\limits_{\mathcal{A}}\int\limits_{\mathcal{S}}\cdots\int\limits_{\mathcal{S}} f_3 ds_1\dots ds_1 da_1\dots da_1 ds_2\dots ds_2 da_2\dots da_2\dots ds_{|T|}\dots ds_{|T|} da_{|T|}\dots da_{|T|}\right) \end{array}\right)\right) \tag{26}$$

Note that, in (26), $\frac{dd\dots ddd\dots df_1}{ds_1 ds_2\dots ds_{|T|} da_1 da_2\dots da_{|T|}}$ (in each term after $c_1, c_2, \dots, c_n$) are different with respect to each other as they may have derivatives of different orders with respect to various states and actions. Here, the multiple integral terms in (26) have one integral (with respect to all state and action variables $s_1, a_1, s_2, a_2, s_3, a_3, \dots, s_{|T|}, a_{|T|}$) less than that of the expression resulting after the application of the by-part integration formula recursively infinite times on $(f_1 f_2)$ in (20) since one integral (with respect to each state and action variable) will be consumed by the expression of f_3(25). As we have performed while moving from (A2) to (A3) in Appendix A, the expression (25) can be reduced to the following:

$$f_3 = \int\limits_{(\mathcal{S}\times\mathcal{A})^{t}} \left(g_{0t} \prod_{u\in\{x\in T|x>1, x\le t\}} (p(s_u|s_{u-1},a_{u-1})p_\pi(a_u|s_u)p_0(s_1)p_\pi(a_1|s_1)) \right) ds_1\dots ds_t da_1\dots da_t \tag{27}$$

Let

$$\begin{aligned} f_4 &= \int\limits_{\mathcal{S}\times\mathcal{A}} \left(g_{0t} \sum_{t\in T} \left(\gamma^{t-1} p_\pi(s_t = s\in\mathcal{S}, a_t = a\in\mathcal{A})\right)\right) dsda = \\ &\frac{1}{(1-\gamma)} \int\limits_{\mathcal{S}\times\mathcal{A}} \left(g_{0t}(1-\gamma) \sum_{t\in T} \left(\gamma^{t-1} p_\pi(s_t = s\in\mathcal{S}, a_t = a\in\mathcal{A})\right)\right) dsda \end{aligned} \tag{28}$$

From DCT and (A4) in Appendix A, we can write the following form of (26) above (similar to (A5) in Appendix A) by interchanging positions of integrals and summation.

$$\begin{array}{rl} M_0(\pi\in\mathcal{P}) & = c_1 \frac{dd\dots ddd\dots df_1}{ds_1 ds_2\dots ds_{|T|} da_1 da_2\dots da_{|T|}} \left(\int\limits_{\mathcal{A}}\cdots\int\limits_{\mathcal{A}}\int\limits_{\mathcal{S}}\cdots\int\limits_{\mathcal{S}} f_4 ds_1\dots ds_1 da_1\dots da_1 ds_2\dots ds_2 da_2\dots da_2\dots ds_{|T|}\dots ds_{|T|} da_{|T|}\dots da_{|T|}\right) \\ & +c_2 \frac{dd\dots ddd\dots df_1}{ds_1 ds_2\dots ds_{|T|} da_1 da_2\dots da_{|T|}} \left(\int\limits_{\mathcal{A}}\cdots\int\limits_{\mathcal{A}}\int\limits_{\mathcal{S}}\cdots\int\limits_{\mathcal{S}} f_4 ds_1\dots ds_1 da_1\dots da_1 ds_2\dots ds_2 da_2\dots da_2\dots ds_{|T|}\dots ds_{|T|} da_{|T|}\dots da_{|T|}\right)+\dots \\ & +c_n \frac{dd\dots ddd\dots df_1}{ds_1 ds_2\dots ds_{|T|} da_1 da_2\dots da_{|T|}} \left(\int\limits_{\mathcal{A}}\cdots\int\limits_{\mathcal{A}}\int\limits_{\mathcal{S}}\cdots\int\limits_{\mathcal{S}} f_4 ds_1\dots ds_1 da_1\dots da_1 ds_2\dots ds_2 da_2\dots da_2\dots ds_{|T|}\dots ds_{|T|} da_{|T|}\dots da_{|T|}\right) \end{array} \tag{29}$$

Let f_5 and the Equation of occupancy measure (A6) in Appendix A above, we can write the following:

$$f_5 = \int\limits_{\mathcal{S}\times\mathcal{A}} \left(g_{0t}(1-\gamma) \sum_{t\in T} \left(\gamma^{t-1} p_\pi(s_t = s\in\mathcal{S}, a_t = a\in\mathcal{A})\right)\right) dsda = \int\limits_{\mathcal{S}\times\mathcal{A}} (g_{0t}\rho(s\in\mathcal{S}, a\in\mathcal{A}))dsda, \forall i\in P \tag{30}$$

Then, we can write (29) naturally as follows, which is the objective of the "The Perturbation function":

$$\begin{aligned} M_0(\pi \in \mathcal{P}) = {} & c_1 \frac{1}{1-\gamma} \frac{dd \dots ddd \dots df_1}{ds_1 ds_2 \dots ds_{|T|} da_1 da_2 \dots da_{|T|}} \left(\int_{\mathcal{A}} \cdots \int_{\mathcal{A}} \int_{\mathcal{S}} \cdots \int_{\mathcal{S}} f_5 ds_1 \dots ds_1 da_1 \dots da_1 ds_2 \dots ds_2 da_2 \dots da_2 \dots ds_{|T|} \dots ds_{|T|} da_{|T|} \dots da_{|T|} \right) \\ & + c_2 \frac{1}{1-\gamma} \frac{dd \dots ddd \dots df_1}{ds_1 ds_2 \dots ds_{|T|} da_1 da_2 \dots da_{|T|}} \left(\int_{\mathcal{A}} \cdots \int_{\mathcal{A}} \int_{\mathcal{S}} \cdots \int_{\mathcal{S}} f_5 ds_1 \dots ds_1 da_1 \dots da_1 ds_2 \dots ds_2 da_2 \dots da_2 \dots ds_{|T|} \dots ds_{|T|} da_{|T|} \dots da_{|T|} \right) + \dots \\ & + c_n \frac{1}{1-\gamma} \frac{dd \dots ddd \dots df_1}{ds_1 ds_2 \dots ds_{|T|} da_1 da_2 \dots da_{|T|}} \left(\int_{\mathcal{A}} \cdots \int_{\mathcal{A}} \int_{\mathcal{S}} \cdots \int_{\mathcal{S}} f_5 ds_1 \dots ds_1 da_1 \dots da_1 ds_2 \dots ds_2 da_2 \dots da_2 \dots ds_{|T|} \dots ds_{|T|} da_{|T|} \dots da_{|T|} \right) \end{aligned} \tag{31}$$

Concavity of the Perturbation Function

From (A7) in Appendix A and (31), "The Perturbation function" itself can be written in its mathematical program below (given that (31) is to be optimized over the set of feasible occupation measures $\mathcal{T}$):

$$P(\xi) \triangleq \max_{\pi \in \mathcal{P}} M_0(\pi \in \mathcal{P}) = \max_{\rho(.) \in \mathcal{T}} M_0(\pi \in \mathcal{P}) \tag{32}$$

Subject to: (6)–(8)

$$\frac{1}{1-\gamma} \int_{\mathcal{S} \times \mathcal{A}} (g_{it} \rho(s \in \mathcal{S}, a \in \mathcal{A})) ds da \leq \xi_i \quad \forall i \in P \tag{33}$$

With the only consideration being the risk-averse objective, the occupation measure set $\mathcal{T}$ (in (4)) being convex and compact (compact means closed and bounded) is directly followed by Theorem 3.1 in [45], which makes us able to write the following:

$$\rho_\mu(s \in \mathcal{S}, a \in \mathcal{A}) = \mu \rho_1(s \in \mathcal{S}, a \in \mathcal{A}) + (1-\mu) \rho_2(s \in S, a \in \mathcal{A}), \forall \mu \in (0,1) \tag{34}$$

Now, for the feasible two policies, $\pi_1 \in \mathcal{P}$ and $\pi_2 \in \mathcal{P}$, which are optimal for the perturbation vectors ξ^1 and ξ^2, respectively, considered above, let their corresponding two occupation measures be $\rho_{'1}(s \in \mathcal{S}, a \in \mathcal{A})$ and $\rho_{'2}(s \in \mathcal{S}, a \in \mathcal{A})$, respectively. Then, from the convexity of the set $\mathcal{T}$, there must be a feasible policy $\pi_\mu \in \mathcal{P}$ whose occupation measure is given by $\rho_\mu(s,a)$. From (A7) in Appendix A, linearity of constraint integrals and constraint satisfaction by $\rho_1(s \in \mathcal{S}, a \in \mathcal{A})$ and $\rho_2(s \in \mathcal{S}, a \in \mathcal{A})$ with slacks $\xi_i^1, \forall i \in P$ and $\xi_i^2, \forall i \in P$, respectively, $\rho_\mu(s \in \mathcal{S}, a \in \mathcal{A})$ will satisfy the constraints with slack $\mu \xi_i^1 + (1-\mu)\xi_i^2, \forall i \in P, \mu \in (0,1)$.

From (A7) in Appendix A and (34), we can write the following:

$$\begin{aligned} \int_{\mathcal{S} \times \mathcal{A}} \big(g_{it} \rho_\mu(s \in \mathcal{S}, a \in \mathcal{A})\big) ds da &= \int_{\mathcal{S} \times \mathcal{A}} (g_{it}(\mu \rho_1(s \in \mathcal{S}, a \in \mathcal{A}) + (1-\mu)\rho_2(s \in \mathcal{S}, a \in \mathcal{A}))) ds da \\ &= \mu \int_{\mathcal{S} \times \mathcal{A}} (g_{it} \rho_1(s \in \mathcal{S}, a \in \mathcal{A})) ds da + (1-\mu) \int_{\mathcal{S} \times \mathcal{A}} (g_{it} \rho_2(s \in \mathcal{S}, a \in \mathcal{A})) ds da, \forall i \in P \end{aligned} \tag{35}$$

For the risk-averse objective, let $f_{\xi^1} = \int_{\mathcal{S} \times \mathcal{A}} (g_{0t} \rho_1(s \in \mathcal{S}, a \in \mathcal{A})) ds da$ and $f_{\xi^2} = \int_{\mathcal{S} \times \mathcal{A}} (g_{0t} \rho_2(s \in \mathcal{S}, a \in \mathcal{A})) ds da$. Also, consider $f_{\xi^\mu} = \int_{\mathcal{S} \times \mathcal{A}} \big(g_{0t} \rho_\mu(s \in \mathcal{S}, a \in \mathcal{A})\big) ds da$. From (34) above, $f_{\xi^\mu} = \mu f_{\xi^1} + (1-\mu) f_{\xi^2}$, and we can write, for "The Perturbation Function" objective, $P\big(\mu \xi^1 + (1-\mu)\xi^2\big)$, corresponding to the perturbation vector $\mu \xi^1 + (1-\mu)\xi^2$ as the following:

$$
\begin{aligned}
&P(\mu\xi^1+(1-\mu)\xi^2)\\
&\geq \begin{pmatrix} c_1\frac{1}{1-\gamma}\frac{dd\ldots ddd\ldots df_1}{ds_1ds_2\ldots ds_{|T|}da_1da_2\ldots da_{|T|}}\left(\int_{\mathcal{A}}\cdots\int_{\mathcal{A}}\int_{\mathcal{S}}\cdots\int_{\mathcal{S}} f_{\xi^\mu}ds_1\ldots ds_1da_1\ldots da_1ds_2\ldots ds_2da_2\ldots da_2\ldots ds_{|T|}\ldots ds_{|T|}da_{|T|}\ldots da_{|T|}\right)\\ +c_2\frac{1}{1-\gamma}\frac{dd\ldots ddd\ldots df_1}{ds_1ds_2\ldots ds_{|T|}da_1da_2\ldots da_{|T|}}\left(\int_{\mathcal{A}}\cdots\int_{\mathcal{A}}\int_{\mathcal{S}}\cdots\int_{\mathcal{S}} f_{\xi^\mu}ds_1\ldots ds_1da_1\ldots da_1ds_2\ldots ds_2da_2\ldots da_2\ldots ds_{|T|}\ldots ds_{|T|}da_{|T|}\ldots da_{|T|}\right)+\ldots\\ +c_n\frac{1}{1-\gamma}\frac{dd\ldots ddd\ldots df_1}{ds_1ds_2\ldots ds_{|T|}da_1da_2\ldots da_{|T|}}\left(\int_{\mathcal{A}}\cdots\int_{\mathcal{A}}\int_{\mathcal{S}}\cdots\int_{\mathcal{S}} f_{\xi^\mu}ds_1\ldots ds_1da_1\ldots da_1ds_2\ldots ds_2da_2\ldots da_2\ldots ds_{|T|}\ldots ds_{|T|}da_{|T|}\ldots da_{|T|}\right)\end{pmatrix}\\
&= \begin{pmatrix} c_1\frac{1}{1-\gamma}\frac{dd\ldots ddd\ldots df_1}{ds_1ds_2\ldots ds_{|T|}da_1da_2\ldots da_{|T|}}\left(\int_{\mathcal{A}}\cdots\int_{\mathcal{A}}\int_{\mathcal{S}}\cdots\int_{\mathcal{S}}\left(\mu f_{\xi^1}+(1-\mu)f_{\xi^2}\right)ds_1\ldots ds_1da_1\ldots da_1\ldots ds_{|T|}\ldots ds_{|T|}da_{|T|}\ldots da_{|T|}\right)\\ +c_2\frac{1}{1-\gamma}\frac{dd\ldots ddd\ldots df_1}{ds_1ds_2\ldots ds_{|T|}da_1da_2\ldots da_{|T|}}\left(\int_{\mathcal{A}}\cdots\int_{\mathcal{A}}\int_{\mathcal{S}}\cdots\int_{\mathcal{S}}\left(\mu f_{\xi^1}+(1-\mu)f_{\xi^2}\right)ds_1\ldots ds_1da_1\ldots da_1\ldots ds_{|T|}\ldots ds_{|T|}da_{|T|}\ldots da_{|T|}\right)+\ldots\\ +c_n\frac{1}{1-\gamma}\frac{dd\ldots ddd\ldots df_1}{ds_1ds_2\ldots ds_{|T|}da_1da_2\ldots da_{|T|}}\left(\int_{\mathcal{A}}\cdots\int_{\mathcal{A}}\int_{\mathcal{S}}\cdots\int_{\mathcal{S}}\left(\mu f_{\xi^1}+(1-\mu)f_{\xi^2}\right)ds_1\ldots ds_1da_1\ldots da_1\ldots ds_{|T|}\ldots ds_{|T|}da_{|T|}\ldots da_{|T|}\right)\end{pmatrix}
\end{aligned}
\tag{36}
$$

