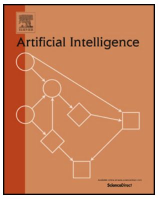
Artificial Intelligence 312 (2022) 103775

edited text

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint



Simplified Risk-aware Decision Making with Belief-dependent Rewards in Partially Observable Domains [☆], [☆☆]

Andrey Zhitnikov [a,*], Vadim Indelman [b]

[a] Technion Autonomous Systems Program (TASP), Haifa, 3200003, Israel
[b] Department of Aerospace Engineering, Technion - Israel Institute of Technology, Haifa, 32000, Israel

ARTICLE INFO

Article history:
Received 14 October 2021
Received in revised form 21 August 2022
Accepted 21 August 2022
Available online 27 August 2022

Keywords:
Artificial intelligence
Decision making under uncertainty
Belief space planning
POMDP

ABSTRACT

With the recent advent of risk awareness, decision-making algorithms' complexity increases, posing a severe difficulty to solve such formulations of the problem online. Our approach is centered on the distribution of the return in the challenging continuous domain under partial observability. This paper proposes a simplification framework to ease the computational burden while providing guarantees on the simplification impact. On top of this framework, we present novel stochastic bounds on the return that apply to any reward function. Further, we consider simplification's impact on decision making with risk averse objectives, which, to the best of our knowledge, has not been investigated thus far. In particular, we prove that stochastic bounds on the return yield deterministic bounds on Value at Risk. The second part of the paper focuses on the joint distribution of a pair of returns given a pair of candidate policies, thereby, for the first time, accounting for the correlation between these returns. Here, we propose a novel risk averse objective and apply our simplification paradigm. Moreover, we present a novel tool called the probabilistic loss (`PLoss`) to completely characterize the simplification impact for *any* objective operator in this setting. We provably bound the cumulative and tail distribution function of `PLoss` using `PbLoss` to provide such a characterization online using only the simplified problem. In addition, we utilize this tool to offer deterministic guarantees to the simplification in the context of our novel risk averse objective. We employ our proposed framework on a particular simplification technique - reducing the number of samples for reward calculation or belief representation within planning. Finally, we verify the advantages of our approach through extensive simulations.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

Autonomous online decision-making is a fundamental aspect of intelligence. In a partially observable setting, which is common in real world scenarios, there is no direct access to the state. Instead, the robot has to maintain a belief over the state and reason about its evolution while accounting for different sources of uncertainty within the decision-making stage. The renowned framework to do so is the Partially Observable Markov Decision Process (POMDP) [18]. Crucial elements

[☆] This research was partially supported by the Israel Science Foundation (ISF), grant No. 371/20 and by a donation from the Zuckerman Fund to the Technion Center for Machine Learning and Intelligent Systems (MLIS).

[☆☆] This paper is part of the Special Issue: "Risk-aware Autonomous Systems: Theory and Practice".

[*] Corresponding author.
E-mail addresses: andreyz@campus.technion.ac.il (A. Zhitnikov), vadim.indelman@technion.ac.il (V. Indelman).

https://doi.org/10.1016/j.artint.2022.103775
0004-3702/© 2022 Elsevier B.V. All rights reserved.

defining the robot's behavior are the random reward and the objective operator applied to the reward distribution. The random nature of the reward arises from the uncertainties in the system.

Solving a POMDP, i.e., calculating the "right decision" in terms of an optimal action sequence or policy, involves anticipating every imaginable turn of future events and computing the *returns* based on the corresponding rewards. One typical example of the return is the future cumulative reward.

There is a large body of algorithms, formulated on top of POMDP, to approximate decision-making under uncertainty. Classical offline methods [21] are trying to find offline a policy that is optimal for all possible beliefs. These methods are based on α-vectors and point based value iteration [24,26,34]. Since the α-vector is the vector of the utility function values, starting from the state realizations or samples from the belief distribution, under the conditional plan, the set of α-vectors, each annotated with an action, can represent the policy for all beliefs. The application of such policy is to find α-vector maximizing the inner product with the belief. Unfortunately, these methods are suitable solely for the discrete state, action, and observation spaces. Some work on extending the α-vectors to continuous spaces has been done by [28]. More recent formulations suitable for continuous spaces are operating on the belief tree.

A necessary factor for planning to be successful is the number of future steps ahead (horizon) that an agent considers in the decision-making process. The belief tree grows exponentially with the horizon. However, the exponential growth with the horizon is not the only problem of the belief tree based approaches. Additionally, the number of possible states grows exponentially with the state space dimension, and consequently, an adequate representation of the belief requires more particles in the setting of non-parametric beliefs. Those last two problems are known as the *curse of history* and the *curse of dimensionality* respectively.

More recently, online methods became successful. Some of them are suitable for continuous state and observation spaces. The output of these methods is an action recommended for the current belief. The algorithm itself is a policy which maps from beliefs to actions online. Prominent examples are POMCP [33] and its various extensions (e.g., [37]), an algorithm designed for large POMDP and based on Monte Carlo tree search. Another popular algorithm, DESPOT [35] [44], focuses on the set of randomly sampled scenarios over the belief tree, avoiding drawbacks of the UCT [22] algorithm used in POMCP.

Standard POMDP formulations consider state-dependent rewards and assume that the belief-dependent reward is *nothing but expectation over the state reward*. POMDP with *belief-dependent rewards* received much less attention, although these rewards are essential in numerous problems, such as information gathering, autonomous navigation, and active sensing. Information theoretic rewards are especially significant for belief space planning (BSP) [17], [10]. Araya et al. [1] introduced ρ-POMDP and extended the exact α-vectors method and a family of point based approximation algorithms to consider convex belief-dependent reward functions. Later Fehr et al. [9] extended their work further to Lipschitz-continuous reward functions. Spaan et al. [36] proposed to augment action space with information-reward actions. Dressel and Kochenderfer [7] proposed an extension of SARSOP [24] to specific forms of belief-dependent rewards. However, these extensions are limited either to a discrete setting or to specific forms of belief dependent rewards. In a general setting, belief-dependent rewards are computationally demanding and prohibitively expensive.

Further, the most popular and widespread objective operator is the expected value of the return. However, the expected value as the objective has inherent flaws. It is oblivious to the distribution of the reward. Meaning it is unable to account for the risk that the selected action or policy is suboptimal and to prevent rare undesirable events. One way to introduce the notion of risk to decision making is to augment the expected return value with chance constraints. This augmentation, however, introduces additional complications which are out of the scope of this paper [31]. With this motivation in mind, we focus on an alternative to the expected value objectives in the context of BSP in continuous domains.

Replacing expected value by other objectives in the context of MDP has been discussed in [6]. Importantly, Defourny et al. [6] discuss risk measures and applicability of Bellman form. Attractive risk averse objectives include Value at Risk (VaR) and Conditional VaR (CVaR) [5]. VaR and CVaR were extensively studied in the context of MDP, whereas in the POMDP planning community, they started to emerge only recently [12,13]. So far, we did not find work considering belief-dependent rewards in the context of decision making under uncertainty with risk averse objectives.

The computational burden incurred by the complexity of POMDP planning inspired many research works to focus on approximations of the problem, e.g., [14]. Typically, approximation based planners show asymptotical guarantees, e.g., the convergence of the algorithms. We take a different path, which is to simplify the original decision-making problem. In other words, instead of approximating the problem, we substitute it with a simpler one. If the order of policies with respect to the original and simplified problems' objective is preserved, such substitution does not affect the decision-making quality. Moreover, suppose we can find online bounds over the original problems' returns/rewards or objective function, utilizing the simplified problem. In that case, it is possible to account for the simplification loss.

Replacement of various parts of the decision making problem to ease the computation burden while preserving the precedence of objectives for potential action plans recently appeared in the literature under the names *simplification* paradigm [39,41,8,16,19] [32,2], *action consistency* [8,19] and *tree consistency* [41]. Yet, these works have limitations. A common assumption is a specific objective operator - expectation. Moreover, Elimelech and Indelman [8], Kitanov and Indelman [19] assume Gaussian distributions and maximum likelihood observations while working in the highly challenging setting of a high dimensional state. Sztyglic et al. [41] consider non-parametric beliefs; however, they build upon a specific belief dependent reward operator.

