Optimizing Stochastic Control through State Transition Separability and Resource-Utility Exchange

Larkin Liu
Technische Universität München
larkin.liu@tum.de

Shiqi Liu
École Polytechnique
shiqi.liu@polytechnique.edu

Matej Jusup
ETH Zürich
mjusup@ethz.ch

ABSTRACT

In the realm of stochastic control, particularly in the fields of economics and engineering, Markov Decision Processes (MDP's) are employed to represent various processes ranging from asset management to transportation logistics. Upon closer examination these constrained MDP's often exhibit specific causal structures concerning the dynamics of transitions and rewards. Thus, leveraging this structure can facilitate computational simplifications for determining the optimal policy. This study introduces a framework, which we denote as SD-MDP, in which we disentangle the causal structure of state transition and reward function dynamics. Through this method, we are able to establish theoretical guarantees on improvements in computational efficiency compared to standard MDP solver (such as linear programming). We further derive error bounds on the optimal value approximation via Monte Carlo simulation for this family of stochastic control problems.

1. INTRODUCTION

Certain stochastic decision processes for optimal control, namely those found in domains of robotics, and logistics, operate with dynamics and do not always require a full MDP formulation. Techniques such as policy and value iteration, deep reinforcement learning, etc., can be used for computing approximate optimal solutions. Nevertheless, disentangling and applying the causal structure of an MDP can improve the computational complexity of MDP solvers via seperability of the search space.

Traditionally, resource allocation problems were tackled through multi-stage stochastic programming or approximating them as MDPs [4]. Yet, these methods struggle with seamless integration with machine learning. To address this gap, we introduce SD-MDP (Section 2), offering a flexible modelling approach for various resource allocation problems and a route to derive theoretical guarantees.

We introduce a construct, termed the *resource-utility* exchange model, akin to energy conservation principles in physics. It allows for generic modelling of MDPs via simulation. This construct also facilitates theoretical guarantees on value function estimates, especially when integrating Monte Carlo approximations with MDP solvers utilizing online learning.

2. THE SD-MDP FRAMEWORK

From the perspective of causal reinforcement learning [3], the SD-MDP effectively partitions the state transition mechanics via the causal relation of the intervening action. This allows the state transition to be modelled separately, and independent of the reward dynamics. The *transition separability* characteristic of the SD-MDP isolates the causal effect of actions $\mathbf{a}^t$ on the state transition $\mathbf{x}_\eta^t \rightarrow \mathbf{x}_\eta^{t+1}$, as illustrated in Fig. 1.

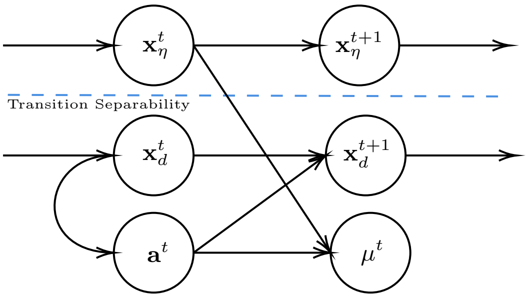

Figure 1: A display of the SD-MDP Markov dynamics.

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, $\mathbf{x} = [\mathbf{x}_\eta, \mathbf{x}_d]^T$. At face value, this model is similar to the restless bandit problem [1], 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 $\mathbf{x}_d$, reward outcomes must be planned over the complete time horizon T, rather than maximizing at each given opportunity, under perfect information or otherwise.

SD-MDP **Definition:** Formally, the SD-MDP is represented as $(\mathcal{X}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \mathbf{x}^1)$, where $\mathcal{X}$ denotes the state space. $\mathcal{A}$ denotes the action space, and is of a fixed dimension. $\mathcal{R} \in \mathbb{R}$ denotes the reward space. $\mathcal{T}$ is the transition function for $\mathbf{x} \in \mathcal{X}$, and $\mathbf{x}^1$ is the initial state.