From (31) and the linearity of the integrals, (36) above can be written as follows:

$$P\left(\mu\xi^1+(1-\mu)\xi^2\right)\geq \mu M_0(\pi_1\in\mathcal{P})+(1-\mu)M_0(\pi_2\in\mathcal{P}) \tag{37}$$

Since $M_0(\pi_1\in\mathcal{P})=P\left(\xi^1\right)$ and $M_0(\pi_2\in\mathcal{P})=P\left(\xi^2\right)$, the following holds

$$P\left(\mu^1\xi^1+(1-\mu)\xi^2\right)\geq \mu P\left(\xi^1\right)+(1-\mu)P\left(\xi^2\right),\forall\mu\in(0,1) \tag{38}$$

which means that "The Perturbation function" is concave, which then implies, from "Proposition 1", that strong duality holds for "Problem 1". □

4. Risk-Averse Markov Decision Processes with Instantaneous Constraints

Derived from "Problem 1", the problem class that we will consider in this section consists of deterministic instantaneous constraints (40) instead of cumulative expectation constraints (9). This problem will be referred to as "Problem 2", and its mathematical program is given as follows:

$$\max_{\pi\in\mathcal{P}} M_0(\pi\in\mathcal{P}) \tag{39}$$

Subject to the following: (6)–(8)

$$\gamma^{t-1}g_{it}\leq 0,\quad \forall i\in P, t\in T \tag{40}$$

The assumption of deterministic constraints is justified in many real-world contexts. Not only do risk constraints [5] make the problem complex, but they are unnecessary in many real-world problems due to the very reason for the constraint utilization itself (in the presence of risk-averseness), which is to completely eliminate extreme realizations of certain stochastic variables regardless of risk-averseness or decision makers' behavioral aspects. In addition, if the robustification of feasible solution space is required using box uncertainty sets, it will make the constraints deterministic in a robust counterpart program.

Strong Duality for "Problem 2"

Now based on Theorem 1 in Section 3, we will prove two theorems (Theorems 2 and 3) for proving strong duality for the "Problem 2" above. Based on these theorems, we will develop a constraint handling mechanism in Section 5 for solving "Problem 2", where we will also perform a theoretical convergence analysis of this algorithm in Section 5.5.

In this regard, we will first assume the following, which we will refer to as "Assumption 1":

Assumption 1. *The feasible policy space $\mathcal{P}$ of "Problem 2" has a non-empty relative interior (Slater's condition holds), and if $\sum_{t\in T}\left(\gamma^{t-1}g_{it}\right)\leq 0,\ \forall i\in P$ holds for any feasible policy $\pi\in\mathcal{P}$, then $\gamma^{t-1}g_{it}\leq 0,\ \forall i\in P, t=T$ also holds for that policy.*

Now, on the basis of our above assumption, we will prove the following:

Theorem 2. *Under Assumption 1, strong duality holds for "Problem 2".*

Proof. Naturally, for any policy $\pi \in \mathcal{P}$, the condition $\gamma^{t-1} g_{it} \leq 0,\ \forall i \in P, t = T$ implies condition $\sum_{t \in T} \left(\gamma^{t-1} g_i(s_t, a_t)\right) \leq 0,\ \forall i \in P$, and from Assumption 1, $\sum_{t \in T} \left(\gamma^{t-1} g_{it}\right) \leq 0,\ \forall i \in P$ implies $\gamma^{t-1} g_{it} \leq 0,\ \forall i \in P, t = T$, which means the following proposition holds

"$\sum_{t \in T} \left(\gamma^{t-1} g_{it}\right) \leq 0,\ \forall i \in P$ if and only if $g_{it} \leq 0,\ \forall i \in P, t = T$".

Let λ' be the Lagrangian vector, with its components $\lambda_{it} \in \mathbb{R} \backslash \mathbb{R}^+, \forall i \in P, t \in T$, then the dual of "Problem 2", which we will refer to as "Dual 2", is as follows:

$$\min_{\lambda_{it} \in \mathbb{R} \backslash \mathbb{R}^+,\ \forall i \in P, t \in T} \max_{\pi \in \mathcal{P}} \left(M_0(\pi \in \mathcal{P}) + \sum_{t \in T} \sum_{i \in P} \left(\gamma^{t-1} \lambda_{it} g_{it} \right) \right) \tag{41}$$

Subject to the following: (6)–(8)

It is important to note here that λ (with its components $\lambda_i,\ \forall i \in P$) is the Lagrangian vector for "Problem 1" while λ' is the Lagrangian vector for the "Problem 2" with its components $\lambda_{it} \in \mathbb{R} \backslash \mathbb{R}^+, \forall i \in P, t \in T$.

As Theorem 1 implies

$$\begin{aligned} P^* = D^* = \min_{\lambda_i \in \mathbb{R} \backslash \mathbb{R}^+,\ \forall i \in P} \max_{\pi \in \mathcal{P}} L(\pi, \lambda) = \\ \min_{\lambda_i \in \mathbb{R} \backslash \mathbb{R}^+,\ \forall i \in P} \max_{\pi \in \mathcal{P}} \left(M_0(\pi \in \mathcal{P}) + \sum_{i \in \mathcal{P}} \left(\lambda_i (V_i(\pi \in P)) \right) \right), \quad \forall \xi \end{aligned} \tag{42}$$

If we consider all the constraint cost functions $g_{it},\ \forall i \in P, t \in T$ in (9) as deterministic and the risk measure $\mathbb{M}$ in (5) satisfies translation invariance (a property of coherent risk measures [46]), we can write (42) above as follows (referring to (3)), which can also be regarded as the min–max objective in "Dual 1" in (11):

$$P^* = \min_{\lambda_{it} \in \mathbb{R} \backslash \mathbb{R}^+,\ \forall i \in P, t \in T} \max_{\pi \in \mathcal{P}(\mathcal{S})} \left(\mathbb{M} \left[g_{0:|P|,1:|T|} \right] \right) \tag{43}$$

As a side note, the reason for making the above form of the dual objective of "Problem 1" is because, during RL training (for optimization of "Problem 2"), our model-free decision-making agent (heuristic) will obtain $g_{0t} + \sum_{i \in P} (\lambda_{it} g_{it}),\ \forall t \in T$ as values of reward signals from the MDP environment (in each time period) and from that value, the RL agent has to stay within the feasible solution region (constraint satisfaction) in addition to maximizing the objective function.

Let $\pi^* \in \mathcal{P}$ be the optimal policy for "Problem 2", then given Lagrangian multipliers $\lambda_{it} \in \mathbb{R} \backslash \mathbb{R}^+,\ \forall i \in P, t \in T$ of its dual, we will write the following (by construction given the translation invariance of the dynamic risk measure $\mathbb{M}[.]$):

$$\begin{aligned} \max_{\pi \in \mathcal{P}} \mathbb{M} \left[g_{0:|P|,1:|T|} \right] \geq \left(\mathbb{M} \left[g_{0:|P|,1:|T|} \middle| \pi^* \right] \right) \geq \\ \left(\mathbb{M} \left[\gamma^{t-1} \sum_{t \in T} g_{it} \middle| s_0, \pi^* \right] \right) + \sum_{t \in T} \sum_{i \in P} \left(\gamma^{t-1} \lambda_{it} (g_{it} | \pi^*) \right) \geq P^* \end{aligned} \tag{44}$$

From the above, we can make the following implication:

$$\min_{\lambda_{it} \in \mathbb{R} \backslash \mathbb{R}^+,\ \forall i \in P, t \in T} \max_{\pi \in \mathcal{P}} \left(\mathbb{M} \left[g_{0:|P|,1:|T|} \right] \right) \geq P^* \tag{45}$$

Suppose λ_{it}^*, $\forall i \in P, t \in T$ are the optimal Lagrangian multipliers in"Dual 2", then we can write as follows given Theorem 1:

$$\max_{\pi \in \mathcal{P}} \left(\mathbb{M}\left[g_{0:|P|,1:|T|} \middle| \lambda_{it}^*, \ \forall i \in P, t \in T \right] \right) = P^* \tag{46}$$

As a result, the following directly follows from the above:

$$\min_{\lambda_{it} \in \mathbb{R} \backslash \mathbb{R}^+, \ \forall i \in P, t \in T} \max_{\pi \in \mathcal{P}} \left(\mathbb{M}\left[g_{0:|P|,1:|T|} \right] \right) \leq P^* \tag{47}$$

Hence, from (45) and (47), the strong duality condition is satisfied for "Problem 2".□

Due to time-varying Lagrangian multipliers $\lambda_{it} \in \mathbb{R} \backslash \mathbb{R}^+$, $\forall i \in P, t \in T$, with the increase in the number of time periods, the number of Lagrangian multipliers will grow, and, because this, memory complexity may become the point of concern. Also, for infinite time horizon problems, time-varying Lagrangian multipliers become infeasible. To counter this effect, we will prove Theorem 3 below.

Note: It is worth noting that the time invariance of Lagrangian multipliers (in the context of Problem 2) directly follows from the fact that we have used $\lambda_{it} = \lambda_i$, $\forall t \in T$ while proving Theorem 2 from Theorem 1 (on page 15) when we moved from (42) to (43). However, if we start from the statement of Theorem 2, we need to prove the time invariance of Lagrangian multipliers $\lambda_{it} \in \mathbb{R} \backslash \mathbb{R}^+$, $\forall i \in P, t \in T$ in "Dual 2", which we will perform in Theorem 3 below (which is a redundant theoretical result).

Theorem 3. *Suppose* $\left(\left\{ \lambda_{it}^*, \forall i \in P, t \in T \right\}, \pi^* \in \mathcal{P} \right)$ *to be the optimal solution for "Dual 2" and suppose* $\left(\left\{ \lambda_i^* \in \mathbb{R} \backslash \mathbb{R}^+, \forall i \in P \right\}, \overline{\pi}^* \in \mathcal{P} \right)$ *to be the optimal solution for the following Dual mathematical program having instantaneous constraint cost functions* $g_{it}, \forall i \in P, t \in T$ *and time-invariant Lagrangian multipliers* $\lambda_i \in \mathbb{R} \backslash \mathbb{R}^+, \forall i \in P$*:*

$$\min_{\lambda_i \in \mathbb{R} \backslash \mathbb{R}^+, \ \forall i \in P} \max_{\pi \in \mathcal{P}} \left(M_0(\pi \in P) + \sum_{t \in T} \left(\gamma^{t-1} \sum_{i \in P} \left(\lambda_i \gamma^{t-1} g_{it} \right) \right) \right) \tag{48}$$

Subject to the following: (6)–(8)

Now, suppose that in "Dual 2", $\lambda_{it} = \lambda_i^*, \forall i \in P, t \in T$, under Assumption 1, $\left(\left\{ \lambda_{it} = \lambda_i^*, \forall i \in P, t \in T \right\}, \pi^* \in \mathcal{P} \right)$ is also optimal for "Problem 2" or "Dual 2".

Proof. By definition (under Assumption 1), for policy $\pi \in \mathcal{P}$, satisfying instantaneous constraints $\gamma^{t-1} g_{it} \leq 0$, $\forall i \in P, t = T$ implies that $\pi \in \mathcal{P}$ will also satisfy cumulative constraints $\sum_{t \in T} \left(\gamma^{t-1} g_{it} \right) \leq 0$, $\forall i \in P$.

Then, by assuming under closed loop dynamics, $s_{t+1} \sim p_z(.|s_t, a_t), \forall t \in T$ and $a_t \sim \pi(s_t), \forall t \in T$, that $\sum_{t \in T} \left(\gamma^{t-1} g_{it} \right) \leq 0$, $\forall i \in P$ implies $\gamma^{t-1} g_{it} \leq 0$, $\forall i \in P, t = T$ (under Assumption 1), and"Problem 1" shares the same feasible domain with "Dual 2" given the dynamic risk measure's translation invariance property on the objective (39) and the fact that constraint cost functions $g_{it}, \forall i \in P, t \in T$ are deterministic. From Theorem 1, strong duality will hold for the problem given above (in the statement of Theorem 3), which then implies with Theorem 2 that the optimal solution for the above-motioned problem, which has time-invariant Lagrangian multipliers, is also the optimal solution for "Dual 2". □

Note: It is important to mention that there can be optimal solutions $(\{\lambda_{it} \in \mathbb{R} \backslash \mathbb{R}^+, \forall i \in P, t \in T\}, \pi^* \in \mathcal{P})$ for "Dual 2", which has time variant Lagrangian multipliers in addition to the optimal solution $\left(\lambda_i^* \in \mathbb{R} \backslash \mathbb{R}^+, \overline{\pi}^* \in \mathcal{P} \right)$, which has time-invariant Lagrangian multipliers as given by Theorem 3 above.