The general simplification paradigm is concerned with carefully replacing the nonessential elements of the decision making problem and quantifying the impact of this relaxation. Specifically, simplification methods are accompanied by stringent guarantees while alleviating the computational burden of the decision making problem. Therefore, previous works formulated the *simplification paradigm* on top of analytical bounds and a conventional expectation operator as the objective. Existing works consider a deterministic belief update; however this is problematic for non-parametric beliefs which cannot be updated in a deterministic way. In the setting of general beliefs, we shall resort to a particle filter [42] which is a stochastic belief update method since it is based on sample approximations.

1.1. Contributions

We study the simplification impact on decision making under uncertainty considering a general objective operator and non-parametric beliefs, which therefore involves the distribution over returns. This distribution conveys all the information about the decision making problem. Our overall goal is to examine how the simplification method influences the performance of the decision-maker while accelerating the decision making process.

To account for the impact of simplification on the distribution over returns, we first relax typical assumptions regarding the belief update along with the operator reward and introduce probabilistic ρ-POMDP, which we denote as $\mathbb{P}\rho$-POMDP. Given a simplification and an objective operator, we utilize bounds over the return to provide performance guarantees in terms of quality of solution with respect to the original (un-simplified) decision making problem. These bounds can be analytical, and thus hold with probability one. Crucially, we also introduce stochastic bounds that are applicable to any reward function, in contrast to analytical bounds that must be derived for each reward function separately.

Further, we consider specifically simplification impact on decision making with risk averse objectives. To the best of our knowledge, this is the first work that investigates simplification in this context. Our key result is the derivation of *deterministic* bounds on the risk averse objective (Value at Risk) using stochastic bounds on the return/reward. Consequently, we obtain solving speedup and provide guarantees.

Moreover, we examine how simplification impacts the joint distribution over the returns for two candidate policies. We believe this distribution conveys previously unaccounted information, as generally the returns for different policies, conditioned on the current belief, are coupled. To the best of our knowledge, this joint distribution has not been studied yet. Towards this end, we propose a novel risk aware objective operator on top of the joint distribution over returns for two candidate policies. This is as opposed to conventional objectives that are based on the marginal distribution of the return given a policy. Furthermore, we develop a method to provide guarantees for the simplification of such an objective. Specifically, we introduce probabilistic loss (`PLoss`) and the corresponding online bound on probabilistic loss (`PbLoss`) to completely characterize the simplification impact on the joint distribution of the rewards given two candidate policies, meaning for any objective operator. We then utilize the latter to provide performance guarantees in terms of deterministic bounds considering the mentioned risk aware objective operator.

Finally, we apply our general formulation considering a specific simplification: reducing the number of samples of the belief for the reward calculation. To be precise, in the setting of an explicitly given belief surface (e.g. Gaussian mixture model), we endow the stochastic bounds with an adaptivity property and show how to take the lowest possible number of samples while remaining action consistent. In the setting of general beliefs represented by particles we lower the number of particles of the belief and provide performance guarantees.

To summarize, our key contributions are as follows. (a) We extend ρ-POMDP to probabilistic ρ-POMDP ($\mathbb{P}\rho$-POMDP) by relaxing the assumption that the reward operator and the belief update are deterministic; (b) We introduce novel stochastic bounds on the return/reward and rigorously formulate the simplification framework on top of general objective operators and returns/rewards; (c) Using our formulations we present simplification of risk averse decision making under uncertainty; (d) We present a novel objective utilizing joint distribution of the rewards corresponding to two candidate policies and describe a method to simplify such decision making while preserving *action consistency*; (e) We introduce the general concept of `PLoss` and provide its online description with `PbLoss` and utilize it to provide guarantees in terms of deterministic bounds; (f) Finally, we exemplify our framework on a particular simplification technique, which is reducing the number of samples within planning.

1.2. Paper structure

This paper is organized as follows. In section 2 we introduce the notations and formulate the problem. In section 3 we provide mathematical foundations for our approach. We then focus on the marginal distribution of the return in section 4 and on the joint distribution of a pair of the returns corresponding to two candidate policies in section 5. We present a specific simplification in section 6. In section 7 we exemplify our findings on the problem of autonomous navigation with light beacons.

2. Notations and problem formulation

Let us denote by $\mathbb{P}$ the probability density function and by P the probability. By lowercase letter we denote a random vector or its realization. For two random variables x and y, we say that they are equal $x = y$ if they are equal as functions on their measurable space. Further, to shorten notations, we shall often use $\square_{k+}$ to denote $\square_{k+1:k+L}$, where L is the planning horizon. By $\equiv$ we denote identity. We summarize important notations used throughout the paper in Table 1.

Table 1
List of important notation.

Nomenclature	
$\mathcal{X}, \mathcal{A}, \mathcal{Z}$	State, Action, and Observation spaces
$x_k \in \mathcal{X}, a_k \in \mathcal{A}, z_k \in \mathcal{Z}$	Momentary state, action, and observation, respectively.
$\mathbf{1}\{\cdot\}$	Indicator function defined on set $\{\}$.
$a \wedge b$	$\min\{a, b\}$ where $a, b \in \mathbb{R}$.
$a \vee b$	$\max\{a, b\}$ where $a, b \in \mathbb{R}$.
ρ_{k+}, $\rho_{k+1:k+L}$	Reward vector from time index $k+1$ until $k+L$ including the ends.
$\breve{\rho}_{k+}$, $\breve{\rho}_{k+1:k+L}$	Simplified reward vector.
g_k	Return calculated from the reward vector, such that $g_k \triangleq f_{g_k}(\rho_{k+1:k+L})$ and f_{g_k} is some deterministic function.
$\breve{g}_k$	Simplified return calculated from simplified reward vector.
z_{k+}, $z_{k+1:k+L}$	Observation vector from time index $k+1$ until $k+L$ including the ends.
ψ	A general method for updating the belief.
π, $\pi_{k:k+L-1}$	Policy sequence from time index k up until $k+L-1$ including the edges.
ν, $\nu_{k:k+L}$	The sequence of simplification operators from time index k up until $k+L$ including the edges.
$\mathcal{H}_{k+L} \triangleq \{b_k, \pi, z_{k+}\}$	The future history at the time index $k+L$.

2.1. POMDP with belief dependent rewards

Let k be an arbitrary time step. ρ-POMDP [1] is an eight tuple

$$\langle \mathcal{X}, \mathcal{A}, \mathcal{Z}, T, O, \rho, \gamma, b_0 \rangle, \tag{1}$$

where $\mathcal{X}, \mathcal{A}, \mathcal{Z}$ are state, action, and observation spaces with $x_k \in \mathcal{X}, a_k \in \mathcal{A}, z_k \in \mathcal{Z}$ the momentary state, action, and observation, respectively, $T(x_k, a_k, x_{k+1}) = \mathbb{P}_T(x_{k+1}|x_k, a_k)$ is the stochastic transition model from the past momentary state x_k to the next x_{k+1} through action a_k, $O(z_k, x_k) = \mathbb{P}_Z(z_k|x_k)$ is the stochastic observation model, $\rho\left(b_{k+1}, z_{k+1}, a_k, b_k\right)$ is a scalar reward operator, $\gamma \in [0, 1]$ is the discount factor, and b_0 is the belief about the initial state (prior). Notably, the infinite horizon planning case necessitates $\gamma < 1$, whereas $\gamma = 1$ is permitted in a finite horizon. In this paper, we focus on the finite horizon setting. Moreover, the reward can be dependent on consecutive beliefs and the elements relating them (e.g., information gain [10]).

