Improved Monte Carlo Planning via Causal Disentanglement for Structurally-Decomposed Markov Decision Processes

Abstract

Markov Decision Processes (MDPs), as a general-purpose framework, often overlook the benefits of incorporating the causal structure of the transition and reward dynamics. For a subclass of resource allocation problems, we introduce the *Structurally Decomposed* MDP (`SD-MDP`), which leverages causal disentanglement to partition an MDP's temporal causal graph into independent components. By exploiting this disentanglement, `SD-MDP` enables dimensionality reduction and computational efficiency gains in optimal value function estimation. We reduce the sequential optimization problem to a fractional knapsack problem with log-linear complexity $\mathcal{O}(T \log T)$, outperforming traditional stochastic programming methods that exhibit polynomial complexity with respect to the time horizon T. Additionally, `SD-MDP` computational advantages are independent of state-action space size, making it viable for high-dimensional spaces. Furthermore, our approach integrates seamlessly with Monte Carlo Tree Search (MCTS), achieving higher expected rewards under constrained simulation budgets while providing a vanishing simple regret bound. Empirical results demonstrate superior policy performance over benchmarks across various logistics and finance domains.

1 INTRODUCTION

While Markov Decision Processes (MDPs) offer a comprehensive framework for many sequential decision-making problems under uncertainty, certain problem structures and assumptions allow for simplified approaches that avoid the full complexity of an standard MDP formulation. For instance, in linear quadratic Gaussian control problems, the optimal control policy has a reduced form due to the equivalence principle [AM07]. In finance, under log utility and geometric Brownian motion asset dynamics, the optimal investment strategy has an explicit closed-form solution in some instances of financial derivatives [ZY03]. Economic models with rational expectations and additive components can often leverage the certainty equivalence principle. This means that separating deterministic and stochastic components can simplify the model, and provide pathways to derive error bounds for Monte Carlo (MC) estimation algorithms. Sampling-based approaches that avoid modelling the full probability distribution can be carefully adopted to provide tractable solutions in various stochastic control applications while retaining key problem characteristics [CF12].

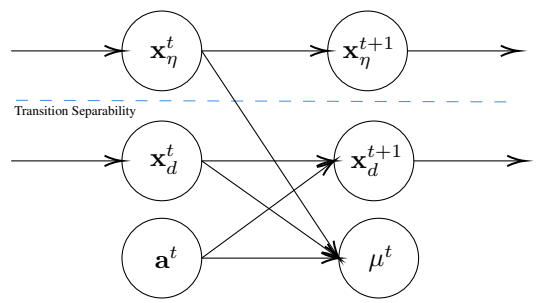


Figure 1: **Causal Structure & Partitioning of the `SD-MDP`:** The `SD-MDP` splits transition dynamics into stochastic component $\mathbf{x}_\eta^t$ and deterministic $\mathbf{x}_d^t$. The reward μ^t is driven by both partitions, and the action $\mathbf{a}^t$.

Specifically, we focus on the problem of resource allocation over a finite time horizon. Traditionally, resource allocation problems were solved using multi-stage stochastic programming or formulating the problem as an MDP and applying some form of MDP solver, such as approximate dynamic programming, for large-scale problems [WW11] [Pow+05] [DV04] [BDG00]. Nevertheless, these traditional methods are often very specific to the problem setting and do not generalize to a class of similar problems—they often require a full reformulation [Kus90] [Pow+05]. Furthermore, they do not take into account the causal structure of the MDP to obtain computational simplifications [ZBD10]. Similar to energy conservation principles in physics, we impose a construct which we denote as a *resource-utility* exchange model

Submitted to the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025). **To be used for reviewing only**.

(defined in Sec. 2.1) [Hau79]. In this model, resources can be converted to utility and vice versa, subject to certain environmental constraints.

Standard approaches to optimal planning include policy iteration, value iteration, approximate dynamic programming, and deep reinforcement learning [Ber11] [Sil+17] [FSM10]. [Meu+98] decompose a large MDP into smaller, independently solvable MDPs under resource constraints. These sub-problems, guided by heuristic solutions, lack theoretical convergence guarantees, potentially leaving an optimality gap. [BL16] introduce a budgeted MDP that partitions a single resource across tasks but does not consider converting multiple resources for a single task. [Car+19] reformulate constrained MDPs (CMDP) by transforming value and Q-functions into two dimensions—reward outcomes and constraint values—solving the problem via expectation maximization. However, incorporating additional Q-functions increases state-action space dimensionality, adding complexity. Furthermore, we aim to design a framework which specifically enables the ease-of-substitution of one resource to another to accomplish a single task (i.e. hybrid vehicles etc.) - where previous CMDP frameworks typically consider the exchange of a single resource to multiple objectives [Meu+98; BL16; Alt21].

Recent research has explored leveraging causal knowledge as side information to uncover the causal structure of MDPs. This involves analyzing how state space components, transitions, and rewards arise from the interaction between the system and the agent over time. By applying *causal disentanglement* to the MDP structure, we can simplify computations for MDP solvers [LMT22; BJS21; RB+22]. Disentangling and simplifying the causal structure of an MDP enhances computational efficiency by enabling separability in the search space [LMT22].

In this paper, we introduce the framework for the rigorous modelling of subclass of MDP's through a structured decomposition via side information corrsponding to the temporal causal behaviour of the system. Contrasting with previous works in CMDP's, our MDP framework is designed to integrate seamlessly into Monte Carlo planning algorithms, such as Monte Carlo tree search (MCTS), while ensuring convergence to the optimal solution. To be specific, the framework first disentangles the stochastic environmentally induced state transitions and deterministic action-driven reward functions, as illustrated in Fig. 1. This separation enables independent optimization of components the agent can model perfectly with lower complexity, while computing expectations over stochastic outcomes separately, improving efficiency and simplifying theoretical guarantees on value approximations [Gen+20] [Tod09]. Moreover, it allows us to provide theoretical guarantees on value function estimates via Monte Carlo (MC) value iteration.

To provide a concise overview of the paper, in Sec. 2.1, we explore how the action and transition dynamics of MDP's can inform us about the structure of the optimal solution, motivating the formal definition of the `SD-MDP`. Next, in Sec. 3, we describe the integration of this knowledge into well-known MC planning algorithms, such as UCT [KS06] and MENTS [Xia+19], and offer theoretical guarantees on the upper bound of simple regret under mild assumptions. Lastly in Sec. 4, we present several empirical results demonstrating the efficacy of our methodology. These results demonstrate that our methodology, achieves higher expected rewards (or lower costs) under fixed simulation budgets compared to vanilla MCTS, while outperforming established instance-dependent baselines across various applied domains in engineering and economics.

We provide a new framework for a subclass of MDP's to reduce computational complexity and improve value approximation results via causal disentanglement. We summarize our contributions as follows:

[colback=green!5!white, colframe=black, arc=4pt, boxsep=0.3pt,]**Contribution 1:** We leverage *causal disentanglement* to partition a compliant MDP's temporal causal graph into independent components to enable dimensionality reduction and computational efficiency gain. [colback=blue!5!white, colframe=black, arc=4pt, boxsep=0.3pt,] **Contribution 2:** We showcase a reduction of sequential optimization under perfect information to a fractional knapsack problem of complexity $\mathcal{O}(T\log(T))$ outperforming traditional stochastic programming methods with polynomial scaling with respect to T. [colback=red!5!white, colframe=black, arc=4pt, boxsep=0.3pt,] **Contribution 3:** We provide a seamless integration with MCTS and theoretical guarantees on a vanishing simple regret bound, supported by empirical benchmarks in logistics, control, and finance problems.

2 PROBLEM DEFINITION

Classical MDP: Let a well-defined general discrete time MDP be represented as $\mathcal{M} = (\mathbf{x}^1, \mathcal{X}, \mathcal{A}, P, \mu)$, where $\mathcal{X}$ is the set of states, $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ...\}$, $\mathcal{A}$ is the set of actions, $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, ...\}$, $\mathbf{x}^1$ is the initial state of the system, P represents the state transition probabilities, $P(\mathbf{x}^{t+1} \mid \mathbf{x}^t, \mathbf{a}^t)$, the probability of transitioning to state $\mathbf{x}^{t+1}$ given action $\mathbf{a}$ at state $\mathbf{x}$, and $\mu(\cdot)$ is the reward function, $\mu(\mathbf{x}, \mathbf{a})$, the immediate reward upon taking action $\mathbf{a}$ at state $\mathbf{x}$ at time t. The objective of our optimization is to obtain a policy π, which maps states to actions, that maximizes the expected cumulative reward,

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \mu(\mathbf{x}^{\mathrm{t}}, \mathbf{a}^{\mathrm{t}}) \middle| \mathbf{x}^1, \pi\right]. \tag{2.1}$$

This objective aims to identify the policy that maximizes the expected sum of rewards over a finite time horizon over randomness induced by the MDP parameters θ, where the expectation is taken over the randomness in the transition dynamics and policy when it is stochastic. Here, $\mathbf{x}^t$ represents the state at time t, $\mathbf{a}^t$ denotes the action taken at time

t, and $\mathbf{x}^{t+1}$ denotes the next state at $t+1$. It is important to note that negative rewards are also possible, especially in problem settings where minimizing costs is the goal.

2.1 THE SD-MDP FRAMEWORK

We define a special class of MDP's, termed *structurally-decomposed MDP* (SD-MDP). This constitutes a stochastic reduction on the classical MDP, partitioning it into various components driven by the *causal dynamics* and Markovian properties. From the perspective of causal reinforcement learning [LMT22], the SD-MDP partitions the state transition dynamics via the causal relation of the intervening action. This side information pertaining to the causal dynamics of the MDP allows us obtain more efficient MC value estimates, suitable for stochastic planning problems. To be specific, this allows the state transition to be modelled separately and independent of the reward dynamics.

Causal Disentanglement: The process of identifying and separating the underlying causal factors that generate observed data enables a clearer understanding of the underlying causal structure [RB+22; Kom+23]. We apply this concept of *causal disentanglement* to the SD-MDP to isolate the causal effect of actions $\mathbf{a}^t$ on the state transition $\mathbf{x}_\eta^t \to \mathbf{x}_\eta^{t+1}$, as illustrated in Fig. 1.

Formal Definition: The SD-MDP is represented as $(\mathcal{X}, \mathcal{A}, \mathcal{R}, P, \mathbf{x}^1)$, where $\mathcal{X}$ denotes the state space; $\mathcal{A}$ denotes the action space, and is of dimension $D \in \mathbb{N}$ (where $\mathbb{N}$ denotes the set of counting numbers); $\mathcal{R} \subseteq \mathbb{R}$ denotes the reward space; P is the transition function for $\mathbf{x} \in \mathcal{X}$, and $\mathbf{x}^1$ is the initial state. The SD-MDP integrates both deterministic ($\mathbf{x}_d$) and environmentally driven ($\mathbf{x}_\eta$) state components, the combination of which defines an MDP state, $\mathbf{x} = [\mathbf{x}_\eta, \mathbf{x}_d]^T$. To standardize notation, $\mathbf{x} \in \mathcal{X}$ is decomposed into $\mathbf{x}_\eta \in \mathcal{X}_\eta$ and $\mathbf{x}_d \in \mathcal{X}_d$. At face value, this model is similar to the restless bandit problem [Git79], aiming to maximize cumulative expected rewards within a finite time frame for environmentally changing state transitions. Unlike a classical restless bandit, due to constraints on $\mathbf{x}_d$ (we later illustrate what such constraints are in Table 1), reward outcomes must be planned over the complete time horizon T, rather than maximizing at each given opportunity, under perfect information or otherwise.

In particular, we partition the state vector representation into a deterministic partition, $\mathbf{x}_d$, and an independent stochastic partition, $\mathbf{x}_\eta$, both exhibiting different properties when subject to an intervention (or action) $\mathbf{a}^t$. The stochastic transitions governed by P are independent of the action taken. The transition probabilities can be expressed as,

$$P(\mathbf{x}_d^{t+1} | \mathbf{a}^t, \mathbf{x}^t) \in \{0, 1\}, \tag{2.2}$$

$$P(\mathbf{x}_\eta^{t+1} | \mathbf{a}^t, \mathbf{x}^t) = P(\mathbf{x}_\eta^{t+1} | \mathbf{x}_\eta^t), \tag{2.3}$$

$$P(\mathbf{x}^{t+1} | \mathbf{a}^t, \mathbf{x}^t) = P(\mathbf{x}_d^{t+1} | \mathbf{a}^t, \mathbf{x}^t) P(\mathbf{x}_\eta^{t+1} | \mathbf{x}_\eta^t), \tag{2.4}$$

where Eq. (2.2) represents if the future deterministic component $\mathbf{x}_d^{t+1}$ is reached by taking action $\mathbf{a}^t$. Eq. (2.3) represents the natural transition of the stochastic partition $\mathbf{x}_\eta^t$ independent of $\mathbf{a}^t$, and Eq. (2.4) represents the combined probability of transition for the SD-MDP. The state dynamics of the SD-MDP are composed of partitionable components, which include both stochastic and deterministic elements. The stochastic components evolve independently of the agent's actions (for example such as the price of certain financial assets). In contrast, the deterministic components evolve causally driven by the agent's actions (for example incremental adjustments to inventory levels).

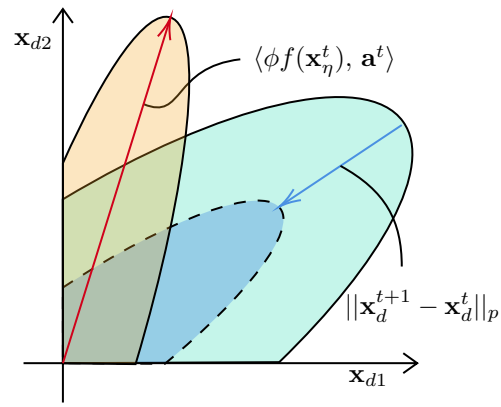


Figure 2: **Norm-Capacity Dynamics:** As the capacity of $\mathbf{x}_d$ shrinks given the constraints of the norm-capacity, the consumption of resource can be transformed into a reward $\langle \phi f(\mathbf{x}_\eta^t), \mathbf{a}^t \rangle$. The blue shading represents shrinkage of the the resource capacity, and the orange shading represents the vector space of possible outcomes, the magnitude of this vector (represented by the red arrow) represents the reward.

Resource Utility Exchange: To allow for a general model of resource consumption and utility exchange, we use $f(\cdot)$ and $g(\cdot)$ to denote coordinate-wise separable functions composed of a series of smooth weakly monotone Lipschitz functions governing the dimension-wise scaling of each dimension when an action is taken by the agent. To be specific $f : \mathbb{R}^D \to \mathbb{R}^D$ and $g : \mathbb{R}^D \to \mathbb{R}^D$ are coordinate-wise separable. For a D dimensional vector, both f and g are any weakly monotonic functions which,

$$f(\mathbf{x}) \equiv [f_1(\mathbf{x}_1), f_2(\mathbf{x}_2), \ldots, f_D(\mathbf{x}_D)]^T, \tag{2.5}$$

$$g(\mathbf{x}) \equiv [g_1(\mathbf{x}_1), g_2(\mathbf{x}_2), \ldots, g_D(\mathbf{x}_D)]^T. \tag{2.6}$$

To motivate, f represents the rate of utility gain, while g represents the rate of resource consumption, both depending on the context $\mathbf{x}_\eta$. A very basic example could be the exchange of fuel to mileage (as illustrated in Sec. 4).

In Table 1, we provide a list of the underlying dynamics that govern the behaviour of the SD-MDP. To begin, an agent may have a particular resource that they are consuming over time (money, fuel, battery etc.). This resource can be converted to rewards for the agent. First, this motivates Dynamic (**D1**), which ensures a valid representation of multi-dimensional resource capacity consumption over time, as illustrated in Fig. 2. We impose the constraint of a strictly

Definition	Expression	
(D1) Positive Action & Capacity Space: We assume strictly positive action and capacity spaces.	$\mathbf{a} > \mathbf{0}, \mathbf{x}_d \geq \mathbf{0}$	(2.7)
(D2) General Linear Reward Dynamics: The reward function $\mu(\mathbf{a}^t, \mathbf{x}^t)$ obeys a linear relationship w.r.t. action $\mathbf{a}^t$ and stochastic state partition $\mathbf{x}^t$.	$\mu(\mathbf{a}^t, \mathbf{x}^t) = \langle \phi f(\mathbf{x}_\eta^t),\, \mathbf{a}^t \rangle$	(2.8)
(D3) Incremental Action Dynamics: We define a linear transformation matrix ϕ', which is anti-parallel to ϕ. To model the expansion and contraction of the capacity $\mathbf{x}_d$, we impose constraints on the transition function acting on $\mathbf{x}_d$ in Eq. (2.9) and Eq. (2.10).	$\|\|\mathbf{x}_d^{t+1} - \mathbf{x}_d^t\|\|_p = \|\|\langle \phi' g(\mathbf{x}_\eta^t),\, \mathbf{a}^t \rangle\|\|_p$ $\underline{\Delta}_a(t) \leq \|\|\mathbf{x}_d^{t+1} - \mathbf{x}_d^t\|\|_p \leq \bar{\Delta}_a(t),\, \forall a \in \mathcal{A}$	(2.9) (2.10)
(D4) Capacity Objective: The accumulation of resources, as measured by $\|\|\langle \phi' g(\mathbf{x}_\eta^t),\, \mathbf{a}^t \rangle\|\|_p$, should meet a predetermined maximum and minimum goals.	$\underline{A} \leq \sum_{t=1}^{T} \left\|\langle \phi' g(\mathbf{x}_\eta^t),\, \mathbf{a}^t \rangle\right\|_p \leq \bar{A}$	(2.11)
(D5) Recency Preference: Ordinal preference of equivalent states w.r.t. to t.	$\mathbf{x}_\eta^t = \mathbf{x}_\eta^{t+\Delta} \implies \mathbf{x}_\eta^t \succ \mathbf{x}_\eta^{t+\Delta}, \quad \Delta \in \mathbb{Z}$	(2.12)

Table 1: Summary of the system dynamics of the `SD-MDP`.

element-wise positive action space, $\mathbf{a} > \mathbf{0}$. Additionally, the capacity space is also subject to a similar constraint, ensuring each component of the capacity vector $\mathbf{x}_d$ is non-negative, i.e., $\mathbf{x}_d \geq \mathbf{0}$.

Dynamic **(D2)** stipulates that the `SD-MDP` obeys a reward function of a general linear form. Action $\mathbf{a}^t$, together with the stochastic state partition $\mathbf{x}_\eta^t$, invokes a deterministic reward outcome with a linear relation, $\mu(\mathbf{a}^t, \mathbf{x}^t)$. Let $\mu(\cdot) : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{R}$ denote a standard map that yields a scalar in $\mathbb{R}$ when provided with inputs $\mathbf{a} \in \mathcal{A}$ and $\mathbf{x}_\eta^t \in \mathcal{X}_\eta$, subject to constraints on the system at time t. Next, we employ a linear transformation on $f(\mathbf{x}_\eta^t)$, with a positive semi-definite matrix ϕ. This homogeneous scaling map allows for both enlargement and shrinking of the vector along the positive dimensions. The reward function results from an inner product between the transformed vector $\phi\, f(\mathbf{x}_\eta^t)$ and $\mathbf{a}^t$, as expressed in Eq. (2.8). Furthermore, the dimension of $\mathcal{A}$ is $\dim(\mathcal{A}) = D$, which must also be equal to the dimension of $\phi f(\mathbf{x}_\eta) \in \phi\mathcal{X}$ where $\dim(\phi\mathcal{X}) = D$.

Dynamic **(D3)** governs resource consumption incrementally. We define a linear transformation matrix ϕ', which is anti-parallel to ϕ. Similarly, we apply function $g : \mathbb{R}^D \to \mathbb{R}^D$, to model the expansion and contraction of the capacity $\mathbf{x}_d$. We impose the transition function acting on $\mathbf{x}_d$ in Eq. (2.9). Where $\Delta_d(t) : \{1, ..., T\} \to \mathbb{R}$ is a natural discrete change on $\mathbf{x}_d^t$ as deterministically determined by the system, and $\langle \phi' g(\mathbf{x}_\eta^t),\, \mathbf{a}^t \rangle)$ is the contribution to the expansion or contraction of $\mathbf{x}_d^t$ based on the agent's action taken at $\mathbf{a}^t$ taken at time t. We impose a constraint on the magnitude of capacity change per time interval via Eq. (2.10), where constraints $\underline{\Delta}_a(t)$ and $\overline{\Delta}_a(t)$ are given by the system.

Dynamic **(D4)** enforces a *path constraint* on the trajectory of actions that the agent can take. We constrain this trajectory by limiting the accumulation of actions measured with p-norms. As defined, the accumulation of resources $\langle \phi' g(\mathbf{x}_\eta^t),\, \mathbf{a}^t \rangle)$ should meet some maximum and minimum goals, as expressed in Eq. (2.11).

Dynamic **(D5)** posits that the value of receiving the exact same reward sooner is more valuable to the agent than receiving it later without the need for an explicit discount factor. To further expound, using a preference to break any ties should any policy lead to the same reward outcome, which is useful for tie-breaking under identical outcomes.

2.2 VALUE ESTIMATION PROPERTIES

We provide an intuitive analysis on the behivour of the optimal policy. In the final state at T, the deterministic property ensures from Dynamic **(D2)**, at the value of the final state T, can be computed via pure exploitation, by taking the maximum allowable action at time T according to the constraints from Dynamics **(D3)** and **(D4)**. Thus, we express the value function as,

$$V(\mathbf{x}^T) = \max_{\mathbf{a} \in \mathcal{A}(T)} \mu(\mathbf{x}^T, \mathbf{a}). \tag{2.13}$$

We examine the problem through induction. Consider that the agent is at time $T-1$ and would like to obtain the value estimate for time T. We express the *conditional value function* $V(\mathbf{x}^T | \mathbf{x}^{T-1})$ as,

$$V(\mathbf{x}^T | \mathbf{x}^{T-1}) = \langle \phi f(\mathbb{E}[\mathbf{x}_\eta^T | \mathbf{x}_\eta^{T-1}]), \mathbf{a}^* \rangle. \tag{2.14}$$

To obtain the optimal value of $\mathbf{a}^*$, ideally the agent performs the optimal action to yield the highest reward at time T. This however, depends on the capacity constraints of the action sequence, which must obey constraints Eq. (2.10) and (2.11). We can thus express the value function at $T-1$, subject to the incremental dynamics, and goal constraints as,

$$V(\mathbf{x}^{T-1}) = \max_{\mathbf{a} \in \mathcal{A}(T-1)} \Big\{ \langle \phi f(\mathbf{x}_\eta^{T-1}), \mathbf{a} \rangle + \int_{\mathbf{x}^T} P_\theta(\mathbf{x}^T | \mathbf{a}, \mathbf{x}^{T-1})\, V(\mathbf{x}^T)\, d\mathbf{x} \Big\}. \tag{2.15}$$

where $\mathcal{A}(T-1)$ represents the set of actions available to the agent at time $T-1$, as governed by Dynamics **(D3)** and **(D4)**. Assuming capacity is available at time T, given special properties of the problem, we can partition, assuming $g_\eta(\mathbf{a})$

is deterministic and consider introduction of $\Delta_d(t)$,

$$V(\mathbf{x}^{T-1}) = \max_{\mathbf{a}\in\mathcal{A}(T-1)} \Big\{ \langle \phi f(\mathbf{x}_\eta^{T-1}), \mathbf{a}\rangle \tag{2.16}$$
$$+ \langle (\mathbf{x}_d^{T-1} + \langle \phi' g(\mathbf{x}_\eta^t),\, \mathbf{a}^t\rangle),\, \mathbb{E}[\phi f(\mathbf{x}_\eta^T)|\mathbf{x}_\eta^{T-1}]\rangle \Big\}.$$

For non trivial solutions to Eq. (C.5), we adhere to the *incremental action dynamic* (**D3**) property of the SD−MDP. The binary structure of the optimal policy becomes apparent at $T-1$. (Please see derivation in Appendix C.1.)