5. Augmented Lagrangian-Based Constraint Handling Mechanism

In this section, based on Theorems 1 and 2, we will develop a risk-averse RL constraint-handling algorithm (Algorithm 1) for optimizing "Problem 2" (risk-averse MDP with instantaneous deterministic constraints). In the end, we will also theoretically prove its convergence. However, to take advantage of Algorithm 1, "Assumption 1" must be satisfied during its execution because Theorem 2 is based on the assumption. While the assumption can be automatically satisfied in the case of some problems, we ensure its satisfaction during the execution of Algorithm 1 by utilizing a so-called "Clipping method" proposed by [47]. This method is described as follows.

5.1. Clipping Method

In this method, values of constraints for the states where constraints are not violated are reduced to zero, i.e., if constraints are satisfied, then we will consider the constraint violation penalty as zero. For an illustration of the technique, consider a set of instantaneous constraints $\gamma^{t-1} g_{it} \leq 0,\ \forall i \in P, t \in T$ as (40) in "Problem 2". Then, to form a reward signal, $\gamma^{t-1} g_{it}, \forall i \in P, t = T$ will only be added as a constraint violation penalty to the reward value $\gamma^{t-1} g_{0t},\ \forall t \in T$ if it is greater than zero; otherwise, it will not be added or considered zero. This can be performed by using functions or mappings which map $\mathbb{R}^+$ to $\mathbb{R}^+$ while mapping $\mathbb{R}^-$ to zero, e.g., $ReLU\left(\gamma^{t-1} g_{0t}\right), \forall i \in P, t = T$ (in the case of which $ReLU(x) = x,\ \forall x \in \mathbb{R}^+ \cup \{0\}$; otherwise, $ReLU(x) = 0,\ \forall x \in \mathbb{R}^-$). Other non-negative activation functions like SoftPlus and Segmoid can also be used [47].

5.2. Surrogate Objective

As RL is mostly used for non-convex multi-stage optimization, in order to handle constraints in this setting, it is better to use augmented Lagrangian instead of the Lagrangian formulation [48], which we will use in our constraint-handling algorithm by first designing a surrogate objective and then a surrogate reward function, from the augmented Lagrangian formulation of "Problem 2".

Let $\lambda_i \in \mathbb{R} \backslash \mathbb{R}^+,\ \forall i \in P$ be the Lagrangian multipliers, which are invariant to time, and $\zeta_i \in \mathbb{R}^+, \forall i \in P$ be the time-invariant quadratic penalty coefficients corresponding to each of the instantaneous constraints (40) in "Problem 2". Based on the strong duality result for "Problem 2" (Theorem 2) and the Lagrangian-based objective (41) of "Dual 2" (with time-invariant Lagrangian multipliers), the augmented Lagrangian-based surrogate objective for "Problem 2" with clipping method (which will be directly optimized by an RL algorithm) is as follows:

$$S(\pi \in \mathcal{P}, \{\lambda_i,\ \forall i \in P\}, \{\zeta_i, \forall i \in P\}) = \mathbb{M}\left[\sum_{t \in T}\left(\gamma^{t-1} g_{0t} + \gamma^{t-1}\sum_{i \in P}(\lambda_i ReLU(g_{it})) - \left(\gamma^{t-1}\right)^2 \sum_{i \in P}\left(\frac{\zeta_i}{2} ReLU(g_{it})^2\right)\right)\right] \quad (49)$$

5.3. Reward Function

Correspondingly, the instantaneous reward function (stage-wise objective component) is given as follows, where the first term is the discounted reward function while the second and third terms are the discounted linear and quadratic penalty terms for constraint violations, respectively, which will discourage the agent from violating hard constraints.

$$s_t(s_t \in \mathcal{S}, a_t \in \mathcal{A}, s_{t+1} \in \mathcal{S}) = \gamma^{t-1} g_{0t} + \gamma^{t-1}\sum_{i \in P}(\lambda_i ReLU(g_{it})) - \left(\gamma^{t-1}\right)^2 \sum_{i \in P}\left(\frac{\zeta_i}{2} ReLU(g_{it})^2\right) \quad (50)$$

In the above Equations (49) and (50), the ReLU function is used for the implementation of the clipping method described above to realize Assumption 1 during the execution of the risk-averse RL algorithm.

5.4. Developed Mechanism

Based on the surrogate reward function (50) above and [47], we have designed "Algorithm 1" for the solution of "Problem 2" in the presence of only deterministic instantaneous constraints as follows.

Algorithm 1: Augmented Lagrangian-based constrained risk-averse RL

1. Take $u_\zeta \in [1, \infty)$ and dual assent step size $\Lambda \in \mathbb{R}^+$, and initialize $\lambda_i^0 \in \mathbb{R} \backslash \mathbb{R}^+, \forall i \in P$ and $\zeta_i^0 \in \mathbb{R}^+, \forall i \in P$
2. Initialize RL policy $\pi^0 \in \mathcal{P}$
3. **For** $e = 0, 1, \ldots, E$ (where, E defines the number of iterations in which Lagrangian multipliers and quadratic penalty coefficients need to be updated)
4. $\quad \pi^e = \underset{\pi^e \in \mathcal{P}}{arg\,max} S(\pi \in \mathcal{P}, \{\lambda_i \in \mathbb{R} \backslash \mathbb{R}^+, \forall i \in P\}, \{\zeta_i \in \mathbb{R}^+, \forall i \in P\})$
5. $\quad \lambda_i^{e+1} \leftarrow f_{step}\left(\lambda_i^e - \Lambda\left(\sum_{t \in T}\left(\gamma^{t-1} Relu(g_{it})\right)\middle|\pi_e\right)\right)$
   $\quad \zeta^{e+1} \leftarrow u_\zeta \zeta^e$
6. **End**

In the above algorithm, $\lambda_i^{e+1} \leftarrow f_{step}\left(\lambda_i^e - \Lambda\left(\sum_{t \in T}\left(\gamma^{t-1} Relu(g_{it})\right)|\pi_e\right)\right)$ is the mapping $f_{step}(.) : \mathbb{R} \rightarrow \mathbb{R}$ such that $f_{step}(x) = 0$ if $x \geq 0$ and $f_{step}(x) = x$ if $x < 0$. $\pi^e = \underset{\pi^e \in \mathcal{P}}{arg\,max} S(\pi \in \mathcal{P}, \{\lambda_i \in \mathbb{R} \backslash \mathbb{R}^+, \forall i \in P\}, \{\zeta_i \in \mathbb{R}^+, \forall i \in P\})$ is the optimization of unconstrained surrogate objective function for "Problem 2", which will be solved by the proposed risk-averse RL. Then, we will update Lagrangian multipliers via dual assent in dual-domain $\lambda_i^{e+1} \leftarrow f_{step}\left(\lambda_i^e - \Lambda\left(\sum_{t \in T}\left(\gamma^{t-1} Relu(g_{it})\right)|\pi_e\right)\right)$, while in tandem, we will also update the quadratic penalty coefficient $\zeta^{e+1} \leftarrow u_\zeta \zeta^e$ such that the penalty from both Lagrangian and quadratic penalty terms will increase with every update of $\lambda_i \in \mathbb{R} \backslash \mathbb{R}^+, \forall i \in P$ and, $\zeta_i \in \mathbb{R}^+, \forall i \in P$. For ease of understanding, the block diagram of the workings of the risk-averse actor–critic algorithm [12] with the proposed augmented lagrangian-based constraint handling mechanism is shown in Figure 2. Here, the upper shaded religion contains all the learning components inside the actor–critic algorithml while the shaded region at the bottom contains the proposed constraint handling mechanism, which updates the Lagrangian multipliers and quadratic penalty factors (inside the environment) after every fixed number of risk-averse actor–critic algorithm's training iterations. For elucidation on conditionally elicitable dynamic risk measure and scoring function mentioned in the following diagram, readers are referred to [12].

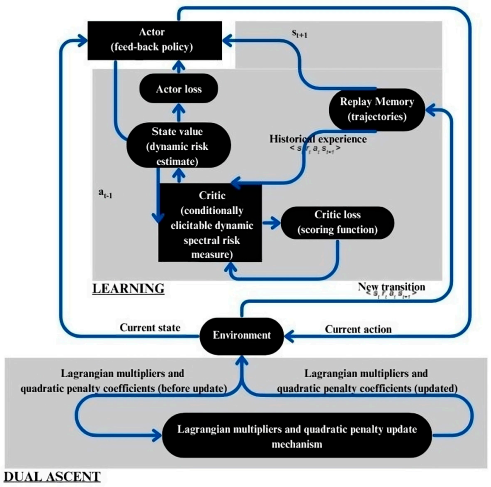


Figure 2. Block diagram of risk-averse actor–critic algorithm with proposed augmented Lagrangian-based constraint handling mechanism.

5.5. Theoretical Results for Convergence

Now, we will perform the above algorithm's convergence analysis by proving Theorem 4, but first, we will prove a proposition as follows, which will be used in proving Theorem 4:

Proposition 2. *Consider following two optimization problems such that both of these have time-invariant Lagrangian multipliers* $\lambda_i \in \mathbb{R}\backslash\mathbb{R}^+, \forall i \in P$ *, and* $\zeta_i \in \mathbb{R}^+$ *, as follows under "Assumption 1":*

Problem A:

$$d_s^1(\lambda_i, \forall i \in P) = \max_{\pi \in \mathcal{P}} \mathbb{M}\Big[g_{0:|P|,1:|T|}\Big] \tag{51}$$

Subject to the following: (6)–(8)

$$\lambda_i = \lambda_{it},\ \forall t \in T \tag{52}$$

Problem B:

$$d_s^2(\{\lambda_i, \forall i \in P\}, \{\zeta_i,\ \forall i \in P\}) = \max_{\pi \in \mathcal{P}} S(\pi \in P, \{\lambda_i, \forall i \in P\}, \{\zeta_i, \forall i \in P\}) \tag{53}$$

Subject to the following: (6)–(8)

Let us fix Lagrangian multipliers for all the constraints, except the constraint $i \in P$, to any value $\lambda_{i'}^f \in \mathbb{R}\backslash\mathbb{R}^+. \forall i' \in P\backslash i$ (where these fixed values can be different for "Problem A" and "Problem B") such that the Lagrangian multiplier λ_i is the only dual variable in both "Problem A" and "Problem B".

Suppose $\lambda_i^{1*} = \underset{\lambda_i \in \mathbb{R}\backslash\mathbb{R}^+}{\operatorname{argmin}} d_s^1\Big(\{\lambda_i,\ i \in P\}, \Big\{\lambda_{i'}^f, \forall i' \in P\backslash i\Big\}\Big)$ and $\lambda_i^{2*} = \underset{\lambda_i \in \mathbb{R}\backslash\mathbb{R}^+}{\operatorname{argmin}} d_s^2\Big(\{\lambda_i,\ i \in P\}, \Big\{\lambda_{i'}^f, \forall i' \in P\backslash i\Big\}, \{\zeta_i, i \in P\}\Big)$. Also, let $\pi^* \in \mathcal{P}$ be an optimal policy for "Problem 2" under which $\gamma^{t-1} g_{it} \leq 0,\ \forall i \in P, t \in T$ holds; then, the following relation will also hold

$$\begin{aligned} &d_s^1\Big(\{\lambda_i^{1*},\ i \in P\}, \Big\{\lambda_{i'}^f, \forall i' \in P\backslash i\Big\}\Big) \leq d_s^2\Big(\{\lambda_i^{2*},\ i \in P\}, \Big\{\lambda_{i'}^f, \forall i' \in P\backslash i\Big\}, \{\zeta_i, i \in P\}\Big) \leq \\ &d_s^2\Big(\{\lambda_i^{1*},\ i \in P\}, \Big\{\lambda_{i'}^f, \forall i' \in P\backslash i\Big\}, \{\zeta_i, i \in P\}\Big) \leq d_s^1\Big(\{\lambda_i^{1*},\ i \in P\}, \Big\{\lambda_{i'}^f, \forall i' \in P\backslash i\Big\}\Big) \end{aligned} \tag{54}$$

Proof. As we know $\zeta_i \geq 0$, the following holds

$$\begin{aligned} &d_s^2\Big(\Big\{\lambda_i, \lambda_{i'}^f \in \mathbb{R}\backslash\mathbb{R}^+,\ \forall i' \in P\backslash i\Big\}, \{\zeta_i \in \mathbb{R}^+,\ i \in P\}\Big) \leq \\ &d_s^2\Big(\Big\{\lambda_i, \lambda_{i'}^f \in \mathbb{R}\backslash\mathbb{R}^+,\ \forall i' \in P\backslash i\Big\}, \{\zeta_i = 0,\ i \in P\}\Big) = d_s^1\Big(\lambda_i, \lambda_{i'}^f \in \mathbb{R}\backslash\mathbb{R}^+, \forall i' \in P\backslash i\Big) \end{aligned} \tag{55}$$