2.2. Belief space planning

The posterior belief at time instant k is given by

$$b_k(x_k) \approx \mathbb{P}\left(x_k | b_0, a_{0:k-1}, z_{1:k}\right). \tag{2}$$

The belief is an efficient way of storing all relevant information that has been obtained so far. The usual assumption is that the belief is a sufficient statistic for decision making objective [3]. However, in practice, the belief requires some representation. In general, this representation is not perfect, e.g., parametric or sampled form; thus, in (2), we used the $\approx$ sign. In a real life scenario

$$b_k = \psi(\psi(\ldots\psi(b_0, a_0, z_1), a_{k-2}, z_{k-1}), a_{k-1}, z_k), \tag{3}$$

where ψ is a method for updating the belief. Denote by π_ℓ policy at time step ℓ such that $\pi_\ell(b_\ell) = a_\ell$ maps belief to the action. It is noteworthy that policy $\pi(b)$ is a random function of the belief in general. For simplicity we assume that policy is deterministic. However, our development is not constrained to deterministic policies. By $\pi \triangleq \pi_{k:k+L-1}$ we denote a vector of policies for L time steps starting from time step k. Let us focus on the finite horizon setting. The general decision making under uncertainty objective function is of the following form

$$V^L(b_k, \pi) = \varphi\left(\mathbb{P}\left(\rho_{k+1:k+L} | b_k, \pi_{k:k+L-1}\right), g_k\right) \tag{4}$$
$$\text{s.t. } b_\ell = \psi(b_{\ell-1}, \pi_{\ell-1}(b_{\ell-1}), z_\ell),$$

where L is the planning horizon, ρ_ℓ is a random immediate reward, φ is an objective operator, and $g_k \triangleq f_{g_k}(\rho_{k+1:k+L})$ is the return [38]. The return is a deterministic function of the realization of $\rho_{k+1:k+L}$. A common choice for φ is expectation over the distribution of future rewards given all data available [6]. The return is some known function of the realization of $\rho_{k+1:k+L}$; as discussed in [6], e.g., it could correspond to the cumulative reward $g_k = \sum_{\ell=1}^{L} \rho_{k+\ell}$. Finally, ψ is a general method for propagating the belief with action and updating it with the received observation.

The objective (4) is ultimately based on the *distribution of the return* given all information available for planning and some selected policy $\mathbb{P}(g_k|b_k, \pi_{k:k+L-1})$, which decomposes via marginalization over future observations $z_{k+} \equiv z_{k+1:k+L}$ as

$$\mathbb{P}(g_k|b_k, \pi) = \int_{z_{k+}} \mathbb{P}(g_k|b_k, \pi_{k:k+L-1}, z_{k+1:k+L}) \cdot \mathbb{P}(z_{k+1:k+L}|b_k, \pi_{k:k+L-1}) dz_{k+1:k+L}. \tag{5}$$

A conventional assumption is that $\mathbb{P}(g_k|b_k, \pi, z_{k+},)$ is a Dirac delta function.

3. Foundations

In this section we extend POMDP with belief-dependent rewards to probabilistic POMDP and rigorously define the *simplification* paradigm. We further continue to the formulation of the general bounds on the reward/return which can be analytical or stochastic. We conclude this section with our key insight.

3.1. Extended setting, probabilistic POMDP with belief dependent reward

Sometimes the belief $b_{\ell-1}$ has a simple parametric form, where $\theta_{\ell-1}$ is a vector of parameters, e.g., a Gaussian belief. In this case, belief update ψ can be deterministic, and is denoted by $\psi_{\mathrm{dt}}(\theta_{\ell-1}, \pi_{\ell-1}(\theta_{\ell-1}), z_\ell)$, where the subscript dt stands for deterministic. In more general and challenging scenarios the belief $b_{\ell-1}$ is given by a set of weighted samples $\{(w^i_{\ell-1}, x^i_{\ell-1})\}_{i=1}^N$. Therefore, ψ is a stochastic method, e.g., a particle filter [42]. Applying multiple times ψ on the same input will yield different sets of samples approximating the same distribution of the posterior belief. We denote the stochastic ψ by $\psi_{\mathrm{st}}(b_{\ell-1}, \pi_{\ell-1}(b_{\ell-1}), z_\ell)$. Thus, ψ_{st} is a random function of the previous belief, an action and the observation. Note also another common situation where $b_{\ell-1}$ is parameterized, but there is no closed form update. In this case, ψ is also a stochastic method. Another form to formulate the above is that the distribution

$$B(b_{\ell-1}, \pi_{\ell-1}(b_{\ell-1}), z_\ell, b_\ell) \triangleq \mathbb{P}_B\left(b_\ell|b_{\ell-1}, \pi_{\ell-1}, z_\ell\right), \tag{6}$$

is not a Dirac delta function. This aspect was disregarded so far, to the best of our knowledge. Note that in a Belief MDP (BMDP) formulation, the assumption is that B is a Dirac delta function.

Similar arguments also hold for the momentary reward operator of the belief and the previous action. In its pure theoretical form, the momentary reward is a deterministic operator of the posterior belief and possibly an action. For example, a common immediate reward is of the form

$$\rho_{\mathrm{dt}}(b) = \mathbb{E}_{x\sim b}\left[f(b(x), x)\right] = \int_x b(x) f(b(x), x) dx, \tag{7}$$

where usually $f(b(x), x) = -\log b(x)$ or some reward on the state $f(b(x), x) = r(x)$, producing differential entropy or mean distance to goal. Unfortunately, an analytical expression for the reward operator $\rho_{\mathrm{dt}}(\cdot)$ is available in only limited scenarios, e.g., if the belief is modeled as Gaussian and the reward is differential entropy. The representation of the beliefs in (6) dictates adequate practical reward operators. Sometimes the deterministic operator can be constructed on top of a particular belief representation. E.g., (6) outputs a set of weighted samples and (7) is adapted to be a deterministic operator of this output [4]. However, it is not always possible. In extremely challenging situations the reward includes modification of the representation of the belief. This could introduce an additional source of stochasticity. We extend (7) to

$$R(b_{\ell-1}, \pi_{\ell-1}(b_{\ell-1}), z_\ell, b_\ell, \rho_\ell) \triangleq \mathbb{P}_R\left(\rho_\ell|b_\ell, z_\ell, \pi_{\ell-1}(b_{\ell-1}), b_{\ell-1}\right), \tag{8}$$

embracing these possibilities. To our knowledge, we are the first who treat these aspects as random.

Before introducing simplification formally and analyzing its impact, we shall account for all potential sources of variability. We remove conventional approximations by extending (1) to a probabilistic reward model R (8) and probabilistic belief update B (6), and introduce

$$M = \langle \mathcal{X}, \mathcal{A}, \mathcal{Z}, T, O, R, \gamma, b_k, B\rangle, \tag{9}$$

which we name probabilistic ρ-POMDP ($\mathbb{P}\rho$-POMDP). The rationale behind these conditional distributions (R and B) is to capture additional sources of stochasticity, such as stochastic belief update, stochastic calculation of a given reward operator or simply not knowing the operator reward in an explicit analytic form.

As discussed earlier, the value function (4) is based on (5). These previously overlooked sources of stochasticity impact the likelihood of the observations

$$\mathbb{P}\left(z_{k+1:k+L}|b_k, \pi\right), \tag{10}$$

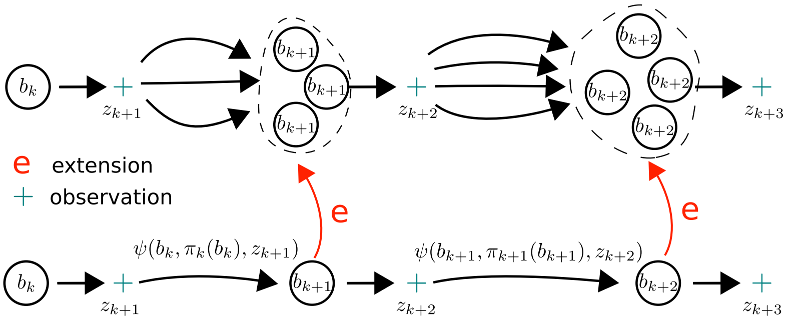


Fig. 1. Illustration of one branch of the extended belief tree. In a conventional setting (bottom), under the policy π, a specific realization of observations $z_{k+1:k+3}$ defines the beliefs along the way. In our extended setting (top), that is not the case, as discussed in text. It is customary to choose the same beliefs used to build the tree to obtain reward distribution or samples from the reward. We decoupled beliefs from the tree and beliefs from the reward calculation. By the red arrow, we denote our extension (red e). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

as well as the joint reward distribution $\mathbb{P}\left(\rho_{k+}|b_k, \pi, z_{k+}\right) \equiv \mathbb{P}\left(\rho_{k+1:k+L}|b_k, \pi_{k:k+L-1}, z_{k+1:k+L}\right)$ given a realization of future observations. The latter can be factorized as

$$\mathbb{P}\left(\rho_{k+}|b_k, \pi, z_{k+}\right) = \\ \int\limits_{b_{k+1}} \mathbb{P}_R\left(\rho_{k+1}|b_{k+1}, z_{k+1}, \pi_k, b_k\right) \mathbb{P}_B\left(b_{k+1}|b_k, \pi_k, z_{k+1}\right) \\ \int\limits_{b_{k+2}} \dots \int\limits_{b_{k+L}} \mathbb{P}_R\left(\rho_{k+L}|b_{k+L}, z_{k+L}, \pi_{k+L-1}, b_{k+L-1}\right) \mathbb{P}_B\left(b_{k+L}|b_{k+L-1}, \pi_{k+L-1}, z_{k+L}\right) \mathrm{d}b_{k+L} \dots \mathrm{d}b_{k+2} \mathrm{d}b_{k+1}. \tag{11}$$

In contrast, in the regular setting of POMDP and ρ-POMDP $\mathbb{P}\left(\rho_{k+}|b_k, \pi, z_{k+}\right)$ is Dirac's delta function. If B is a Dirac function, a sample from (10) uniquely defines the corresponding posterior beliefs $b_{k+1:k+L}$. This, therefore, corresponds to the classical belief tree (R could still be non a Dirac function). In contrast, our $\mathbb{P}\rho$-POMDP (9), corresponds to an *extended* belief tree, which, due to (6), allows many samples of the beliefs $b_{k+1:k+L}$ for each sample of $z_{k+1:k+L}$ from (10). We illustrate this in Fig. 1.