2.3 STRUCTURE OF THE OPTIMAL POLICY

Let $\tau_a \equiv (\mathbf{a}^{i=1}, \mathbf{a}^{i=2}, \mathbf{a}^{i=3}, \ldots, \mathbf{a}^{i=t})$ denote a sequence of $\mathbf{a}$ from 1 to t. Further, let us denote the operators,

$$\underline{\aleph}^t[\tau_a] \equiv \tilde{T}\underline{\Delta}_a(t) + \sum_i^{t-1} ||\langle \phi' g(\mathbf{x}_\eta^i),\, \mathbf{a}^i\rangle||_p - \bar{A} \tag{2.17}$$

$$\overline{\aleph}^t[\tau_a] \equiv \tilde{T}\bar{\Delta}_a(t) + \sum_i^{t-1} ||\langle \phi' g(\mathbf{x}_\eta^i),\, \mathbf{a}^i\rangle||_p - \underline{A} \tag{2.18}$$

Where $\tilde{T}$ represents $T-t+1$. Intuitively, $\overline{\aleph}^t[\tau_a]$ and $\underline{\aleph}^t[\tau_a]$ represent the maximum and minimum allowable consumption under the path constraint in Eq. (2.11) at time t. Moving forward let, $\mathcal{A}(t)$ denote the action set at time t, given the constraints from equations Eq. (2.10) and (2.11), such that the expression $\mathbf{a} \in \mathcal{A}(t)$ encapsulates the constraints from all action dynamics pertaining to the SD−MDP.

$$\mathcal{A}(t) \equiv \Big\{ \mathbf{a} : \underline{\mathfrak{A}}(t) \le ||\langle \phi' g(\mathbf{x}_\eta^t),\, \mathbf{a}^t\rangle||_p \le \overline{\mathfrak{A}}(t) \Big\} \tag{2.19}$$

$$\underline{\mathfrak{A}}(t) = \max\Big\{ \underline{\aleph}^t[\tau_a],\ \underline{\Delta}_a(t) \Big\} \tag{2.20}$$

$$\overline{\mathfrak{A}}(t) = \min\Big\{ \overline{\aleph}^t[\tau_a],\ ||\mathbf{x}_d^t||_p,\ \bar{\Delta}_a(t) \Big\} \tag{2.21}$$

Intuitively, $\underline{\Delta}_a(t)$ and $\bar{\Delta}_a(t)$ constitute the minimum and maximum incremental capacity specified by the system. To note in Eq. (2.19), as $||\mathbf{x}_d^t||_p$ lower bounded by 0, and we can omit 0 from the set. The *incremental action dynamic* (**D3**) forms a constraint on the capacity from the deterministic component of the SD−MDP. Along with the goal constraint of the system, $\underline{\aleph}^t[\tau_a]$ and $\overline{\aleph}^t[\tau_a]$ form a bound on the admissible actions at time t, denoted as $\mathcal{A}(t)$.

$$\{\mathbf{a}^+\} = \operatorname*{argmax}_{\mathbf{a}\in\mathcal{A}(t)} ||\mathbf{a}||_p,\ \{\mathbf{a}^-\} = \operatorname*{argmin}_{\mathbf{a}\in\mathcal{A}(t)} ||\mathbf{a}||_p \tag{2.22}$$

Given that ϕ and ϕ' are antiparallel linear maps on $\mathbf{a}$, the solutions of $\{\mathbf{a}^+\}$ and $\{\mathbf{a}^-\}$ constitute a linear optimization problem. $\{\mathbf{a}^+\}$ corresponds to the solution which exploits the maximum achievable reward at time t as expressed in Eq. (2.8), and $\{\mathbf{a}^-\}$ expresses the action conserves the minimizes the consumption of the capacity for the future. Given $\mathcal{A}(t)$, at any time t, there exists two sets $\{\mathbf{a}^+\}$, and $\{\mathbf{a}^-\}$ which either maximizes allowable reward, or maximally reduces consumption of resource $\mathbf{x}_d$. Let us denote $\mathbf{a}^+[\mathbf{x}_\eta]$ and $\mathbf{a}^-[\mathbf{x}_\eta]$ as the following,

$$\mathbf{a}^+[\mathbf{x}_\eta] = \operatorname*{argmax}_{\mathbf{a}\in\{\mathbf{a}^+\}} \langle \phi f(\mathbf{x}_\eta),\, \mathbf{a}\rangle, \tag{2.23}$$

$$\mathbf{a}^-[\mathbf{x}_\eta] = \operatorname*{argmax}_{\mathbf{a}\in\{\mathbf{a}^-\}} \langle \phi f(\mathbf{x}_\eta),\, \mathbf{a}\rangle. \tag{2.24}$$

In Lem. 2.1, we show that the optimal policy consists of an action, represented as a vector, corresponding to one of two sets $\{\mathbf{a}^+\}$ or $\{\mathbf{a}^-\}$, each with dimension D. A continuous action space MDP thereby reduces to a sequential discrete action decision problem, where the action space forms a finite dimension subspace with a maximum cardinality of with at most $2D$.

Lemma 2.1. ***Finite and Bounded Action Space for the SD−MDP:*** *For the SD−MDP, for any action taken in the finite time horizon, optimal policy lies to the union of 2 subspaces, that is $\mathbf{a}^* \subset \{\mathbf{a}^+\} \cup \{\mathbf{a}^-\} \subset \mathcal{A}(t) \subseteq \mathcal{A}$, for all time steps t. The cardinality of the dimension of the optimal solution space is upper bounded by $2D$. (Proof in Appendix A.1.)*

Let $\tau_\eta \equiv \{\mathbf{x}_\eta^t, \mathbf{x}_\eta^{t+1}, \mathbf{x}_\eta^{t+2}, \ldots, \mathbf{x}_\eta^T\}$ denote a sequence of stochastic outcomes, the expectation of which is denoted as $\mathbb{E}[\tau_\eta] \equiv \{\mathbb{E}[\mathbf{x}_\eta^t], \mathbb{E}[\mathbf{x}_\eta^{t+1}], \mathbb{E}[\mathbf{x}_\eta^{t+2}], \ldots, \mathbb{E}[\mathbf{x}_\eta^T]\}$. We define the $\text{Top}_k(\mathbb{E}[\tau_\eta])$ for a series of multidimensional vectors be defined as,

$$\text{Top}_{k=T}(\tau_\eta) = (\mathbb{E}[\mathbf{x}_\eta^{k=1}], \mathbb{E}[\mathbf{x}_\eta^{k=2}], \ldots, \mathbb{E}[\mathbf{x}_\eta^{k=T}]) \tag{2.25}$$

where define with shorthand, $\widetilde{\tau} \equiv \text{Top}_{k=T}(\tau)$, such that,

$$\phi\, f(\widetilde{\tau}^1]) \succ \phi f(\widetilde{\tau}^2]), \cdots \succ \phi\, f(\widetilde{\tau}^T]) \tag{2.26}$$

where we index $\widetilde{\tau}^i \equiv \mathbb{E}[\mathbf{x}_\eta^i] \in \text{Top}_{k=T}(\tau_\eta)$.

Sketch of Proof: First we demonstrate the separability of $\mathbb{E}[\tau_\eta]$ with respect to any deterministic action sequence. The solution is therefore to find a maximizing solution for each $\mathbb{E}[\mathbf{x}_\eta^t] \in \mathbb{E}[\tau_\eta]$, which is possible under full information. Under incremental dynamics, $\underline{\Delta}_a(t) \le ||\langle \phi' g(\mathbf{x}_\eta^t),\, \mathbf{a}^t\rangle||_p \le \bar{\Delta}_a(t)$, only a limited amount of resources can be dedicated to maximizing each $\mathbb{E}[\tau_\eta]$, thus, the problem reduces to a fractional knapsack problem. We show that when we majorize over $\mathbb{E}[\tau_\eta]$, the optimal sequence, $(\mathbf{a}^*)$, to the sequence, $\mathbb{E}[\tau_\eta]$, is an order preserving union of two sequences, one corresponding to a maximizing vector $\mathbf{a}^+[\mathbf{x}_\eta]$ and one corresponding to a minimizing vector $\mathbf{a}^-[\mathbf{x}_\eta]$.

Lemma 2.2. ***Solving for Optimal Value via Top K Allocation:*** *For the SD−MDP, the optimal value can be obtained by solving the dual problem, which involves finding the value of k in $\text{Top}_k(\mathbb{E}[\tau_\eta])$ over $k \in \{1, \ldots, T\}$ possibilities. (Proof in Appendix A.2.)*

Sketch of Proof: Given the separability of $\mathbb{E}[\tau_\eta]$ with respect to any deterministic action sequence. We show that when we majorize over $\mathbb{E}[\tau_\eta]$, to produce an ordered set of sequences, we simply select the top k vectors in this ordered list which satisfies the norm maximization constraints for the resource allocation. For the rest of the $\mathbb{E}[\tau_\eta]$ we allo-

cate minimum resources within the constraints. The solution involves therefore simply finding the value of k which maximizes the value function in Eq. (2.27) subject to constraints derived from Eq. (2.19).

$$V_k(\mathbf{x}^t) = \sum_{i=1}^{k} \langle \phi f\left(\widetilde{\tau}^i\right), \mathbf{a}^+[\widetilde{\tau}^i] \rangle + \sum_{i=k+1}^{T} \langle \phi f\left(\widetilde{\tau}^i\right), \mathbf{a}^-[\widetilde{\tau}^i] \rangle \tag{2.27}$$

As shown by Lemmas 2.1 and 2.2, under perfect information where $\mathbb{E}[\tau_\eta]$ is known, the problem reduces from a complex sequential optimization to a fractional knapsack problem with computational complexity $\mathcal{O}(T\log(T))$ for exact solutions. This is a significant improvement over stochastic programming methods which scale polynomially with T^s, for number of scenarios $s > 1$ always [DS06; SN05]. In addition, our approach is independent of the state-action space size, which potentially be infinite, offering a strong advantage for many real-world problems over value or policy iteration, whose complexity scales polynomially with the state-action space [Win04; Sut18].

Monte Carlo Value Estimation for the `SD-MDP`: Thm. 1 states that any value function estimate using N MC estimate, will have a best case estimate error on the order of $\mathcal{O}(1/\sqrt{N})$. The key underlying approach is that we compute the expectation over the trajectory, $\mathbb{E}[\tau_\eta]$, via MC simulation. Via this approximation, we solve for an approximate optimal policy, which constitutes a pure strategy (deterministic policy) in a limited action space as a consequence of Lem. 2.1. This policy only depends on the discrete time position, allowing MC simulations to take place to estimate $V_k(\mathbf{x}^t)$.

Theorem 1. ***Upper bound on the Monte Carlo Value Estimation for the `SD-MDP`:*** *For the `SD-MDP` abstraction partitionable MDP, the optimal policy, where the value function is upper bounded by $|\hat{V}_N - V^*(\mathbf{x})| \leq \mathcal{O}((\delta\sqrt{N})^{-1})$, with probability $1-\delta$. Where $\hat{V}_N$ is the Monte Carlo simulation estimate of the value function under N iterations. (Proof in Appendix A.4.)*

Sketch of Proof: Any naturally evolving time series has an expected outcome which can be computed $\mathbb{E}[\tau_\eta]$, and thus the problem reduces to an allocation problem which can be solved using the dual formulation, in solving for $\text{Top}_k(\cdot)$ in Lem. 2.2. Via Hoeffding's inequality, we can upper bound the approximation error from Monte Carlo sampling.

3 MONTE CARLO PLANNING WITH VALUE FUNCTION APPROXIMATION

To yield improvements to MCTS within the `SD-MDP` framework, in the *Expansion* phase, knowledge from Lem. 2.1 is used to restrict the action space away from suboptimal actions. In the *Rollout* phase, drawing from knowledge from Lem. 2.2, a more efficient value function estimator is employed, obviating the need for an uninformed (typically uniform) rollout policy. Moreover, as shown by Thm. 1, we guarantee that for any $\hat{V}_N(\mathbf{x})$, as the simulation budget increases, $N \to \infty$, the approximation error $|\hat{V}_N(\mathbf{x}) - V^*(\mathbf{x})| \to 0$. We employ two variants of MCTS: Upper Confidence Tree (UCT) [KS06] and Maximum Entropy Monte Carlo Planning (MENTS) [Xia+19].

UCT: Leveraging bandit algorithms like UCB, as discussed in [CM07] and [KS06], MCTS efficiently approximates $\pi^*(\mathbf{x}, a)$ by navigating the state-action space via the UCT metric. Note $\bar{\mu}(t)$ is the reward for a trajectory, prior and after using uniform rollout.

$$\pi_{\text{UCT}}(\mathbf{x}) = \max_{a \in A} \bar{Q}(\mathbf{x}, \mathbf{a}) + c\sqrt{\frac{\log N(\mathbf{x})}{N(\mathbf{x}, \mathbf{a})}}, \tag{3.1}$$

$$\bar{Q}(\mathbf{x}^t, \mathbf{a}^t) \leftarrow \bar{Q}(\mathbf{x}^t, \mathbf{a}^t) + \frac{\bar{\mu}(t) - \bar{Q}(\mathbf{x}^t, \mathbf{a}^t)}{N(\mathbf{x}^t, \mathbf{a}^t) + 1}. \tag{3.2}$$

Softmax Entropy Policies (MENTS): Nevertheless, an alternative to UCT are softmax style Boltzmann policies, where the key difference is that a stochastic selection policy is used for action selection versus a deterministic policy. This encourages exploration and has been shown to be faster to converge compared to UCT. [Gri+20] elaborates on the use of AlphaZero and UCT, offering theoretical bounds such as $\hat{\pi} \leq \bar{\pi}$, where $\hat{\pi}$ represents empirical policy, $\bar{\pi}$ is a softmaxed policy balancing an empirical Q function, and π_θ denotes the supervised learning policy. The effectiveness of regularized MCTS, particularly in low sample count scenarios, is highlighted by [Gri+20].sample efficiency in bandit-based regret minimization algorithms. [Haz11] showed its effectiveness, achieving regret of $O(\log(T))$.

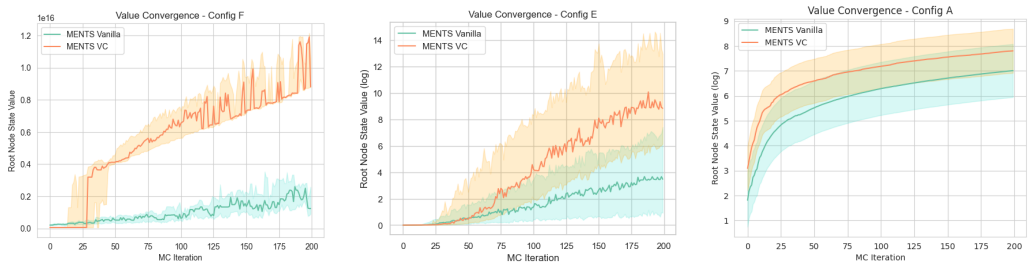
Maximum Entropy Monte Carlo Planning (MENTS): [Xia+19] proposes the use of a convex regularizer, which upper-bounds the value function estimate to improve the sampling efficiency of MCTS. In *maximum entropy MCTS* (MENTS), entropy is used to enhance exploration and convergence to the optimal policy for MDP planning. Furthermore, theoretical guarantees are also provided in [Xia+19] with respect to the suboptimality of the alogrithm over time. Let us define our *approximate Bellman update function*,

$$\mathcal{G}_P(\mathbf{x}, \mathbf{a}, V(\cdot)) = \mu(\mathbf{x}, \mathbf{a}) + \sum_{\mathcal{S}(\mathbf{x}', \mathbf{a})} \left(\frac{N(\mathbf{x}')}{N(\mathbf{x}, \mathbf{a})} V(\mathbf{x}') \right) \tag{3.3}$$

Given a tree with visited states P, and a value function estimator based on uniform rollout, $V_s(\cdot)$, in MENTS, there are two modes of updates,

$$Q_{\text{sft}}(\mathbf{x}^t, \mathbf{a}^t) = \begin{cases} \mathcal{G}_P(\mathbf{x}, \mathbf{a}, V_{\text{sft}}(\cdot)) & \text{if non-terminal in } P \\ \mathcal{G}_P(\mathbf{x}, \mathbf{a}, V_s(\cdot)) & \text{else} \end{cases} \tag{3.4}$$

Where $\mathcal{G}_P(\mathbf{x}, \mathbf{a}, V_{\text{sft}}(\cdot))$ represents the softmax value Q-update based on the softmax value function, and $\mathcal{G}_P(\mathbf{x}, \mathbf{a}, V_s(\cdot))$ which is the value function estimate obtained from a uniform rollout policy. The softmax value function, $V_{\text{sft}}(\cdot)$ is updated by a regularized function of the



(a) Maritime (b) Hybrid Fuel. (c) Financial Options.

Figure 3: We illustrate the convergence to the optimal value function as a function of the number of MC iterations for the MENTS algorithm [Xia+19]. We demonstrate that MENTS VC yields stronger value convergence properties compared to vanilla MENTS.

softmax Q function $Q_{\text{sft}}(\cdot)$.

$$V_{\text{sft}}(\mathbf{x}^t) \leftarrow \alpha \log \sum_{\mathbf{a}\in\mathcal{A}} \exp\left(\frac{1}{\alpha} Q_{\text{sft}}(\mathbf{x},\mathbf{a})\right) \tag{3.5}$$

$$\pi_{\text{M}}(\mathbf{a}|\mathbf{x}) = (1-\lambda_s)\frac{1}{\alpha}\left(Q_{\text{sft}}(\mathbf{s},\mathbf{a}) - V_{\text{sft}}(\mathbf{s})\right) + \frac{\lambda_s}{|\mathbf{a}|} \tag{3.6}$$

MENTS uses a soft Bellman update for $\hat{Q}_{\text{sft}}(\cdot)$, unlike the rollout policy via UCT. [Pai+24] suggests that MENTS is not consistent, meaning it will not always converge, and there may exist MDP's where MENTS fails to converge. Decaying entropy MCTS (DENTS) is an MCTS algorithm that guarantees convergence as $t \to \infty$, but it lacks strong guarantees regarding the probability of taking suboptimal actions ($P(\mathbf{a}^* \neq \mathbf{a})$), needed for proving Thm. 2.

3.1 VALUE CLIPPING

We study the problem of optimal Monte Carlo planning under perfect information. From Sec. 2.2 we can leverage the MC estimation properties of the value function approximator (VFA) to enhance any MDP solver. As we draw more samples our estimate of the model parameters increases, nevertheless the MC VFA still relies on a simulation budget to estimate the value function. We can further leverage the theoretical guarantees in Thm. 1 with respect to the MC estimation error of the optimal value function. $\overline{V}^*$ represents the maximum value estimate under perfect information, for actions in *hindsight*. $\underline{V}^*$ represents the value estimate under perfect information for the *anticipative* solution on expectation. Let us define,

$$\Delta_V(\mathbf{x}) = \overline{V}^*(\mathbf{x}) - \underline{V}^*(\mathbf{x}), \quad V^*(\mathbf{x}) \in [\underline{V}^*(\mathbf{x}), \overline{V}^*(\mathbf{x})] \tag{3.7}$$

Where the true solution $V^*(\mathbf{x})$ belongs somewhere between the optimal value in hindsight and the anticipative solution on expectation. We leverage $\overline{V}(\mathbf{x})^*$ and $\underline{V}(\mathbf{x})^*$ to improve the convergence rate of our planning algorithm. Therefore, given $\Delta_V(\mathbf{x})$ we can clip the outcome of any rollout policy in MCTS by $\Delta_V(\mathbf{x})$. Typically, when a new node is added to the search tree P, a uniform rollout policy or neural network is implemented to provide an initial estimate of the $V^*(\cdot)$.

Guarantees on Simple Regret: We can provide a guarantee on simple regret based on the structure of the `SD-MDP`, and extending the work of [Xia+19] for UCT and [KS06] for MENTS. Let simple regret be defined as,

$$\text{reg}(T) = \sup_{\mathbf{x}\in\mathcal{X}} \left(V^*(\mathbf{x}) - V^T(\mathbf{x})\right) \tag{3.8}$$

Where $V^T(\mathbf{x})$ is the value estimate of value function estimator after T samples. We provide high probability bounds on the simple regret for MCTS-MENTS, $\text{reg}_M(T)$ and MCTS-UCT $\text{reg}_U(T)$, where there exists some constant $C \in \mathbb{R}^+$ to bound the simple regret.

Theorem 2. ***Simple Regret:*** *Given a Monte Carlo planning algorithm, $\mathcal{M}$, where $\tilde{p}(T) = P(\mathbf{a}^* \neq \mathbf{a})$, where $\lim_{T\to\infty} \tilde{p}(T) = 0$ and $\tilde{p}(T)$ is asymptotically bounded above by $\mathcal{O}(\frac{1}{T})$ for T samples, when running $\mathcal{M}$ over the `SD-MDP`, the the simple regret $\text{reg}(T)$, as defined in Eq. (3.8), is bounded by $C_k T\tilde{p}(T)$, for some value $C_k \in \mathbb{R}^+$. (Proof in Appendix A.6.)*

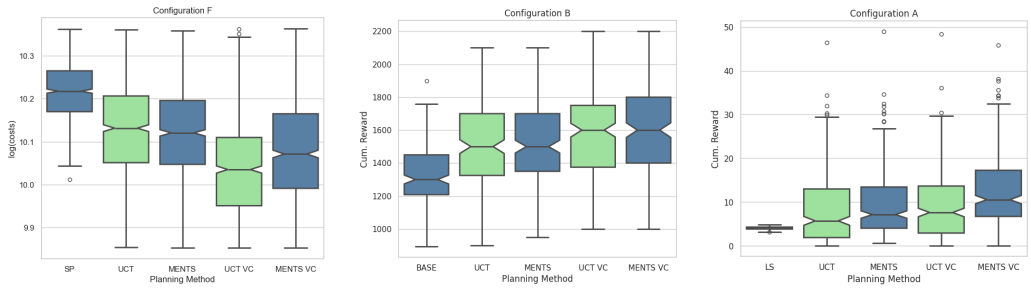
Sketch of Proof: The proof of $\text{reg}(T)$ bounds for MENTS and UCT for the `SD-MDP` begins with the identification that the optimal action $\mathbf{a}^*$ belongs to a discrete set. The $\text{reg}(T)$ is then analyzed based on the probability of action swaps, with the worst-case regret per swap denoted by $\tilde{\Delta}_a$. By quantifying the expectation over simple regret, we bound it using binomial probabilities and an upper-bounding polynomial function. As $T \to \infty$, the expectation over simple regret is shown to be upper-bounded by $\mathcal{O}(T^2 \exp(T/\log^3(T)))$ for MENTS and $\mathcal{O}(T^{1-\rho})$ for UCT.

As a consequence of Thm. 2, when running MENTS or UCT over the `SD-MDP`, the simple regret, $\text{reg}(T)$, is bounded by $\mathcal{O}(T^2 \exp(T/\log^3(T)))$ for MENTS, and $\mathcal{O}(T^{1-\rho})$ for UCT, as expressed in Eq. (3.9) and Eq. (3.10) respectively. Eq. (3.9) is a consequence of Thm. 5 from [Xia+19] and Eq. (3.10) is a consequence of Thm. 5 from [KS06].

$$\text{reg}_{\text{MENTS}}(T) \leq CT^2 \exp\left(T/\log^3(T)\right) \tag{3.9}$$

$$\text{reg}_{\text{UCT}}(T) \leq CT^{1-\rho} \tag{3.10}$$

The upper bound on the simple regret of UCT converges to zero as $T \to \infty$ for $\rho > 1$. Likewise, the simple regret of MENTS also vanishes asymptotically. A detailed outline



(a) Maritime (Cost Based.) (b) Hybrid Fuel. (c) Financial Options.

Figure 4: We compare empirical results based on cost reduction or reward maximization. The leftmost boxplot presents an instance-dependent baseline for reference. Evidently, MCTS value clipping within the `SD-MDP` framework improves expected cost/reward performance over vanilla MCTS, as shown for both UCT and MENTS variants.

of the value clipping implementation in combination with MENTS and UCT can be found in Appendix D.

4 EMPIRICAL RESULTS

We provide a series of empirical experiments to justify the efficacy of our algorithm. We further impose a computational constraint of the power of MCTS, such that the number of MCTS iterations $N \leq K_c(2D)^T,\ K_c = 0.1$, that is it is only possible to explore at most K_c percentage of all possible trajectories using MCTS, before making a decision. This constraint prevents us from overpowering MCTS in such a way that would allow it to brute force search all possible combinations, and must rely on efficient exploration. For all experiments, we compare our solution with an instance-dependent baseline solution, traditionally used to solve such problems. For MCTS we apply the selection strategies of both UCT, or MENTS, to provide comparative study. We record empirical results on cumulative reward/cost optimization, as well as improvement of the of the optimal value estimate at the root node. (The details for all experiments are outlined in Appendices E, F, & G respectively.)

Maritime Bunkering: Maritime bunkering is a logistical challenge in seaborn transportation aimed at minimizing fuel costs for fleets like ships by optimizing refuelling policies at ports or at sea. The planning problem is traditionally addressed through stochastic programming, which considers factors like fuel consumption, tank capacity, and price variability to find cost-effective refuelling strategies under fixed schedules. The solution via multi-stage stochastic programming is often computationally challenging.

Hybrid Fuel Consumption: In hybrid vehicles, a driving system dynamically switches between power sources based on driving conditions, with regenerative braking replenishing the battery during deceleration. The challenge lies in optimizing fuel allocation over a journey under stochasticity. We model such conditions as a Hidden Markov Process, where the vehicle must balance immediate efficiency with future resource availability to maximize overall mileage. Solutions often involve belief-state MDP's which is applied as a benchmark against our `SD-MDP` MCTS framework.

Financial Options Pricing: In financial trading, American options, which allow holders to exercise at any time before maturity, present an optimal stopping problem where the goal is to maximize profit by deciding when to exercise the option. Serving as a baseline, the Longstaff-Schwartz algorithm is a standard approach, using Monte Carlo simulations and polynomial regression to estimate the continuation value of holding the option, comparing it to the immediate exercise value to determine the optimal strategy. We provide a comparison of the baseline performance against our our `SD-MDP` MCTS framework.