We can imply the following from the above, where instead of λ_i, we used λ_i^{1*}:

$$d_s^2\Big(\Big\{\lambda_i^{1*},\ i \in P\Big\}, \Big\{\lambda_{i'}^f, \forall i' \in P\backslash i\Big\}, \{\zeta_i, i \in P\}\Big) \leq d_s^1\Big(\Big\{\lambda_i^{1*},\ i \in P\Big\}, \Big\{\lambda_{i'}^f, \forall i' \in P\backslash i\Big\}\Big) \tag{56}$$

Also, by definition

$$d_s^2\Big(\Big\{\lambda_i^{2*},\ i \in P\Big\}, \Big\{\lambda_{i'}^f, \forall i' \in P\backslash i\Big\}, \{\zeta_i, i \in P\}\Big) \leq d_s^2\Big(\Big\{\lambda_i^{1*},\ i \in P\Big\}, \Big\{\lambda_{i'}^f, \forall i' \in P\backslash i\Big\}, \{\zeta_i = 0, i \in P\}\Big) \tag{57}$$

By assuming $\pi^* \in P$ isthe policy which is optimal for "Problem 2" and, thus, feasible, we can write the following for "Problem A" and "Problem B" (given strong duality holds for "Problem 2"):

$$\begin{aligned} &d_s^2\Big(\{\lambda_i^{2*},\ i \in P\}, \Big\{\lambda_{i'}^f, \forall i' \in P\backslash i\Big\}, \{\zeta_i, i \in P\}\Big) \geq \\ &S\Big(\pi^* \in P, \{\lambda_i^{1*},\ i \in P\}, \Big\{\lambda_{i'}^f, \forall i' \in P\backslash i\Big\}, \{\zeta_i, i \in P\}\Big) = \\ &\qquad M_0(\pi^* \in P) = d_s^1\Big(\{\lambda_i^{1*},\ i \in P\}, \Big\{\lambda_{i'}^f, \forall i' \in P\backslash i\Big\}\Big) \end{aligned} \tag{58}$$

From (56)–(58), we can write the following:

$$\begin{array}{c} d_s^1\Big(\{\lambda_i^{1*},\ i\in P\},\Big\{\lambda_{i'}^f,\forall i'\in P\backslash i\Big\}\Big)\le d_s^2\Big(\{\lambda_i^{2*},\ i\in P\},\Big\{\lambda_{i'}^f,\forall i'\in P\backslash i\Big\},\{\zeta_i,i\in P\}\Big)\le \\ d_s^2\Big(\{\lambda_i^{1*},\ i\in P\},\Big\{\lambda_{i'}^f,\forall i'\in P\backslash i\Big\},\{\zeta_i,i\in P\}\Big)\le d_s^1\Big(\{\lambda_i^{1*},\ i\in P\},\Big\{\lambda_{i'}^f,\forall i'\in P\backslash i\Big\}\Big) \end{array} \tag{59}$$

Since we have taken no assumption on the constraint $i \in P$ (whose Lagrangian multiplier (dual variable) is not fixed), the relation (59) above is true for all constraints which are indexed on the set P, i.e.,

$$\begin{array}{c} d_s^1\Big(\{\lambda_i^{1*},\ i\in P\},\Big\{\lambda_{i'}^f,\forall i'\in P\backslash i\Big\}\Big)\le d_s^2\Big(\{\lambda_i^{2*},\ i\in P\},\Big\{\lambda_{i'}^f,\forall i'\in P\backslash i\Big\},\{\zeta_i,i\in P\}\Big)\le \\ d_s^2\Big(\{\lambda_i^{1*},\ i\in P\},\Big\{\lambda_{i'}^f,\forall i'\in P\backslash i\Big\},\{\zeta_i,i\in P\}\Big)\le d_s^1\Big(\{\lambda_i^{1*},\ i\in P\},\Big\{\lambda_{i'}^f,\forall i'\in P\backslash i\Big\}\Big),\ \forall i\in P \end{array} \tag{60}$$

By direct implication from (60) above, we proved (54). □

Theorem 4. *Suppose* $\lambda_i^{2*} = \underset{\lambda_i\in\mathbb{R}\backslash\mathbb{R}^+}{arg\ min} d_s^2(\{\lambda_i,\forall i\in P\},\{\zeta_i,\ \forall i\in P\})$ *subject to: (6)–(8).* $\pi^*\in\mathcal{P}$ *is the optimal solution for "Problem 2" under which* $\gamma^{t-1}g_i\le 0,\ \forall i\in P, t=T$, *and the optimal policy of "Problem B" is* $\pi_s^{2*}\left(\lambda_i^{2*},\ \forall i\in P\right)$, *then as* $\zeta_i\to\infty,\ \forall i\in P$, *the following condition holds*

$$\left\|M_0(\pi^*)-M_0\Big(\pi_s^{2*}\Big(\lambda_i^{2*},\ \forall i\in P\Big)\Big)\right\|_2\to 0 \tag{61}$$

Proof. Again, let us fix Lagrangian multipliers for all the constraints to any value $\lambda_{i'}^f\in\mathbb{R}\backslash\mathbb{R}^+$, $\forall i'\in P\backslash i$ except $i\in P$ such that the Lagrangian multiplier $\lambda_i\in\mathbb{R}\backslash\mathbb{R}^+$ for the constraint $i\in P$ is the only dual variables, while ζ_i can evolve.

Suppose there is a policy $\pi_s^{2*}\left(\lambda_i^{2*},\ i\in P\right)\in P$ optimal for "Problem B" with some $\zeta_i>0,\ i\in P$. Also, presume that the policy is infeasible for "Problem 1":

$$V_i\Big(\pi_s^{2*}\Big(\lambda_i^{2*},\ i\in P\Big)\Big)>0,\quad \forall i\in P \tag{62}$$

Here, we will consider two cases to prove our proposition.

Consider the objective function $\mathbb{M}\Big[g_{0,1:|T|}\Big]$ of "Problem 1". There are only two cases possible for the constraint $i\in P$:$\lambda_i^{2*}<\lambda_i^{1*}$ or $\lambda_i^{2*}\ge\lambda_i^{1*}$.

Suppose $\lambda_i^{2*}<\lambda_i^{1*}$ holds. As $\pi_s^{2*}\left(\lambda_i^{2*},\ i\in P\right)$ is the policy optimal for "Problem B", then the following expression can be written

$$\begin{array}{l} d_s^1\Big(\{\lambda_i^{1*},\ i\in P\},\Big\{\lambda_{i'}^f,\forall i'\in P\backslash i\Big\}\Big)\ge\mathbb{M}\Big[g_{0:|P|,1:|T|}\Big|\pi_s^{2*}\left(\lambda_i^{2*},\ i\in P\right)\Big]= \\ \mathbb{M}\Big[g_{0,1:|T|}\Big|\pi_s^{2*}\left(\lambda_i^{2*},\ i\in P\right)\Big]+\lambda_i^{1*}V_i\left(\pi_s^{2*}\left(\lambda_i^{2*},\ i\in P\right)\right)> \\ \mathbb{M}\Big[g_{0,1:|T|}\Big|\pi_s^{2*}\left(\lambda_i^{2*},\ i\in P\right)\Big]+\lambda_i^{2*}V_i\left(\pi_s^{2*}\left(\lambda_i^{2*},\ i\in P\right)\right)> \\ d_s^2\Big(\{\lambda_i^{2*},\ i\in P\},\Big\{\lambda_{i'}^f,\forall i'\in P\backslash i\Big\},\{\zeta_i,i\in P\}\Big) \end{array} \tag{63}$$

The result from case 1 contradicts Proposition 2 above.

Hence, it is proved that $\lambda_i^{2*}\ge\lambda_i^{1*}$ for constraint $i\in P$.

As we do not assume anything for constraint $i\in P$, for any combination of the values of $\Big\{\lambda_{i'}^f\in\mathbb{R}\backslash\mathbb{R}^+,\forall i'\in P\backslash i\Big\}$, $\lambda_i^{2*}\ge\lambda_i^{1*}$ holds $\forall i\in P$.

Given the above, let $\zeta_i'>0,\ \forall i\in P$ such that the following holds

$$\mathbb{M}\Big[g_{0,1:|T|}\Big|\pi_s^{2*}\Big(\lambda_i^{2*},\ \forall i\in P\Big)\Big]-\sum_{i\in P}\left(\frac{\zeta_i'}{2}ReLU(g_i)^2\Big|\pi_s^{2*}\Big(\lambda_i^{2*},\ \forall i\in\mathcal{P}\Big)\right)<\mathbb{M}\Big[g_{0,1:|T|}\Big|\pi^*\in P\Big] \tag{64}$$

For Proposition 2 and the strong duality of "Problem 2", the following is implied

$$d_s^2\left(\left\{\lambda_i^{2*},\ \forall i \in P\right\}, \left\{\zeta_i', \forall i \in P\right\}\right) = \mathbb{M}\left[g_{0,1:|T|}\middle|\pi^* \in P\right] \tag{65}$$

and

$$\begin{aligned}
&\mathbb{M}\left[g_{0,1:|T|}\middle|\pi^* \in P\right] > \\
&\mathbb{M}\left[g_{0,1:|T|}\middle|\pi_s^{2*}\left(\lambda_i^{2*},\ \forall i \in P\right)\right] - \sum_{i\in P}\left(\frac{\zeta_i'}{2}ReLU(g_i)^2\middle|\pi_s^{2*}\left(\lambda_i^{2*},\ \forall i \in P\right)\right) \geq \\
&\mathbb{M}\left[g_{0,1:|T|}\middle|\pi_s^{2*}\left(\lambda_i^{2*},\ \forall i \in P\right)\right] + \sum_{i\in P}\left(\lambda_i^{2*}\sum_{t\in T}ReLU(g_{it})\middle|\pi_s^{2*}\left(\lambda_i^{2*},\ \forall i \in P\right)\right) - \\
&\qquad \sum_{i\in P}\left(\frac{\zeta_i'}{2}ReLU(g_{it})^2\middle|\pi_s^{2*}\left(\lambda_i^{2*},\ \forall i \in P\right)\right)
\end{aligned} \tag{66}$$

From this, we can say that $\pi_s^{2*}\left(\lambda_i^{2*}, \forall i \in P\right)$ becomes suboptimal for "Problem B" as $\zeta_i \to \zeta_i'$, $\forall i \in P$.

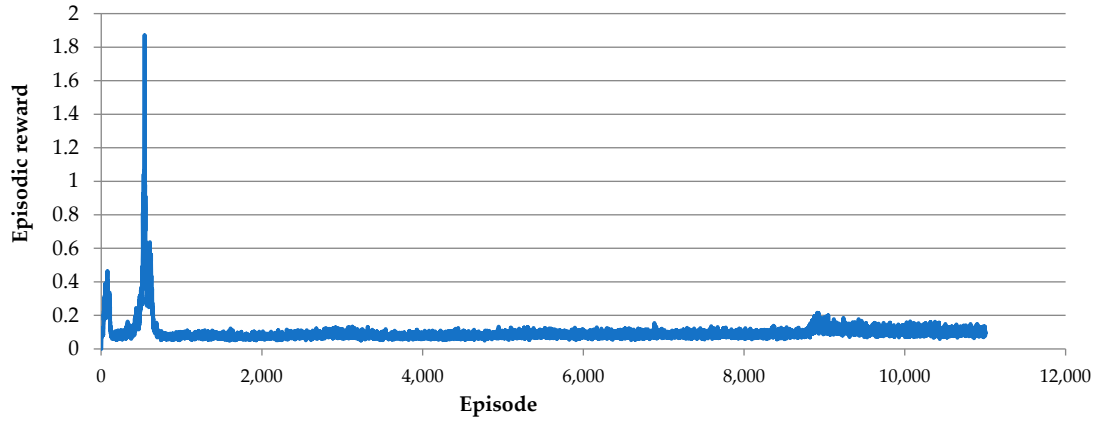
The above implies that as $\zeta_i \to \infty$, $\forall i \in P$, any infeasible policy for "Problem 2"becomes suboptimal for "Problem B".