3.2. Simplification formulation

To formally define the simplification procedure, we augment the $\mathbb{P}\rho$-POMDP tuple (9) with a simplification operator ν,

$$M_\nu = \langle \mathcal{X}, \mathcal{A}, \mathcal{Z}, T, O, R, \gamma, b_k, B, \nu \rangle, \quad \nu \triangleq \nu_k, \dots, \nu_{k+L}. \tag{12}$$

This general operator defines any possible modification of the original problem defined by (9) alongside with (4) to a new, simpler to solve, problem. The definition (12) allows us to retain the connection to the original nonsimplified problem (9) and examine the impact of the simplification on (9). Further, we also define a novel decision making problem, undergoing simplification to ease the computational burden. The operator ν can be for example, sparsification of the initial belief b_k [8], substitution of the operator differential entropy by a simpler operator, e.g., trace of covariance matrix, discarding the normalizer in the differential entropy operator [30], replacing the reward by its topological signature [19], direct calculation of lightweight reward bounds [40], selecting a subset of hypotheses in a hybrid or mixture belief [32]. In Section 6, we consider a specific simplification of taking less samples for reward calculation considering parametric and non-parametric beliefs.

Generally, M and M_ν are different decision making problems. We shall be interested in working online with the latter while providing the guarantees with respect to the former.

To distinguish a simplified reward from the original reward, we denote the former by $\breve{\rho}$ instead of ρ; similarly, we denote the simplified belief by $\breve{b}$ instead of b. Note the operator ν can be stochastic, as discussed below.

Specifically, belief simplification is described by the distribution

$$\mathbb{P}(\breve{b}_\ell|b_\ell; \nu_\ell^b). \tag{13}$$

In general, the distribution (13) over the simplified belief $\breve{b}_\ell$ corresponds to a stochastic simplification operator ν_ℓ^b. This is the case, for example, when b_ℓ is represented by a set of N weighted samples and ν_ℓ^b is the operation of subsampling n samples according to weights; i.e., applying this operation on b_ℓ multiple times leads to different sets of n samples, each representing another realization of $\breve{b}_\ell$ from (13). Overall there are $\binom{N}{n}$ such combinations. For a deterministic operator ν^b_ℓ, (13) is a Dirac function.

Further, there are several cases of how a simplification affects belief update (6) from time $\ell - 1$ to ℓ.

1. Without any simplification we have $\mathbb{P}_B(b_\ell | b_{\ell-1}, \pi_{\ell-1}, z_\ell)$ from (6).
2. Given a simplified belief $\breve{b}_{\ell-1}$, while keeping the original stochastic belief update ψ_{st}, we have

   $$\mathbb{P}_B(\breve{b}_\ell | \breve{b}_{\ell-1}, \pi_{\ell-1}, z_\ell),$$

   where each realization of $\breve{b}_\ell$ is obtained via ψ_{st}. Thus, given $\breve{b}_{\ell-1}$, this distribution is not a function of ν.
3. We can also simplify the belief update operator, ψ_{st}, to $\breve{\psi}_{\text{st}}$. Denoting the corresponding simplification operator ν^ψ_ℓ, this yields

   $$\mathbb{P}_{\breve{B}}(\breve{b}_\ell | \breve{b}_{\ell-1}, \pi_{\ell-1}, z_\ell; \nu^\psi_\ell).$$

4. Finally, one can decide at time ℓ to apply simplification on the belief (determined by ν^b_ℓ) via (13). The corresponding belief update can be written as

   $$\mathbb{P}_{\breve{B}}(\breve{b}_\ell | \breve{b}_{\ell-1}, \pi_{\ell-1}, z_\ell; \nu^b_\ell, \nu^\psi_\ell) = \int\limits_{\tilde{b}_\ell} \mathbb{P}(\breve{b}_\ell | \tilde{b}_\ell; \nu^b_\ell) \mathbb{P}_{\breve{B}}(\tilde{b}_\ell | \breve{b}_{\ell-1}, \pi_{\ell-1}, z_\ell; \nu^\psi_\ell) \mathrm{d}\tilde{b}_\ell,$$

   where $\tilde{b}_\ell$ is the integration variable.

We combine these cases and write

$$\breve{B}\left(\breve{b}_{\ell-1}, \pi_{\ell-1}, z_\ell, \breve{b}_\ell; \nu\right) \triangleq \mathbb{P}_{\breve{B}}(\breve{b}_\ell | \breve{b}_{\ell-1}, \pi_{\ell-1}, z_\ell; \nu^b_\ell, \nu^\psi_\ell). \tag{14}$$

Similarly, reward simplification could be, in general, stochastic, leading to the distribution

$$\mathbb{P}(\breve{\rho}_\ell | \rho_\ell; \nu^\rho_\ell). \tag{15}$$

Thus, given a simplified belief $\breve{b}_\ell$ and $\breve{b}_{\ell-1}$, and recalling (8), the distribution over $\breve{\rho}_\ell$ is

$$\mathbb{P}_{\breve{R}}(\breve{\rho}_\ell | \breve{b}_\ell, z_\ell, \pi_{\ell-1}(\breve{b}_{\ell-1}), \breve{b}_{\ell-1}; \nu) = \int\limits_{\tilde{\rho}_\ell} \mathbb{P}(\breve{\rho}_\ell | \tilde{\rho}_\ell; \nu^\rho_\ell) \mathbb{P}_R(\tilde{\rho}_\ell | \breve{b}_\ell, z_\ell, \pi_{\ell-1}(\breve{b}_{\ell-1}), \breve{b}_{\ell-1}) \mathrm{d}\tilde{\rho}_\ell,$$

which we denote as the simplified reward model,

$$\breve{R}(\breve{b}_\ell, z_\ell, \pi_{\ell-1}(\breve{b}_{\ell-1}), \breve{\rho}_\ell; \nu) \triangleq \mathbb{P}_{\breve{R}}\left(\breve{\rho}_\ell | \breve{b}_\ell, z_\ell, \pi_{\ell-1}(\breve{b}_{\ell-1}), \breve{b}_{\ell-1}; \nu\right). \tag{16}$$

Throughout the document we assume that operator ν does not affect the observations likelihood. In other words, the measurements are sampled as in the original problem as in (10).