5 CONCLUSION

Certain stochastic decision processes in optimal control and economics demonstrate remarkable efficacy when coupled with specific assumptions and advanced approximation techniques, particularly in value approximation. Disentangling the causal structure of Monte Carlo (MDP's) not only yields unique insights but also significantly simplifies problem-solving. However, traditional methods for addressing resource allocation problems often struggle to seamlessly integrate with Monte Carlo planning techniques. In response, we propose `SD-MDP`, an innovative framework for solving structurally decomposed MDPs, offering a versatile modeling approach alongside robust theoretical guarantees. Inspired by fundamental energy conservation principles, we introduce a resource-utility exchange model, which not only enhances computational efficiency but also reduces planning problem complexity. Moreover, we showcase the effective disentanglement of Monte Carlo sampling from the planning process within the `SD-MDP` framework, facilitating the derivation of Monte Carlo value estimates for both upper and lower bounds of the MDP problem at each state. By seamlessly integrating this approach into MCTS, we not only establish theoretical guarantees but also provide empirical evidence of its efficacy in addressing well-known problems in economic logistics. As future avenues, we envision extending this tool to tackle a broader spectrum of economic problems while delving deeper into the learning setting, where the parameters of the MDP must be learned rather than given.

REFERENCES

[Alt21] Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

[AM07] Brian Do Anderson and John B Moore. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

[BDG00] Craig Boutilier, Richard Dearden, and Moisés Goldszmidt. "Stochastic dynamic programming with factored representations". In: *Artificial intelligence* 121.1-2 (2000), pp. 49–107.

[Ber11] Dimitri P Bertsekas. "Approximate policy iteration: A survey and some new methods". In: *Journal of Control Theory and Applications* 9.3 (2011), pp. 310–335.

[BJS21] Ioana Bica, Daniel Jarrett, and Mihaela van der Schaar. "Invariant causal imitation learning for generalizable policies". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 3952–3964.

[BKP16] Berit Dangaard Brouer, Christian Vad Karsten, and David Pisinger. "Big data optimization in maritime logistics". In: *Big data optimization: Recent developments and challenges*. Springer, 2016, pp. 319–344.

[BL16] Craig Boutilier and Tyler Lu. "Budget Allocation using Weakly Coupled, Constrained Markov Decision Processes." In: *UAI*. 2016.

[Car+19] Nicolas Carrara et al. "Budgeted reinforcement learning in continuous state space". In: *Advances in neural information processing systems* 32 (2019).

[CF12] Giuseppe C Calafiore and Lorenzo Fagiano. "Robust model predictive control via scenario optimization". In: *IEEE Transactions on Automatic Control* 58.1 (2012), pp. 219–224.

[CLP01] Emmanuelle Clément, Damien Lamberton, and Philip Protter. *An analysis of the Longstaff-Schwartz algorithm for American option pricing*. Tech. rep. Cornell University Operations Research and Industrial Engineering, 2001.

[CM07] Pierre-Arnaud Coquelin and Rémi Munos. "Bandit Algorithms for Tree Search". In: *CoRR* abs/cs/0703062 (2007). arXiv: `cs/0703062`. URL: `http://arxiv.org/abs/cs/0703062`.

[Day72] Peter W Day. "Rearrangement inequalities". In: *Canadian Journal of Mathematics* 24.5 (1972), pp. 930–943.

[DS06] Martin Dyer and Leen Stougie. "Computational complexity of stochastic programming problems". In: *mathematical programming* 106 (2006), pp. 423–432.

[DV04] Daniela Pucci De Farias and Benjamin Van Roy. "On constraint sampling in the linear programming approach to approximate dynamic programming". In: *Mathematics of operations research* 29.3 (2004), pp. 462–478.

[Fel91] William Feller. *An introduction to probability theory and its applications, Volume 2*. Vol. 81. John Wiley & Sons, 1991.

[FSM10] Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. "Error propagation for approximate policy and value iteration". In: *Advances in Neural Information Processing Systems* 23 (2010).

[Gen+20] Sinong Geng et al. "Deep PQR: Solving inverse reinforcement learning using anchor actions". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 3431–3441.

[Git79] John C Gittins. "Bandit processes and dynamic allocation indices". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 41.2 (1979), pp. 148–164.

[Gri+20] Jean-Bastien Grill et al. "Monte-Carlo tree search as regularized policy optimization". en. In: (2020).

[Hau79] Mark P Haugan. "Energy conservation and the principle of equivalence". In: *Annals of Physics* 118.1 (1979), pp. 156–186.

[Haz11] E Hazan. *The convex optimization approach to regret minimization, Optimization for Machine Learning (S. Sra, S. Nowozin, and S. Wright, eds.)* 2011.

[HLP52] G.H. Hardy, J.E. Littlewood, and G. Pólya. *Inequalities*. Cambridge Mathematical Library. Cambridge University Press, 1952. ISBN: 9780521358804. URL: `https://books.google.com.au/books?id=t1RCSP8YKt8C`.

[Jam03] Peter James. *Option theory*. John Wiley & Sons, 2003.

[Kom+23] Aneesh Komanduri et al. "Learning causally disentangled representations via the principle of independent causal mechanisms". In: *arXiv preprint arXiv:2306.01213* (2023).

[KS06] Levente Kocsis and Csaba Szepesvári. "Bandit Based Monte-Carlo Planning". In: *Machine Learning: ECML 2006*. Ed. by Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 282–293. ISBN: 978-3-540-46056-5.

[Kus90] Harold J Kushner. "Numerical methods for stochastic control problems in continuous time". In: *SIAM Journal on Control and Optimization* 28.5 (1990), pp. 999–1048.

[LMT22] Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. "Efficient reinforcement learning with prior causal knowledge". In: *Conference on Causal Learning and Reasoning*. PMLR. 2022, pp. 526–541.

[LS01] Francis A Longstaff and Eduardo S Schwartz. "Valuing American options by simulation: a simple least-squares approach". In: *The review of financial studies* 14.1 (2001), pp. 113–147.

[Meu+98] Nicolas Meuleau et al. "Solving very large weakly coupled Markov decision processes". In: *AAAI/IAAI* 8 (1998), p. 2.

[Pai+24] Michael Painter et al. "Monte Carlo Tree Search with Boltzmann Exploration". In: *Advances in Neural Information Processing Systems* 36 (2024).

[Pow+05] Warren B Powell et al. "Approximate dynamic programming for high dimensional resource allocation problems". In: *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* Vol. 5. IEEE. 2005, pp. 2989–2994.

[RB+22] Abbavaram Gowtham Reddy, Vineeth N Balasubramanian, et al. "On causally disentangled representations". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 7. 2022, pp. 8089–8097.

[Sil+17] David Silver et al. "Mastering the game of go without human knowledge". In: *Nature* 550.7676 (2017), pp. 354–359.

[SN05] Alexander Shapiro and Arkadi Nemirovski. "On complexity of stochastic programming problems". In: *Continuous optimization: Current trends and modern applications* (2005), pp. 111–146.

[Sut18] Richard S Sutton. "Reinforcement learning: An introduction". In: *A Bradford Book* (2018).

[Tod09] Emanuel Todorov. "Efficient computation of optimal actions". In: *Proceedings of the national academy of sciences* 106.28 (2009), pp. 11478–11483.

[Win04] David Wingate. *Solving large mdps quickly with partitioned value iteration*. Brigham Young University, 2004.

[WML13] Shuaian Wang, Qiang Meng, and Zhiyuan Liu. "Bunker consumption optimization methods in shipping: A critical review and extensions". In: *Transportation Research Part E: Logistics and Transportation Review* 53 (2013), pp. 49–62.

[WW11] Jean-Paul Watson and David L Woodruff. "Progressive hedging innovations for a class of stochastic mixed-integer resource allocation problems". In: *Computational Management Science* 8 (2011), pp. 355–370.

[Xia+19] Chenjun Xiao et al. "Maximum entropy monte-carlo planning". In: *Advances in Neural Information Processing Systems* 32 (2019).

[YNL12] Zhishuang Yao, Szu Hui Ng, and Loo Hay Lee. "A study on bunker fuel management for the shipping liner services". In: *Computers & Operations Research* 39.5 (2012), pp. 1160–1172.

[ZBD10] Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. "Modeling Interaction via the Principle of Maximum Causal Entropy". en. In: (2010).

[ZY03] Xun Yu Zhou and George Yin. "Markowitz's mean-variance portfolio selection with regime switching: A continuous-time model". In: *SIAM Journal on Control and Optimization* 42.4 (2003), pp. 1466–1482.

Improved Monte Carlo Planning via Causal Disentanglement for Structurally-Decomposed Markov Decision Processes (Supplementary Material)

A PROOFS AND THEORY

A.1 PROOF OF LEM. 2.1

Finite and Bounded Action Space for the SD-MDP: Under the SD-MDP framework, when $\phi f(\cdot)$ is a perfect antisymetric reflection of $\phi f(\cdot)$, the optimal policy belongs to the union of 2 subspaces, that is $\mathbf{a}^* \subset \{\mathbf{a}^+\} \cup \{\mathbf{a}^-\} \subset \mathcal{A}$, for all time steps t. The cardinality of the dimension of the optimal solution space is upper bounded by $2D$ where D is the dimension of $\mathbf{a}$.

Proof. We use proof by induction. Suppose we start at any time t. The reward function,

$$\mu(\mathbf{x}^t, \mathbf{a}) = \langle \phi f(\mathbf{x}_\eta^t),\, \mathbf{a} \rangle \tag{A.1}$$

To remind the reader, we denote $\{\mathbf{a}^+\}$ and $\{\mathbf{a}^-\}$ as the following,

$$\{\mathbf{a}^+\} = \underset{\mathbf{a} \in \mathcal{A}(t)}{\operatorname{argmax}} ||\mathbf{a}||_p, \qquad \{\mathbf{a}^-\} = \underset{\mathbf{a} \in \mathcal{A}(t)}{\operatorname{argmin}} ||\mathbf{a}||_p \tag{A.2}$$

Given $\mathcal{A}(t)$, at any time t, there exists two sets $\{\mathbf{a}^+\}$, and $\{\mathbf{a}^-\}$ which either maximizes allowable reward, or maximally reduces consumption of $\mathbf{x}_d$. Let us denote $\mathbf{a}^+[\mathbf{x}_\eta]$ and $\mathbf{a}^-[\mathbf{x}_\eta]$ as the following,

$$\mathbf{a}^+[\mathbf{x}_\eta] = \underset{\mathbf{a} \in \{\mathbf{a}^+\}}{\operatorname{argmax}} \langle\, \phi f(\mathbf{x}_\eta),\, \mathbf{a} \rangle, \qquad \mathbf{a}^-[\mathbf{x}_\eta] = \underset{\mathbf{a} \in \{\mathbf{a}^-\}}{\operatorname{argmax}} \langle \phi f(\mathbf{x}_\eta),\, \mathbf{a} \rangle \tag{A.3}$$

We know that the dimension of $|\mathbf{a}|$, $|\mathbf{x}_d|$, and $|\mathbf{x}_\eta|$ are equal, denoted as D. From the fundamental theorem of linear programming there can be at most D solutions at the vertices of the convex hulls in Eq. (A.3). Depending on the scaling property of $\phi(\cdot)$, the sets $\mathbf{a}^+[\mathbf{x}_\eta]$ and $\mathbf{a}^-[\mathbf{x}_\eta]$ could partially overlap or be disjoint, nevertheless the number of unique solutions would to the either maximization problem would be at most D.

The solutions of $\{\mathbf{a}^+\}$ and $\{\mathbf{a}^-\}$ constitute disjoint sets, where $\mathbf{a}^+[\mathbf{x}_\eta]$ is the action which myopically maximize the reward obtained at time t, and $\mathbf{a}^-[\mathbf{x}_\eta]$ is the action which maximizes the potential for the future (or conserves curretn resources). In fact, when $p = 1$ the solution to $\mathbf{a}^+$ is unique due when the feasible region is bounded, and the objective function is linear and convex in the feasible region.

Value function w.r.t. time: Under perfect information, we can naturally compute the expectation over τ_η, leading to the expected stochastic outcome at each discrete time step, denoted as $\mathbb{E}[\tau_\eta] \equiv \{\mathbf{x}_\eta^t, \mathbb{E}[\mathbf{x}_\eta^{t+1}], \mathbb{E}[\mathbf{x}_\eta^{t+2}], \ldots, \mathbb{E}[\mathbf{x}_\eta^T]\}$. This evolution transitions independently w.r.t. to any action sequence $(\mathbf{a}^t, \ldots, \mathbf{a}^T)$, due to the properties of the SD-MDP. Thus, for any stochastic trajectory τ_η, constituting a single scenario outcome, there exists a solution which optimizes the cumulative

rewards from t to T.

Next let us denote the top k elements of $\mathbb{E}[\tau_\eta]$ a set of vectors with size k, as $\text{Top}_k(\tau_\eta)$. Should the dimension of $\mathbf{x}_\eta^t$ be greater than 1, in order to compare various values of $\mathbf{x}_\eta^t$, we majorize over the sequence provided by $\phi f \odot \mathbb{E}[\tau_\eta]$ to ensure a ranking,

$$\text{Top}_{k=T}(\mathbb{E}[\tau_\eta]) = (\mathbb{E}[\mathbf{x}_\eta^{k=1}], \mathbb{E}[\mathbf{x}_\eta^{k=2}], \mathbb{E}[\mathbf{x}_\eta^{k=3}], \ldots, \mathbb{E}[\mathbf{x}_\eta^{k=T}]) \tag{A.4}$$

Such that,

$$\phi f(\mathbb{E}[\mathbf{x}_\eta^{k=1}]) \succ \phi f(\mathbb{E}[\mathbf{x}_\eta^{k=2}]) \succ \phi f(\mathbb{E}[\mathbf{x}_\eta^{k=3}]) \cdots \succ \phi f(\mathbb{E}[\mathbf{x}_\eta^{k=T}]) \tag{A.5}$$

Also expressed as,

$$\phi f(\widetilde{\tau}^{k=1}) \succ \phi f(\widetilde{\tau}^{k=2}) \succ \phi f(\widetilde{\tau}^{k=3}) \cdots \succ \phi f(\widetilde{\tau}^{k=T}) \tag{A.6}$$

The $\text{Top}_k(\tau_\eta)$ operator will truncate over this ordered sequence to only include the top k elements. And should we generalize for any $t \in T$, the optimal value function can be expressed as,

$$V_k(\mathbf{x}^t) = \sum_{i=1}^{k} \langle \phi f\left(\widetilde{\tau}^i\right), \mathbf{a}^+[\widetilde{\tau}^i] \rangle + \sum_{i=k+1}^{T} \langle \phi f\left(\widetilde{\tau}^i\right), \mathbf{a}^-[\widetilde{\tau}^i] \rangle \tag{A.7}$$

Let sequence $(\mathbf{a}^+) \equiv (\mathbf{a}^+[\mathbf{x}_\eta^1], \ldots, \mathbf{a}^+[\mathbf{x}_\eta^t])$, and $(\mathbf{a}^-) \equiv (\mathbf{a}^-[\mathbf{x}_\eta^1], \ldots, \mathbf{a}^-[\mathbf{x}_\eta^t])$. Let $\tilde{\mathbf{a}} \equiv (\mathbf{a}^+)\tilde{\cup}(\mathbf{a}^-)$, where $\tilde{\cup}$ represents an order preserving union of sequences. As an aside, we can also express Eq. (A.7) therefore as,

$$V_k(\mathbf{x}^t) = \phi f \odot \text{Top}_{k=T}(\tau_\eta) \odot \tilde{\mathbf{a}} \tag{A.8}$$

Suppose that the optimal policy consists of the of the expression in Eq. (A.7), we demonstrate that there exists no way of achieving a higher value should k be optimal.

- Under the norm constraints, suppose a reduction in a vector belonging to $(\mathbf{a}^+)$ occurs, and is transferred to a vector in $(\mathbf{a}^-)$, the maximization of $V_k(\mathbf{x}^t)$ would yield a lesser result. This is due to the explicit assumption that ϕf is a strict orthogonal reflection of $\phi' g$.
- By definition the sequence $\tilde{\mathbf{a}}$ is majorized, and by extension of the Hardy-Littlewold-Polya Theorem [HLP52] [Day72], the sum of Hadmard product of any ranked majorized sequence with another ranked majorized sequence is always maximizing. We can infer that by swapping any element from $(\mathbf{a}^+)$ with $(\mathbf{a}^-)$ would result in a sub-optimal value.

We know that the dimension of $|\mathbf{a}|$, $|\mathbf{x}_d|$, and $|\mathbf{x}_\eta|$ are equal of dimension D, and by Eq. (A.7), the solution over trajectory $\mathbb{E}[\tau_\eta]$ consists of either $\mathbf{a}^+[\mathbf{x}_\eta^t]$ or $\mathbf{a}^-[\mathbf{x}_\eta^t]$ at each time interval t. Therefore the structure of the optimal policy posits that the optimal solution will fall into any of the two sets, $\mathbf{a}^+[\mathbf{x}_\eta] \in \{\mathbf{a}^+\}$ or $\mathbf{a}^-[\mathbf{x}_\eta] \in \{\mathbf{a}^-\}$ both with dimension D, and therefore the maximum number of unique solutions is $2D$, when $\{\mathbf{a}^+\}$ and $\{\mathbf{a}^-\}$ are disjoint.

□

A.2 PROOF OF LEM. 2.2

Solving for Optimal Value via Top K Allocation: For the `SD-MDP`, the optimal value can be obtained by solving the dual problem, which involves finding the value of k in $\text{Top}_k(\mathbb{E}[\tau_\eta])$ over $k \in \{1, \ldots, T\}$ possibilities.

Proof. The optimal value problem can be solved, by solving the dual problem, selecting the value of k in $\text{Top}_k(\mathbb{E}[\tau_\eta])$ over $k \in \{1, \ldots, T\}$ possibilities. As the $\text{Top}_k(\tau_\eta)$ operator truncates over an ordered sequence to only include the top k elements, we generalize for any $t \in T$, the optimal value function. Let $||\mathbf{a}^+||_p$ and $||\mathbf{a}^-||_p$ be shorthand for $||\langle \phi' g(\mathbf{x}_\eta^t),\, \mathbf{a}^+\rangle||_p$ and $||\langle \phi' g(\mathbf{x}_\eta^t),\, \mathbf{a}^-\rangle||_p$ respectively, and these are norm equivalent for any value of $\mathbf{x}_\eta$ when considering $||\mathbf{a}^+||_p$ or $||\mathbf{a}^-||_p$ specifically. The optimal value function can be expressed as,

$$V^*(\mathbf{x}^t) = \max_{k \in \{1,\ldots,T\}} V_k(\mathbf{x}^t) \tag{A.9}$$

$$\text{where, } \underline{A} \leq \sum_k \left\|\mathbf{a}^+[\mathbf{x}_\eta]\right\|_p + (T-k)\left\|\mathbf{a}^-[\mathbf{x}_\eta]\right\|_p \leq \bar{A} \tag{A.10}$$

Next, the application of the $\text{Top}_k(\cdot)$ constitutes the *non-anticipative* solution for the planning problem under perfect information. Under unconstrained incremental dynamics, one could set $\|\mathbf{a}^+[\mathbf{x}_\eta]\|_p = \bar{A}$, and $K = 1$, as the trivial solution, which represents the optimal stopping problem. But suppose incremental dynamics do exist, and $\underline{\Delta}_a \leq \|\mathbf{a}[\mathbf{x}_\eta]\|_p \leq \bar{\Delta}_a$, then because $f(\cdot)$ is a strictly monotonic function the optimizing solution to Eq. (A.9) occurs at $\{\mathbf{a}^+\} \cup \{\mathbf{a}^-\}$.

Solving for K: We propose the dual problem from the primal problem in Eq. (A.9). K intuitively controls the maximum capacity allowable. Let us postulate and we seek to maximize K,

$$\frac{\underline{A}}{T} \leq \frac{k}{T}\left(\left\|\mathbf{a}^+[\mathbf{x}_\eta]\right\|_p - \left\|\mathbf{a}^-[\mathbf{x}_\eta]\right\|_p\right) + T\left\|\mathbf{a}^-[\mathbf{x}_\eta]\right\|_p \leq \frac{\bar{A}}{T} \tag{A.11}$$

Given the fixed span $\|\mathbf{a}^+[\mathbf{x}_\eta]\|_p - \|\mathbf{a}^-[\mathbf{x}_\eta]\|_p$, we therefore can compute K from Eq. (A.11), where K and $\|\mathbf{a}^-[\mathbf{x}_\eta]\|_p$ are free variables subject to certain constraints, and $\|\mathbf{a}^+[\mathbf{x}_\eta]\|_p$ and T are given. Next, the application of the $\text{Top}_k(\cdot)$ constitutes the *non-anticipative* solution for the planning problem under perfect information. We propose the dual problem from the primal problem in Eq. (A.9),

$$k^* = \underset{k \in \{1,\ldots,T\}}{\text{argmax}}\ k \tag{A.12}$$

$$\text{where,} \quad k\left(\left\|\mathbf{a}^+[\mathbf{x}_\eta]\right\|_p - \left\|\mathbf{a}^-[\mathbf{x}_\eta]\right\|_p\right) + T^2\left\|\mathbf{a}^-[\mathbf{x}_\eta]\right\|_p \leq \bar{A} \tag{A.13}$$

$$k\left(\left\|\mathbf{a}^+[\mathbf{x}_\eta]\right\|_p - \left\|\mathbf{a}^-[\mathbf{x}_\eta]\right\|_p\right) + T^2\left\|\mathbf{a}^-[\mathbf{x}_\eta]\right\|_p \geq \underline{A} \tag{A.14}$$

To determines the maximizing k, we simply fix k, and search for a maximizing solution to the linear program represented in Eq. (A.12). This LP formulation, assuming the constraints produce a feasible set, will generate one unique solution for $\|\mathbf{a}^+[\mathbf{x}_\eta]\|_p$, and $\|\mathbf{a}^-[\mathbf{x}_\eta]\|_p$.

□

A.3 TECHNICAL NOTE ON CONCENTRATION BOUNDS

Given any sub-Gaussian random variable X_i, we can express this inequality,

$$\text{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq 2\exp\left(-\frac{ct^2}{\sum_{i=1}^n \|X_i\|_{\psi_2}^2}\right) \tag{A.15}$$

Where $\|X_i\|_{\psi_2}$ is the sub-Gaussian norm of X_i, defined as, $\|X\|_{\psi_2} := \inf\left\{c \geq 0 : \text{E}\left(e^{X^2/c^2}\right) \leq 2\right\}$. The sub-Gaussian norm of a random variable X is defined as: $|X|_{\psi_2} = \inf\left\{c > 0 : \mathbb{E}\left[e^{X^2/c^2}\right] \leq 2\right\}$.

The sub-Gaussian norm is $|X|_{\psi_2} = \frac{b-a}{2}$. The variance proxy is $\sigma^2 = \frac{(b-a)^2}{4}$ and $\sum_{i=1}^n (b_i - a_i)^2 = 4\sum_{i=1}^n \sigma_i^2$.

Relating Norm and Variance: For a bounded random variable X with $a \leq X \leq b$, the sub-Gaussian norm is related to the variance proxy $\sigma^2 = \frac{(b-a)^2}{4}$ as 2: $|X|_{\psi_2} = \sqrt{\sigma^2} = \frac{b-a}{2}$.

A.4 PROOF OF THM. 1

Upper bound on the Monte Carlo Value Estimation for the `SD-MDP`: For the `SD-MDP` abstraction partitionable MDP, the optimal policy, where the value function is upper bounded by $|\hat{V}_N - V^*(\mathbf{x})| \leq \mathcal{O}((\delta\sqrt{N})^{-1})$, with probability $1-\delta$. Where $\hat{V}_N$ is the Monte Carlo simulation estimate of the value function under N iterations.