In particular, we divvy up 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 $\mathbf{a}^t$. θ denotes the parameters which govern the dynamics of a specific SD-MDP.

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

$$P(\mathbf{x}_\eta^{t+1}|\mathbf{a}^t, \mathbf{x}^t) = P_\theta(\mathbf{x}_\eta^{t+1}|\mathbf{x}_\eta^t) \tag{2}$$

$$P(\mathbf{x}^{t+1}|\mathbf{a}^t, \mathbf{x}^t) = P(\mathbf{x}_d^{t+1}|\mathbf{a}^t, \mathbf{x}^t) P_\theta(\mathbf{x}_\eta^{t+1}|\mathbf{x}_\eta^t) \tag{3}$$

2.1 Dynamics of the SD-MDP

To align the formulation with a concrete application domain and justify the partitioning the MDP into stochastic-deterministic partitions (SD-MDP), we model a specific MDP using the *resource-utility exchange* principle, providing abstraction for sequential decision making across various domains. State transitions are driven by the environment, denoted $\mathbf{x}_\eta$, while rewards depend on actions and the entire state space, denoted $\mu_\theta(\mathbf{a}^t, \mathbf{x}^t)$. The SD-MDP partitions state transitions based on causal relations of actions, allowing separate modelling of transition and reward dynamics.

Applications: Concrete examples of the SD-MDP framework include an agent liquidating portfolio of assets over discrete time periods while being subject to restrictions on the amount of assets exercisable. Another example is the refuelling ofa maritime liner at different ports-of-calls, subject to stochastic fuel costs. In both scenarios, the agent is subject to some form of capacity and action constraint(s), while converting resources for utility over a finite time horizon to maximize their respective cumulative utility.

(D1): Positive Action Space: In the first assumption, we impose the constraint of a strictly element-wise positive action space, wherein each component of the action vector is greater than 0, $\mathbf{a} > \mathbf{0}$. 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., $\mathbf{x}_d \geq \mathbf{0}$. This constraint is necessary to represent a multi-dimensional capacity quantity that is consumed over discrete time increments, as depicted in Fig. 2.

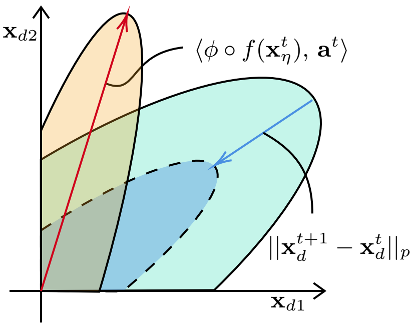
$\mathbf{x}_{d2}$
$\langle \phi \circ f(\mathbf{x}_\eta^t),\, \mathbf{a}^t \rangle$
$||\mathbf{x}_d^{t+1} - \mathbf{x}_d^t||_p$
$\mathbf{x}_{d1}$

Figure 2: A 2D display of resource-utility exchange subject to the dynamics of the SD-MDP.

(D2): General Linear Reward Dynamics: The SD-MDP obeys a reward function of a general linear form. Referencing Fig. 1, state variable $\mathbf{x}_d^t$, and intervention variable $\mathbf{a}_d^t$ are deterministic. Similarly, we also define the causal effect of the contemporaneous state-action space on the reward. Together with the stochastic partition $\mathbf{x}_\eta^t$, it invokes a causal relationship for the outcome at time t, $\mu_\theta(\mathbf{a}^t, \mathbf{x}^t)$. The reward function is defined as a deterministic function of $\mathbf{a}^t$ and $\mathbf{x}^t$, denoted as $\mu_\theta(\mathbf{a}^t, \mathbf{x}^t)$.