3.2.1. Joint distribution of simplified and the original reward given the candidate policy and the observations

Consequently, the models (14) and (16) impact (11), and lead to several alternatives for the original and the simplified joint reward distribution given a realization of the future observations. The first alternative is to simplify the initial belief b_k to $\breve{b}_k$ and apply the update method ψ_{st} on the simplified belief

$$\begin{aligned}
\mathbb{P}\left(\rho_{k+}, \breve{\rho}_{k+} | b_k, \pi, z_{k+}, \nu\right) = & \int\limits_{\breve{b}_k} \mathbb{P}(\breve{b}_k | b_k; \nu^b_k) \int\limits_{b_{k+1}} \int\limits_{\breve{b}_{k+1}} \mathbb{P}_{\breve{B}}\left(\breve{b}_{k+1} | \breve{b}_k, \pi_k, z_{k+1}; \nu\right) \mathbb{P}_B\left(b_{k+1} | b_k, \pi_k, z_{k+1}\right) \\
& \cdot \mathbb{P}_{\breve{R}}\left(\breve{\rho}_{k+1} | \breve{b}_{k+1}, z_{k+1}, \pi_k, \breve{b}_k; \nu\right) \mathbb{P}_R\left(\rho_{k+1} | b_{k+1}, z_{k+1}, \pi_k, b_k\right) \int\limits_{b_{k+2}} \int\limits_{\breve{b}_{k+2}} \dots \\
& \int\limits_{b_{k+L}} \int\limits_{\breve{b}_{k+L}} \mathbb{P}_{\breve{B}}\left(\breve{b}_{k+L} | \breve{b}_{k+L-1}, \pi_{k+L-1}, z_{k+L}; \nu\right) \mathbb{P}_B\left(b_{k+L} | b_{k+L-1}, \pi_{k+L-1}, z_{k+L}\right) \\
& \mathbb{P}_{\breve{R}}\left(\breve{\rho}_{k+L} | \breve{b}_{k+L}, z_{k+L}, \pi_{k+L-1}, \breve{b}_{k+L-1}; \nu\right) \mathbb{P}_R\left(\rho_{k+L} | b_{k+L}, z_{k+L}, \pi_{k+L-1}, b_{k+L-1}\right) \mathrm{d}b_{k+L} \mathrm{d}\breve{b}_{k+L} \dots \\
& \mathrm{d}b_{k+2} \mathrm{d}\breve{b}_{k+2} \mathrm{d}b_{k+1} \mathrm{d}\breve{b}_{k+1} \mathrm{d}\breve{b}_k.
\end{aligned} \tag{17}$$

The second alternative is to maintain/update the original belief. In this situation, we maintain and update the original belief and then use it to determine a simplified belief to calculate the simplified reward. This is in contrast to updating based on simplified beliefs from previous times as in (17). Thus,

$$\begin{aligned}
&\mathbb{P}\left(\rho_{k+}, \breve{\rho}_{k+} | b_k, \pi, z_{k+}, \nu\right) = \\
&\int\limits_{\breve{b}_k} \mathbb{P}(\breve{b}_k | b_k; \nu^b_k) \int\limits_{b_{k+1}} \int\limits_{\breve{b}_{k+1}} \mathbb{P}_{\breve{R}}(\breve{\rho}_{k+1} | \breve{b}_{k+1}, z_{k+1}, \pi_k(\breve{b}_k), \breve{b}_k; \nu) \mathbb{P}_R\left(\rho_{k+1} | b_{k+1}, z_{k+1}, \pi_k, b_k\right) \\
&\mathbb{P}(\breve{b}_{k+1} | b_{k+1}; \nu^b_{k+1}) \mathbb{P}_B\left(b_{k+1} | b_k, \pi_k, z_{k+1}\right) \int\limits_{b_{k+2}} \int\limits_{\breve{b}_{k+2}} \dots \\
&\int\limits_{b_{k+L}} \int\limits_{\breve{b}_{k+L}} \mathbb{P}_{\breve{R}}(\breve{\rho}_{k+L} | \breve{b}_{k+L}, z_{k+L}, \pi_{k+L-1}(\breve{b}_{k+L-1}), \breve{b}_{k+L-1}; \nu) \mathbb{P}_R\left(\rho_{k+L} | b_{k+L}, z_{k+L}, \pi_{k+L-1}, b_{k+L-1}\right) \\
&\mathbb{P}(\breve{b}_{k+L} | b_{k+L}; \nu^b_{k+L}) \mathbb{P}_B\left(b_{k+L} | b_{k+L-1}, \pi_{k+L-1}, z_{k+L}\right) \mathrm{d}b_{k+L} \mathrm{d}\breve{b}_{k+L} \dots \mathrm{d}b_{k+2} \mathrm{d}\breve{b}_{k+2} \mathrm{d}b_{k+1} \mathrm{d}\breve{b}_{k+1} \mathrm{d}\breve{b}_k.
\end{aligned} \tag{18}$$

Having introduced the two alternatives above, we are ready to go through their differences. The simplification approach defined by (17) uses only the observations from the belief tree. In the sequel, we explain why it is advantageous. In this setting, in addition to maintaining/updating b_ℓ for constructing the extended belief tree, we also have to maintain/update the simplified version $\breve{b}_\ell$. Nevertheless, the advantage is that by definition of (17), we nullify covariance between simplified and the original reward/return as opposed to the equation (18). In other words, considering (17) one can write,

$$\mathbb{P}\left(\rho_{k+}, \breve{\rho}_{k+} | b_k, \pi, z_{k+}, \nu\right) = \mathbb{P}\left(\rho_{k+} | b_k, \pi, z_{k+}\right) \mathbb{P}\left(\breve{\rho}_{k+} | b_k, \pi, z_{k+}, \nu\right) \tag{19}$$

Alternatively, if maintaining/updating a simplified belief is not desirable, e.g., belief is given as a surface but no closed form solution for reward exists, the update of a simplified belief $\breve{b}_\ell$ can be avoided. Towards this end we use (18). This is in contrast to updating based on simplified beliefs from previous times as in (17). Such simplification does not necessarily require to sample the original beliefs again. One could utilize original beliefs already present in the belief tree for simulating the observations. In the next section we delve into the subject of the bounds. Importantly, from structure of (18) we see that this distribution cannot be broken down to the multiplication of the marginals as in (19). In particular, the correlation is present through the component $\mathbb{P}(\breve{b}_\ell | b_\ell; \nu^b_\ell)$.

Note, sometimes estimators of the reward, e.g., [4] require a specific connection between two consecutive beliefs.

3.3. Online stochastic and analytical bounds

While thus far we considered the joint distribution over original and simplified rewards, $\mathbb{P}\left(\rho_{k+}, \breve{\rho}_{k+} | b_k, \pi, z_{k+}, \nu\right)$, in an online setting we do not have access to the original rewards as calculating them explicitly defeats the purpose of simplification. Instead, we shall now utilize simplification to provide bounds over the original rewards. These bounds can be used to provide performance guarantees, and should be cheaper to calculate than the original rewards.

Further, the bounds can be analytical as in previous simplification approaches, e.g., [8]. The bounds can be obtained via a simplified reward or directly as in [39]. For example, the authors of [39] proposed lightweight analytical adaptive bounds, calculated from the original belief, such that $l_\ell \leq \rho_\ell \leq u_\ell$ is always true by definition. If the bounds are calculated directly, we skip (18) and have instead the following.

$$\begin{aligned}
&\mathbb{P}\left(\rho_{k+}, l_{k+}, u_{k+} | b_k, \pi, z_{k+}, \nu\right) = \int\limits_{\breve{b}_k} \mathbb{P}(\breve{b}_k | b_k; \nu^b_k) \int\limits_{b_{k+1}} \int\limits_{\breve{b}_{k+1}} \\
&\mathbb{P}(\rho_{k+1}, l_{k+1}, u_{k+1} | b_{k+1}, \breve{b}_{k+1}, z_{k+1}, \pi_k(b_k), b_k, \breve{b}_k; \nu) \mathbb{P}(\breve{b}_{k+1} | b_{k+1}; \nu^b_{k+1}) \mathbb{P}_B\left(b_{k+1} | b_k, \pi_k, z_{k+1}\right) \\
&\int\limits_{b_{k+2}} \int\limits_{\breve{b}_{k+2}} \dots \int\limits_{b_{k+L}} \int\limits_{\breve{b}_{k+L}} \mathbb{P}(\rho_{k+L}, l_{k+L}, u_{k+L} | b_{k+L}, \breve{b}_{k+L}, z_{k+L}, \pi_{k+L-1}(b_{k+L-1}), b_{k+L-1}, \breve{b}_{k+L-1}; \nu) \\
&\mathbb{P}(\breve{b}_{k+L} | b_{k+L}; \nu^b_{k+L}) \mathbb{P}_B\left(b_{k+L} | b_{k+L-1}, \pi_{k+L-1}, z_{k+L-1}\right) \mathrm{d}\breve{b}_{k+L} \mathrm{d}b_{k+L} \dots \mathrm{d}\breve{b}_{k+2} \mathrm{d}b_{k+2} \mathrm{d}\breve{b}_{k+1} \mathrm{d}b_{k+1} \mathrm{d}\breve{b}_k.
\end{aligned} \tag{20}$$