Proof. For each stochastic trajectory τ_η, we denote as as a sequence $(\cdot)$,

$$\tau_\eta \equiv (\mathbf{x}_\eta^{t=1}, \mathbf{x}_\eta^{t=2}, ..., \mathbf{x}_\eta^{T}) \tag{A.16}$$

We define a sequence of actions as $\tau_{a|\eta}$,

$$\tau_{a|\eta} \equiv (\mathbf{a}^{t=1}, \mathbf{a}^{t=2}, ..., \mathbf{a}^{T}) \tag{A.17}$$

There must exists at least one optimal solution that for the optimal policy π^*, which we also denote as a sequence $\tau^*_{a|\eta}$, where $\tau^*_{a|\eta}$ is a sequence of deterministic actions, which yield the optimal solution for each trajectory τ_η. Given a trajectory runs from $1 \rightarrow T$, and as consequence of Lem. 2.1, there exists at most $(2\mathrm{D})^T$ possibly optimal permutations of $\tau^*_{a|\eta}$, where D is the dimension of the action space $\mathcal{A}$. Let $\tau^*_{a|\eta}$ denote an optimal solution such that,

$$\tau^*_{a|\eta} = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \sum_{x_\eta^t \in \tau_\eta} \mathbf{a}^t \cdot \phi f(\mathbf{x}_\eta^t), \quad s.t. \quad \underline{A} \leq \sum_{t=1}^{T} ||\mathbf{a}^t||_p \leq \bar{A} \tag{A.18}$$

Where we know x_η^t is a stochastic variable belonging to a trajectory τ_η. Based on the *recency preference* assumption, illustrated by Eq. (2.11).

$$\mathbf{x}_\eta^t = \mathbf{x}_\eta^{t+\Delta} \implies \mathbf{x}_\eta^t \succ \mathbf{x}_\eta^{t+\Delta}, \quad \Delta \in \mathbb{Z} \tag{A.19}$$

Under this assumption, for each τ_η there must be only one unique $\tau^*_{a|\eta}$, constituting a surjection from $\tau_\eta \mapsto \tau^*_{a|\eta}$.

$$\tau_\eta \mapsto \tau^*_{a|\eta} \tag{A.20}$$

Since by Lem. 2.1 we have a finite ation space per discrete time period, in finite time the cardinality of $\tau^*_{a|\eta}$ is finite. Nevertheless, the cardinality of τ_η is infinite. Thus by the pigeonhole principle, we can conclude that a $\tau^*_{a|\eta}$ can inversely map to potentially more than one τ_η, we denote this set of corresponding sequences as $\tilde{\tau}_\eta$.

$$\tau_\eta \mapsto \tau^*_{a|\eta}, \quad \tilde{\tau}_\eta \leftrightarrow \tau^*_{a|\eta} \tag{A.21}$$

Furthermore, let $\{\tau^*_{a|\eta}\}$ denote all valid sequences of $\tau^*_{a|\eta}$ corresponding to τ_η, by consequence of Lem. 2.1, the cardinality of the injective map's image $\{\tau^*_{a|\eta}\}$ bounded by,

$$1 \leq \{\tau^*_{a|\eta}\} \leq (2D)^T \tag{A.22}$$

Most crucially, $\{\tilde{\tau}_\eta\}$ denotes a set of sequences, of which are *indifferent* from each other in terms of their corresponding

optimal policy $\tau^*_{a|\eta}$, constituting the optimal solution to the trajectory, for $\mathbf{x}_\eta \subset \mathbb{R}^{\mathrm{D}}$. Although given parameters of stochastic generator θ, the probability of observing τ_η is infinitesimal, the probability of $\tau_\eta \in \{\tilde{\tau}_\eta\}$ is quantifiable, given the functional form and parameters of the stochastic generator. That is,

$$P_\theta(\tau_\eta \in \{\tilde{\tau}_\eta\}|\mathcal{F}_t) > 0 \tag{A.23}$$

Let the operator $\{\tau_\eta\}$ define the set of all sequence of trajectories τ_η, we define an operator $\mathcal{Q}(\cdot) : \{\tau_\eta\} \mapsto \{\tilde{\tau}_\eta\}$ which partitions $\{\tau_\eta\}$ into into *indifference sets*, $\tilde{\tau}_\eta$. In fact, from the relation in Eq. (A.21) and Lem. 2.1 the cardinality of this partition function can be deduced as,

$$1 \leq |\mathcal{Q}(\{\tau_\eta\})| \leq \binom{T}{k} D^T \leq (2D)^T \tag{A.24}$$

Where $k \leq T$ is the capacity maximizing number governed by solution to the dual problem posed in Eq. (A.9). Let $\{\mathbf{x}_\eta^{t\to T}\}$ denote the set of all possible trajectories for $\mathbf{x}_\eta$ between t to T. The expectation thereof for $\mathbb{E}[\{\mathbf{x}_\eta^{t\to T}\}]$ can be calculated via stochastic calculus, or approximated by simulation. Given two sequences of vectors, $\mathbf{x}^{t\to T}$ and $\mathbf{y}^{t\to T}$, let us define a an operator $\langle\langle\cdot,\cdot\rangle\rangle$ as,

$$\langle\langle \mathbf{x}^{t\to T}, \mathbf{y}^{t\to T}\rangle\rangle = \sum_{i=1}^{T} \langle \mathbf{x}_i, \mathbf{y}_i\rangle \tag{A.25}$$

We can now examine the value function,

$$\begin{aligned}
V^*(\mathbf{x}^t) &= \mathbb{E}\Big[\langle\langle \phi f(\mathbf{x}_\eta^{t\to T}), \tau^*_{a|\eta}\rangle\rangle\Big] && \text{(A.26)}\\
&= \int_{\tau_\eta} P_\theta(\tau_\eta|\mathcal{F}_t)\ \langle\langle \phi f(\tau_\eta), \tau^*_{a|\eta}\rangle\rangle\, d\tau_\eta, \quad 1 < |\mathcal{Q}(\{\tau_\eta\})| < (2D)^T && \text{(A.27)}\\
&\leq \sum_{\mathcal{Q}(\{\tau_\eta\})} P_\theta(\tau_\eta \in \{\tilde{\tau}_\eta\}|\mathcal{F}_t)\ \langle\langle\ \bar{\tau}_s, \phi f(\tau^*_{a|\eta})\ \rangle\rangle, \quad 1 < |\mathcal{Q}(\{\tau_\eta\})| < (2D)^T && \text{(A.28)}
\end{aligned}$$

Eq. (A.26) represents the optimal value function as the expectation over the reward/cost of the unravelling of the stochastic partition of the `SD-MDP`, $\phi f(\mathbf{x}_\eta^{t\to T})$, assuming an optimal deterministic policy $\tau^*_{a|\eta}$ in adhered to by the agent. Eq. (A.27) expresses the value function as an integral over the probability of each possible outcome $\phi f(\tau_\eta)$. In principle, $\{\mathbf{x}_\eta^{t\to T}\}$ can be partitioned in to indifference sets $\{\tilde{\tau}_\eta\}$, via the $\mathcal{Q}(\cdot)$ partition operator, where each indifference maps to a unique $\tau^*_{a|\eta}$. And thus we can make the expectation in Eq. (A.26) decomposable. This presents a key advantage as any upper bounding value function $\bar{V}^*(\mathbf{x}^t) \geq V^*(\mathbf{x}^t)$ can be expressed as the summation of the product of two terms.

$$\begin{aligned}
V^*(x^t) \leq \bar{V}^*(x^t) &= \sum_{\mathcal{Q}(\{\tau_\eta\})} \bar{g}_{\tau_\eta}(\{\tilde{\tau}_\eta\})\ \bar{g}_s(\tau^*_{a|\eta}) && \text{(A.29)}\\
\text{where,}\quad \bar{g}_{\tau_\eta}(\cdot) &\geq P_\theta(\tau_\eta \in \{\tilde{\tau}_\eta\}|\mathcal{F}_t), \quad \forall \tau_\eta \in \{\tilde{\tau}_\eta\} && \text{(A.30)}\\
\bar{g}_s(\cdot) &\geq \langle\langle\ \phi f(\tau_\eta),\ \tau^*_{a|\eta}\ \rangle\rangle, \quad \forall \tau^*_{a|\eta} \in \tau_a && \text{(A.31)}
\end{aligned}$$

Finding the upper-bound to $\bar{V}^*(x^t)$ is now decomposed into a problem involving finding an two tight upper bounding functions $\bar{g}_{\tau_\eta}(\cdot)$ and $\bar{g}_s(\cdot)$. For *incremental action dynamics* as defined for the `SD-MDP` framework, there exists a limit on the capacity of actions as defined in Eq. (2.11).

If we define an operator over $\{\tau_\eta\}$ such that,

$$\text{Top}_k(\tau_\eta) : \{\tau_\eta\} \times K \mapsto \mathbb{R}^{|\mathcal{X}|\times T}, \quad \text{where, } K \in \mathbb{Z}^+ \tag{A.32}$$

Where the image, $\text{Im}(\text{Top}_k(\cdot)) \in \mathbb{R}^{|\mathcal{X}|\times T}$. We know that the $\text{Top}_k(\tau_\eta)$ computation can occur with complexity $\mathcal{O}(T)$, as it involves sorting over elements. Therefore, we can denote an expression for $\langle\langle \phi f(\tau_\eta), \tau^*_{a|\eta}\rangle\rangle$ by writing,

$$\langle\langle \phi f(\tau_\eta), \tau^*_{a|\eta}\rangle\rangle = \max_{\mathbf{a}\in\{\mathbf{a}^+\}} \langle\langle \phi f(\text{Top}_K(\{\tau_\eta\})), \mathbf{a}\rangle\rangle + \max_{\mathbf{a}\in\{\mathbf{a}^-\}} \langle\langle \phi f(/\text{Top}_K(\{\tau_\eta\})), \mathbf{a}\rangle\rangle, \quad \text{where, } K = \left\lfloor \frac{\bar{A}}{T} \right\rfloor \tag{A.33}$$

Note that Eq. (A.33) is equivalent to Eq. (A.7) from Lem. 2.2, and $\tau_\eta \in \{\tilde{\tau}_\eta\}$. So now fundamentally, we can express the computation of the value function as a decomposition,

$$\begin{aligned} V^*(\mathbf{x}^t) &= \langle\langle \mathbb{E}[\phi f(\tau_\eta)], \mathbb{E}[\tau^*_{a|\eta}]\rangle\rangle && \text{(A.34)} \\ &= \mathbb{E}[\langle\langle \phi f(\tau_\eta), \tau^*_{a|\eta}\rangle\rangle] && \text{(A.35)} \\ &= \max_{\mathbf{a}\in\{\mathbf{a}^+\}} \mathbb{E}[\langle\langle \phi f(\text{Top}_K(\{\tau_\eta\})), \mathbf{a}\rangle\rangle] + \max_{\mathbf{a}\in\{\mathbf{a}^-\}} \mathbb{E}[\langle\langle \phi f(/\text{Top}_K(\{\tau_\eta\})), \mathbf{a}\rangle\rangle] && \text{(A.36)} \end{aligned}$$

From Eq. (A.36) we can see that an approximation of $V^*(\mathbf{x}^t)$ can be computed by simulating over the stochastic trajectory τ_η, then applying a deterministic function, $\text{Top}_K(\cdot)$, over it.

Bounding the Approximation Error: To bound the expression on Eq. (A.36) we can take advantage of the $\mathcal{Q}(\cdot)$ operator.

$$V^*(\mathbf{x}^t) = \sum_{\mathcal{Q}(\tau_\eta)} P_\theta(\tau_\eta \in \tilde{\tau}_\eta | \mathcal{F}_t) \left(\max_{\mathbf{a}\in\{\mathbf{a}^+\}} \langle\langle \phi f(\mathbb{E}[\text{Top}_K(\tilde{\tau}_\eta)]), \mathbf{a}\rangle\rangle + \max_{\mathbf{a}\in\{\mathbf{a}^-\}} \langle\langle \phi f(\mathbb{E}[/\text{Top}_K(\tilde{\tau}_\eta)]), \mathbf{a}\rangle\rangle \right) \tag{A.37}$$

The next major challenge likes in the computation of $P_\theta(\tau_\eta \in \{\tilde{\tau}_\eta\}|\mathcal{F}_t)$ which is the probability that any trajectory τ_η belongs to the indifference set $\{\tilde{\tau}_\eta\}$, given parameters of the stochastic generator θ and filtration $\mathcal{F}_t$. Nevertheless, this formulation allow us to separate the reward outcome, which is deterministically computable from the probability of it occurring. To compute the conditional expectation of $\mathbb{E}[\text{Top}_K(\tilde{\tau}_\eta)]$, we can apply Bayes rule to conditional expectations and obtain a closed from expression for $\mathbb{E}[\text{Top}_K(\tilde{\tau}_\eta)]$ in Eq. (A.38).

$$\mathbb{E}[\text{Top}_K(\tilde{\tau}_\eta)] = \frac{1}{P(\tau_\eta \in \tilde{\tau}_\eta)} \int_{\tau_\eta \in \tilde{\tau}_\eta} \text{Top}_K(\tilde{\tau}_\eta) \; \mathbb{1}[\tau_\eta \in \tilde{\tau}_\eta] \, d\tau_\eta \tag{A.38}$$

Therefore, we can then apply a Monte Carlo approach to solve the problem of value estimation. But key advantages are imposed, first is *hindsight* optimality. From Eq. (A.37) we can solve an optimal policy using the $\text{Top}_K(\cdot)$ operator over τ_η because there exists no causal relationship between action and state transition, as defined in the `SD-MDP` dynamic. We also know, that every trajectory has a unique deterministic optimal value outcome by the relation in Eq. (A.21). The same argument applies to $\mathbb{E}[/\text{Top}_K(\tilde{\tau}_\eta)]$. Moving forward, for convienience, let us define,

$$\mathcal{H}(\tilde{\tau}_\eta) \equiv \max_{\mathbf{a}\in\{\mathbf{a}^+\}} \langle\langle \phi f(\mathbb{E}[\text{Top}_K(\tilde{\tau}_\eta)]), \mathbf{a}\rangle\rangle + \max_{\mathbf{a}\in\{\mathbf{a}^-\}} \langle\langle \phi f(\mathbb{E}[/\text{Top}_K(\tilde{\tau}_\eta)]), \mathbf{a}\rangle\rangle \tag{A.39}$$

We can see that $\mathcal{H}(\tilde{\tau}_\eta)$ serves as a deterministic function for each input $\tilde{\tau}_\eta$. As we know that $\mathcal{Q}(\cdot)$ will produce at most $(2D)^T$ partitions of the trajectory space $\{\tau_\eta\}$, we can treat this as the approximation of a multinomial distribution via MC. Let our estimator simply be,

$$\hat{P}_\theta^N(\tau_\eta \in \tilde{\tau}_\eta | \mathcal{F}_t) = \frac{1}{N} \sum_i^N \mathbb{1}[\tau_\eta \sim \theta \in \tilde{\tau}_\eta] \tag{A.40}$$

Secondly, our value approximator would be defined as,

$$\hat{V}_N^*(\mathbf{x}^t) = \sum_{\mathcal{Q}(\{\tau_\eta\})} \hat{P}_\theta^N(\tau_\eta \in \tilde{\tau}_\eta | \mathcal{F}_t)\ \mathcal{H}(\tilde{\tau}_\eta), \quad \text{where, } K = \left\lfloor \frac{\bar{A}}{T} \right\rfloor \tag{A.41}$$

Convergence: By law of large numbers $||\hat{P}_\theta^N(\cdot) - P|| \to 0$, and we argue consequently $||\hat{V}_N^*(\cdot) - V^*(\cdot)||$ when $N \to \infty$. To demonstrate convergence, we can apply concentration bounds to quantify the approximation error, via Hoeffding's inequality. Thus we seek the probability that $\tau_\eta \in \tilde{\tau}_\eta$ by taking N Monte Carlo samples,

$$N(\tilde{\tau}_\eta) = \sum^N \mathbb{1}[\tau_\eta \in \{\tilde{\tau}_\eta\}] \tag{A.42}$$

$$\mathbb{E}[N(\tilde{\tau}_\eta)] = N \cdot p_\theta(\tilde{\tau}_\eta) \tag{A.43}$$

Where $p_\theta(\tilde{\tau}_\eta)$ is the finite multinomial probability that the event $\tau_\eta \in \tilde{\tau}_\eta$ occurs. Thus the total variance of each counts on each $\tau_\eta \in \tilde{\tau}_\eta$ is,

$$\text{Var}[N(\tilde{\tau}_\eta)] = N \cdot p_\theta(\tau_\eta \in \tilde{\tau}_\eta) \cdot (1 - p_\theta(\tau_\eta \in \tilde{\tau}_\eta)) \tag{A.44}$$

$$\text{Var}[P(\tilde{\tau}_\eta)] = p_\theta(\tau_\eta \in \tilde{\tau}_\eta) \cdot (1 - p_\theta(\tau_\eta \in \tilde{\tau}_\eta)) \tag{A.45}$$

We can now apply Hoeffding's inequality, let shorthand $p_\theta(\tau_\eta) \equiv p_\theta(\tau_\eta \in \tilde{\tau}_\eta | \mathcal{F}_t)$,

$$P\left(\left|\mathbb{E}[\hat{P}_\theta^N(\tilde{\tau}_\eta) - P_\theta(\tilde{\tau}_\eta)]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2N\epsilon^2}{p_\theta(\tau_\eta) \cdot (1 - p_\theta(\tau_\eta))}\right) \tag{A.46}$$

$$P\left(\left|\mathbb{E}[\hat{P}_\theta^N(\tilde{\tau}_\eta) - P_\theta(\tilde{\tau}_\eta)]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2N\epsilon^2}{\sigma^2}\right) \tag{A.47}$$

Therefore, with probability $1 - \delta$,

$$P\left(\left|\mathbb{E}[\hat{P}_\theta^N(\tilde{\tau}_\eta) - P_\theta(\tilde{\tau}_\eta)]\right| \geq \epsilon\right) \leq \delta \tag{A.48}$$

$$2\exp\left(-\frac{2N\epsilon^2}{\sigma^2}\right) \leq \delta \tag{A.49}$$

$$\log\left(2\exp\left(-\frac{2N\epsilon^2}{\sigma^2}\right)\right) \leq \log\delta \tag{A.50}$$

$$\epsilon \geq \sigma\sqrt{\frac{1}{2N}\log\frac{2}{\delta}} \tag{A.51}$$

Thus we have, with probability $1 - \delta$,

$$\left|\mathbb{E}[\hat{P}_\theta^N(\tilde{\tau}_\eta) - P_\theta(\tilde{\tau}_\eta)]\right| \leq \sigma\sqrt{\frac{1}{2N}\log\frac{2}{\delta}} = \epsilon(\tilde{\tau}_\eta) \tag{A.52}$$

Therefore, we can express the value function as,

$$V^*(x^t) = \sum_{\mathcal{Q}(\{\tilde{\tau}_\eta\})} P_\theta(\tilde{\tau}_\eta)\ \mathcal{H}(\tilde{\tau}_\eta) \tag{A.53}$$

Where with probability $1-\delta$,

$$\begin{aligned} V^*(x^t) &\leq \sum_{\mathcal{Q}(\{\tilde{\tau}_\eta\})} \Big(P_\theta(\tilde{\tau}_\eta) + \epsilon(\tilde{\tau}_\eta)\Big)\ \mathcal{H}(\tilde{\tau}_\eta) && \text{(A.54)} \\ &= \sum_{\mathcal{Q}(\{\tilde{\tau}_\eta\})} P_\theta(\tilde{\tau}_\eta)\mathcal{H}(\tilde{\tau}_\eta) + \sum_{\mathcal{Q}(\{\tilde{\tau}_\eta\})} \epsilon(\tilde{\tau}_\eta)\mathcal{H}(\tilde{\tau}_\eta) && \text{(A.55)} \\ &= V^*(x^t) + \sum_{\mathcal{Q}(\{\tilde{\tau}_\eta\})} \epsilon(\tilde{\tau}_\eta)\mathcal{H}(\tilde{\tau}_\eta) && \text{(A.56)} \end{aligned}$$

Which has the implication that,

$$V^*(x^t) \leq \sum_{\mathcal{Q}(\{\tilde{\tau}_\eta\})} \hat{P}_\theta^N(\tilde{\tau}_\eta)\ \mathcal{H}(\tilde{\tau}_\eta) \implies \bar{V}^*(x^t) - V^*(x^t) \leq \sum_{\mathcal{Q}(\{\tilde{\tau}_\eta\})} \epsilon_{\tau_\eta}\mathcal{H}(\tilde{\tau}_\eta) \tag{A.57}$$

We can write this also in vector notation as,

$$\Delta_V^+ = \mathbf{C}_{\mathcal{Q}(\cdot)} \cdot \mathcal{H}(\tilde{\tau}_\eta) \tag{A.58}$$

Where, for $q \in \{1, \ldots, |Q|\}$ partitions,

$$\begin{aligned} \mathbf{C}_{\mathcal{Q}(\cdot)} &= [c_1(\tau_\eta, \delta) \ldots c_Q(\tau_\eta, \delta)]^T && \text{(A.59)} \\ c_q(\tau_\eta, \delta) &= \sqrt{\frac{p_q(\tilde{\tau}_\eta)(1 - p_q(\tilde{\tau}_\eta))}{2} \log\left(\frac{2}{\delta}\right)}, \quad \forall q \in \mathcal{Q}(\{\tilde{\tau}_\eta\}) && \text{(A.60)} \end{aligned}$$

Δ_V^+ constitutes the maximum possible over estimation due to misspecification. This is fixed and determined by the properties of the stochastic generator θ, and due to misspecification, the upperbound on the value function decreases with rate $\frac{1}{\sqrt{N}}$. Thus we have an upper bound on the value,

$$\bar{V}^*(x^t) - V^*(x^t) \leq \sum_{\mathcal{Q}(\{\tilde{\tau}_\eta\})} \epsilon(\tilde{\tau}_\eta)\, \mathcal{H}(\tilde{\tau}_\eta) = \frac{\Delta_V^+}{\sqrt{N}}. \tag{A.61}$$

Without loss of generality, the opposite can be argued for a lower bound, as $\epsilon(\tilde{\tau}_\eta)$ adheres to a symmetric relation,

$$V^*(x^t) - \underline{V}^*(x^t) \leq \frac{\Delta_V^+}{\sqrt{N}}. \tag{A.62}$$

□

A.5 LEM. A.1 AND PROOF

Lemma A.1. ***Binomial Sum Simplification:*** *From Feller (1991) (Pg. 151), the following inequality holds,*

$$P(\Sigma_i \geq k) = P(\Sigma_i = k)\frac{k(1-p)}{k-Tp} \tag{A.63}$$

Proof. Let $X_1, \cdots, X_T$ be a sequence of variables with Bernoulli distribution $X_T \sim B(T,p)$, assuming $k > Tp$,

$$P(\sum_{i=1}^{T} X_i \geq k) \leq P(\sum_{i=1}^{T} X_i = k)\frac{k(1-p)}{k-Tp} \tag{A.64}$$

We can express,

$$\frac{P(\sum_{i=1}^{T} X_i = k+1)}{P(\sum_{i=1}^{T} X_i = k)} = \frac{\frac{T!}{(k+1)!(T-k-1)!}p^{k+1}(1-p)^{T-k-1}}{\frac{T!}{k!(T-k)!}p^{k}(1-p)^{T-k}} = \frac{(T-k)p}{(k+1)(1-p)} \tag{A.65}$$

For $j \geq k$,

$$P(\sum_{i=1}^{T} X_i = j) = \frac{P(\sum_{i=1}^{T} X_i = k+1)}{P(\sum_{i=1}^{T} X_i = k)} \cdots \frac{P(\sum_{i=1}^{T} X_i = j)}{P(\sum_{i=1}^{T} X_i = j-1)} P(\sum_{i=1}^{T} X_i = k) \tag{A.66}$$

$$\leq P(\sum_{i=1}^{T} X_i = k)\left(\frac{(T-k)p}{(k+1)(1-p)}\right)^{j-k} \tag{A.67}$$

We therefore obtain,

$$P(\sum_{i=1}^{T} X_i \geq k) = \sum_{j=k}^{T} P(\sum_{i=1}^{T} X_i = j) \leq P(\sum_{i=1}^{T} X_i = k)\sum_{j=k}^{T}\left(\frac{(T-k)p}{(k+1)(1-p)}\right)^{j-k} \tag{A.68}$$

$$= P(\sum_{i=1}^{T} X_i = k)\frac{1-\left(\frac{(T-k)p}{(k+1)(1-p)}\right)^{T-k+1}}{1-\frac{(T-k)p}{(k+1)(1-p)}} \tag{A.69}$$

$$\leq P(\sum_{i=1}^{T} X_i = k)\frac{1}{1-\frac{(T-k)p}{(k+1)(1-p)}} = P(\sum_{i=1}^{T} X_i = k)\frac{(k+1)(1-p)}{k-Tp+1-p} \tag{A.70}$$

$$< P(\sum_{i=1}^{T} X_i = k)\frac{k(1-p)}{k-Tp} \tag{A.71}$$

□

A.6 PROOF OF THM. 2

Simple Regret: Given a Monte Carlo planning algorithm, $\mathcal{M}$, where $\tilde{p}(T) = P(\mathbf{a}^* \neq \mathbf{a})$, where $\lim_{T\to\infty}\tilde{p}(T) = 0$ and $\tilde{p}(T)$ is asymptotically bounded above by $\mathcal{O}(\frac{1}{T})$ for T samples, when running $\mathcal{M}$ over the `SD-MDP` framework, the the simple regret $\mathrm{reg}(T)$, as defined in Eq. 3.8, is bounded by $C_k T\tilde{p}(T)$, for some value $C_k \in \mathbb{R}^+$.