Let $\mu(\cdot) : \mathbb{R}^{|\mathcal{X}_\eta|} \times \mathbb{R}^{|\mathcal{A}|} \mapsto \mathbb{R}$ denote a standard map that yields a single reward value in $\mathbb{R}$ when provided with inputs $\mathbf{a} \in \mathcal{A}$ and $\mathbf{x}_\eta^t \in \mathcal{X}_\eta$, subject constraints on the system at time t. $f(\cdot) : \mathbb{R}^D \to \mathbb{R}^D$ is a coordinate-wise separable kernel functions composed of a series of smooth positively monotone Lipschitz functions, governing the dimension-wise non-linear scaling corresponding to $\mathbf{x}_\eta$. 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 a inner product between the transformed $\phi\, f(\mathbf{x}_\eta^t)$ and $\mathbf{a}^t$, as expressed in in Equation (4).

$$\mu_\theta(\mathbf{a}^t, \mathbf{x}^t) = \langle \phi \circ f(\mathbf{x}_\eta^t),\, \mathbf{a}^t \rangle \tag{4}$$

(D3): Linear Incremental Action Dynamics: We define a linear transformation matrix ϕ', which provides and anti-parallel transform in comparison to ϕ. Similarily we introduce function $g(\cdot) : \mathbb{R}^D \to \mathbb{R}^D$, which similar to $f(\cdot)$, is also a coordinate-wise separable function composed of a series of smooth positively monotone Lipschitz kernel functions. To model the expansion and contraction of the capacity $\mathbf{x}_d$, we impose the linear transition function acting on $\mathbf{x}_d$ in Eq. (5).

$$\underline{\Delta}_a(t) \leq ||\mathbf{x}_d^{t+1} - \mathbf{x}_d^t||_p = ||\underbrace{\Delta_d(t)}_{\text{System}} + \underbrace{\phi'\, g(\mathbf{a}^t)}_{\text{Agent}}||_p \leq \bar{\Delta}_a(t) \tag{5}$$

Where $\Delta_d(t)$ is a natural discrete change on $\mathbf{x}_d^t$ as deterministically determined by the system, and $\phi' g(\mathbf{a}^t)$ is the contribution to the expansion or contraction of $\mathbf{x}_d^t$ based on the agent's action taken at $\mathbf{a}^t$. We impose a constraint on the magnitude of capacity change per time interval via Eq. (9), where constraints $\underline{\Delta}_a(t)$ and $\bar{\Delta}_a(t)$ are given by the system.

(D4): Capacity Objective: We provide a constraint on the trajectory of actions, and this constraint is expressed in the form of a path constraint (accumulation). This *path constraint* on the action space restricts the path of the action sequence the agent takes. As defined, the accumulation of resources $\phi' g(\mathbf{a})$ should meet some maximum and minimum goals, as expressed in Eq. (6).

$$\underline{A} \leq \sum_{t=1}^{T} ||\phi'\, g(\mathbf{a}^t)||_p \leq \bar{A} \tag{6}$$

3. OPTIMAL POLICY STRUCTURE

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

$$\underline{\aleph}^t[(\mathbf{a})^t] \equiv (T - t + 1)\underline{\Delta}_a(t) + \sum_{i=1}^{t-1} ||\phi' g(\mathbf{a}^i)||_p - \bar{A} \tag{7}$$

$$\overline{\aleph}^t[(\mathbf{a})^t] \equiv (T - t + 1)\bar{\Delta}_a(t) + \sum_{i=1}^{t-1} ||\phi' g(\mathbf{a}^i)||_p - \underline{A} \tag{8}$$

Intuitively, $\overline{\aleph}^t[(\mathbf{a})^t]$ and $\underline{\aleph}^t[(\mathbf{a})^t]$ represent the maximum and minimum allowable consumption under the path constraint in Eq. (6). Correspondingly, $\underline{\Delta}_a(t)$ and $\bar{\Delta}_a(t)$ constitute the minimum and maximum incremental capacity constraints specified by the system. Moving forward let, $\mathcal{A}(t)$ denote the action set at time t, given the constraints from equations Eq. (5) and (6), such that the expression $\mathbf{a} \in \mathcal{A}(t)$ encapsulates the constraints pertaining to the SD-MDP dynamics.