Hence, as $\zeta_i \to \infty\ \forall i \in P$, $\left\|M_0(\pi^*) - M_0\left(\pi_s^{2*}\right)\right\|_2 \to 0$. □

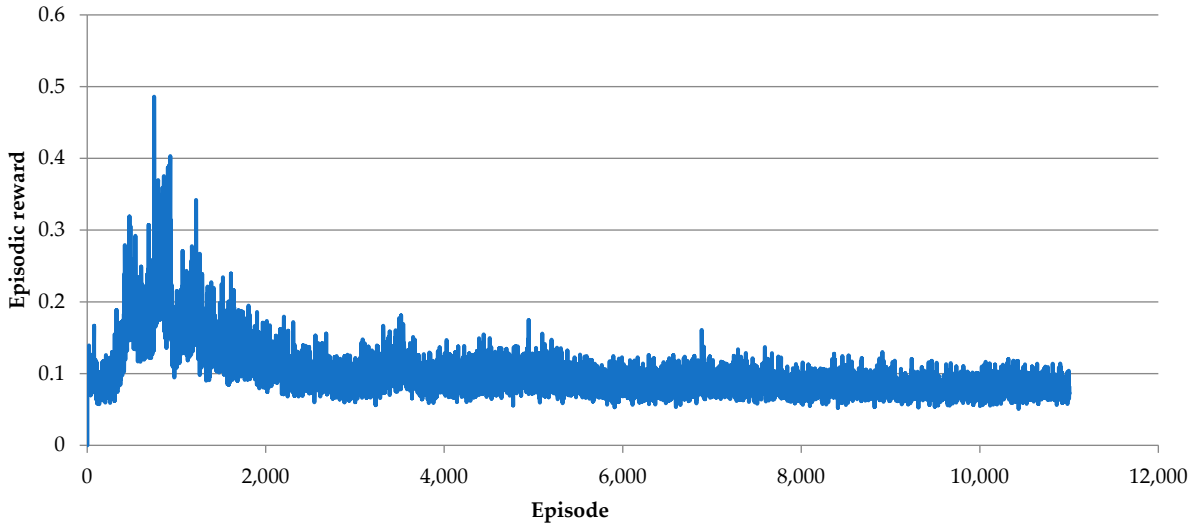
6. Numerical Example

In this section, we showcase the implications of our theoretical results. For this, all the computational experiments were conducted with a 78.16 GB hard disk, 12.68 GB RAM, Intel(R) Xeon(R) CPU @ 2.00GHz dual-core processor, Ubuntu operating system, and Python 3.10. For deep learning implementation, feed-forward neural networks were used for both actor and critic with two and six layers, respectively. They were implemented in Pytorch and trained on the Tesla T4 (CUDA Version: 12.0) graphical processing unit (GPU). In this regard, a non-convex, non-differentiable, multi-stage disaster response relief allocation problem [49] with three affected areas was considered as a numerical example, the detailed formulation of which is given in Appendix B. It was assumed that all the affected areas have **uniform** demand stochasticity, and relief is supplied to them from a capacitated local response center. Dynamic conditional value at risk (DCVaR) is used as a risk-averse stochastic ordering for the stochastic objective. The problem has an upper bound (constraint) on the deprivation level of the third affected area. This is because this affected area is assumed to be relatively more vulnerable to disruption than the rest [49,50]. These disruptions can be in the form of secondary disasters inside the affected areas or logistical blockages hindering relief supply, which will have devastating effects (increase in deprivation from life-sustaining resources) on the affected population. So, as a preparation tactic against these disruptions, deprivation levels are kept low a priori by using the constraint (A18) in Appendix B, the form of which is assumed to be $\left[\text{Re}LU\left(g_{it}'\right)\middle|\pi\right] \leq 0$, $\forall i \in P$, where $g_{it}' = s_{i\omega|T|+1}' - T_i$, $\forall i \in P$ (given at the end of Appendix B). Therefore, if we assume the quadratic penalty coefficient to be zero, the classical Lagrangian dual method will be obtained [32,34–36]. In this regard, using our formulation, we train the agent to be risk-averse in the face of uncertainty while also requiring it to satisfy the hard constraint (A18) in the Appendix via Algorithm 1. All the Lagrangian multipliers $\lambda_i^0 \in \mathbb{R}\backslash\mathbb{R}^+, \forall i \in P$ and quadratic penalty factors $\zeta_i^0 \in \mathbb{R}^+, \forall i \in P$ are initialized as "0". Before every dual ascent (with the dual ascent step size $\Lambda = 0.005$), we find the locally optimal solution using a conditionally elicitable risk-averse actor–critic algorithm from [12]. By setting $u_\zeta = 1.1$, we rapidly increase the quadratic penalty coefficient $\zeta_1 \in \mathbb{R}^+$ as the dual ascent iteration number grows. We update $\lambda_1 \in \mathbb{R}\backslash\mathbb{R}^+$ and $\zeta_1 \in \mathbb{R}^+$ every 25 algorithmic epochs (each epoch has 666,675 MDP episodes) of the risk-averse actor–critic algorithm. As we increase $\lambda_1 \in \mathbb{R}\backslash\mathbb{R}^+$ and $\zeta_1 \in \mathbb{R}^+$, we see that the CVaR of episodic rewards converges, as shown in Figure 3a–c, and the constraint violations

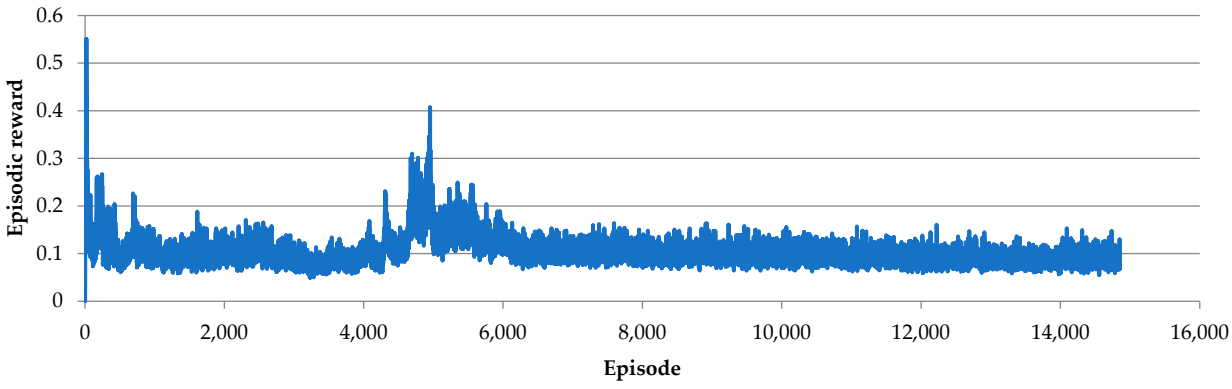
decrease until they vanish, as shown in Figure 4a–c), in the case of all values of risk-aversion confidence levels (0.6, 0.8 and 0.99) considered in our simulations.



(**a**)

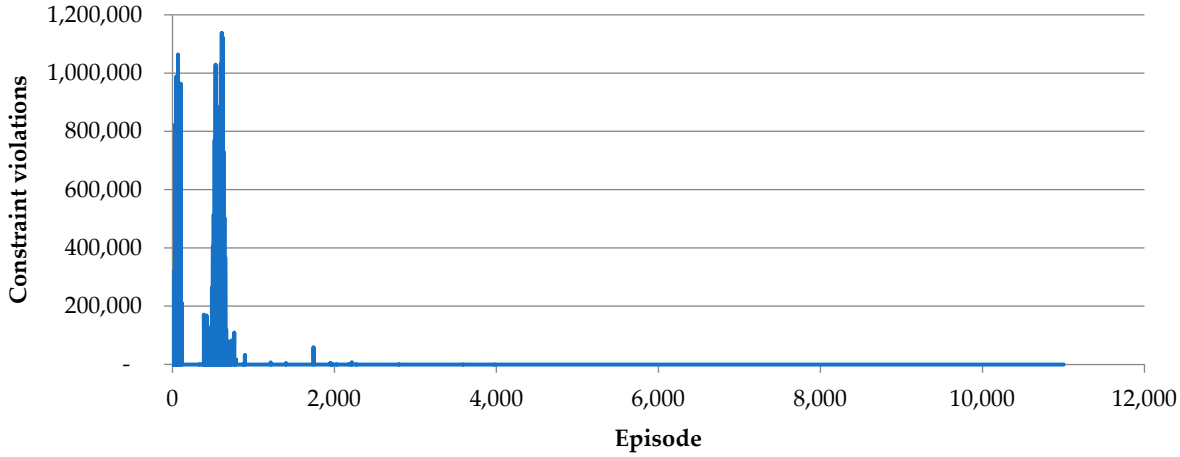


(**b**)

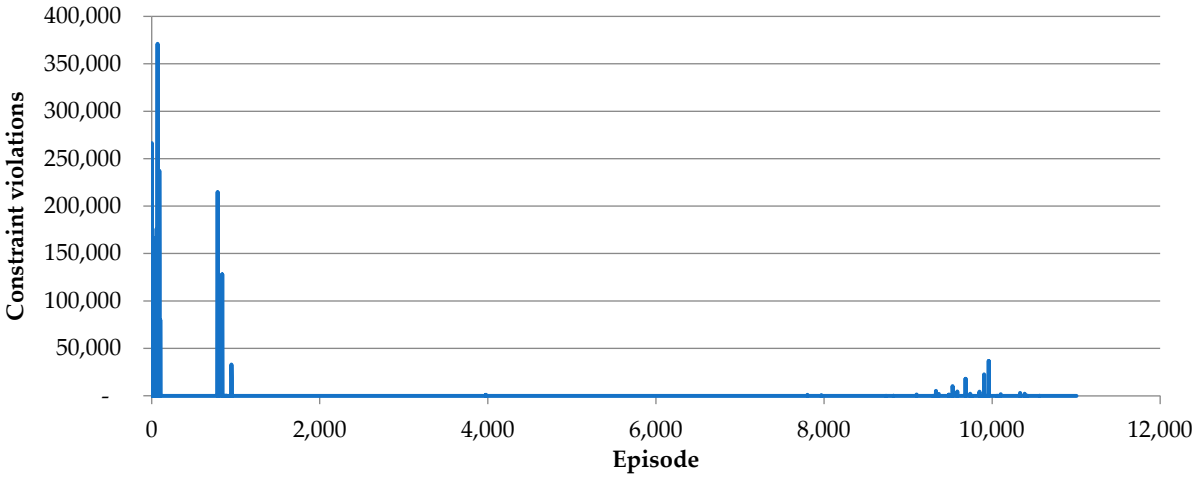


(**c**)

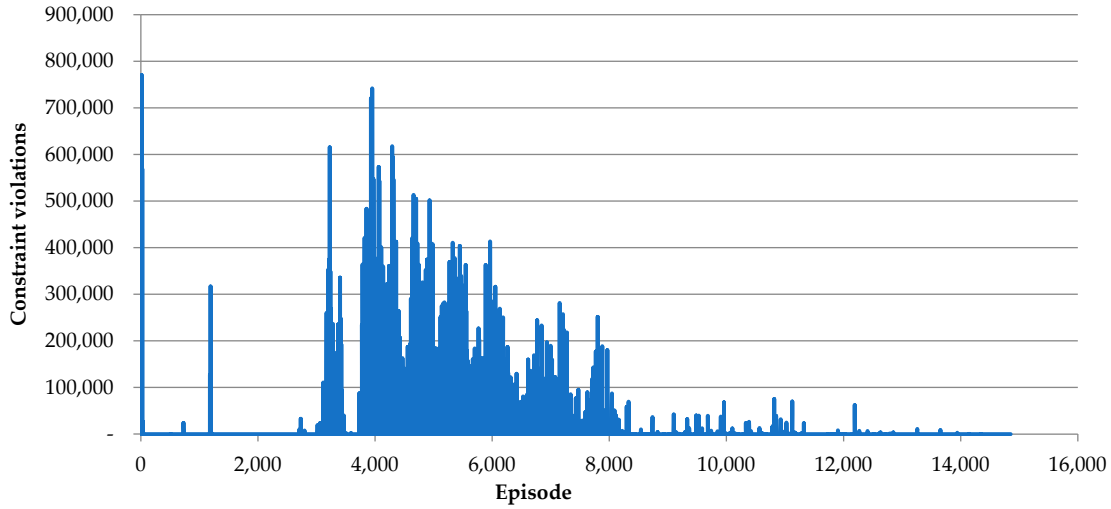
Figure 3. Convergence curves of CVaR of episodic rewards with the augmented Lagrangian-based constraint-handling approach using 0.6 (**a**), 0.8 (**b**), and 0.99 (**c**) risk-aversion confidence levels.



(**a**)



(**b**)



(**c**)

Figure 4. Vanishing constraint violations as training progresses with the increase in the number of training iterations while using 0.6 (**a**), 0.8 (**b**), and 0.99 (**c**) risk-aversion confidence levels.

We also show constraint violations and convergence curves (with a risk-averse confidence level of 0.99) when we used negative reward scheming during training. Here, the fixed constraint violation penalty factor is multiplied by constraint violation and then added to the reward function [51–53]. As we can see in Figure 5, for negative reward scheming, in case of a larger fixed penalty factor of "10^0", the constraint violations vanish to zero but, as shown in Figure 6, the CVaR episodic rewards converge to a higher cost (red curve) relative to that of the augmented Lagrangian-based approach (blue curve). This is because we are not optimizing over the fixed penalty factors, which makes these penalty factors not the optimal Lagrangian coefficients of the dual problem. Consequently, the optimal solution of the regularized unconstrained problem is different from that of the primal problem. Also, larger constraint violation factors earlier in training may hinder the RL agent's exploration leading it to converge to the local optima of primal and/or regularized unconstrained problem. On the other hand, in the case of a lower fixed penalty factor of 10^{-1}, although the reward value converges to lower costs, as shown in Figure 7, constraint violations do not reach zero, as shown in Figure 8. This issue is resolved by our proposed constraint-handling mechanism where at the beginning of the training, the action space is less constrained and the agent can freely learn, but as the training progresses, the constraint violations gradually accentuate with the increase in the Lagrangian multipliers and quadratic penalty factors in (49) and (50), respectively.