Proof. From Lem. 2.1 we ascertain that as t grows $t \to T$ at each time increment the optimal action $\mathbf{a}^*$ lines in a discrete set $\{\mathbf{a}^+\} \cup \{\mathbf{a}^-\}$. Let $\tilde{p}(t) = P(\mathbf{a}^* \neq \mathbf{a})$, given a trajectory, where the agent has acted perfectly, we first quantify the regret accumulated from making k *swaps*, that is instead of performing $\mathbf{a}^t = \mathbf{a}^* \in \{\mathbf{a}^-\}$, the agent instead performs $\mathbf{a}^t \in \{\mathbf{a}^+\}$ and (vice-versa). Logically it follows, for k swaps of any given capacity, the agent would play the $\mathbf{a}^t \in \{\mathbf{a}^+\}$ at another point in time. When each of these swaps occur, the worst case instantaneous regret (or gap) is defined as $\tilde{\Delta}_a$,

$$\tilde{\Delta}_a = \max_{\mathbf{a}\in\{\mathbf{a}^+\}} \langle \phi f(\mathbf{x}_s), \mathbf{a}\rangle - \max_{\mathbf{a}\in\{\mathbf{a}^-\}} \langle \phi f(\mathbf{x}_s), \mathbf{a}\rangle \tag{A.72}$$

Let us next determine the consequences of each outcome, should i swaps occur.

$$\text{reg}(T|\tau_\eta) = \begin{cases} k\tilde{\Delta}_a, & k \leq i \leq T \\ i\tilde{\Delta}_a, & i < k \leq T \end{cases} \tag{A.73}$$

Assuming $k > T\tilde{p}$, which is possible as we state that $\tilde{p}$ is asymptotically bounded above by $\mathcal{O}(1/T)$, we can therefore bound the expectation over simple regret as,

$$\mathbb{E}[\text{reg}(T)] \leq \tilde{\Delta}_a \sum_{i=1}^{k-1} i \binom{T}{i} \tilde{p}^i (1-\tilde{p})^{T-i} + k\tilde{\Delta}_a \sum_{i=k}^{T} \binom{T}{i} \tilde{p}^i (1-\tilde{p})^{T-i} \tag{A.74}$$

$$= \tilde{\Delta}_a \sum_{i=1}^{k-1} i \binom{T}{i} \tilde{p}^i (1-\tilde{p})^{T-i} + k\tilde{\Delta}_a P(\Sigma_i \geq k) \tag{A.75}$$

Which we can simplify further as (from Lem. A.1),

$$\mathbb{E}[\text{reg}(T)] \leq \tilde{\Delta}_a \sum_{i=1}^{k-1} i \binom{T}{i} \tilde{p}^i (1-\tilde{p})^{T-i} + k\tilde{\Delta}_a P(\Sigma_i = k)\Big(\frac{k(1-\tilde{p})}{k-T\tilde{p}}\Big) \tag{A.76}$$

$$= \tilde{\Delta}_a \sum_{i=1}^{k-1} i \underbrace{\binom{T}{i} \tilde{p}^i (1-\tilde{p})^{T-i}}_{F_{\tilde{p}}^T(i)} + k\tilde{\Delta}_a \underbrace{\binom{T}{k} \tilde{p}^k (1-\tilde{p})^{T-k}}_{F_{\tilde{p}}^T(k)} \Big(\frac{k(1-\tilde{p})}{k-T\tilde{p}}\Big) \tag{A.77}$$

Taking a close up of the limit on the fractional term, $\frac{k(1-\tilde{p})}{k-T\tilde{p}}$,

$$\lim_{\tilde{p}\to 0} \frac{k(1-\tilde{p})}{k-T\tilde{p}} = 1 \tag{A.78}$$

Let $F_{\tilde{p}}^T(x)$ represent the binomial probability, with respect to x successes, given T attempts, and a success probability $\tilde{p}$.

$$F_{\tilde{p}}^T(x) = \binom{T}{x} \tilde{p}^x (1-\tilde{p})^{T-x} \tag{A.79}$$

We can then represent Eq. (A.76) as,

$$\mathbb{E}[\text{reg}(T)] \leq \tilde{\Delta}_a \sum_{i=1}^{k-1} i F_{\tilde{p}}^T(i) + k\tilde{\Delta}_a F_{\tilde{p}}^T(k) \tag{A.80}$$

From Eq. (A.80) the right hand side is vanishing as $T \to \infty$, as $F_{\tilde{p}}^T(k) \to 0$. What is more important is we find an upper-bound to the right hand side term, $\tilde{\Delta}_a \sum_{i=1}^{k-1} iF_{\tilde{p}}^T(i)$. To accomplish this, let us impose an upper-bounding function, $\mathcal{U}^T(x, \tilde{p})$ which we define as a polynomial,

$$\mathcal{U}^T(x,\tilde{p}) = \alpha x(x-\beta) + \gamma \geq F_{\tilde{p}}^T(x), \qquad \forall x \in \mathbb{Z},\, 1 \leq x \leq k, \quad \forall \tilde{p} \in [0,1] \tag{A.81}$$

By adjusting the terms $\alpha, \beta, \gamma \in \mathbb{R}$, we can always construct some upperbounding function which upper-bounds the function $F_{\tilde{p}}^T(x)$ on the interval $\forall x \in \mathbb{Z},\, 1 \leq x \leq k, \forall \tilde{p} \in [0,1]$, this is due to the concave nature of $F_{\tilde{p}}^T(x)$. That is for any given $T \in [1,\infty)$ we can select some combination of α, β, γ to satisfy Eq. (A.81). Once again we remind the reader that $\tilde{p}(t)$ is actually a dynamic feature of time t, as t increments from $1 \to T$. We can see that α, β are invariant of $\tilde{p}(t)$ for $t \in [1,T]$, should α, β be properly selected. Although we are free to select γ as we desire to form $\mathcal{U}^T(x)$, it will become evident later that we must select a $\gamma(T)$ which is a function of T, that approaches 0 as $T \to \infty$. To accomplish this, we first acknowledge that for the binomial distribution, $F_{\tilde{p}}^T(x)$ for small values of $\tilde{p}$, which is the case when $T \to \infty$,

$$F_{\tilde{p}}^T(x=0) \geq F_{\tilde{p}}^T(x=1) \geq F_{\tilde{p}}^T(x=2) \geq \cdots \geq F_{\tilde{p}}^T(x=T) \tag{A.82}$$

We solve the inequality for the polynomial $\mathcal{U}^T(x=0,\tilde{p}) \geq F_{\tilde{p}}^T(x=1)$, where from Eq. (A.82), $\gamma(T)$ serves as an upper-bound for all $F_{\tilde{p}}^T(x)$, $\forall x \in 1 \ldots T$ by induction.

$$\mathcal{U}^T(x=0,\tilde{p}) = \alpha x(x-\beta) + \gamma(T) = \gamma(T) \geq F_{\tilde{p}}^T(x=1) \tag{A.83}$$

Given any fixed value of k which is determined by the capacity constraint of the `SD-MDP` framework, for any arbitrarily large value of T, there exists some α, and β independent of T which, in combination with a properly selected $\gamma(T)$ will serve to upper-bound $F_{\tilde{p}}^T(x=1)$. We enforce the assumption that that $\tilde{p}(T) \to 0$ as $T \to \infty$, we therefore utilize the inequality $\gamma(T) \geq F_{\tilde{p}}^T(x=1)$,

$$\begin{aligned}
\lim_{T\to\infty} \gamma(T) &\geq \lim_{T\to\infty} F_{\tilde{p}}^T(x=1) && \text{(A.84)}\\
&= \lim_{T\to\infty} \frac{T!}{1!(T-1)!}\tilde{p}(T)^1(1-\tilde{p}(T))^{T-1} && \text{(A.85)}\\
&= \lim_{T\to\infty} T\tilde{p}(T)(1-\tilde{p}(T))^{T-1} && \text{(A.86)}\\
&= \lim_{T\to\infty} T\tilde{p}(T) \lim_{T\to\infty} (1-\tilde{p}(T))^{T-1} && \text{(A.87)}\\
&= \lim_{T\to\infty} T\tilde{p}(T) && \text{(A.88)}
\end{aligned}$$

This allows us to select the upper bound on $F_{\tilde{p}}^T(\cdot)$, as,

$$T\tilde{p}(T) \geq F_{\tilde{p}}^T(x),\, \forall x \in 0 \ldots T \tag{A.89}$$

An identical result can also alternatively be obtained by applying Lem. A.2. To study simple regret behaviour as $T \to \infty$, we have a return look at Eq. (A.80).

$$\lim_{T\to\infty} \mathbb{E}[\text{reg}(T)] \leq \lim_{T\to\infty} \tilde{\Delta}_a \sum_{i=1}^{k-1} iF_{\tilde{p}}^T(i) + k\tilde{\Delta}_a F_{\tilde{p}}^T(k)\frac{k}{T} \tag{A.90}$$

$$= \lim_{T\to\infty} \tilde{\Delta}_a \sum_{i=1}^{k-1} iF_{\tilde{p}}^T(i) \tag{A.91}$$

$$\leq \lim_{T\to\infty} \tilde{\Delta}_a \sum_{i=1}^{k-1} i(\alpha_k i(i-\beta_k) + \gamma(T)) \tag{A.92}$$

Where for any particular value of k, we select $\alpha_k < 0$ and $\beta_k > 0$ such that for some decreasing value $\gamma(T)$, the term $\alpha_k i(i-\beta_k) + \gamma(T)$ upper bounds $F_{\tilde{p}}^T(i)$. We wish to investigate the scenario for large values of T, as $T \to \infty$, therefore, we can assume that the binomial distribution, for any value $F_{\tilde{p}}^T(x > 1) < F_{\tilde{p}}^T(x = 1)$ for sufficiently large values of T, as $F_{\tilde{p}}^T(x)$ enters a regime of monotonically decreasing tail behaviour for increasing values of $x > 1$, and small values of $\tilde{p}$. Therefore, in that regime, we can consider $\alpha_k = 0$, and we can upper bound $F_{\tilde{p}}^T(x)$ simply with $\gamma(T)$. Therefore, in this regime, for large values of T, we can say,

$$\lim_{T\to\infty} \mathbb{E}[\text{reg}(T)] \leq \lim_{T\to\infty} \tilde{\Delta}_a \sum_{i=1}^{k-1} i\gamma(T) \tag{A.93}$$

Effectively we have k constants independently of T, which in summation can serve to bound the simple regret in the asymptotic regime. And we can combine the additive terms into some crude upper-bounding additive term C_k.

$$\lim_{T\to\infty} \mathbb{E}[\text{reg}(T)] \leq \lim_{T\to\infty} \sum_{i=1}^{k-1} C_i\gamma(T) \tag{A.94}$$

$$\leq \lim_{T\to\infty} C_k\gamma(T) \tag{A.95}$$

In summary, by coming results from Eq. (A.95) with Eq. (A.88) for large values of $T \to \infty$, we have the following upper-bound on simple regret,

$$\mathbb{E}[\text{reg}(T)] \leq C_k T\tilde{p}(T) \tag{A.96}$$

□

A.7 BOUND ON THE BINOMIAL DISTRIBUTION FOR LARGE T AND SMALL p

Lemma A.2. *For a binomial distribution with large values of T and small values of p, it holds that,*

$$\lim_{T\to\infty} \sum_{i=1}^{T} i\binom{T}{i}p^i(1-p)^{T-i} \leq Tp$$

Proof. We wish to show the following by observing the equivalence between the binomial and Poisson distributions, for small values of $\tilde{p}$ and large values of T. We can see that the first part of Eq. (A.77), is understood as the expectation over binomial distribution, for any integer i. A binomial distribution can be converted to a Poisson distribution as follows,

$$\sum_{i=1}^{k-1} e^{(-Tp)}\frac{(Tp)^i}{i!} \tag{A.97}$$

We then further express as,

$$Tpe^{-Tp}\sum_{i=0}^{k-2}\frac{(Tp)^i}{i!} = Tp\frac{\Gamma(k-1,Tp)}{(k-2)!} \tag{A.98}$$

By the property of an incomplete Gamma function,

$$Tp\frac{\Gamma(k-1,Tp)}{(k-2)!} < Tp \tag{A.99}$$

As we know $\Gamma(k-1) = (k-2)!$ for integer k, and $\Gamma(k,Tp)$ is the upper incomplete gamma function, let us express,

$$\gamma(s,x) = x^s\sum_{i=0}^{\infty}\frac{(-x)^i}{i!(s+i)} \tag{A.100}$$

For a better approximation for small Tp, we can use the series,

$$\gamma(s,x) = \left(\frac{x^s}{s} - \frac{x^{s+1}}{s+1} + \frac{x^{s+2}}{2!(s+2)} - \frac{x^{s+3}}{3!(s+3)} + \cdots\right) \tag{A.101}$$

We thus obtain,

$$\Gamma(s,x) \approx \Gamma(s) - \left(\frac{x^s}{s} - \frac{x^{s+1}}{s+1} + \frac{x^{s+2}}{2!(s+2)} - \frac{x^{s+3}}{3!(s+3)} + \cdots\right)$$

As Tp approaches 0, then we ignore the second terms onward, we can approximate with decreasing approximation error,

$$\sum_{i=1}^{k-1} i\binom{T}{i}\tilde{p}^i(1-\tilde{p})^{T-i} \approx Tp - \frac{(Tp)^k}{(k-1)!} \tag{A.102}$$

Therefore, we can infer the asymptotic performance, as $T \to \infty$ and $p \to 0$, as,

$$\lim_{T\to\infty}\sum_{i=1}^{k-1} i\binom{T}{i}\tilde{p}^i(1-\tilde{p})^{T-i} \leq Tp$$

□

B TECHNICAL NOTES

C GEOMETRIC BROWNIAN MOTION (GBM)

The Geometric Brownian Motion (GBM) is a stochastic process used in various fields including finance and physics, to model the random movement of a variable over time. The stochastic differential equation (SDE) is written as ,

$$dS^t = \mu S^t \, dt + \sigma S^t \, dW_t, \tag{C.1}$$

Where S^t as the value of the process at time t, μ as the drift coefficient, σ is the volatility coefficient, and W_t is a standard Brownian motion. The variable evolves with a mean growth rate μ, and volatility σ determining the magnitude of fluctuations. The solution for the above SDE is written as,

$$S^t = S^1 \exp\left(\left(\mu - \frac{1}{2}\sigma^2\right) t + \sigma W_t\right), \tag{C.2}$$

where S^1 is the initial value of the process at time $t = 0$. The term $\left(\mu - \frac{1}{2}\sigma^2\right)$ represents the adjusted drift.

C.1 DERIVATION OF RECURSIVE VALUE FUNCTION FORMULATION

The derivation of the value function $V(\mathbf{x}^{T-1})$ begins by expressing the optimization problem over the action space $\mathcal{A}(T-1)$ at time $T-1$. The goal is to maximize the sum of an immediate reward term and an expected future reward term. Below is an outline of the derivation steps:

$$V(\mathbf{x}^{T-1}) = \max_{\mathbf{a}\in\mathcal{A}(T-1)} \left\{ \langle \phi f(\mathbf{x}_\eta^{T-1}), \mathbf{a} \rangle + \int_{\mathbf{x}_\eta} P_\theta(\mathbf{x}_\eta^T | \mathbf{x}_\eta^{T-1}) \, \langle (\mathbf{x}_d^{T-1} + \langle \phi' g(\mathbf{x}_\eta^t), \, \mathbf{a}^t \rangle), \phi f(\mathbf{x}_\eta^T) \rangle \, d\mathbf{x}_\eta^T \right\} \tag{C.3}$$

$$= \max_{\mathbf{a}\in\mathcal{A}(T-1)} \left\{ \langle \phi f(\mathbf{x}_\eta^{T-1}), \mathbf{a} \rangle + \langle (\mathbf{x}_d^{T-1} + \langle \phi' g(\mathbf{x}_\eta^t), \, \mathbf{a}^t \rangle), \int_{\mathbf{x}_\eta} P_\theta(\mathbf{x}_\eta^T | \mathbf{x}_\eta^{T-1}) \, \phi f(\mathbf{x}_\eta^T) \rangle \, d\mathbf{x}_\eta^T \right\} \tag{C.4}$$

$$= \max_{\mathbf{a}\in\mathcal{A}(T-1)} \left\{ \langle \phi f(\mathbf{x}_\eta^{T-1}), \mathbf{a} \rangle + \langle (\mathbf{x}_d^{T-1} + \langle \phi' g(\mathbf{x}_\eta^t), \, \mathbf{a}^t \rangle), \, \mathbb{E}[\phi f(\mathbf{x}_\eta^T) | \mathbf{x}_\eta^{T-1}] \rangle \right\}. \tag{C.5}$$

The area of integration over $\mathbf{x}_\eta$ denotes the set of all possible stochastic states reachable from $\mathbf{x}_\eta^{T-1}$ to $\mathbf{x}_\eta^T$ via a stochastic transition, expressed with probability $P_\theta(\mathbf{x}_\eta^T | \mathbf{x}_\eta^{T-1})$ for the stochastic component. Should the deterministic component of the $\mathbf{x}_d^T$ not be reachable via the operation $\mathbf{x}_d^{T-1} + \langle \phi' g(\mathbf{x}_\eta^t), \, \mathbf{a}^t \rangle), \phi f(\mathbf{x}_\eta^T$, it is not included in $\mathbf{x}_\eta$.

D ALGORITHMS

D.1 MCTS UCT BELLMAN VC

Algorithm 1 MCTS Bellman UCT Value Clipping Algorithm

Input: Initialize state (chance node) $\mathbf{x}_0$.
Output: Best action $\mathbf{a}^*$
Define: clip$(v, \underline{v}, \bar{v})$ = min(max$(v, \underline{v}), \bar{v}), \forall v \in \mathbb{R}$
while Max iterations not exceeded. **do**
  $\mathbf{x}_s \rightarrow$ Selection$(\mathbf{x}, \pi_{UCT}(\cdot))$
  $v' \leftarrow 0$
  if $\alpha_c >$ Uniform$[0, 1]$ **then**
    for each possible state $\mathbf{x}'$ extending from inducing action $\mathbf{a}_s$ and state $\mathbf{x}_s$ **do**
      $v \leftarrow$ Simulation$(\mathbf{x}', \pi_s(\cdot))$
      $\underline{v} \leftarrow \underline{V}_k(\mathbf{x}')$ From Eq. (2.27)
      $\bar{v} \leftarrow$ Hindsight perfect solution.
      $V(\mathbf{x}') \leftarrow$ clip$(v, \underline{v}, \bar{v})$
      $v' \leftarrow v' + P(\mathbf{x}'|\mathbf{a}, \mathbf{x})V(\mathbf{x}')$
    end for
  else
    for each possible state $\mathbf{x}'$ extending from inducing action $\mathbf{a}_s$ and state $\mathbf{x}_s$ **do**
      $V(\mathbf{x}') \leftarrow$ Simulation$(\mathbf{x}', \pi_s(\cdot))$
      $v' \leftarrow v' + P(\mathbf{x}'|\mathbf{a}, \mathbf{x})V(\mathbf{x}')$
    end for
  end if
  $Q(\mathbf{x}, \mathbf{a}) \leftarrow \mu(\mathbf{x}, \mathbf{a}) + v'$
  Perform Q-Update according to Eq. (3.2).
  Backpropagation($\mathbf{x}_s$, $Q(\mathbf{x}_s, \mathbf{a})$)
end while
$\mathbf{a}^* \leftarrow \underset{\mathbf{a} \in \mathcal{A}}{\operatorname{argmax}}\, Q(\mathbf{x}_0, \mathbf{a})$
return $\mathbf{a}^*$

D.2 MCTS MENTS BELLMAN VALUE CLIPPING

Algorithm 2 MCTS Bellman MENTS Value Clipping Algorithm

Input: Initialize state (chance node) $\mathbf{x}_0$.
Output: Best action $\mathbf{a}^*$
Define: clip$(v, \underline{v}, \bar{v})$ = min(max$(v, \underline{v}), \bar{v}), \forall v \in \mathbb{R}$
while Max iterations not exceeded. **do**
  $\mathbf{x}_s \rightarrow$ Selection$(\mathbf{x}, \pi_M(\cdot))$
  $v' \leftarrow 0$
  if $\alpha_c >$ Uniform$[0, 1]$ **then**
    for each possible state $\mathbf{x}'$ extending from inducing action $\mathbf{a}_s$ and state $\mathbf{x}_s$ **do**
      $v \leftarrow$ Simulation$(\mathbf{x}', \pi_s(\cdot))$
      $\underline{v} \leftarrow \underline{V}_k(\mathbf{x}')$ From Eq. (2.27)
      $\bar{v} \leftarrow$ Hindsight perfect solution.
      $V_s(\mathbf{x}') \leftarrow$ clip$(v, \underline{v}, \bar{v})$
      $v' \leftarrow v' + P(\mathbf{x}'|\mathbf{a}, \mathbf{x})V_s(\mathbf{x}')$
    end for
  else
    for each possible state $\mathbf{x}'$ extending from inducing action $\mathbf{a}_s$ and state $\mathbf{x}_s$ **do**
      $V_s(\mathbf{x}') \leftarrow$ Simulation$(\mathbf{x}', \pi_s(\cdot))$
      $v' \leftarrow v' + P(\mathbf{x}'|\mathbf{a}, \mathbf{x})V_s(\mathbf{x}')$
    end for
  end if
  $Q(\mathbf{x}, \mathbf{a}) \leftarrow \mu(\mathbf{x}, \mathbf{a}) + v'$
  Perform Value and Q-Update according to Eq. (3.4) and Eq. (3.5).
  Backpropagation($\mathbf{x}_s$, $Q_{\text{sft}}(\mathbf{x}_s, \mathbf{a})$)
end while
$\mathbf{a}^* \leftarrow \operatorname*{argmax}_{\mathbf{a} \in \mathcal{A}} Q_{\text{sft}}(\mathbf{x}_0, \mathbf{a})$
return $\mathbf{a}^*$

E MARITIME BUNKERING

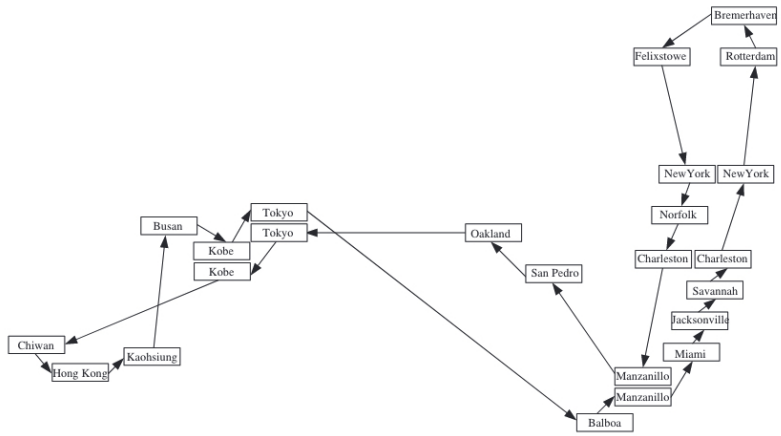


Figure 5: Atlantic Pacific Express (APX) liner route. [YNL12]

Maritime refuelling, also known as *bunkering*, is a problem in the field of transportation logistics that involves finding the optimal policy to refuel a fleet of vehicles, such as trucks or ships, while they are in operation. The goal is to minimize the total cost of refuelling, which includes the cost of the fuel itself. It is a common practice in the shipping industry, as ships require large amounts of fuel to power their engines and systems during long voyages. Bunkering is typically done at ports, where the ship can be moored and connected to a fuel supply by hoses or pipelines. Bunkering can also be done at sea, using smaller vessels known as bunker barges to transfer the fuel to the ship. Solving the bunkering problem can allow companies in the transportation industry reduce costs and improve the efficiency of their operations by minimizing fuel costs.

A real-world examples is the Atlantic Pacific Express (APX) liner routes (see Fig. 5). In the liner scenario, a schedule of port visits is prearranged. That is the liner must only determine how much to refuel at each port of call, and at which speed to travel to the next port. But, for simplicity, we omit the speed component. The liner knows which location they are at, how much fuel they possess, how much minimum fuel they need, and the prices of the fuel at all ports in the schedule. We assume that prices follow a known probability distribution. In the normal macroeconomic conditions that is a reasonable assumption since the prices can be estimated from the historical data, while during the turbulent periods designing stress price scenarios might be the best we can do.

One way to approach the bunkering problem is to use mathematical optimization techniques, such as mixed integer programming, to determine the optimal refuelling amounts, given a fixed schedule or flexible schedule[1]. This may involve considering factors such as the capacity of the vessel, the distance they need to travel, the availability of fuel at different locations, and any constraints on when and where the vehicles can be refuelled. A liner is a type of shipping vessel that operates on a regular schedule between specified ports, carrying cargo and sometimes passengers. Liner shipping refers to the use of these vessels to transport cargo on a regular basis between predetermined ports of call. Liners are typically owned and operated by shipping companies, and they follow a set route, stopping at a predetermined list of ports to load and unload cargo. Liners are an important part of the global shipping industry, as they provide a reliable and efficient way to transport a wide variety of goods, including consumer goods, raw materials, and manufactured products. They play a vital role in supporting international trade and the movement of goods around the world.

Ship Route (Schedule): In this experiment, we assume that the ship route (i.e. order of the ports) is fixed, and that the fuel is available at all ports on the route. The objective is finding the optimal refuelling policy. Upon arrival at each port, the price is revealed and we need to determine the refuelling amount. In practice, bunkering purchasing problems typically involve big data (i.e., 500 vessels, 40,000 ports of call, governed by 750 contracts [BKP16]).

Fuel prices: Fuel prices can be governed by various stochastic processes – in this work, we assume that the fuel prices obey a discrete time geometric Brownian motion, as described in Sec. C. We discretize the GBM over many different price outcomes, stemming from realizations of the stochastic process. Therefore, the problem in principle can be solved approximately via stochastic programming, stochastic programming [WML13; YNL12]. Nevertheless, as the scale of the problem increases, i.e. more price scenarios arise, the problem becomes intractable via an increase in stochastic scenarios. Each liner is constrained by predetermined schedules and must ensure that there is enough fuel to travel between each

[1] A ship with a flexible schedule is referred to as a tramper.

intermediate port of call. The liner can choose the refuel amount and the speed of the ship, which affects fuel consumption. The maritime bunkering model is subject to the following simplifying model assumptions,

i. Fuel prices are subject to global stochastic variation.
ii. Fuel consumption is deterministic.
iii. Distance, to and from each port, is fixed and deterministic.
iv. No possibility of service disruptions.