The simplification type depicted by (20) is an extension of *in-place simplification* described in [39] to an extended setting. Ultimately for each realization of the return we are interested in the following relation

$$l \leq g_k \leq u. \tag{21}$$

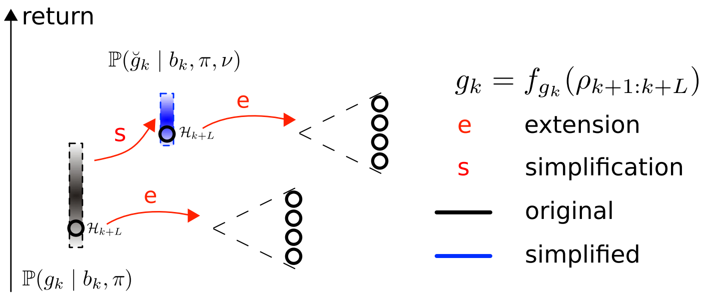

return
$\mathbb{P}(\breve{g}_k \mid b_k, \pi, \nu)$
s
$\mathcal{H}_{k+L}$
e
e
$\mathcal{H}_{k+L}$
$\mathbb{P}(g_k \mid b_k, \pi)$
$g_k = f_{g_k}(\rho_{k+1:k+L})$
e extension
s simplification
— original
— simplified

Fig. 2. Our extended setting permits variability of the reward given the present and a realization of the future. On the contrary, in a conventional setting, (23) is always a Dirac delta function. Our extension reflects on the original distribution of the return as well as the simplified.

One way to do that is to develop analytical bounds, which will hold for any possible observation $z_{k+1:k+L}$ received and any realization of return, e.g., as in [39].

In this section we show that there is another way to find more lenient bounds. Let $\mathcal{H}_{k+L} \triangleq \{b_k, \pi, z_{k+}\}$ be future history at the time index $k+L$. Our extension allows R and B, as well as $\breve{R}$ and $\breve{B}$ to be any distributions.

They can remain Dirac functions, e.g., if belief update and the reward calculation have a closed form

$$\mathbb{P}\left(\rho_\ell|b_{\ell-1},a_{\ell-1},z_\ell\right)=\delta(\rho_\ell-\rho_{\rm dt}(\psi_{\rm dt}(\theta_{\ell-1},a_{\ell-1},z_\ell))). \tag{22}$$

Successively, $\mathbb{P}(g_k|b_k,\pi,z_{k+},)$ remains Dirac delta. However, in the more general case, following our extension, there is a joint distribution of original and simplified returns given a realization of the future and the present

$$\mathbb{P}(g_k,\breve{g}_k|\mathcal{H}_{k+L},\nu), \tag{23}$$

as illustrated in Fig. 2. As we observe in Fig. 2, given the history $\mathcal{H}_{k+L}$, the return g_k as well as the simplified return $\breve{g}_k$ has variability, in contrast to the conventional approach. Ordinarily, the belief update is commenced once and treated as deterministic. So as the rewards and return do not have variance given the history of the actions and the observations. Since (23) is no longer a Dirac function, we can use knowledge about this distribution to design bounds, which will hold with *some* probability. In Section 6, we show that it is possible to harness the structure of (23) to design the mentioned more lenient online bounds.

Our framework permits to detach the process of estimation of the bounds from the realization of the reward and truly use all accessible information in a simplified problem. For example, one way to design probabilistic bounds is to find online a random variable or deterministic scalar ϵ such that the probability

$$P(|g_k-\breve{g}_k|\leq\epsilon|\mathcal{H}_{k+L},\nu) \tag{24}$$

is bounded from below. The corresponding probabilistic lower and upper bounds will be $l=\breve{g}_k-\epsilon$ and $u=\breve{g}_k+\epsilon$, respectively. We, therefore, refer to l and u as random variables. In our setting, even if the bounds actually bound with very low probability, it is still possible to analyze the quality of the simplification. Moreover, analytical bounds, designed in a conventional setting, can be used in our extended setting without any revision. In our extended setting, they will bound with probability one.

Having introduced the novel stochastic bounds, we proceed to the formulation of the constraints, that these bounds shall fulfill to be meaningful. The following conditional $\mathbb{P}(g_k,\breve{g}_k,l,u|\mathcal{H}_{k+L},\nu)$ encloses all the variables situated in the problem. Let the parameter controlling the confidence level be $\alpha\in[0,1)$. For every possible sample $\breve{g}_k$ we do not know which sample g_k one could obtain in the original problem. However, if the bounds are designed such that $\mathbb{P}(g_k,l,u|\mathcal{H}_{k+L},\nu)$ render

$$1-\alpha\leq P\left(\mathbf{1}\{l\leq g_k\leq u\}=1|\mathcal{H}_{k+L},\nu\right) \tag{25}$$

these bounds can be useful. Notably, the above equation does not involve simplified return, so is applicable also in the case bounds are directly formulated (and not via a simplified return). However, in this case the bounds are analytical and $\alpha=0$. To summarize, there are three types of online reward/return bounds:

1. Deterministic bounds. These analytical bounds exist in case of a closed form belief update $\psi_{\rm dt}$ and a deterministic operator reward $\rho_{\rm dt}(b)$ from (7), e.g., belief is a Gaussian and the reward is differential entropy. In this case, even in our extended setting R and B remain Dirac functions.
2. Stochastic bounds that hold with probability one, namely $\alpha=0$. These are also analytical bounds. In our extended setting R and B are no longer Dirac functions. However, these bounds hold for any realization of sample approximation, as stated around (21).
3. Stochastic bounds that hold at least with probability $1-\alpha$. They exist only in our extended setting when R and B are not Dirac functions.

3.4. Key insight - characterization of the return using stochastic bounds

Let us recite that our goal is to accelerate the decision making. We recall the notion of "usual stochastic order" and interpret the definition within our context.

Usual stochastic order implies, that if for three random variables l, g_k, u given b_k, π holds $l \leq g_k$ and $g_k \leq u$ for $\forall \omega \in \Omega$, so $\forall \xi \in (-\infty, \infty)$

$$P(l > \xi | b_k, \pi, \nu) \leq P(g_k > \xi | b_k, \pi) \leq P(u > \xi | b_k, \pi, \nu). \tag{26}$$

Let us present out main theorem which we will extensively use further.

Theorem 1 *(Characterization of the return using stochastic bounds). Fix $\alpha \in [0, 1)$. Assume that (25) holds. This implies that $\forall \xi \in (-\infty, \infty)$*

$$(P(l > \xi | b_k, \pi, \nu) - \alpha)(1 - \alpha) \leq P(g_k > \xi | b_k, \pi) \leq \frac{P(u > \xi | b_k, \pi, \nu)}{1 - \alpha} + \alpha. \tag{27}$$

For the detailed proof please refer to Appendix A.1. Let us further improve the bounds as such

$$\mathcal{LB}_\alpha(\xi) = 0 \vee \big(P(l > \xi | b_k, \pi, \nu) - \alpha\big)(1 - \alpha) \leq P(g_k > \xi | b_k, \pi), \tag{28}$$

where $\vee$ is a maximum operator.

$$P(g_k > \xi | b_k, \pi) \leq 1 \wedge \left(\frac{P(u > \xi | b_k, \pi, \nu)}{1 - \alpha} + \alpha\right) = \mathcal{UB}_\alpha(\xi), \tag{29}$$

where $\wedge$ is the minimum operator.

4. Simplification impact on a marginal distribution of the return

Previously, we defined a simplification procedure that results in a corresponding new decision making problem that should be easier to solve. From $\mathbb{P}\left(\rho_{k+} | b_k, \pi, z_{k+}, \nu\right)$ and $\mathbb{P}\left(\breve{\rho}_{k+} | b_k, \pi, z_{k+}, \nu\right)$ we arrive at the distribution of the original as well as simplified returns $\mathbb{P}(g_k | b_k, \pi)$ and $\mathbb{P}(\breve{g}_k | b_k, \pi, \nu)$ for the evaluated candidate policy. In this section, we show how the stochastic bounds can be utilized in the context of known risk aware objectives such as VaR. Notably, this section presents a discussion concerning the marginal distribution of the pair of the returns - original and simplified given a candidate policy.