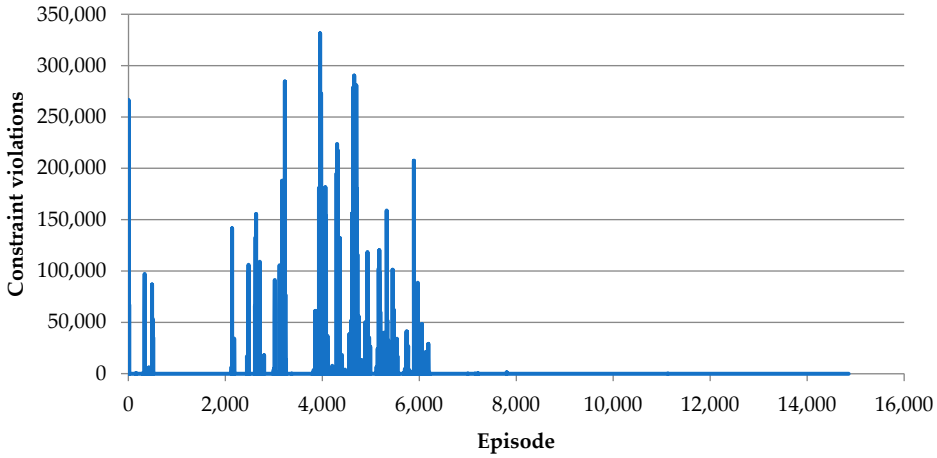


Figure 5. Constraint violations during training with fixed penalty factor of "10^0".

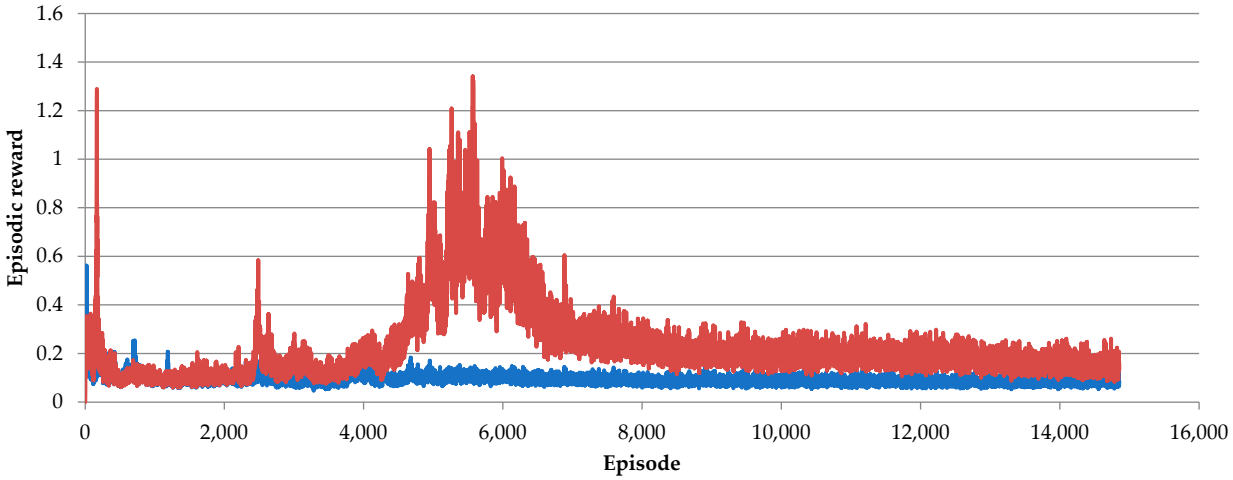


Figure 6. Convergence to a higher CVaR of episodic costs (red curve) with a fixed penalty factor of "10^0" relative to that of the augmented Lagrangian-based approach (blue curve).

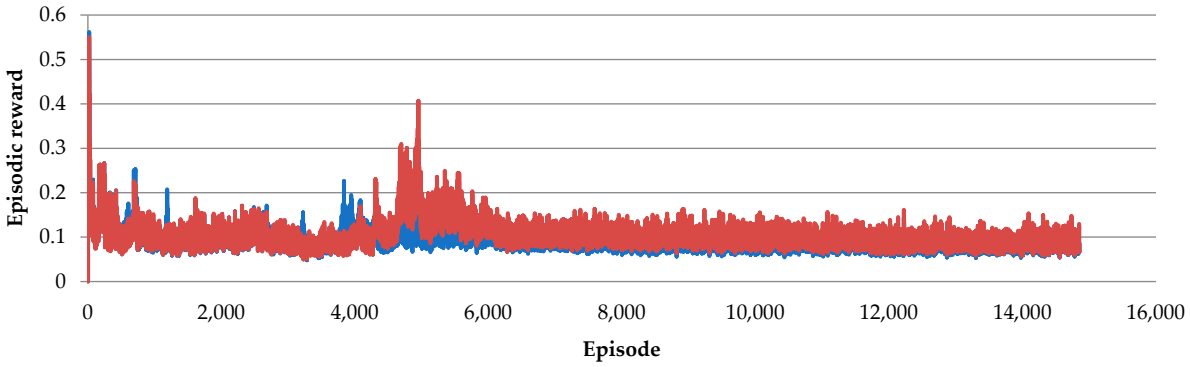


Figure 7. Convergence curves of CVaR of episodic rewards in case of fixed constraint violation penalty factor of "10^{-1}" and the augmented Lagrangian-based approach.

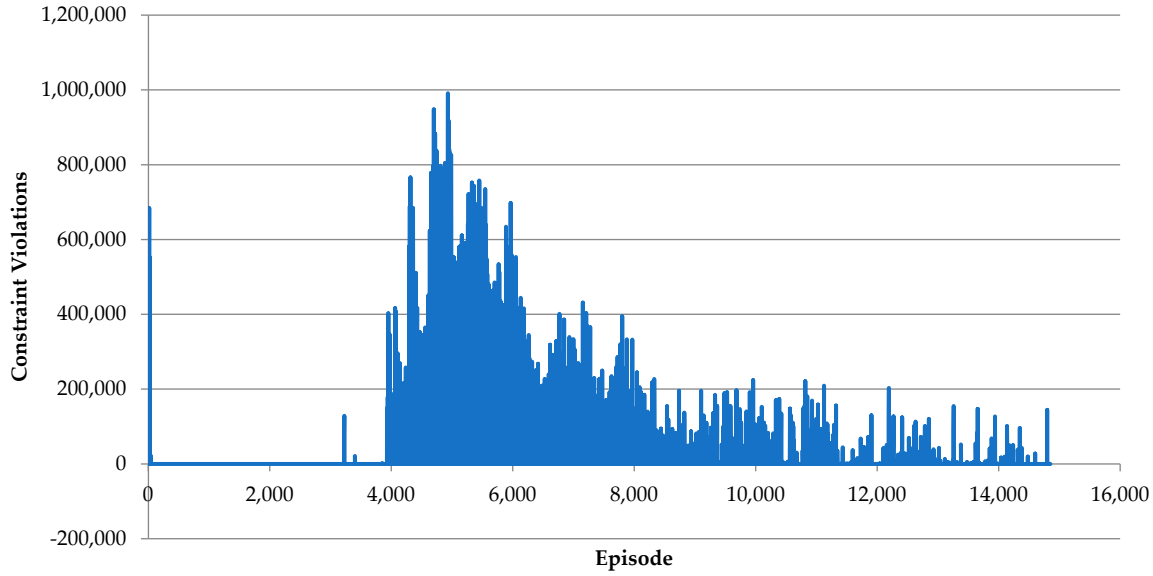


Figure 8. Constraint violations with fixed penalty factor of "10^{-1}".

7. Discussion

Throughout this work, we have developed a duality theory for risk-averse reinforcement learning problems with non-risk constraints. Specifically, in the case of an arbitrary policy space, the duality gap is zero, and thus, the primal problem is perfectly solvable in its dual domain, where it becomes convex in terms of dual variables. This is significant as it establishes that the constrained problem and the regularized unconstrained problem chase the same Prato optimal front and, therefore, are equivalent. However, despite these theoretical results, these problems may be computationally intractable. This is due to the requirement of the consideration of arbitrary policy in cases in which the evaluation of the dual function becomes practically infeasible. In place of arbitrary policy spaces, linear approximations or neural networks are used, which may cause weak duality or even widen the duality gap. Despite this approximation issue, the theoretical results, the resulting algorithm, and the empirical demonstrations provide a way to solve the constrained risk-averse policy optimization problem without the need to conduct an exhaustive search over the penalty weights (or factors) for constraint violation as was performed in [54–56].

Methodical Contribution of This Research

This research develops a model-free feed-back policy optimization method (Algorithm 1), which makes it now possible to optimize non-convex versions of constrained risk-averse multi-stage mathematical programs. This is useful for robust decision-making in high-

stakes environments [13]. Previously, only linear or convex versions of these programs weresolvable [8,9,17,42], but now convergence to the optimal solution is possible for non-convex and non-smooth versions as well using our developed model-free method, which we have empirically validated as well in Section 6. In addition, our method can also handle constraints with state variables in them which cannot be handled by existing methods like risk-averse stochastic dual dynamic programming [57].

8. Conclusions

In this work, we have combined two embryonic fields of constrained and risk-averse RL. In this regard, a duality theory is developed for MDPs with non-risk cumulative and instantaneous constraints, and spectral risk objectives. We have established strong duality with certain assumptions on the risk measures given on page 7–9 (related to the derivative and integral of the risk aversion function) and on page 13 (translation invariance of risk measures). The assumptions are not so restrictive in the context of problems considered in dynamic programming literature [5–7,23,58]. Based on these theoretical results, we have proposed an augmented Lagrangian-based constraint handling mechanism for risk-averse RL along with its theoretical guarantees and empirical demonstration for its viability. This mechanism can be used with any risk-averse RL algorithm for the solution of risk-averse MDPs with non-risk constraints.

However, our theoretical results do not imply that the proposed approach in Section 5 will always be able to solve all the considered classes of problems. This can be due to the intractability of the dual function, making it impossible to evaluate and/or the fact that our theoretical results are for a general policy space, which is impossible to consider realistically. In place of general policy, exploration usually takes place in linear or neural networks-based parameterized (restricted) (sub-) policy spaces, which makes the problem different from the original one and, consequently, the duality gap might not be zero. Therefore, our future direction of investigation will be to theoretically and/or empirically investigate how to further improve the algorithm's convergence behavior in case of different parameterizations of the policy space. Lastly, the need for constraint satisfaction might come directly from within the RL algorithm, e.g., in the case of trust region policy optimization [59], where the divergence of the policy is constrained by a threshold. Therefore, strong duality must also be investigated under these constraints.

Author Contributions: Conceptualization, M.A.; methodology, M.A. and M.B.R.; validation, M.A. and M.O.; investigation, M.A. and M.S.H.; resources, M.B.R. and M.S.H.; data curation, M.A., M.O. and M.S.H.; writing—original draft preparation, M.A. and M.S.H.; writing—review and editing, M.A. and M.S.H.; supervision, M.B.R. and M.S.H.; project administration, M.S.H.; funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2022S1A5C2A04092540).

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to the excessive length of the data.

Acknowledgments: We would like to thank Dr. Santiago Paternain (Assistant Professor at the Rensselaer Polytechnic Institute, New York, USA) for his helpful remarks.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Expectation-Based Constraints

Although we have followed [34] for the manipulations of expectation-based constraints in Theorem 1, we have given these manipulations in terms of our notations in this appendix for the ease of readers.

First, we will start forming an optimization program by first writing all the constraints (9) in Section 3 indexed on the set P (not the risk-averse objective (5) in Section 3) in "Problem 1" in their linear form as follows:

$$V_i(\pi \in \mathcal{P}) = \int\limits_{(\mathcal{S}\times\mathcal{A})^{|T|}} \left(\sum_{t\in T} \left(\gamma^{t-1} g_{it} \right) p_\pi \left(d^{trj} \in D^{trj} \right) \right) ds_1 da_1 ds_2 da_2 \dots ds_{|T|} da_{|T|}, \quad \forall i \in P \tag{A1}$$

where $p_\pi\left(d^{trj} \in D^{trj}\right)$ is the probability of a particular trajectory $d^{trj} \in D^{trj}$ of the form $s_1, a_1, s_2, a_2, \dots, s_{|T|}, a_{|T|}$, given a policy $\pi \in \mathcal{P}$, and $|T|$ being the cardinality of the set T (number of periods in the planning horizon).

Since the cost functions $g_{it}, \forall i \in P, t \in T$ are bounded (by the assumption of Theorem 1), the dominance convergence theorem (DCT) holds, allowing us to alter the order of integral and summation in (A1). With the use of conditional probabilities and Markov's property of the MDP, we can rewrite (A1) as follows:

$$V_i(\pi \in \mathcal{P}) = \sum_{t\in T} \left(\begin{array}{c} \gamma^{t-1} \int\limits_{(\mathcal{S}\times\mathcal{A})^{|T|}} \left(g_{it} \prod\limits_{u\in T\setminus\{1\}} (p(s_u|s_{u-1}, a_{u-1}) p_\pi(a_u|s_u)) p_0(s_1) p_\pi(a_1|s_1) \right) \\ ds_1 da_1 ds_2 da_2 \dots ds_{|T|} da_{|T|} \end{array} \right), \forall i \in P \tag{A2}$$

In (A2), $p(s_u|s_{u-1}, a_{u-1})$ represents the probability of state s_u given the previous state s_{u-1} and action a_{u-1}. $p_\pi(a_u|s_u)$ is the probability of the action a_u given that the current state s_u and policy $\pi \in \mathcal{P}$, and $p_0(s_1)$ is the probability of the initial state s_1.