E.1 STOCHASTIC PROGRAMMING

Table 2: **Notation and variable descriptions.**

Notation	Description
N	Number of ports of call
S	Number of scenarios
K	Number of stochastic events at each trip step t
$\delta_{n,t}$	Fuel price change change in the port n at the trip step t
β^s	Probability of the scenario $s \in S$ occurring
$X^s_{n,1}$	Fuel level when arriving at port n in scenario s
$X^s_{n,2}$	Fuel level when departing from port n in scenario s
$X_{n,1}$	Fuel level when arriving at port n
$X_{n,2}$	Fuel level when departing from port n
$d(n, n+1)$	Distance function from the port n until the next port $n+1$
$f_c(n, n+1)$	Fuel consumption function from the port n until the next port $n+1$ given the distance $d(n, n+1)$
$\mathcal{N}$	Continuous fuel price percentage change probability distribution
P^s_n	Price of fuel at port n in scenario s
P_n	Expected price of fuel at port n
$Y^s_n \in \{0, 1\}$	Indicator for bunkering decision at port n in scenario s
$Y_n \in \{0, 1\}$	Indicator for bunkering decision at port n
B^s_n	Fixed bunkering cost at port n in scenario s
B_n	Fixed bunkering cost at port n
M	Liner fuel tank capacity

We formulate the liner bunkering problem as a stochastic program. The stochastic factors affect the price of fuel through the price percentage changes expressed by S scenarios. We optimize the actions to take at each port of call. Concretely, at each port n the liner must decide on a refueling amount, denoted as the amount of fuel leaving port n in scenario s, $X^s_{n,2}$, subtracted by the amount of fuel arriving at port n in scenario s, $X^s_{n,1}$. The fuel consumption, i.e., the difference $X^s_{n,2} - X^s_{n+1,1}$, is deterministic given the distance until the next port $d(n, n+1)$ which affects a deterministic fuel consumption $f_c(n, n+1)$. In addition, there is also a fixed bunkering cost B^s_n in case we decide to refuel non-negative amount at port n in scenario s. We indicate the bunkering action by a binary variable Y^s_n. We assume that the liner has the empty fuel tank at the initial port and we constrain its fuel tank capacity by the upper bound M.

Given fuel price percentage change scenarios $s \in S$, which determine the fuel prices P^s_n we have:

$$\min_X \quad C^{SP} = \sum_{s \in S} \beta^s \sum_{n \in N} P^s_n (X^s_{n,2} - X^s_{n,1}) + B^s_n Y^s_n + \tau(T^k_{n+1}) \tag{E.1a}$$

$$\text{subject to} \quad X^s_{n+1,1} = X^s_{n,2} - f_c(n, n+1, V^k_n), n \in N, s \in S \tag{E.1b}$$

$$X^s_{n,2} \geq X^s_{n,1}, n \in N, s \in S \tag{E.1c}$$

$$f_c(n, n+1, V^k_n) \geq 0, n \in N \tag{E.1d}$$

$$Y^s_n \in \{0, 1\}, n \in N, s \in S \tag{E.1e}$$

$$X^s_{n,i} \geq 0, n \in N, s \in S, i \in \{1, 2\} \tag{E.1f}$$

$$X^s_{n,i} \leq M, n \in N, s \in S, i \in \{1, 2\} \tag{E.1g}$$

$$X^s_{1,1} = 0, s \in S \tag{E.1h}$$

We have the stochastic value objective represented by (E.1a), which is the sum of the refuelling costs per port denoted as $P^s_n(X^s_{n,2} - X^s_{n,1})$, the fixed bunkering costs denoted as $B^s_n Y^s_n$, and the time window penalties for arriving late or early to the destination denoted as $\tau(T^k_{n+1})$. (E.1b) is the consumption balance constraint, (E.1c) prevents negative refuelling amounts, (E.1d) states that the consumption is non-negative, (E.1e) is an indicator variable for the fixed bunkering cost, (E.1f) and (E.1g) are lower and upper fuel tank capacities, while (E.1h) imposes the empty fuel tank at the initial port. Each

of these constraints hold for every scenario $s \in S$.

E.1.1 Modelling Fuel Price

We simulated $N_{\text{GBM}} = 200,000$ price trajectories using the Geometric Brownian Motion (GBM) model as illustrated in Sec. C, starting with an initial stock price S_0. Each trajectory was generated by iterating over T time steps, applying the GBM formula $S^t = S^{t-1} \cdot \exp\big((\mu - 0.5\sigma^2)\Delta t + \sigma \Delta W_t\big)$, where ΔW_t represents the increments of a Wiener process. This process involved parameters for drift (μ), volatility (σ), under a fixed discrete time increment (Δt), ensuring the randomness and variability in the simulated price paths. To estimate the probability density function of the simulated prices, we flattened the simulation results and created a histogram with a specified number of bins, and thereafter calculate the probability density for any given price value based on the histogram data. We can this assign probabilities to each simulated price outcome, ensuring a comprehensive probability distribution across the entire range of simulated trajectories.

E.2 TABLE OF PARAMETERS

Table 3: **GBM Stochastic price parameters.**

Config.	**Initial Price** (S^1)	**Price Volatility** (σ)	**Price Drift** (μ)
A	1000	0.9	1.0
B	1000	0.5	1.0
C	100	0.9	1.0
D	1000	0.5	0.5
E	1000	0.9	0.5
F	100	0.9	0.5

Table 4: **Shared parameters across all experimental configurations.**

Description	**Value**
N_{GBM} GBM simulated trajectories.	200,000
N_{H} Number of histogram bins.	20,000
N_{sim} MCTS Number of Iterations	1×10^5
N_{depth} MCTS Depth Limit	500
λ_s MENTS Decay Rate	2×10^9
Number of ports-of-call.	8
Fuel capacity.	50 Units

Table 5: Distance between each port-of-call represented by a distance matrix.

	1	2	3	4	5	6	7	8
1	0	12	7	15	12	18	3	4
2	12	0	25	8	10	15	6	14
3	7	25	0	30	20	16	12	10
4	15	8	30	0	19	25	30	8
5	12	10	20	19	0	9	18	13
6	18	15	16	25	9	0	21	10
7	3	6	12	30	18	21	0	17
8	4	14	10	8	13	10	17	0

Fuel consumption rate is simplified to 1 unit of fuel consumed to 1 unit of distance travelled.

Table 6: **MCTS Dynamic Parameters.**

Config.	N_{sim}	MCTS Exploration Constant (α)
A	1000	0.9
B	1000	0.5
C	100	0.9
D	1000	0.5
E	1000	0.9
F	100	0.9

E.3 EMPIRICAL RESULTS - COST COMPARISON

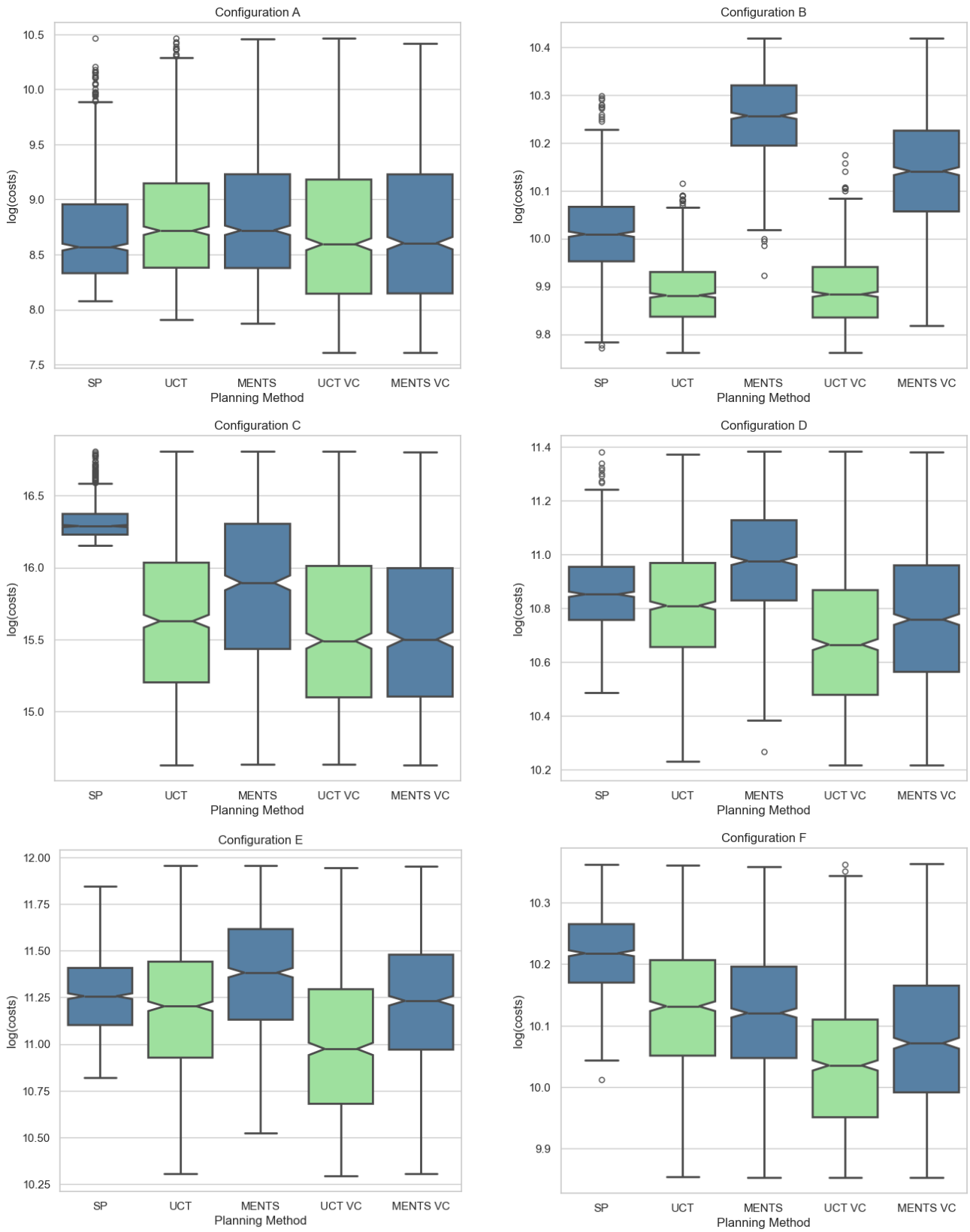


Figure 6: Cost performance of the maritime logistics simulation.

E.4 EMPIRICAL RESULTS - VALUE CONVERGENCE

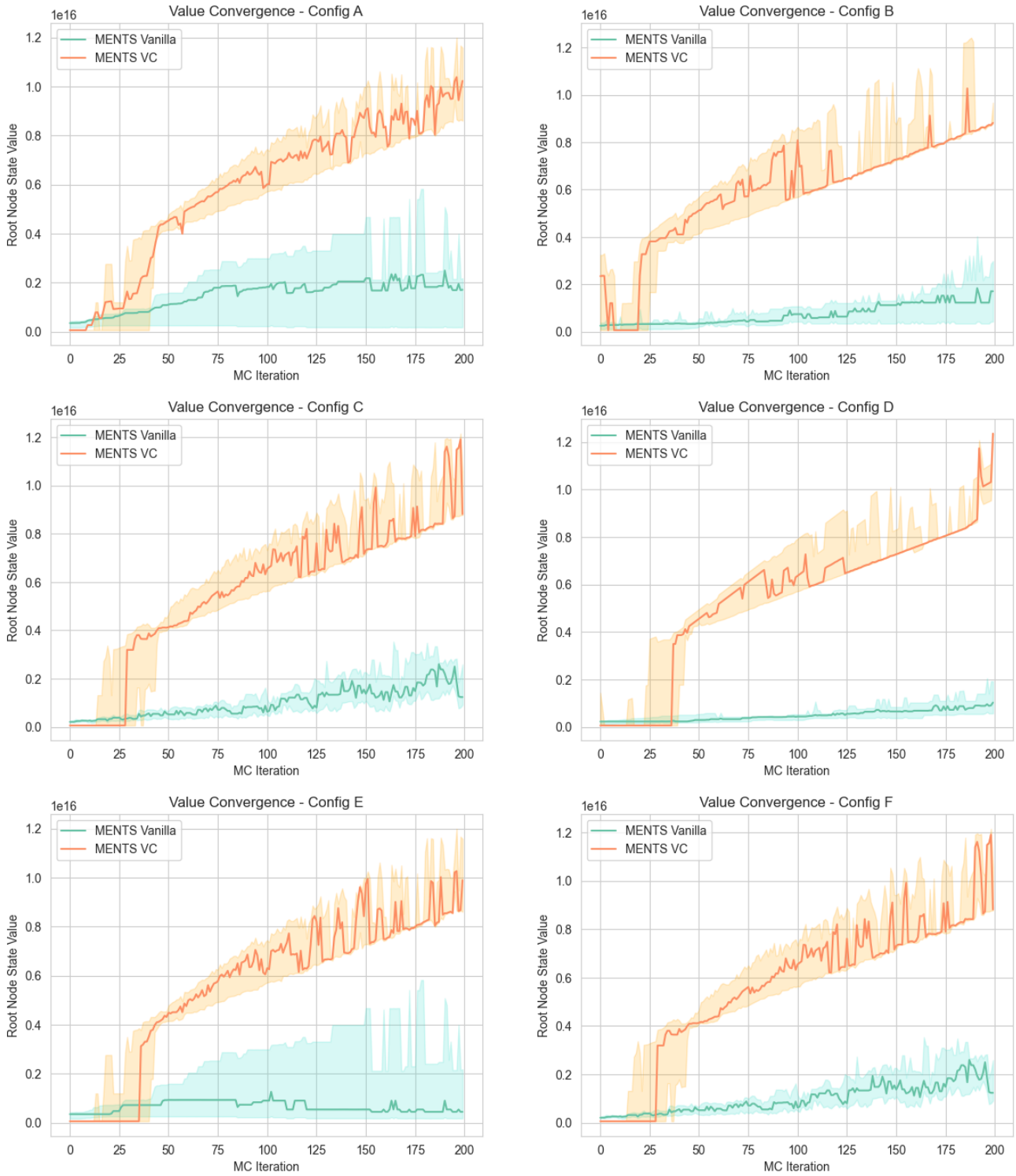


Figure 7: Value convergence performance of the maritime logistics simulation.

F HYBRID FUEL EXAMPLE

Hybrid fuel systems are an increasingly common feature in modern transportation, offering a flexible approach to energy utilization by combining different types of fuel or power sources to optimize efficiency, cost, and environmental impact. In the context of vehicles, this concept is well-illustrated by hybrid electric cars that use a combination of internal combustion engines and electric motors to achieve superior fuel economy and lower emissions. Similarly, hybrid systems have found applications in marine vessels, where multi-fuel engines can switch between liquefied natural gas (LNG), diesel, and even renewable energy sources such as wind power, depending on operational needs and fuel availability.

Problem statement: At each increment of time, an amount of $\underline{\Delta}_a(t) \leq ||\mathbf{x}_d^{t+1} - \mathbf{x}_d^t||_p \leq \bar{\Delta}_a(t)$ of resources are to be consumed. This is a function of a linear combination of the consumption rates $\langle \phi' g(\mathbf{x}_\eta^t),\, \mathbf{a}^t \rangle$. We say that some quanta of interchangeable quantity is expended at each time increment $||\mathbf{x}_d^{t+1} - \mathbf{x}_d^t||_p = \Delta_a(t)$. The controller will decide at each point in time t the convex combination of resources to expend, given the current stochastic scneario $\mathbf{x}_\eta^t$ and a model depicting future scenarios $\mathbf{x}_\eta^{\mathbb{t}}$. To make this problem non-trivial to solve, we impose limitations on each resource.

F.1 DUAL POWER HYBRID VEHICLE

Electric and gasoline use dual power for increased efficiency as some power sources are more beneficial than others in certain scenarios. For example, lower speeds prefer electric power, and higher speeds prefer gasoline, in terms of resource efficiency. Also, one can blend power from both sources. So here the concept is that two (or more) capacities are available (fuel tank, battery), and either of this resource can be converted into some utility.

Problem setting (Hybrid Vehicle): Over the course of a journey a hybrid vehicle has two possible power sources, `electric` and `gasoline`. At each decision period, a well-defined quanta of resource is consumed denoted as $\Delta_a(t) \in \{\underline{\Delta}_a(t), \bar{\Delta}_a(t)\}$. We set $p = 1$, such that $\Delta_a(t) = \langle \phi' g(\mathbf{x}_\eta^t),\, \mathbf{a}^t \rangle$. This implies some convex combination of fuels must be consumed. To facilitate regenerative braking, $\underline{\Delta}_a(t)$ represents the fuel consumption when the vehicle is braking and regenerating, and $\bar{\Delta}_a(t)$ represents the fuel consumption when the vehicle is not braking. The path constraints denote the capacity, $\bar{A}$, and we set trivially, $\underline{A} = \sum \underline{\Delta}_a(t)$.

Utility Model: We define a utility model, typically we could say this is *distance travelled* (and alternative definitions could be environment impact in terms of emissions etc.). Thus we have a stochastic scenario, with mileage $f(\mathbf{x}_\eta^t)$ evolving naturally, and resource space of dimension D, if there are D types of fuel which can be consumed by the agent. Given the rate of conversion to utility, $f(\mathbf{x}_\eta^t)$, can vary, the rate of consumption $\langle \phi' g(\mathbf{x}_\eta^t),\, \mathbf{a}^t \rangle$ is constrained.

Suppose a $D = 2$ resource constraint, the agent decides which resource to expend at each turn. In the fuel consumption example the decision could be made at each unit of time in the journey, which conversely can $\phi f(\mathbf{x}_\eta^t)$ can represent the pro rata mileage per fuel. We allow for potentially a combination of resources to be used, and we assume that within $\Delta_a(t)$ the rate of resource consumption, based on the environmental parameters are constant. The rates of consumption could be different per resource, but a specified amount of resources must be consumed per quanta of time.

Regenerative Braking: We also add to the mix regenerative braking. Thus there is an $\mathbf{x}_\eta$ which tells us that the vehicle is braking. In this case regenerative braking can be activated to replenish the battery capacity, consequently limiting and fixing $\langle \phi' g(\mathbf{x}_\eta^t),\, \mathbf{a}^t \rangle = i$ to some negative constant, as we regain resources while automatically replenishing.

Experiment: Let D = 2 types of fuel, `gas` and `electric`. In our example a vehicle is traversing a predetermined trajectory, and in in this trajectory the vehicle can experience distinct stochastic phases which affect the mileage of the vehicle. Let $\langle \phi f(\mathbf{x}_\eta^t),\, \mathbf{a}^t \rangle$ denote mileage, and $\langle \phi' g(\mathbf{x}_\eta^t),\, \mathbf{a}^t \rangle$ represent fuel consumption, which is subject to a fixed schedule independent our the agent's decision process. As each discrete time period, the agent (onboard computer) must decide how much of, and which type of, fuel to expend for a fixed increment of fuel consumed governed by Dynamic (**D3**). The transition of $\mathbf{x}_\eta^t$ is expressed as a Markovian process.

Transition between modes: Let the modes be M_1, M_2, M_3. The transition between modes is governed by a transition matrix P, where:

$$P = \begin{bmatrix} P(M_1 \to M_1) & P(M_1 \to M_2) & P(M_1 \to M_3) \\ P(M_2 \to M_1) & P(M_2 \to M_2) & P(M_2 \to M_3) \\ P(M_3 \to M_1) & P(M_3 \to M_2) & P(M_3 \to M_3) \end{bmatrix},$$

and $\sum_{j=1}^{3} P(M_i \to M_j) = 1$ for all $i \in \{1, 2, 3\}$.

Sampling from modes: When the system is in mode M_i, a state $\mathbf{x}_\eta^t$ is sampled from the corresponding distribution $p_i(\mathbf{x}_\eta)$:

$$\mathbf{x}_\eta^t \sim p_i(\mathbf{x}_\eta), \quad \text{if the mode at time } t \text{ is } M_i.$$

In a Hidden Markov Model (HMM), the probability of transitioning from one latent state to the next is determined by the transition matrix P, and the observation at each step depends on the current state. The probability of transitioning from state M^t at time t to state m^{t+1} at time $t+1$ is:

$$P(M^{t+1} \mid M^t) = P(M^t \to M^{t+1}),$$

where $P(M^t \to M^{t+1})$ is the (M^t, M^{t+1}) entry of the transition matrix P. The joint probability of the sequence of hidden states $\{M^1, M^2, \ldots, M^T\}$ is given by:

$$P(M^1, M^2, \ldots, M^T) = P(M^1) \prod_{t=1}^{T-1} P(M^{t+1} \mid M^t),$$

where $P(M^1)$ is the initial state probability. The observation sequence $\{x^1, x^2, \ldots, x^T\}$ is conditionally independent given the hidden states, and the emission probability is:

$$P(X^t \mid M^t) = p_{M^t}(x^t),$$

where $p_{M^t}(x^t)$ is the distribution of x^t given the latent state M^t, also known as the *emission probability*. Combining transitions and emissions, the full joint probability is:

$$P(M^1, x^1, \ldots, M^T, x^T) = P(M^1) f_{M^1}(x^1) \prod_{t=1}^{T-1} P(M^{t+1} \mid M^t) p_{M^{t+1}}(x^{t+1}).$$

For a single step transition, the probability simplifies to:

$$P(M^{t+1} \mid M^t) \cdot p_{M^{t+1}}(x^{t+1}).$$

Joint process: The process alternates between mode transitions and state sampling, resulting in a sequence $(M^t, \mathbf{x}_\eta)$ where:

- $M^t \in \{M^1, M^2, M^3\}$ is the mode at time t, determined by the transition matrix P.
- $\mathbf{x}_\eta$ is the observed state, sampled from the distribution associated with M^t.

The joint probability of a sequence $\{M^1, \mathbf{x}_\eta^1, M^2, \mathbf{x}_\eta^2, \ldots, M^T, \mathbf{x}_\eta^T\}$ is given by:

$$P(M^1, \mathbf{x}_\eta^1, \ldots, M^T, \mathbf{x}_\eta^T) = P(M^1) P(\mathbf{x}_\eta^1 | M^1) \prod_{t=2}^{T} P(M^t | M^{t-1}) P(\mathbf{x}_\eta^t | M^t),$$

where $P(M^1)$ is the initial probability distribution over modes, $P(\mathbf{x}_\eta^t | M^t)$ is the probability density (or mass) of $\mathbf{x}_\eta^t$ under the mode M^t, and $P(M^t | M^{t-1})$ is the transition probability between modes. In a structure resembling a Hidden Markov Model (HMM), the observed sequence is $\{\mathbf{x}_\eta^1, \mathbf{x}_\eta^2, \ldots, \mathbf{x}_\eta^T\}$, while the modes $\{M^1, M^2, \ldots, M^T\}$ are latent variables.

Objective: We would therefore like to solve the equation that maximizes the mileage (or distance travelled) by the hybrid vehicle given allocation of the fuel type and mount of fuel over an evolving trajectory, with modes M_1, M_2, M_3, where the agent observes $\mathbf{x}_\eta^t$ at time t. Each of the 3 modes represents different regimes, where it is either gas efficient, electric efficient, or regenerative braking. An example could be,

$$\mathbb{E}[f(\mathbf{x}_\eta) | M_1] = \begin{bmatrix} 12 \\ 3 \end{bmatrix}, \quad \mathbb{E}[f(\mathbf{x}_\eta) | M_2] = \begin{bmatrix} 2 \\ 9 \end{bmatrix}, \quad \mathbb{E}[f(\mathbf{x}_\eta) | M_3] = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{F.1}$$

We see from Eq. (F.1) that there could exist multiple modes for $\mathbb{E}[f(\mathbf{x}_\eta)]$, depending on M, where the first row of $f(\mathbf{x}_\eta)$

represents the mileage for gasoline, and second row mileage for electricity. Here, the 3 distinct modes, represents a mode where gasoline is the more efficient power source, and second where electricity is most efficient, and third where regenerative braking occurs and no fuel should be consumed as a result.

Non-Greedy Optimal Resource Consumption: A stochastic model governs the transition of $\mathbf{x}_\eta^t$. We maintain a fixed incremental fuel consumption, $\bar{\Delta}_a = \underline{\Delta}_a$, to formulate the assignment problem. When $\mathbf{x}_\eta$ is revealed to us, we know which fuel to use to allow for the best mileage, but we cannot always act myopically, as this exploit is limited. It should be the case that neither fuel alone, or each individual maximum expenditure of a single fuel, can allow the vehicle to reach its destination, but a combination of both fuels must be used to complete the journey. Suppose $\bar{\mathbf{a}}_i$ represents the use of a single fuel throughout the journey. We impose that,

$$\sum_{t=1}^{T} \langle \phi' g(\mathbf{x}_\eta^t),\, \bar{\mathbf{a}}_i \rangle \leq \underline{\mathrm{A}}, \quad \exists i \in D \tag{F.2}$$

This indicates that there exists at least one resource type, which cannot be exclusively used through the course of the journey, and the design of the problem should reflect that. This implies that applying the resource that is the most efficient at the current time, may not be the most optimal solution overall (as the agent may deplete such resources for future use) - in other words for at least one resource cannot be repeatedly used indefinitely, there exists a limit.