4.1. Distributions affected by the simplification

In this section we decompose the distribution of interest. To grasp the simplification impact we shall assess the relation between simplified and original returns portrayed by the following distribution

$$\mathbb{P}(g_k, \breve{g}_k | b_k, \pi_{k:k+L-1}, \nu) = \int\limits_{z_{k+}} \mathbb{P}(g_k, \breve{g}_k | b_k, \pi, z_{k+}, \nu) \cdot \mathbb{P}(z_{k+} | b_k, \pi) dz_{k+}. \tag{30}$$

Recall that $z_{k+1:k+L} \equiv z_{k+}$. In general the simplification operator ν can affect also the observation likelihood (10). We leave it to future research.

4.2. Decision making

Algorithm 1 Generic *simplified with performance guarantees* sampling based decision making algorithm with challenging rewards and objectives (note that the theory is formulated for policies but here we discuss discrete space of action sequences $\mathcal{A}$).

```
**Input:** belief $b_k$, action space $\mathcal{A}$.
**for** action sequence $a^i$ from all possible action sequences $i \in 1 : |\mathcal{A}|$ **do**
    Sample returns and calculate interval $\mathcal{LB}^i, \mathcal{UB}^i$.
**end for**
Set optimal action sequence $a^{i^*}$ by $i^* = \arg\max_{i=1:|\mathcal{A}|} \mathcal{LB}^i$
Find $j^\star = \arg\max_{j=1:|\mathcal{A}|\setminus i^*} \mathcal{UB}^j$. Define loss incurred by the simplification as follows $\max\{0, \mathcal{UB}^{j^\star} - \mathcal{LB}^{i^*}\}$;
In case that absolute loss doesn't have meaning, define relative loss as follows $\frac{\max\{0, \mathcal{UB}^{j^\star} - \mathcal{LB}^{i^*}\}}{\min\{|\mathcal{LB}^{i^*}|, |\mathcal{UB}^{j^\star}|\}} \geq \frac{V^* - V(b_k, a^{i^*})}{|V^*|}$, where $V^*$ is true optimal value.
```

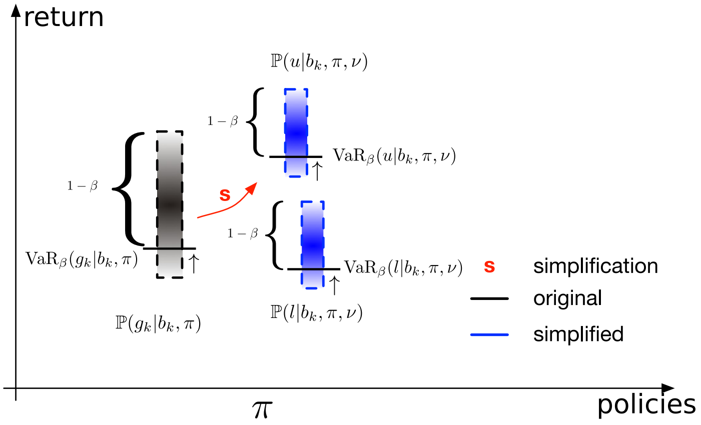

return
$\mathbb{P}(u|b_k,\pi,\nu)$
$1-\beta$
$\mathrm{VaR}_\beta(u|b_k,\pi,\nu)$
$1-\beta$
s
$1-\beta$
$\mathrm{VaR}_\beta(g_k|b_k,\pi)$
$\mathrm{VaR}_\beta(l|b_k,\pi,\nu)$
s simplification
original
simplified
$\mathbb{P}(l|b_k,\pi,\nu)$
$\mathbb{P}(g_k|b_k,\pi)$
π
policies

Fig. 3. Illustration of Value at Risk. Probability that the return will be under VaR is β. In other words VaR is β-quantile of the return.

In this section, we apply our findings in order to bound popular objectives with lightweight bounds to accelerate the decision making mechanism. We skip the most common objective operator, the expectation, since it is *oblivious* to the distribution of the returns. Therefore the expectation is not a risk averse objective. Motivated by this assertion, we consider more versatile objectives. Since it is not clear whether or not these objectives conform to Bellman form we incorporate simplification with generic decision making (Algorithm 1). In this algorithm, we traverse the loop over the possible action sequences. We calculate the interval defined by the upper and lower bound on the value function in each loop iteration. Finally, we select the action sequence with the highest lower bound and compare it with the highest upper bound corresponding to all other action sequences. We report a relative loss. Using analytical or stochastic bounds on the return our goal is to bound the value function V as such

$$\mathcal{LB} \le V \le \mathcal{UB} \tag{31}$$

to accelerate the decision making. To our knowledge there are no attempts to simplify risk aware decision making under uncertainty through maintaining guarantees on the *simplification* impact.

Risk averse objectives One possible risk averse probabilistic objective for POMDP is

$$\mathbb{E}[\mathbf{1}\{g_k > a\}|b_k,\pi] = P(g_k > a|b_k,\pi), \tag{32}$$

where $a \in \mathbb{R}$. Maximizing this objective can be thought as maximizing the probability of achieving the target a. If we choose belief dependent reward to be the negative entropy $I(b_k) = -\mathcal{H}(b_k)$, set $a = I(b_k)$ and the return to be terminal reward $g_k = \rho_L$, or $g_k = \frac{1}{L}\sum_{\ell=k+1}^{k+L}\rho_\ell$ [21], such objective quantifies the probability that information gain is positive; and such decision making prefers the action which maximizes probability of positive information gain. We can further control amount of most probable information gain by setting $a = c \cdot I(b_k)$, where c is a factor larger than one. Once optimal action is obtained we are confident that $g_k > a$ with probability $P(g_k > a|b_k,\pi)$. Substituting ξ by a the bounds from (28) and (29) hold for this objective.

Another objective is reward variant of Value at Risk (VaR) (Fig. 3)

$$\mathrm{VaR}_\beta(g_k|b_k,\pi) = \sup\{\xi \text{ s.t } P(g_k > \xi|b_k,\pi) \ge 1-\beta\}. \tag{33}$$

This objective articulates that we are interested in the maximal worst case return. Meaning maximal return such that probability mass to be above this return is larger than $1-\beta$. Notably, if $g_k|b_k,\pi$ has a strictly increasing Cumulative Distribution Function (CDF), the VaR is its β-quantile $\mathrm{VaR}_\beta(g_k|b_k,\pi) = P^{-1}(g_k \le \beta|b_k,\pi)$. The CDF of a continuous random variable is strictly increasing if it does not have intervals on a real line happening with probability zero. In the case of symmetrical distributions, the expected value overlaps with median, which is VaR with confidence level $\beta = 0.5$. Using again usual stochastic order, we can bound this objective with analytical return bounds as well as with stochastic return bounds. We start from analytical bounds. Let us focus on the lower bound. We want to show that

$$\mathrm{VaR}_\beta(l|b_k,\pi,\nu) \le \mathrm{VaR}_\beta(g_k|b_k,\pi). \tag{34}$$

Since the bounds are analytical we use (26) to behold

$$\{\xi \text{ s.t. } P(l > \xi|b_k,\pi,\nu) \ge 1-\beta\} \subseteq \{\xi \text{ s.t. } P(g_k > \xi|b_k,\pi) \ge 1-\beta\}. \tag{35}$$

We know that maximum on the containing set is above or equal to maximum on the contained. This argument yields

$$\sup\{\xi \text{ s.t. } P(l>\xi|b_k,\pi)\geq 1-\beta\}\leq \sup\{\xi \text{ s.t. } P(g_k>\xi|b_k,\pi)\geq 1-\beta\}. \tag{36}$$

Switching roles of l to g_k and g_k to u we have that

$$\mathrm{VaR}_{\beta}(g_k|b_k,\pi,\nu)\leq \mathrm{VaR}_{\beta}(u|b_k,\pi,\nu). \tag{37}$$

Now let us bound the objective (33) using stochastic bounds.

Theorem 2 *(Deterministic bound of Value at Risk using stochastic bounds on the return). Assume that* (25) *holds. Let* $0\leq\alpha<1$, $0\leq\beta<1$. *Assume that* $\alpha(2-\alpha)\leq\beta\leq 1-\alpha$.

$$\mathrm{VaR}_{\frac{\beta-\alpha(2-\alpha)}{1-\alpha}}(l|b_k,\pi,\nu)\leq\mathrm{VaR}_{\beta}(g_k|b_k,\pi)\leq\mathrm{VaR}_{\beta+\alpha(2-\beta-\alpha)}(u|b_k,\pi,\nu). \tag{38}$$

The reader can find the detailed proof in Appendix A.2. Let us mention that the above bounds hold for theoretical objectives. In practice, however, the sample approximations are sufficiently close to the theoretical values.