Since every integral concerning state–action pair (s_u, a_u) in (A2) results in unity for all $u \geq t$, (by probabilityaxioms [60]), we can again rewrite the above expression as follows:

$$V_i(\pi \in \mathcal{P}) = \sum_{t\in T} \left(\begin{array}{c} \gamma^{t-1} \int\limits_{(\mathcal{S}\times\mathcal{A})^{t}} \left(g_{it} \prod\limits_{u\in\{x\in T|x>1, x\leq t\}} (p(s_u|s_{u-1}, a_{u-1}) p_\pi(a_u|s_u)) p_0(s_1) p_\pi(a_1|s_1) \right) \\ ds_1 \dots ds_t da_1 \dots da_t \end{array} \right), \forall i \in P \tag{A3}$$

From the dominance convergence theorem (DCT) and the fact that we can write (A4), we compactly rewrite the above expressions (A3) as (A5) below:

$$p_\pi(s_t \in \mathcal{S}, a_t \in \mathcal{A}) = \int\limits_{(\mathcal{S}\times\mathcal{A})^{t-1}} \left(\prod_{u\in\{x\in T|x>1, x\leq t\}} (p(s_u|s_{u-1}, a_{u-1}) p_\pi(a_u|s_u)) p_0(s_1) p_\pi(a_1|s_1) \right) ds_1 \dots ds_{t-1} da_1 \dots da_{t-1} \tag{A4}$$

$$V_i(\pi \in \mathcal{P}) = \int\limits_{\mathcal{S}\times\mathcal{A}} \left(g_{it} \sum_{t\in T} \left(\gamma^{t-1} p_\pi(s_t = s \in \mathcal{S}, a_t = a \in \mathcal{A}) \right) \right) dsda, \forall i \in P \tag{A5}$$

where $g_{it} = g_i(s_t = s \in \mathcal{S}, a_t = a \in \mathcal{A})$, and $p_\pi(s_t = s \in \mathcal{S}, a_t = a \in \mathcal{A}) = p_\pi(s \in \mathcal{S}, a \in \mathcal{A})$ is the probability of the occurrence of the state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$ in a period $t \in T$ given the policy $\pi \in \mathcal{P}$.

The occupation measure for the MDP (regardless of its objective or reward function) can be written as

$$\rho(s \in \mathcal{S}, a \in \mathcal{A}) = (1-\gamma) \sum_{t\in T} \left(\gamma^{t-1} p_\pi(s_t = s \in \mathcal{S}, a_t = a \in \mathcal{A}) \right) \tag{A6}$$

From the previous two Equations (A5) and (A6), the following follows:

$$V_i(\pi \in \mathcal{P}) = \frac{1}{1-\gamma} \int\limits_{\mathcal{S}\times\mathcal{A}} (g_{it} \rho(s \in \mathcal{S}, a \in \mathcal{A})) dsda, \quad \forall i \in P \tag{A7}$$

Appendix B

In this appendix, the multi-stage optimization problem used for numerical simulation is rigorously defined. This multi-stage optimization problem is for relief allocation among a set of affected areas during disaster response and is adapted from the second nonlinear program (NLP2) in [49].

The problem in the form of a diagram is shown as follows:

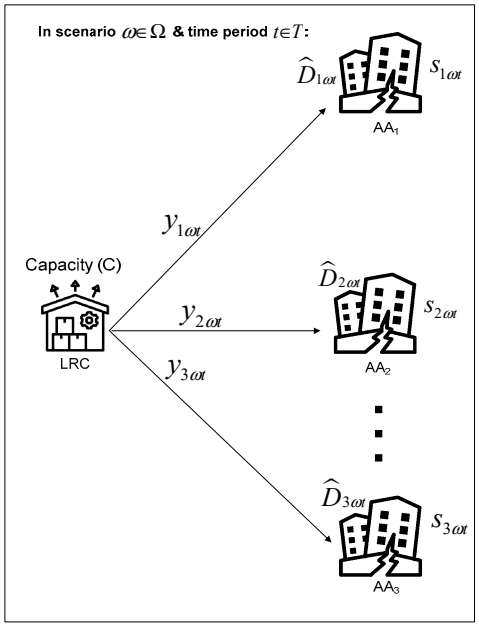


Figure A1. Visual description of resource allocation problem in disaster response.

Before discussing the abstract mathematical programming model, we will first give its notation as follows:

Sets:

K: Index set for affected areas (AAs)

K': Index set for affected areas the terminal deprivation level of which is constrained

Ω: Index set for scenarios

T: Index set for periods (or stages)

Note: Size or cardinality of the above sets will be represented by placing vertical bars around the name of the respective set, e.g., for the set of periods, its size will be represented as $|T|$.

Indices:

k: Index in K

k': Index in K'

ω: Index in Ω

t: Index in T

Decision Variables:

$y_{k\omega t}$: Amount of relief resources being allocated from the LRC location to AA $k \in K$ in scenario $\omega \in \Omega$ and period $t \in T$.

State Variables:

$s_{k\omega t}$: Deprivation level at AA $k \in K$ in scenario $\omega \in \Omega$ and period $t \in T$

Parameters:

a_{kt}: per unit accessibility-based delivery cost for transporting unit quantity of relief resource from the LRC to AA $k \in K$ in period $t \in T$ in all scenarios

$\hat{D}_{k\omega t}$: Demand for relief resource at AA $k \in K$ in scenario $\omega \in \Omega$ and period $t \in T$, it is considered as a stochastic variable

$s_{k\omega 1}$: Initial deprivation level (at the start of the planning horizon) at AA $k \in K$ and period 1, for all $\omega \in \Omega$

C:Capacity (considered as flexible) of the LRC

p_k: Deprivation parameter at AA $k \in K$ (used in ADC and TPC functions)

q_k: Deprivation parameter at AA $k \in K$ (used in ADC and TPC functions)

$T_{k\omega}$: Threshold above which final robustified state variable values $s_{k\omega|T|+1}$, $\forall k \in K'$, $\omega \in \Omega$ must not exceed to ensure population survival beyond the planning horizon.

Len: One time period's length

weg_1: Weight for logistics cost for single objective formulation

weg_2: Weight for deprivation cost for single objective formulation

weg_3: Weight for TPC for single objective formulation

Functions:

$H(s_{k\omega t})$: ADC function with state value of AA $k \in K$ in scenario $\omega \in \Omega$ and period $t \in T$

$\mathrm{I}\left(s_{k\omega|T|+1}\right)$: TPC function with state value of AA $k \in K$ in scenario $\omega \in \Omega$ at the end of the planning horizon

Objectives:

Three objectives are considered corresponding to efficiency, effectiveness, and equity criteria as discussed below. As an efficiency metric, logistics cost (LC) is used [16].

$$\min_{y_{k\omega t}, \forall k \in K, \omega \in \Omega, t \in T} \sum_{t \in T} \sum_{k \in K} (a_{kt} y_{k\omega t}), \quad \forall \omega \in \Omega \tag{A8}$$

Secondly, the sum of AAs' ADCs for all time periods and relief items are considered, as shown as follows:

$$\min_{y_{k\omega t}, \forall k \in K, \omega \in \Omega, t \in T} \sum_{t \in T} \sum_{k \in K} H(s_{k\omega t}), \quad \forall \omega \in \Omega \tag{A9}$$

Thirdly, the sum of all AAs' TPCs for all time periods and relief items are considered, as shown as follows:

$$\min_{y_{k\omega t}, \forall k \in K, \omega \in \Omega, t \in T} \sum_{k \in K} \mathrm{I}(s_{k\omega|T|+1}), \quad \forall \omega \in \Omega \tag{A10}$$

Here, deprivation cost function in (A9) at AA $k \in K$ in period $t \in T$ is represented by $H(s_{k\omega t})$ which is defined as a piecewise non-convex function as shown below:

$$H(s_{k\omega t}) = \begin{cases} e^{p_k}(e^{q_k}-1)(e^{q_i})^{s_{k\omega t}}, & s_{k\omega t} \geq 0 \\ 0, & s_{k\omega t} < 0 \end{cases}, \quad \forall k \in K, \omega \in \Omega, t \in T \tag{A11}$$

TPC function in (A10) at AA $k \in K$ at the end of the planning horizon is represented by $\mathrm{I}(s_{k\omega|T|+1})$ which is also defined as a piecewise non-convex function as shown below:

$$I(s_{k\omega|T|+1}) = \begin{cases} e^{p_k}(e^{q_k}-1)(e^{q_k})^{s_{k\omega|T|+1}}, & s_{k\omega|T|+1} \geq 0 \\ 0, & s_{k\omega|T|+1} < 0 \end{cases}, \quad \forall k \in K, \omega \in \Omega \tag{A12}$$

Here, derivation cost (A9) is the effectiveness metric. TPC (A10) is used to ensure AAs' population survival beyond the planning horizon (in a particular scenario) to ensure equity [61]. It depends on the deprivation amount at the end of the planning horizon.

Single Objective function:

In the above, there are three objectives, namely, logistics cost (A8), deprivation cost (A9), and TPC (A10). To make the problem a single objective, we will use the weighted sum method (WSM) by which the corresponding single objective is as follows:

$$\min_{y_{k\omega t}, \forall k \in K, \omega \in \Omega, t \in T} \left(weg_1 \sum_{t \in T} \sum_{k \in K} (c_{kt} y_{k\omega t}) + weg_2 \sum_{t \in T} \sum_{k \in K} H(s_{k\omega t}) + weg_3 \sum_{k \in K} \mathrm{I}(s_{k\omega|T|+1}) \right), \quad \forall \omega \in \Omega \tag{A13}$$

Stochastic Ordering:

In our problem, objectives(A9) and (A10) are stochastic due to stochastic demand, which requires us to use some stochastic ordering. As we are investigating risk-averse decision-making situations, any spectral risk measure such as CVaR or mean–CVaR M_0 in (A14) as a dynamic risk measure can be used in (A13) as follows:

$$\min_{y_{k\omega t}, \forall k \in K, \omega \in \Omega, t \in T} M_0 \left(weg_1 \sum_{t \in T} \sum_{k \in K} (c_{kt} y_{k\omega t}) + weg_2 \sum_{t \in T} \sum_{k \in K} H(s_{k\omega t}) + weg_3 \sum_{k \in K} \mathrm{I}(s_{k\omega |T|+1}) \right), \quad \forall \omega \in \Omega \tag{A14}$$

Transition Function:

In a period $t \in T$ and $\omega \in \Omega$ scenario, there are state variables $s_{k\omega t}$, $\forall k \in A, \omega \in \Omega$ on which the feedback policy $\pi \in \mathcal{P}$ is conditioned to make decisions dynamically. These variables have an in-going component $s_{k\omega t}$, which is used by the transition function to determine its out-going component $s_{k\omega t+1}$ in period $t \in T$ as given in (A15) below:

$$s_{k\omega t+1} = s_{k\omega t} - y_{k\omega t} + \hat{D}_{k\omega t}, \quad \forall k \in K, \omega \in \Omega, t \in T \tag{A15}$$

Constraints:

First are the capacity constraints (functional) in which solution space (decision space) is constrained, such that all allocations of relief items are only from LRCs that are installed, and total allocation (combined allocation of all the relief items to all AAs) from each of LRC is less than their capacity in a particular period, as shown by the following inequality:

$$\sum_{k \in K} y_{k\omega t} \leq C, \quad \forall \omega \in \Omega, t \in T \tag{A16}$$

The second set of constraints is on the robustified value of terminal state variables $s'_{k\omega |T|+1}$, $\forall k \in K', \omega \in \Omega$ of the affected areas in the set K', such that their value will not exceed their threshold T_k, $\forall k \in K'$ [62]. Here, $s'_{k\omega |T|+1}$, $\forall k \in K', \omega \in \Omega$ is to be determined using the following equation.

$$s'_{k\omega |T|+1} = s_{k\omega |T|} - y_{k\omega t} + \hat{D}^{\max support}_{k\omega t}, \quad \forall k \in K', \omega \in \Omega, t \in T \tag{A17}$$

where $\hat{D}^{\max support}_{k\omega t}$, $\forall k \in K, \omega \in \Omega, t \in T$ is the maximum value in support of demand (stochastic variable) $\hat{D}_{k\omega t}$, $\forall k \in K, \omega \in \Omega, t \in T$. These sets of constraints are essential to ensure that those affected populations particularly vulnerable to logistical disruptions or secondary disasters will not be left with huge deprivation at the end of response operations as it would be unethical.

$$s'_{k\omega |T|+1} \leq T_k, \quad \forall k \in K', \omega \in \Omega \tag{A18}$$

The third set of constraints ensures that all decision variables are positive:

$$y_{k\omega t} \geq 0, \quad \forall k \in K, \omega \in \Omega, t \in T \tag{A19}$$

References

1. Wang, D.; Yang, K.; Yang, L. Risk-averse two-stage distributionally robust optimisation for logistics planning in disaster relief management. *Int. J. Prod. Res.* **2023**, *61*, 668–691. [CrossRef]
2. Habib, M.S.; Maqsood, M.H.; Ahmed, N.; Tayyab, M.; Omair, M. A multi-objective robust possibilistic programming approach for sustainable disaster waste management under disruptions and uncertainties. *Int. J. Disaster Risk Reduct.* **2022**, *75*, 102967. [CrossRef]
3. Habib, M.S. Robust Optimization for Post-Disaster Debris Management in Humanitarian Supply Chain: A Sustainable Recovery Approach. Ph.D. Thesis, Hanyang University, Seoul, Republic of Korea, 2018.
4. Hussain, A.; Masood, T.; Munir, H.; Habib, M.S.; Farooq, M.U. Developing resilience in disaster relief operations management through lean transformation. *Prod. Plan. Control* **2023**, *34*, 1475–1496. [CrossRef]
5. Gu, S.; Yang, L.; Du, Y.; Chen, G.; Walter, F.; Wang, J.; Yang, Y.; Knoll, A. A Review of Safe Reinforcement Learning: Methods, Theory and Applications. *arXiv* **2022**, arXiv:2205.10330.