$$\mathcal{A}(t) \equiv \Big\{ \mathbf{a} : \underline{\|\mathbf{a}(t)\|} \leq ||\phi' g(\mathbf{a}^t)||_p \leq \overline{\|\mathbf{a}(t)\|} \Big\} \tag{9}$$

$$\underline{\|\mathbf{a}(t)\|} = \max \Big\{ \underline{\aleph}^t[(\mathbf{a})^t],\; \underline{\Delta}_a(t) \Big\} \tag{10}$$

$$\overline{\|\mathbf{a}(t)\|} = \min \Big\{ \overline{\aleph}^t[(\mathbf{a})^t],\; ||\mathbf{x}_d^t||_p,\; \bar{\Delta}_a(t) \Big\} \tag{11}$$

$||\mathbf{x}_d^t||_p$ forms a constraint on the capacity from the determinsitic component of the SD-MDP. Along with $\underline{\aleph}^t[(\mathbf{a})^t]$ and $\overline{\aleph}^t[(\mathbf{a})^t]$, they together form a bound on the admissible action space, denoted as $\mathcal{A}(t)$. We denote $\{\mathbf{a}^+\}$ and $\{\mathbf{a}^-\}$ as the following,

$$\{\mathbf{a}^+\} = \underset{\mathbf{a}\in\mathcal{A}(t)}{\operatorname{argmax}} ||\mathbf{a}||_p,\ \{\mathbf{a}^-\} = \underset{\mathbf{a}\in\mathcal{A}(t)}{\operatorname{argmin}} ||\phi' g(\mathbf{a}^t)||_p \quad (12)$$

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 $\mathbf{x}_d$. In Lemma 3.1, we show that the optimal policy consists of a action, represented as a vector, corresponding to one of two sets $\{\mathbf{a}^+\}$ or $\{\mathbf{a}^-\}$.

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

$$\text{Top}_k(\mathbb{E}[\tau_s]) = (\mathbb{E}[\mathbf{x}_\eta^{i=1}], \mathbb{E}[\mathbf{x}_\eta^{i=2}], \ldots, \mathbb{E}[\mathbf{x}_\eta^{i=k}]) \quad (13)$$

Such that,

$$\phi\, f(\mathbb{E}[\mathbf{x}_\eta^{i=1}]) \succeq \phi \circ f(\mathbb{E}[\mathbf{x}_\eta^{i=2}]) \cdots \succeq \phi\, f(\mathbb{E}[\mathbf{x}_\eta^{i=k}]) \quad (14)$$

To note, the series returned by $\text{Top}_k(\mathbb{E}[\tau_s])$ could have a smaller cardinality than $\mathbb{E}[\tau_s]$ due to truncation. We use $/\text{Top}_k(\mathbb{E}[\tau_s])$ to the denote the set of elements not in $\text{Top}_k(\mathbb{E}[\tau_s])$ but in $\mathbb{E}[\tau_s]$. Where $|\text{Top}_k(\mathbb{E}[\tau_s])| + |/\text{Top}_k(\mathbb{E}[\tau_s])| = |\mathbb{E}[\tau_s]|$.

LEMMA 3.1. ***Bounded Action Space for the SD-MDP:*** *Under the SD-MDP framework, for any action taken in the the finite time horizon, optimal policy lines to the union of 2 subspaces, that is* $\mathbf{a}^* \subset \{\mathbf{a}^+\}^t \cup \{\mathbf{a}^-\}^t \subset \mathcal{A}(t) \subseteq \mathcal{A}$.