F.2 BASELINE SOLUTION - VALUE ITERATION WITH BELIEF MDP

As a baseline solution method we apply a standard value iteration approach given observed states $\mathbf{x}_\eta$, without considering the hidden state M. The action space $\mathbf{a}$ is discretized into discrete levels, indicating the amount of fuel to consume. Since the underlying state is not directly observable, the problem can be reformulated as a *belief-state Markov Decision Process*. The belief state represents a probability distribution over states and evolves according to Bayesian filtering. The value function is then defined over the belief space rather than individual states.

The belief-state value function is given by the equation:

$$V^*(b) = \max_{a \in A} \sum_{M \in \mathcal{M}} b(s) \sum_{M^{t+1} \in \mathcal{M}} P(M^{t+1} \mid M^t, a) \sum_{\mathbf{x}_\eta \in \mathcal{X}} P(\mathbf{x}_\eta^{t+1} \mid M^{t+1}) \left[\langle \phi f(\mathbf{x}_\eta^{t+1} | M^{t+1}), \mathbf{a} \rangle + \gamma V^*(b') \right], \tag{F.3}$$

where the transition probability function, observation probability function, and reward function govern the dynamics of the system. The updated belief state after taking an action and receiving an observation is computed using Bayes rule:

$$b'(M^{t+1}) = \frac{P(\mathbf{x}_\eta^{t+1} \mid M^{t+1}) \sum_{M \in \mathcal{M}} P(M^{t+1} \mid M^t, \mathbf{a}) b(M^t)}{P(\mathbf{x}_\eta^{t+1} \mid b, \mathbf{a})}, \tag{F.4}$$

where the probability of observing a particular outcome given the belief and action is:

$$P(\mathbf{x}_\eta^{t+1} \mid b, \mathbf{a}) = \sum_{M^{t+1} \in \mathcal{M}} P(\mathbf{x}_\eta^{t+1} \mid M^{t+1}) \sum_{M \in \mathcal{M}} P(M^{t+1} \mid M^t, \mathbf{a}) b(M^t). \tag{F.5}$$

Value iteration proceeds iteratively by updating the belief-state value function as follows:

$$V_{k+1}(b) = \max_{\mathbf{a} \in \mathcal{A}} \sum_{M \in \mathcal{M}} b(M^t) \left[R(M^t, \mathbf{a}) + \gamma \sum_{\mathbf{x}_\eta^{t+1} \in \mathcal{X}} P(\mathbf{x}_\eta^{t+1} \mid b, \mathbf{a}) V_k(b^{t+1}) \right]. \tag{F.6}$$

F.3 EXPERIMENTAL CONFIGURATIONS

The following are two configuration matrices, for the transition of modes, Configurations `A`, `B`, and `C` adheres to $\mathbf{T}_1$ and Configurations `D`, `E`, and `F` adhere to $\mathbf{T}_2$.

$$\mathbf{T}_1 = \begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0.1 & 0.7 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{bmatrix} \quad \mathbf{T}_2 = \begin{bmatrix} 0.4 & 0.4 & 0.2 \\ 0.4 & 0.4 & 0.2 \\ 0.4 & 0.4 & 0.2 \end{bmatrix}$$

Config.	Mileage Matrices	T	Δ_a	Δ_d
A	$[[10,8]^\top,[8,9]^\top,[8,8]^\top]$	10	4	-2
B	$[[5,2]^\top,[2,5]^\top,[2,2]^\top]$	10	4	-2
C	$[[6,3]^\top,[3,7]^\top,[3,3]^\top]$	15	5	-3
D	$[[4,2]^\top,[2,6]^\top,[2,2]^\top]$	20	3	-1
E	$[[8,3]^\top,[3,5]^\top,[3,3]^\top]$	12	6	-4
F	$[[30,10]^\top,[10,40]^\top,[10,10]^\top]$	16	5	-5

Table 7: Hybrid vehicle mileage and regenerative braking configurations.

Table 8: **Shared MCTS parameters across all experimental configurations.**

Description	Value
No. of Simulations N_{sim}	1000
Exploration Constant (C)	1.0
Simulation Depth Limit (N_{depth})	10
Discount Factor (γ)	0.9
MENTS Temperature (T)	0.7
MENTS Epsilon (ϵ)	0.2

F.4 EMPIRICAL RESULTS - REWARD COMPARISON

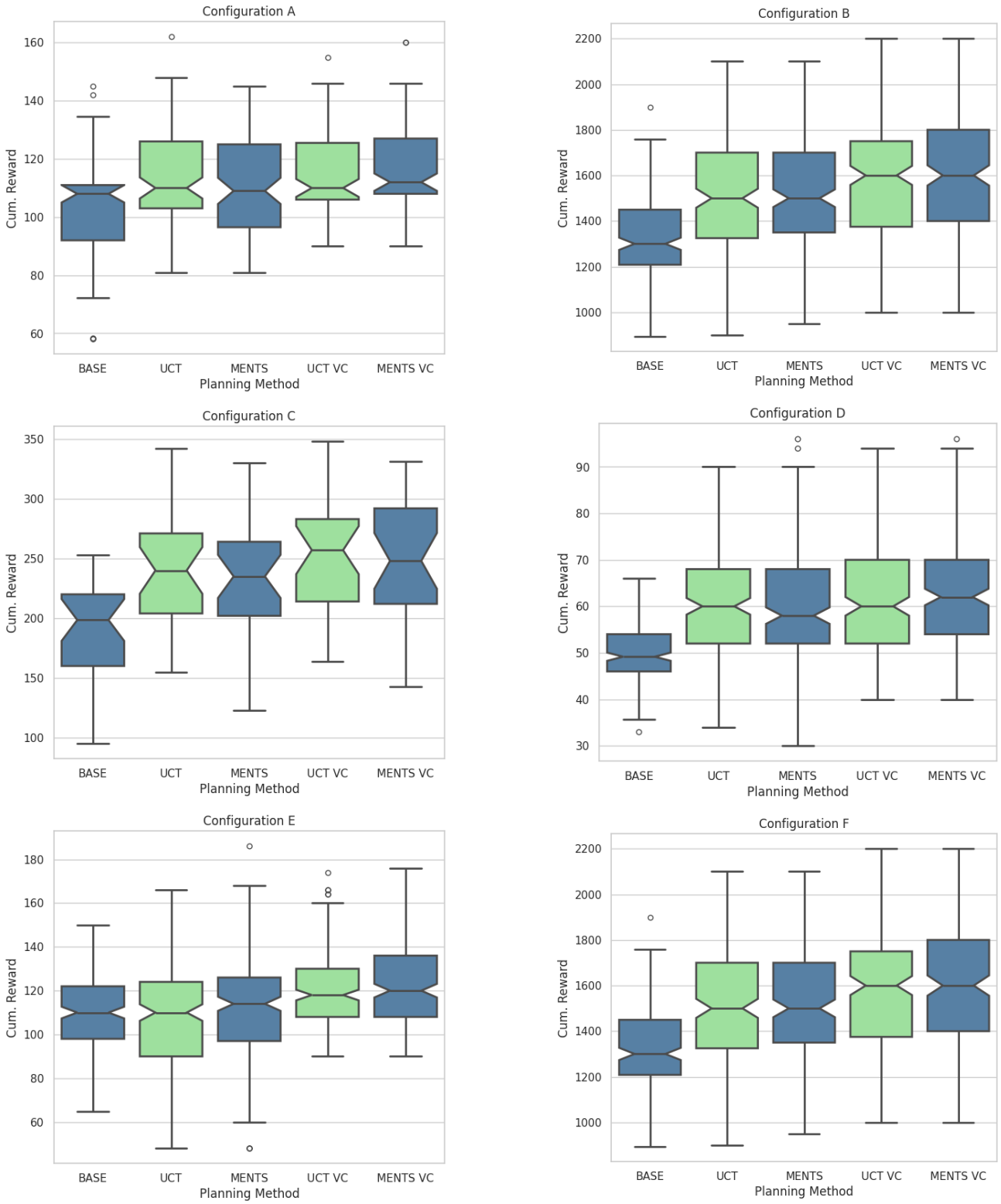


Figure 8: Distance travelled on fuel constraints (reward) per policy. Baseline solution constitutes simple value iteration solution over the MDP.

F.5 EMPIRICAL RESULTS - VALUE CONVERGENCE

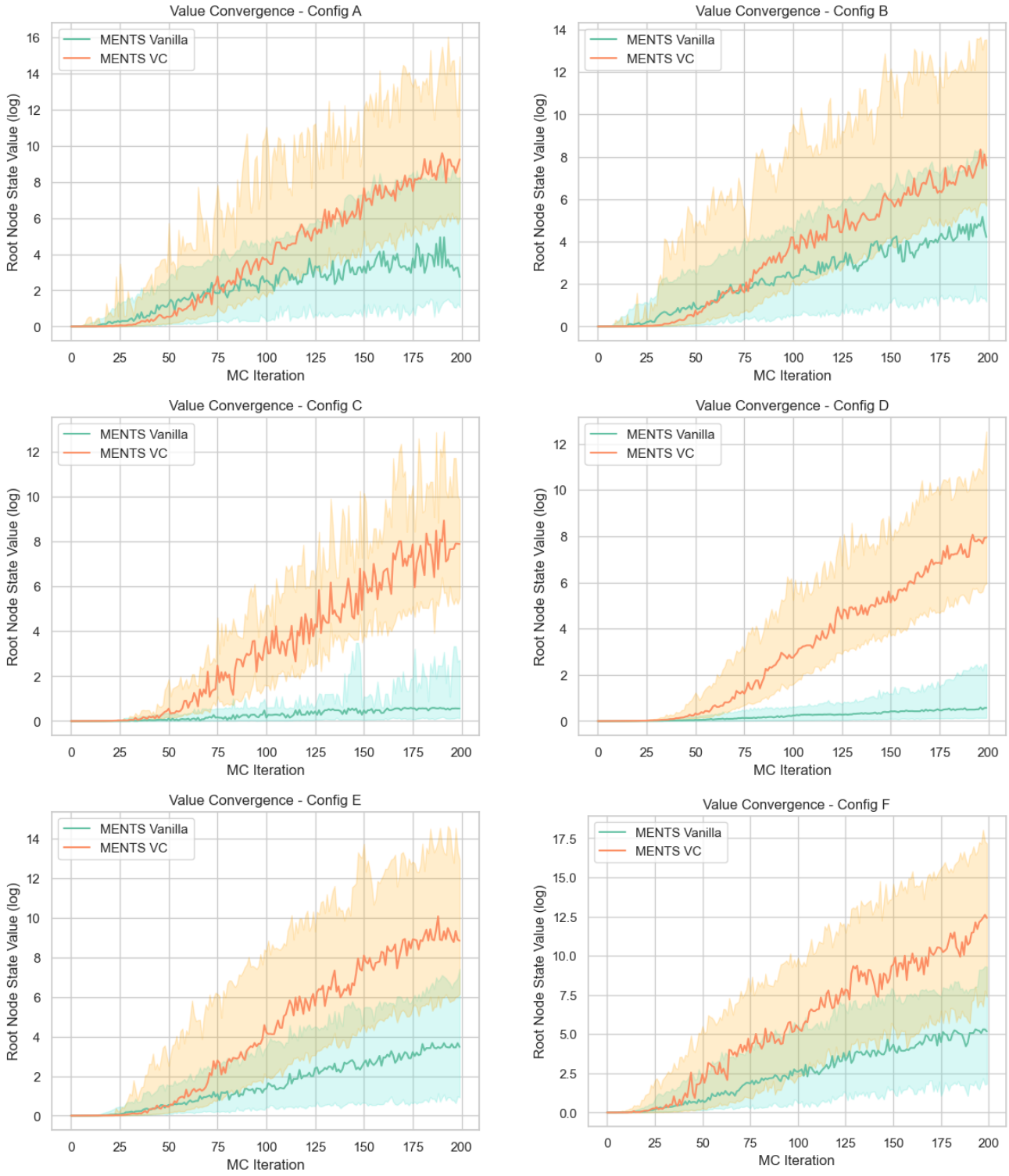


Figure 9: Value convergence performance of the hybrid fuel simulation.

F.6 HYBRID FUEL EXPERIMENTAL CONFIGURATIONS - HIGHER DIMENSION

As an extension of the 2-energy source (gas, electric) hybrid fuel example, we model a hybrid energy system as an expanded finite-state Markov process with seven distinct operational modes: three gasoline types which operate efficiently in the gasoline-efficient states (`ge`, `ge2`, `ge3`), three battery types which operate efficiently in electricity efficient states (`ee`, `ee2`, `ee3`), and one regenerative braking state (`rb`). Each state represents an optimal operating regime in which the system achieves maximum efficiency for its respective energy source. These parameters are specified in Table 9.

Config.	**Mileage Matrices**	T	Δ_a	Δ_d
A	$[[10,8]^\top, [11,8]^\top, [12,8]^\top, [8,9]^\top, [8,10]^\top, [8,11]^\top, [8,8]^\top]$	10	4	-2
B	$[[5,2]^\top, [6,2]^\top, [7,2]^\top, [2,3]^\top, [2,4]^\top, [2,5]^\top, [2,2]^\top]$	10	4	-2
C	$[[10,4]^\top, [15,4]^\top, [20,4]^\top, [4,15]^\top, [4,10]^\top, [4,5]^\top, [4,4]^\top]$	15	5	-3
D	$[[22,4]^\top, [26,4]^\top, [32,4]^\top, [4,15]^\top, [4,16]^\top, [4,17]^\top, [4,4]^\top]$	20	3	-1
E	$[[1,1]^\top, [3,1]^\top, [6,1]^\top, [1,6]^\top, [1,9]^\top, [1,12]^\top, [1,1]^\top]$	12	6	-4
F	$[[30,10]^\top, [40,10]^\top, [50,10]^\top, [10,15]^\top, [10,20]^\top, [10,30]^\top, [10,10]^\top]$	16	5	-5

Table 9: Hybrid vehicle mileage and regenerative braking configurations: The first three vectors represent mileage values for gasoline-efficient operating modes (`ge`, `ge2`, `ge3`), the next three vectors correspond to mileage values for electricity-efficient operating modes (`ee`, `ee2`, `ee3`), and the final vector represents the regenerative braking mode (`rb`).

State transitions follow a discrete-time Markov chain characterized by the transition matrix $\mathbf{T} \in \mathbb{R}^{7\times 7}$, where T_{ij} denotes the transition probability from state i to state j. The system uses two distinct configuration matrices: Configurations `A`, `B`, and `C` follow $\mathbf{T}_1$, while Configurations `D`, `E`, and `F` follow $\mathbf{T}_2$.

$$\mathbf{T}_1 = \begin{array}{c} \\ \texttt{ge} \\ \texttt{ge2} \\ \texttt{ge3} \\ \texttt{ee} \\ \texttt{ee2} \\ \texttt{ee3} \\ \texttt{rb} \end{array} \begin{array}{c} \begin{array}{ccccccc} \texttt{ge} & \texttt{ge2} & \texttt{ge3} & \texttt{ee} & \texttt{ee2} & \texttt{ee3} & \texttt{rb} \end{array} \\ \begin{pmatrix} 0.45 & 0.20 & 0.10 & 0.10 & 0.05 & 0.05 & 0.05 \\ 0.20 & 0.45 & 0.10 & 0.05 & 0.10 & 0.05 & 0.05 \\ 0.10 & 0.20 & 0.45 & 0.05 & 0.05 & 0.10 & 0.05 \\ 0.10 & 0.05 & 0.05 & 0.45 & 0.20 & 0.10 & 0.05 \\ 0.05 & 0.10 & 0.05 & 0.20 & 0.45 & 0.10 & 0.05 \\ 0.05 & 0.05 & 0.10 & 0.10 & 0.20 & 0.45 & 0.05 \\ 0.20 & 0.20 & 0.10 & 0.20 & 0.20 & 0.10 & 0.00 \end{pmatrix} \end{array}$$

$$\mathbf{T}_2 = \begin{array}{c} \\ \texttt{ge} \\ \texttt{ge2} \\ \texttt{ge3} \\ \texttt{ee} \\ \texttt{ee2} \\ \texttt{ee3} \\ \texttt{rb} \end{array} \begin{array}{c} \begin{array}{ccccccc} \texttt{ge} & \texttt{ge2} & \texttt{ge3} & \texttt{ee} & \texttt{ee2} & \texttt{ee3} & \texttt{rb} \end{array} \\ \begin{pmatrix} 0.45 & 0.20 & 0.10 & 0.10 & 0.05 & 0.05 & 0.05 \\ 0.20 & 0.45 & 0.10 & 0.05 & 0.10 & 0.05 & 0.05 \\ 0.10 & 0.20 & 0.45 & 0.05 & 0.05 & 0.10 & 0.05 \\ 0.10 & 0.05 & 0.05 & 0.45 & 0.20 & 0.10 & 0.05 \\ 0.05 & 0.10 & 0.05 & 0.20 & 0.45 & 0.10 & 0.05 \\ 0.05 & 0.05 & 0.10 & 0.10 & 0.20 & 0.45 & 0.05 \\ 0.20 & 0.20 & 0.10 & 0.20 & 0.20 & 0.10 & 0.00 \end{pmatrix} \end{array}$$

F.7 EMPIRICAL RESULTS - REWARD COMPARISON (EXPANDED FUEL SELECTION)

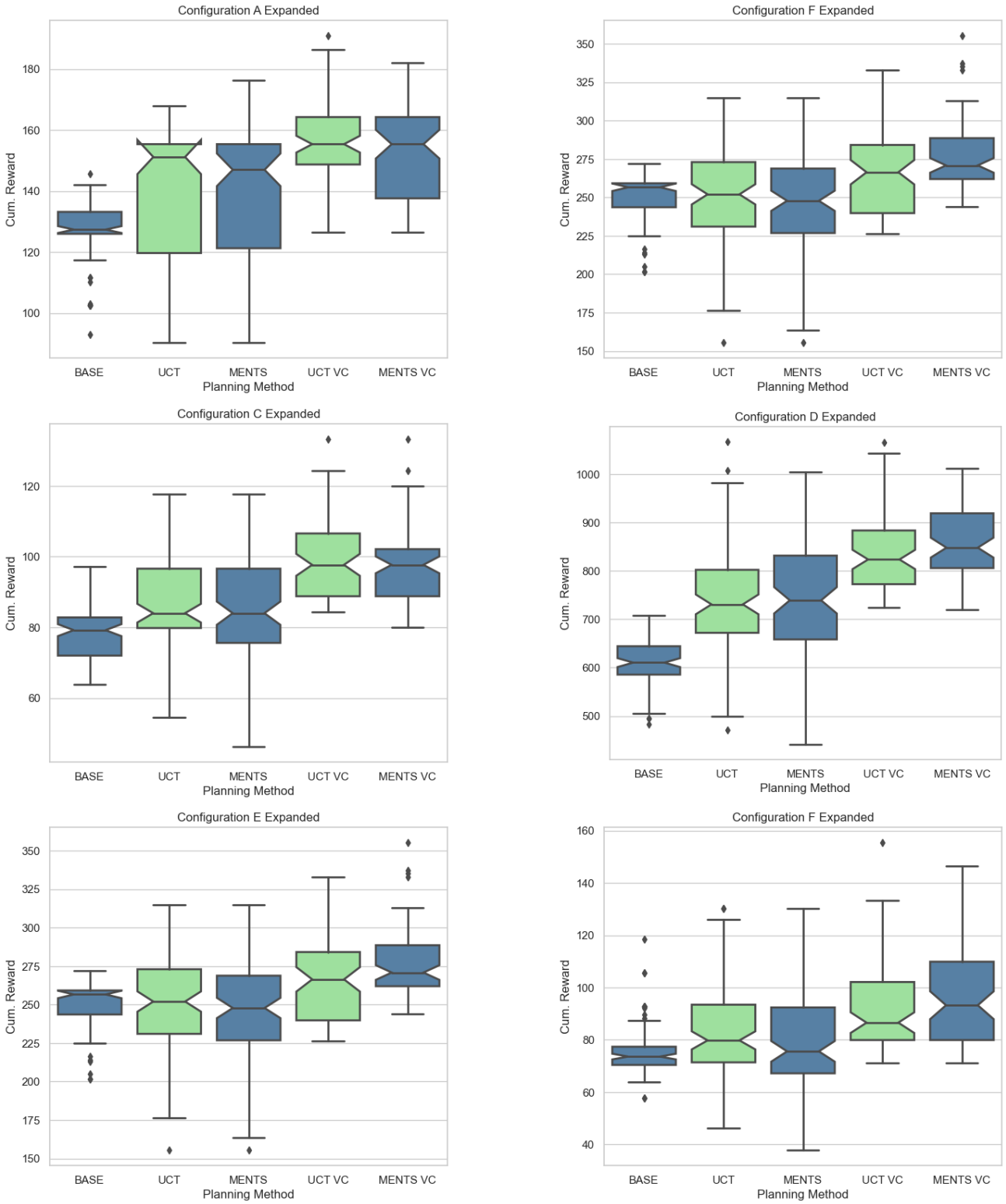


Figure 10: Distance travelled on fuel constraints (reward) per policy. Baseline solution constitutes simple value iteration solution over the MDP.

F.8 EMPIRICAL RESULTS - VALUE CONVERGENCE (EXPANDED FUEL SELECTION)

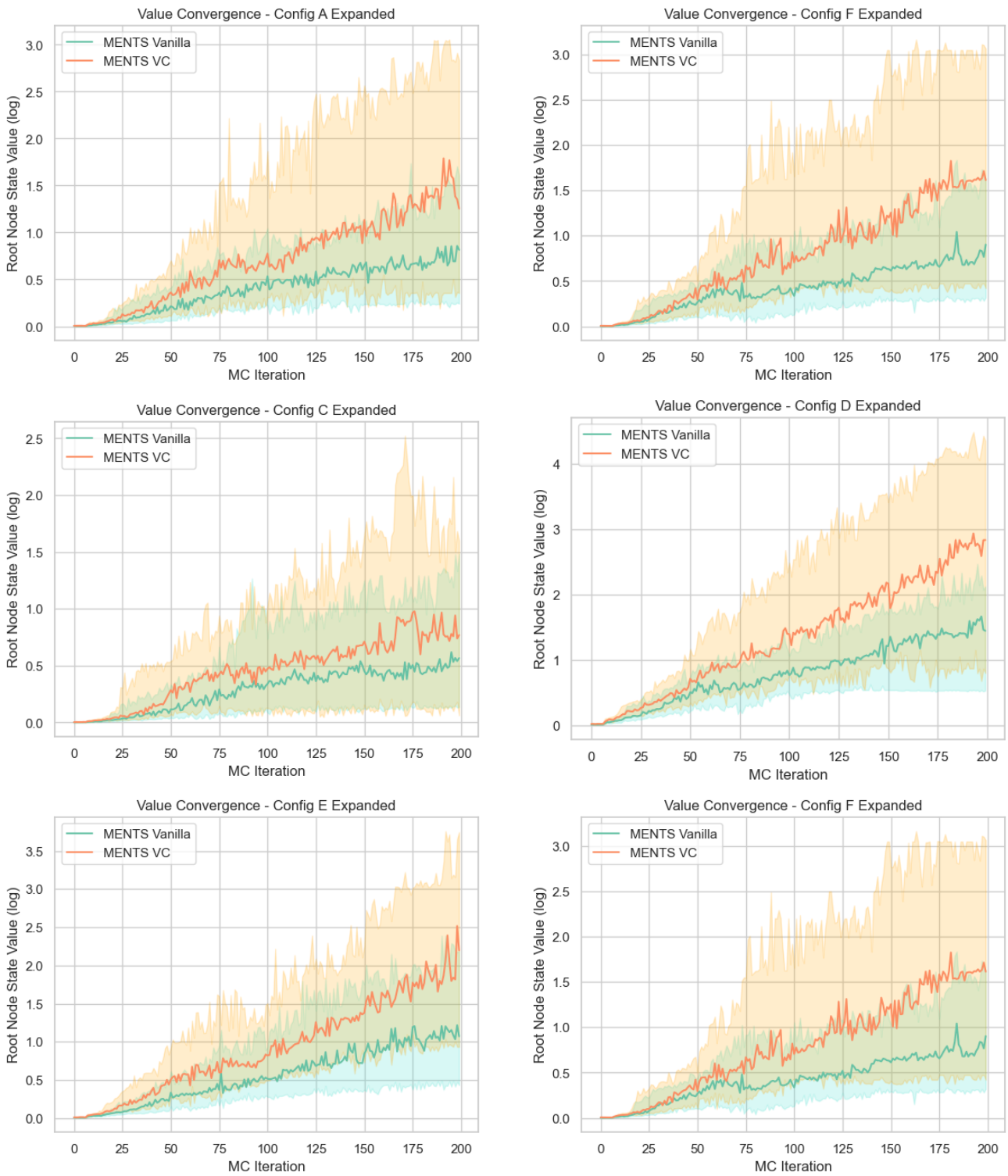


Figure 11: Value convergence performance of the hybrid fuel simulation.