6. Wang, Y.; Zhan, S.S.; Jiao, R.; Wang, Z.; Jin, W.; Yang, Z.; Wang, Z.; Huang, C.; Zhu, Q. Enforcing Hard Constraints with Soft Barriers: Safe Reinforcement Learning in Unknown Stochastic Environments. In Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research, Honolulu, HI, USA, 23–29 July 2023; pp. 36593–36604. Available online: https://proceedings.mlr.press/v202/wang23as.html (accessed on 26 May 2024).
7. Yang, Q.; Simão, T.D.; Tindemans, S.H.; Spaan, M.T.J. Safety-constrained reinforcement learning with a distributional safety critic. *Mach. Learn.* **2023**, *112*, 859–887. [CrossRef]
8. Yin, X.; Büyüktahtakın, İ.E. Risk-averse multi-stage stochastic programming to optimizing vaccine allocation and treatment logistics for effective epidemic response. *IISE Trans. Healthc. Syst. Eng.* **2022**, *12*, 52–74. [CrossRef]
9. Morillo, J.L.; Zéphyr, L.; Pérez, J.F.; Lindsay Anderson, C.; Cadena, Á. Risk-averse stochastic dual dynamic programming approach for the operation of a hydro-dominated power system in the presence of wind uncertainty. *Int. J. Electr. Power Energy Syst.* **2020**, *115*, 105469. [CrossRef]
10. Yu, G.; Liu, A.; Sun, H. Risk-averse flexible policy on ambulance allocation in humanitarian operations under uncertainty. *Int. J. Prod. Res.* **2021**, *59*, 2588–2610. [CrossRef]
11. Escudero, L.F.; Garín, M.A.; Monge, J.F.; Unzueta, A. On preparedness resource allocation planning for natural disaster relief under endogenous uncertainty with time-consistent risk-averse management. *Comput. Oper. Res.* **2018**, *98*, 84–102. [CrossRef]
12. Coache, A.; Jaimungal, S.; Cartea, Á. Conditionally Elicitable Dynamic Risk Measures for Deep Reinforcement Learning. *SSRN Electron. J.* **2023**, *14*, 1249–1289. [CrossRef]
13. Zhuang, X.; Zhang, Y.; Han, L.; Jiang, J.; Hu, L.; Wu, S. Two-stage stochastic programming with robust constraints for the logistics network post-disruption response strategy optimization. *Front. Eng. Manag.* **2023**, *10*, 67–81. [CrossRef]
14. Habib, M.S.; Sarkar, B. A multi-objective approach to sustainable disaster waste management. In Proceedings of the International Conference on Industrial Engineering and Operations Management, Paris, Farance, 26–27 July 2018; pp. 1072–1083.
15. Shapiro, A.; Tekaya, W.; da Costa, J.P.; Soares, M.P. Risk neutral and risk averse Stochastic Dual Dynamic Programming method. *Eur. J. Oper. Res.* **2013**, *224*, 375–391. [CrossRef]
16. Yu, L.; Zhang, C.; Jiang, J.; Yang, H.; Shang, H. Reinforcement learning approach for resource allocation in humanitarian logistics. *Expert Syst. Appl.* **2021**, *173*, 114663. [CrossRef]
17. Ahmadi, M.; Rosolia, U.; Ingham, M.; Murray, R.; Ames, A. Constrained Risk-Averse Markov Decision Processes. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
18. Lockwood, P.L.; Klein-Flügge, M.C. Computational modelling of social cognition and behaviour—A reinforcement learning primer. *Soc. Cogn. Affect. Neurosci.* **2020**, *16*, 761–771. [CrossRef] [PubMed]
19. Collins, A.G.E. Reinforcement learning: Bringing together computation and cognition. *Curr. Opin. Behav. Sci.* **2019**, *29*, 63–68. [CrossRef]
20. Zabihi, Z.; Moghadam, A.M.E.; Rezvani, M.H. Reinforcement Learning Methods for Computing Offloading: A Systematic Review. *ACM Comput. Surv.* **2023**, *56*, 17. [CrossRef]
21. Liu, P.; Zhang, Y.; Bao, F.; Yao, X.; Zhang, C. Multi-type data fusion framework based on deep reinforcement learning for algorithmic trading. *Appl. Intell.* **2023**, *53*, 1683–1706. [CrossRef]
22. Shavandi, A.; Khedmati, M. A multi-agent deep reinforcement learning framework for algorithmic trading in financial markets. *Expert Syst. Appl.* **2022**, *208*, 118124. [CrossRef]
23. Basso, R.; Kulcsár, B.; Sanchez-Diaz, I.; Qu, X. Dynamic stochastic electric vehicle routing with safe reinforcement learning. *Transp. Res. Part E Logist. Transp. Rev.* **2022**, *157*, 102496. [CrossRef]
24. Lee, J.; Lee, K.; Moon, I. A reinforcement learning approach for multi-fleet aircraft recovery under airline disruption. *Appl. Soft Comput.* **2022**, *129*, 109556. [CrossRef]
25. Shi, T.; Xu, C.; Dong, W.; Zhou, H.; Bokhari, A.; Klemeš, J.J.; Han, N. Research on energy management of hydrogen electric coupling system based on deep reinforcement learning. *Energy* **2023**, *282*, 128174. [CrossRef]
26. Venkatasatish, R.; Dhanamjayulu, C. Reinforcement learning based energy management systems and hydrogen refuelling stations for fuel cell electric vehicles: An overview. *Int. J. Hydrogen Energy* **2022**, *47*, 27646–27670. [CrossRef]
27. Demizu, T.; Fukazawa, Y.; Morita, H. Inventory management of new products in retailers using model-based deep reinforcement learning. *Expert Syst. Appl.* **2023**, *229*, 120256. [CrossRef]
28. Wang, K.; Long, C.; Ong, D.J.; Zhang, J.; Yuan, X.M. Single-Site Perishable Inventory Management Under Uncertainties: A Deep Reinforcement Learning Approach. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 10807–10813. [CrossRef]
29. Waubert de Puiseau, C.; Meyes, R.; Meisen, T. On reliability of reinforcement learning based production scheduling systems: A comparative survey. *J. Intell. Manuf.* **2022**, *33*, 911–927. [CrossRef]
30. Hildebrandt, F.D.; Thomas, B.W.; Ulmer, M.W. Opportunities for reinforcement learning in stochastic dynamic vehicle routing. *Comput. Oper. Res.* **2023**, *150*, 106071. [CrossRef]
31. Dalal, G.; Dvijotham, K.; Vecerík, M.; Hester, T.; Paduraru, C.; Tassa, Y.J.A. Safe Exploration in Continuous Action Spaces. *arXiv* **2018**, arXiv:1801.08757.
32. Altman, E. *Constrained Markov Decision Processes*; Routledge: London, UK, 1999.
33. Borkar, V.S. An actor-critic algorithm for constrained Markov decision processes. *Syst. Control Lett.* **2005**, *54*, 207–213. [CrossRef]
34. Paternain, S.; Chamon, L.F.O.; Calvo-Fullana, M.; Ribeiro, A. Constrained reinforcement learning has zero duality gap. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Curran Associates Inc.: New York, NY, USA, 2019; p. 679.

35. Chow, Y.; Ghavamzadeh, M.; Janson, L.; Pavone, M. Risk-constrained reinforcement learning with percentile risk criteria. *J. Mach. Learn. Res.* **2017**, *18*, 6070–6120.
36. Chow, Y.; Nachum, O.; Duenez-Guzman, E.; Ghavamzadeh, M. A lyapunov-based approach to safe reinforcement learning. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 2–8 December 2018.
37. Chen, X.; Karimi, B.; Zhao, W.; Li, P. On the Convergence of Decentralized Adaptive Gradient Methods. *arXiv* **2021**, arXiv:2109.03194. Available online: https://ui.adsabs.harvard.edu/abs/2021arXiv210903194C (accessed on 26 May 2024).
38. Rao, J.J.; Ravulapati, K.K.; Das, T.K. A simulation-based approach to study stochastic inventory-planning games. *Int. J. Syst. Sci.* **2003**, *34*, 717–730. [CrossRef]
39. Dinh Thai, H.; Nguyen Van, H.; Diep, N.N.; Ekram, H.; Dusit, N. Markov Decision Process and Reinforcement Learning. In *Deep Reinforcement Learning for Wireless Communications and Networking: Theory, Applications and Implementation*; Wiley-IEEE Press: Hoboken, NJ, USA, 2023; pp. 25–36.
40. Bakker, H.; Dunke, F.; Nickel, S. A structuring review on multi-stage optimization under uncertainty: Aligning concepts from theory and practice. *Omega* **2020**, *96*, 102080. [CrossRef]
41. Liu, K.; Yang, L.; Zhao, Y.; Zhang, Z.-H. Multi-period stochastic programming for relief delivery considering evolving transportation network and temporary facility relocation/closure. *Transp. Res. Part E Logist. Transp. Rev.* **2023**, *180*, 103357. [CrossRef]
42. Kamyabniya, A.; Sauré, A.; Salman, F.S.; Bénichou, N.; Patrick, J. Optimization models for disaster response operations: A literature review. *OR Spectr.* **2024**, *46*, 1–47. [CrossRef]
43. Rockafellar, R.T. *Convex Analysis*; Princeton University Press: Princeton, NJ, USA, 1997. (In English)
44. Dowd, K.; Cotter, J. Spectral Risk Measures and the Choice of Risk Aversion Function. *arXiv* **2011**, arXiv:1103.5668.
45. Borkar, V.S. A convex analytic approach to Markov decision processes. *Probab. Theory Relat. Fields* **1988**, *78*, 583–602. [CrossRef]
46. Nguyen, N.D.; Nguyen, T.T.; Vamplew, P.; Dazeley, R.; Nahavandi, S. A Prioritized objective actor-critic method for deep reinforcement learning. *Neural Comput. Appl.* **2021**, *33*, 10335–10349. [CrossRef]
47. Li, J.; Fridovich-Keil, D.; Sojoudi, S.; Tomlin, C.J. Augmented Lagrangian Method for Instantaneously Constrained Reinforcement Learning Problems. In Proceedings of the 2021 60th IEEE Conference on Decision and Control (CDC), Austin, TX, USA, 14–17 December 2021; pp. 2982–2989.
48. Boland, N.; Christiansen, J.; Dandurand, B.; Eberhard, A.; Oliveira, F. A parallelizable augmented Lagrangian method applied to large-scale non-convex-constrained optimization problems. *Math. Program.* **2019**, *175*, 503–536. [CrossRef]
49. Yu, L.; Yang, H.; Miao, L.; Zhang, C. Rollout algorithms for resource allocation in humanitarian logistics. *IISE Trans.* **2019**, *51*, 887–909. [CrossRef]
50. Rodríguez-Espíndola, O. Two-stage stochastic formulation for relief operations with multiple agencies in simultaneous disasters. *OR Spectr.* **2023**, *45*, 477–523. [CrossRef]
51. Zhang, L.; Shen, L.; Yang, L.; Chen, S.; Wang, X.; Yuan, B.; Tao, D. Penalized Proximal Policy Optimization for Safe Reinforcement Learning. *arXiv* **2022**, arXiv:2205.11814, 3719–3725.
52. Ding, S.; Wang, J.; Du, Y.; Shi, Y. Reduced Policy Optimization for Continuous Control with Hard Constraints. *arXiv* **2023**, arXiv:2310.09574.
53. Wang, Z.; Shi, X.; Ma, C.; Wu, L.; Wu, J. *CCPO: Conservatively Constrained Policy Optimization Using State Augmentation*; IOS Press: Amsterdam, The Netherlands, 2023.
54. Peng, X.B.; Abbeel, P.; Levine, S.; Panne, M.V.D. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Trans. Graph.* **2018**, *37*, 143. [CrossRef]
55. Tamar, A.; Castro, D.D.; Mannor, S. Policy gradients with variance related risk criteria. In Proceedings of the 29th International Coference on International Conference on Machine Learning, Edinburgh, UK, 26 June–1 July 2012.
56. Tamar, A.; Mannor, S. Variance Adjusted Actor Critic Algorithms. *arXiv* **2013**, arXiv:1310.3697.
57. Dowson, O.; Kapelevich, L. SDDP.jl: A Julia Package for Stochastic Dual Dynamic Programming. *INFORMS J. Comput.* **2021**, *33*, 27–33. [CrossRef]
58. Boda, K.; Filar, J.A. Time Consistent Dynamic Risk Measures. *Math. Methods Oper. Res.* **2006**, *63*, 169–186. [CrossRef]
59. Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; Moritz, P. Trust Region Policy Optimization. In Proceedings of the 32nd International Conference on Machine Learning, Proceedings of Machine Learning Research, Lille, France, 6–11 July 2015. Available online: https://proceedings.mlr.press/v37/schulman15.html (accessed on 26 May 2024).
60. Gillies, A.W. Some Aspects of Analysis and Probability. *Phys. Bull.* **1959**, *10*, 65. [CrossRef]
61. Van Wassenhove, L.N. Humanitarian aid logistics: Supply chain management in high gear. *J. Oper. Res. Soc.* **2006**, *57*, 475–489. [CrossRef]
62. Yu, L.; Zhang, C.; Yang, H.; Miao, L. Novel methods for resource allocation in humanitarian logistics considering human suffering. *Comput. Ind. Eng.* **2018**, *119*, 1–20. [CrossRef]