Sketch of Proof: First we demonstrate the separability of $\mathbb{E}[\tau_s]$ with respect to any deterministic sequence of actions. The solution therefore involves the finding the maximizing $\mathbb{E}[\mathbf{x}_\eta^t] \in \mathbb{E}[\tau_s]$ for each $t \in (1 \ldots T)$. Under incremental dynamics, $||\mathbf{a}|| \leq \bar{\Delta}_a(t)$, only limited resources can be dedicated to maxmizing each $\mathbb{E}[\tau_s]$ via the inner product from Eq. (4). We show that when we majorize over $\mathbb{E}[\phi f(\mathbf{x}_\eta^t)]$, $\forall \mathbb{E}[\mathbf{x}_\eta^t] \in \mathbb{E}[\tau_s]$, the optimal solution to the sequence $(\mathbf{a}^*)$ is an order preserving union of two sequences, comprising of $\mathbf{a}^+ \in \{\mathbf{a}^+\}^t$ and a norm minimizing vector $\mathbf{a}^- \in \{\mathbf{a}^-\}^t$ which are independently computed.

LEMMA 3.2. ***Solving for Optimal Value via Top K Allocation:*** *Under the SD-MDP framework, the optimal value can be obtained by solving the dual problem, which involves the optimization of the value of k in $Top_k(\mathbb{E}[\tau_s])$ over $k \in \{1, \ldots, T\}$ possibilities.*

Sketch of Proof: We show that when we majorize over $\mathbb{E}[\tau_s]$, to produce an ordered set of sequences according to Eq. (13) 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 $\mathbb{E}[\tau_s]$ we allocate minimum resources within the constraints. Via an argument based on an extension of the Hardy-Littlewold-Polya Theorem [2], the solution involves simply finding the value of k which maximizes the value function expressed in Eq. (15), subject to constraints SD-MDP dynamics (Section 2.1).

$$V_k(\mathbf{x}^t) = \sum_k \phi \circ f \odot \text{Top}_k(\mathbb{E}[\tau_s]) \odot \mathbf{a}^+[\mathbb{E}[\mathbf{x}_\eta]] + \sum_{T-k} \phi \circ f \odot /\text{Top}_k(\mathbb{E}[\tau_s]) \odot \mathbf{a}^-[\mathbb{E}[\mathbf{x}_\eta]] \quad (15)$$

COROLLARY 3.1. ***Polynomial Time Solution:*** *There exists a in solution in polynomial time for any well-defined SD-MDP.*

Sketch of Proof: As the constraints are linear, the entire system can be formulated and solved as a block form sparse linear program with T constraints, and therefore a polynomial time solution exists.

THEOREM 1. ***Upper bound on the Monte Carlo Value Estimation for the SD-MDP:*** *For the SD-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.*

Sketch of Proof: Any naturally evolving time series has an expected outcome which can be computed $\mathbb{E}[\tau_s]$, 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 Lemma 3.2. Via Hoeffding's inequality, we can upper bound the approximation error from Monte Carlo sampling by treating each outcome as a random sample.

4. REMARKS

In contrast to linear programming (LP), aimed at achieving anticipative solutions, this streamlined approach enables a reduction in computational complexity, should the appropriate conditions arise when the SD-MDP can be applied. This paradigm shift entails the vector majorization over stochastic outcomes $\mathbb{E}[\tau_s]$ and allocation of $\mathbf{a}^+$ and $\mathbf{a}^-$, which can reduce the complexity from an LP-based polynomial time solution, to that of a sorting-based logarithmic time solution. The exact computational speed-up may be problem specific, and left for future investigation. This methodology can also seamlessly integrate with simulation-based optimization and learning algorithms, particularly those leveraging Monte Carlo simulation techniques such as Monte Carlo Tree Search or Thompson Sampling. This technical note aims disseminate exploratory ideas to disseminate information, with plans for detailed exposition in future work.

Acknowledgements: We thank Jalal Etesami for contributing engaging commentary and discussions to refine this work.

References

[1] John C Gittins. 1979. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41, 2, 148–164.

[2] G.H. Hardy, J.E. Littlewood, and G. Pólya. 1952. *Inequalities. Cambridge Mathematical Library.* Cambridge University Press. ISBN: 9780521358804.

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

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