G FINANCIAL OPTIONS TRADING

G.1 INTRODUCTION

Options are derivative financial instruments that grant the holder the right, but not the obligation, to buy or sell an underlying asset at a predetermined price within a specified time frame [Jam03]. American options differ from their European counterparts in that they can be exercised at any point prior to expiration, whereas European options may only be exercised at maturity. The pricing of American options presents a significant computational challenge due to the need to determine the optimal exercise strategy. This flexibility makes American options a canonical example of an optimal stopping problem. A systematic decision criterion is required to assess, at each timestep, whether immediate exercise maximizes the expected payoff [CLP01]. To address this, we propose a novel approach leveraging Monte Carlo planning to optimize the exercise policy.

Problem Statement: The objective is to optimize the decision-making process for American option holders by learning an optimal policy under realistic stock price dynamics. Specifically, the goal is to determine, at each timestep, whether to continue holding or to exercise the option in order to maximize the expected return. As a baseline, we employ the Longstaff-Schwartz algorithm, comparing its performance against our proposed Monte Carlo planning approach.

G.2 AMERICAN OPTION

Table 10: **Notation and variable descriptions for American option pricing.**

Notation	Description
S_0	Initial stock price at the beginning of the option period.
K	Strike price, the fixed price at which the option holder can buy or sell the underlying asset.
T	Time to maturity, the total duration of the option in years.
r	Risk-free interest rate, representing the theoretical return on a risk-free investment.
σ	Volatility of the underlying asset, indicating the asset's price fluctuations.
Δt	Time step, representing the interval used in the simulation (e.g., $1/10$ means 10 intervals within the option duration).
q	Dividend yield, the annual dividend expressed as a percentage of the stock price.
`option_type`	Type of option, either "call" for a call option or "put" for a put option.

G.2.1 Modelling option price

Two commonly used models to simulate the stock price dynamics are the Binomial Model and the Geometric Brownian Motion (GBM).

Binomial Model: The binomial model approximates stock price movements over time using a discrete-time framework:

$$S_{t+\Delta t} = \begin{cases} S_t \cdot u, & \text{with probability } p, \\ S_t \cdot d, & \text{with probability } 1-p, \end{cases}$$

where $u = e^{\sigma\sqrt{\Delta t}}$ represents the upward movement factor, $d = \frac{1}{u}$ denotes the downward movement factor, and $p = \frac{e^{r\Delta t}-d}{u-d}$ is the risk-neutral probability.

Geometric Brownian Motion: The GBM model for american option is written as:

$$S_{t+\Delta t} = S_t \exp\left[\left(r - 0.5\sigma^2\right)\Delta t + \sigma\sqrt{\Delta t}Z\right],$$

where S_t represents the stock price at time t, r is the risk-free interest rate, σ denotes the stock's volatility, and Z is a standard normal random variable distributed as $\mathcal{N}(0,1)$.

In the setup, we consider a finite time horizon, where a stochastic price is calculated at each predetermined time step using one of the aforementioned methods. Both methods have been tested in practice and demonstrate great performance in accurately estimating stock prices.

G.2.2 Markov Decision Process

The action space in an Markov Decision Process (MDP) consists of two actions: hold and execute. The state of the MDP is defined by the time step t, which is a discrete representation of the time to maturity incremented by Δt at each step, the asset price S_t, which represents the price of the underlying asset at time t, and the terminal status, a boolean variable indicating whether the option has reached maturity or has been exercised. The state is represented as:

$$s_t = \{t, S_t, \texttt{is_terminal}\},$$

where $t \in [0, T]$ and $S_t \geq 0$.

G.3 TABLE OF PARAMETERS

Table 11: **American Option Parameters.**

Config.	S_0	K	T	r	σ	Δt	q	**Type**
A	40	36	1	0.1	0.2	0.1	0	Call
B	12	10	1.5	0.08	0.25	0.1	0.03	Call
C	36	40	0.5	0.05	0.3	0.05	0.05	Put
D	10	14	1	0.12	0.35	0.05	0.05	Put
E	8	5	1.5	0.07	0.2	0.1	0	Call
F	5	8	1	0.1	0.4	0.1	0.05	Put

G.4 LONGSTAFF-SCHWARTZ ALGORITHM

The Longstaff-Schwartz algorithm [LS01] is a widely recognized and industry-standard for pricing American options.

The mathematical problem of American option is an optimal stopping problem. The Longstaff-Schwartz algorithm addresses this by working backward from the option's expiration date, determining at each step whether to exercise the option or continue holding it. Using regression, it estimates the continuation value, which is the expected future payoff of holding the option and compares it to the immediate exercise value to determine the optimal strategy. The continuation value is written as,

$$C(t_i) = \mathbb{E}_Q \left[\frac{V(t_{i+1})}{B(t_{i+1})} \,\middle|\, \mathcal{F}_{t_i} \right] B(t_i). \tag{G.1}$$

It can also be expressed as,

$$F(\omega; t_k) = \mathbb{E}^Q \left[\sum_{j=k+1}^{L} \exp\Big(- \int_{t_k}^{t_j} r(\omega, s)\, \mathrm{d}s \Big)\, C\big(\omega, t_j;\, t_k, T\big) \,\middle|\, \mathcal{F}_{t_k} \right]. \tag{G.2}$$

To estimate the continuation value, N possible paths of the underlying asset are simulated using Monte Carlo methods based on geometric Brownian motion (GBM). At each time step, only in-the-money paths are considered, as these represent scenarios where early exercise might be optimal. Second-degree polynomial regression is then applied to model the continuation value as a function of the asset's price, using the discounted future cash flows as the dependent variable. The polynomial regression can be expressed as,

$$y(i, j) = \beta(0) + \beta(1) \cdot S(i, j) + \beta(2) \cdot S(i, j)^2. \tag{G.3}$$

By comparing this estimated continuation value with the immediate exercise value, the algorithm determines the optimal strategy for each path, and the option price is obtained by averaging the discounted cash flows across all simulated paths.

G.4.1 Empirical Results - Reward Comparison

Longstaff–Schwartz (LS) has a lower variance because it applies a regression-based method, fitting a model across many simulated paths at each time step to approximate the option's continuation value. This aggregated regression smooths out

randomness and yields relatively stable estimates. In contrast, Monte Carlo planning explores one path at a time, making the outcome heavily dependent on whether sampled paths are especially favorable or unfavorable, which leads to higher variance across multiple runs.

G.4.2 Empirical Results - Value Convergence

Table 12: **Shared parameters across all experimental configurations.**

Description	**Value**
No. of Simulations N_{sim}	1000
Exploration Constant (C)	1.0
Simulation Depth Limit (N_{depth})	100
Discount Factor (γ)	0.9
MENTS Temperature (T)	0.7
MENTS Epsilon (ϵ)	0.2

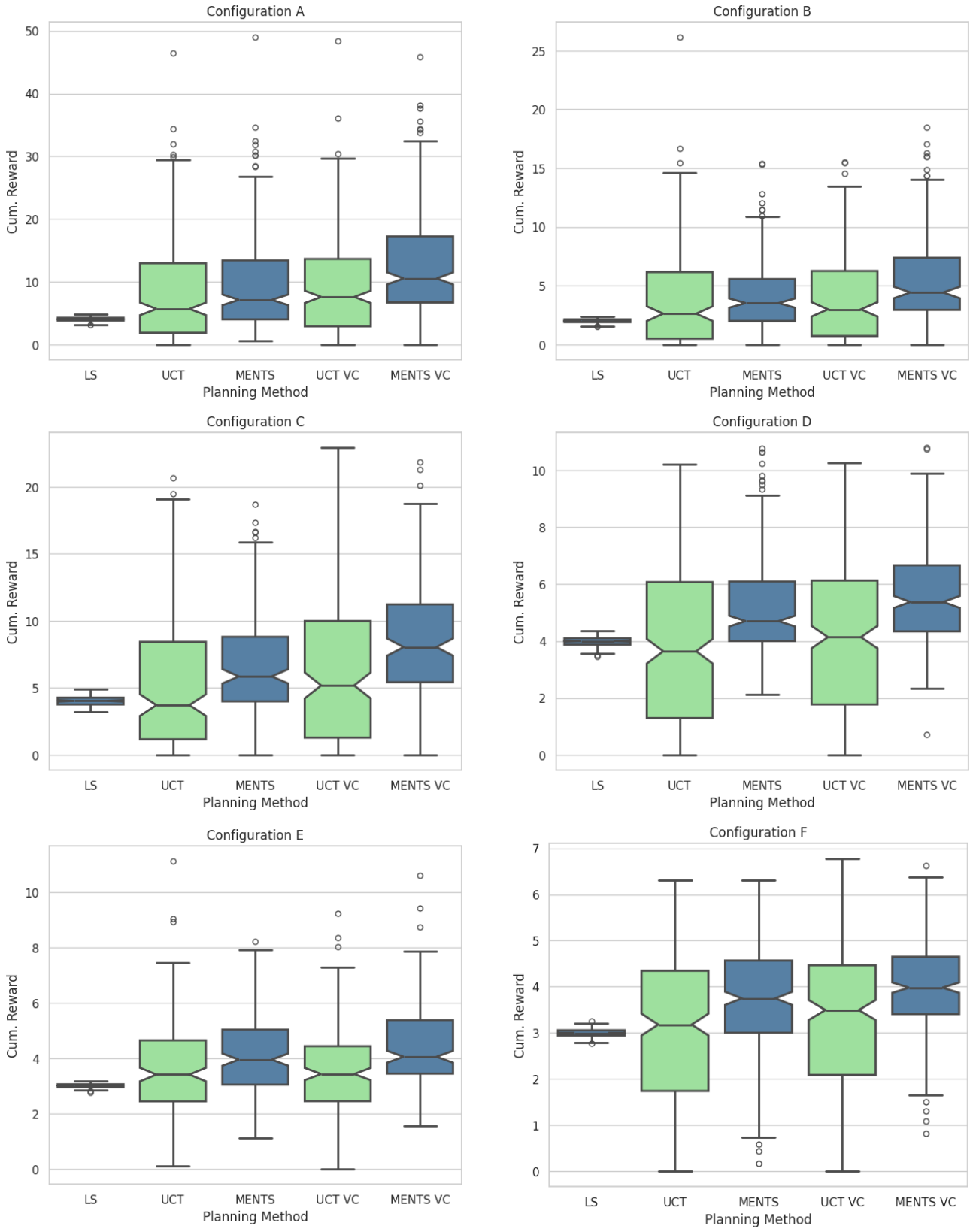


Figure 12: Reward comparison of different American option configurations.

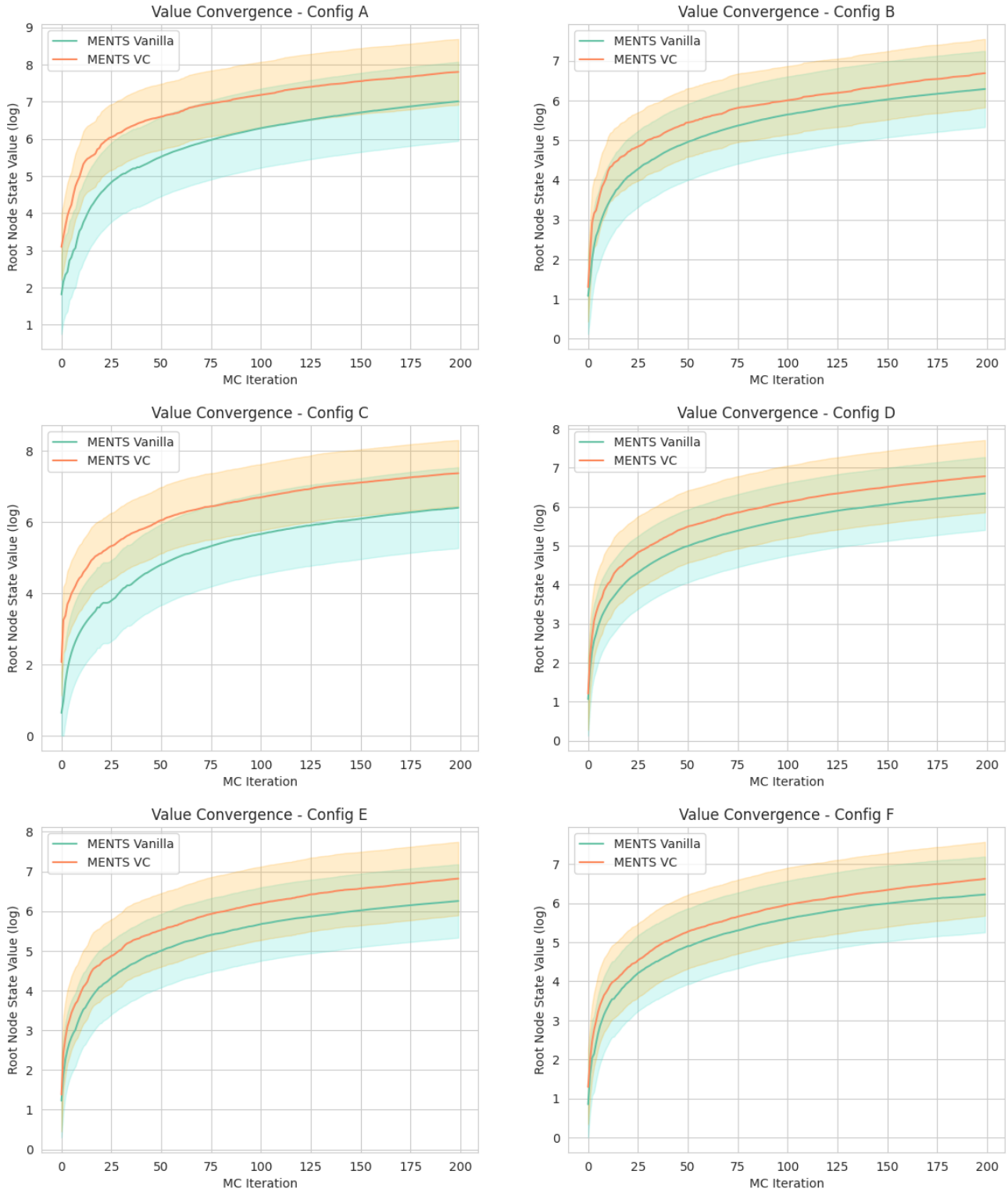


Figure 13: Value convergence performance of different American option configurations.

G.5 FINANCIAL OPTIONS TRADING - EXPANDED

We extend American option pricing to a multivariate setting in which multiple correlated options must be managed simultaneously. The combinatorial complexity arises from asset correlations and constraints on simultaneous exercises (e.g., at most $\bar{\Delta}_a$ options exercisable per period). This framework addresses practical needs in portfolio optimization and algorithmic trading. Asset prices follow a Multivariate Brownian Motion (MVBM) with covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{1,2} & \cdots & \sigma_1\sigma_c\rho_{1,c} \\ \sigma_2\sigma_1\rho_{1,2} & \sigma_2^2 & \cdots & \sigma_1\sigma_c\rho_{2,c} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_c\sigma_1\rho_{1,c} & \vdots & \ddots & \sigma_c^2 \end{bmatrix}$$

where σ_i is volatility asset i and $\rho_{i,j}$ is the correlation between assets i and j.

Basket of Options: The agent manages a portfolio of financial options consisting of *call options* (the right to buy an asset at a fixed strike price) and *put options* (the right to sell at a strike price). For call options, the payoff upon exercise is $\max(S_t - K, 0)$, where S_t is the asset price and K is the strike price; for put options, it is $\max(K - S_t, 0)$. The agent must determine an exercise policy that maximizes the cumulative discounted payoff over the time horizon. Exercises are performed individually per option, subject to constraints (e.g., expiration dates and American-style exercise rules). The total reward is the sum of payoffs from all exercised options.

Dimensionality Increase: For a basket of B options, the action space expands significantly compared to the single-option case. Specifically, a binary vector $\mathbf{x} \in \{0, 1\}^B$ represents hold/exercise decisions, with $0 \leq \|\mathbf{x}_d^{t+1} - \mathbf{x}_d^t\|_p \leq \bar{\Delta}_a$. From dynamic constraint (**D3**), $\bar{\Delta}_a$ denotes the maximum number of options exercisable at each time step. This effectively increases the action space dimensionality from 2 to 2^B.

G.6 FINANCIAL OPTIONS EXPERIMENTAL CONFIGURATIONS - HIGHER DIMENSION

Config	$(S_0,\ K,\ \sigma,\ q,$ **Type**$)$	T	r	dt	$\bar{\Delta}_a$
A	(40, 36, 0.20, 0.00, Call), (12, 10, 0.25, 0.03, Call), (8, 5, 0.20, 0.00, Call)	1.0	0.08	0.02	3
B	(25, 20, 0.30, 0.02, Call), (30, 28, 0.35, 0.01, Put), (15, 16, 0.25, 0.04, Call), (20, 18, 0.40, 0.03, Put)	1.5	0.06	0.05	3
C	(26, 24, 0.27, 0.02, Call), (15, 16, 0.30, 0.03, Put), (38, 35, 0.25, 0.015, Call)	1.3	0.06	0.05	2
D	(40, 42, 0.25, 0.01, Call), (35, 36, 0.30, 0.015, Put), (50, 48, 0.28, 0.02, Call), (45, 47, 0.26, 0.01, Put), (60, 58, 0.27, 0.015, Call)	1.5	0.05	0.025	3
E	(18, 20, 0.26, 0.015, Put), (27, 25, 0.32, 0.02, Call), (22, 24, 0.29, 0.025, Put), (31, 30, 0.33, 0.01, Call)	1.4	0.065	0.04	2
F	(40, 35, 0.20, 0.01, Call), (25, 28, 0.25, 0.02, Put), (30, 25, 0.30, 0.01, Call), (20, 22, 0.22, 0.015, Put)	1.0	0.05	0.05	2

Table 13: Multi-option configurations.

G.7 EMPIRICAL RESULTS - REWARD COMPARISON

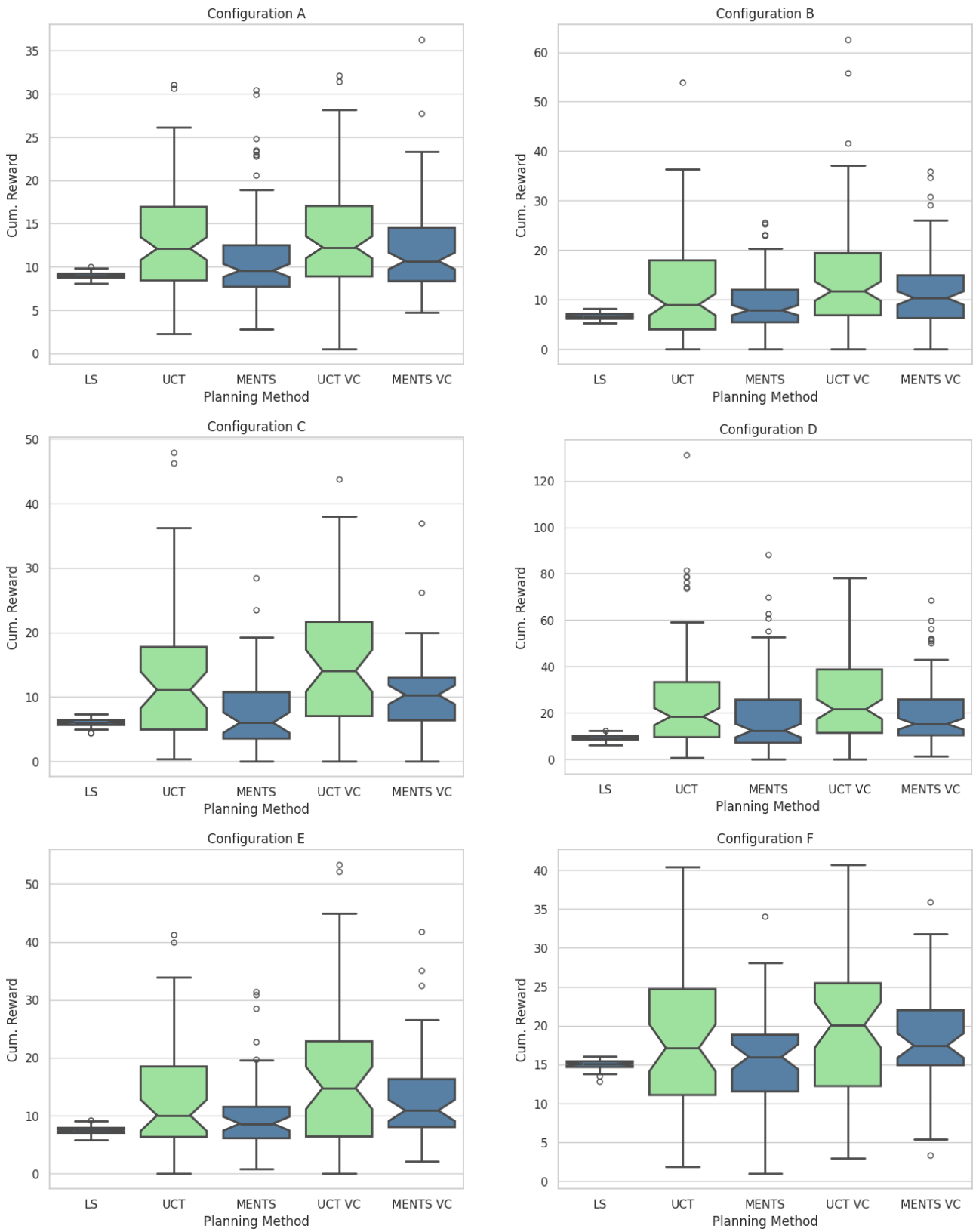


Figure 14: Reward comparison of expanded financial option simulations.

G.8 EMPIRICAL RESULTS - VALUE CONVERGENCE

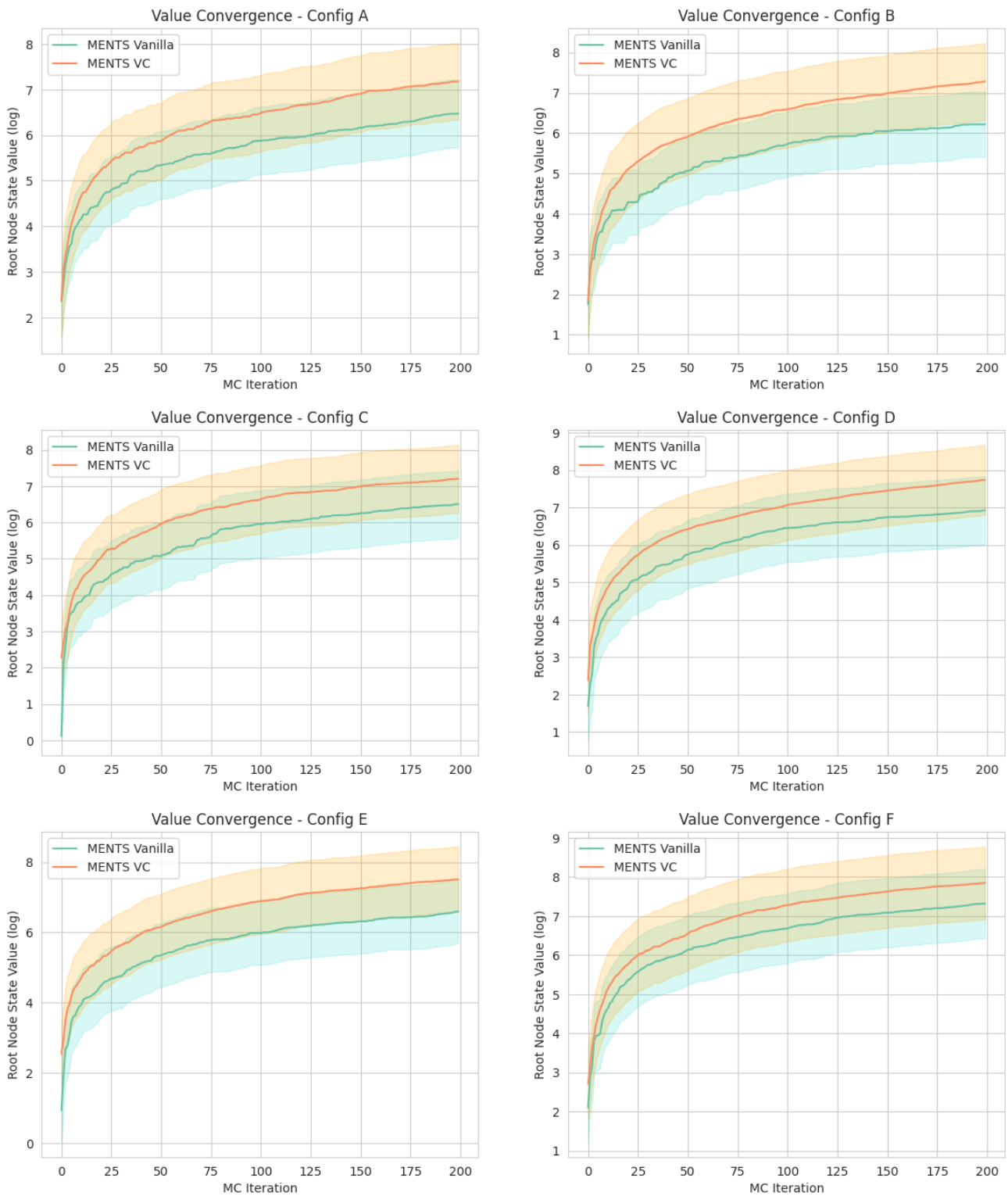


Figure 15: Value convergence performance of the expanded financial option simulation.