5. Simplification impact on the joint distribution of the returns given two policies

So far, we analyzed marginal distributions over the returns/rewards corresponding to a candidate policy in the context of known risk aware objectives. Interestingly, if we consider the joint distribution over the returns corresponding to two candidate policies, as we further show, we can define novel objectives and harness the information encoded in the joint distribution.

We start by showing that $\mathbb{P}(g_k,g'_k|b_k,\pi,\pi')\neq\mathbb{P}(g_k|b_k,\pi)\cdot\mathbb{P}(g'_k|b_k,\pi')$. The source for correlation is the mutual likelihood of observations:

$$\mathbb{P}(g_k,g'_k|b_k,\pi,\pi')=\int\limits_{\substack{z_{k+}\\ z'_{k+}}}\mathbb{P}(g_k,g'_k|b_k,\pi,\pi',z_{k+},z'_{k+})\cdot\mathbb{P}(z_{k+},z'_{k+}|b_k,\pi,\pi')dz_{k+}dz'_{k+}= \tag{39}$$

$$\int\limits_{\substack{z_{k+}\\ z'_{k+}}}\mathbb{P}(g_k|b_k,\pi,z_{k+})\mathbb{P}(g'_k|b_k,\pi' z'_{k+})\cdot\mathbb{P}(z_{k+},z'_{k+}|b_k,\pi,\pi')dz_{k+}dz'_{k+} \tag{40}$$

Let us observe the joint likelihood of observations given the belief at present time and two candidate policies $\mathbb{P}\left(z_{k+1:k+L},z'_{k+1:k+L}|b_k,\pi,\pi'\right)$ breaks down using chain rule as follows

$$\begin{aligned}
&\mathbb{P}\left(z_{k+1:k+L},z'_{k+1:k+L}|b_k,\pi,\pi'\right)=\mathbb{P}\left(z_{k+1},z'_{k+1}|b_k,\pi_k,\pi'_k\right)\\
&\int\limits_{b_{k+1}}\int\limits_{b'_{k+1}}\mathbb{P}\left(z_{k+2}|b_{k+1},\pi_{k+1}\right)\mathbb{P}\left(z'_{k+2}|b'_{k+1},\pi'_{k+1}\right)\mathbb{P}_B\left(b_{k+1}|b_k,\pi_k,z_{k+1}\right)\mathbb{P}_B\left(b'_{k+1}|b'_k,\pi'_k,z'_{k+1}\right)\\
&\int\limits_{b_{k+2}}\int\limits_{b'_{k+2}}\cdots\int\limits_{b_{k+L-1}}\int\limits_{b'_{k+L-1}}\mathbb{P}\left(z_{k+L}|b_{k+L-1},\pi_{k+L-1}\right)\mathbb{P}\left(z'_{k+L}|b'_{k+L-1},\pi'_{k+L-1}\right)\cdot\\
&\mathbb{P}_B\left(b_{k+L}|b_{k+L-1},\pi_{k+L-1},z_{k+L}\right)\mathbb{P}_B\left(b'_{k+L}|b'_{k+L-1},\pi'_{k+L-1},z'_{k+L}\right)\mathrm{d}b_{k+L-1}\mathrm{d}b'_{k+L-1}\ldots\mathrm{d}b_{k+2}\mathrm{d}b'_{k+2}\mathrm{d}b_{k+1}\mathrm{d}b'_{k+1}
\end{aligned} \tag{41}$$

The myopic observations $\mathbb{P}\left(z_{k+1},z'_{k+1}|b_k,\pi_k,\pi'_k\right)$ are correlated through the present time belief b_k. To see this explicitly we marginalize over the propagated states and employ the observation and motion models

$$\begin{aligned}
&\mathbb{P}\left(z_{k+1},z'_{k+1}|b_k,\pi_k,\pi'_k\right)=\\
&\int\limits_{\substack{x_{k+1}\\ x'_{k+1}\\ x_k}}\mathbb{P}_Z(z_{k+1}|x_{k+1})\mathbb{P}_Z(z'_{k+1}|x'_{k+1})\mathbb{P}_T(x_{k+1}|\pi_k(b_k),x_k)\mathbb{P}_T(x'_{k+1}|\pi'_k(b_k),x_k)b_k(x_k)\mathrm{d}x_{k+1}\mathrm{d}x'_{k+1}\mathrm{d}x_k.
\end{aligned} \tag{42}$$

This insinuates that to base decision making on marginal means to lose this correlation which we aim to exploit. Prompted by this insight we suggest an objective uncovered in the next section.

Remark: Note that when the belief is parametric, we do not have a way to jointly parametrically propagate the belief with a pair of actions as in (42). However, we can always sample the parametric belief. So even if the belief b_k is parametric we still can account for correlation in (42) by switching to samples.

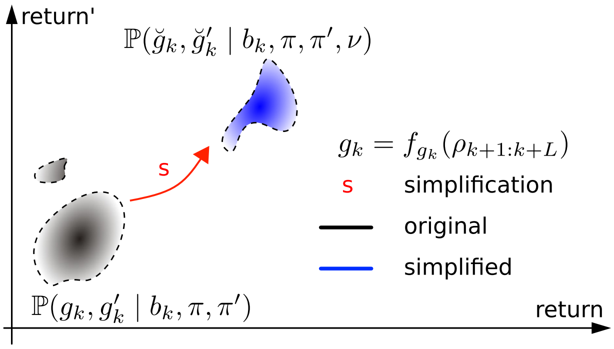

return'
$\mathbb{P}(\breve{g}_k, \breve{g}'_k \mid b_k, \pi, \pi', \nu)$
s
$g_k = f_{g_k}(\rho_{k+1:k+L})$
s simplification
original
simplified
$\mathbb{P}(g_k, g'_k \mid b_k, \pi, \pi')$
return

Fig. 4. This figure shows alteration of the distribution of joint returns g_k and g'_k of two candidate policies π and π' as a result of simplification. Color intensity denotes distribution values. This is a conceptual illustration, i.e., we do not imply higher/lower rewards or change of support due to simplification.

Extension of the decision making objective Let us define the objective to be maximized involving two candidate policies.

$$J^L(b_k,\pi,\pi') = \varphi\Big(\mathbb{P}\left(\rho_{k+1:k+L}, \rho'_{k+1:k+L}|b_k,\pi,\pi'\right), (g_k, g'_k)\Big) \quad (43)$$
$$\text{s.t. } b_\ell = \psi(b_{\ell-1}, \pi_{\ell-1}(b_{\ell-1}), z_\ell).$$

5.1. Distributions affected by simplification

Now we stipulate on the quality of the simplification for two candidate policies π and π'. To quantify the impact of the simplification procedure, we shall concentrate on the *joint* distribution of the pair of simplified and original returns appropriate for two candidate polices $\mathbb{P}(g_k, g'_k, \breve{g}_k, \breve{g}'_k|b_k,\pi,\pi',\nu)$. Our goal is to examine how the simplification procedure alters the joint distribution $\mathbb{P}(g_k, g'_k|b_k,\pi,\pi')$ towards $\mathbb{P}(\breve{g}_k, \breve{g}'_k|b_k,\pi,\pi',\nu)$. These two distributions are illustrated in Fig. 4. In the context of (43) we focus on

$$\mathbb{P}(g_k, g'_k, \breve{g}_k, \breve{g}'_k|b_k,\pi,\pi',\nu), \quad (44)$$

i.e., the joint distribution over original and simplified returns of both policies. This distribution decomposes via marginalization over future observations $z_{k+} \equiv z_{k+1:k+L}$ and $z'_{k+} \equiv z'_{k+1:k+L}$ as

$$\mathbb{P}(g_k, g'_k, \breve{g}_k, \breve{g}'_k|b_k,\pi,\pi',\nu) = \int\limits_{\substack{z_{k+}\\ z'_{k+}}} \mathbb{P}(g_k, g'_k, \breve{g}_k, \breve{g}'_k|b_k,\pi,\pi',\nu,z_{k+},z'_{k+}) \cdot \mathbb{P}(z_{k+},z'_{k+}|b_k,\pi,\pi')dz_{k+}dz'_{k+}, \quad (45)$$

which, according to (6), (8) and (14)-(16), decomposes to

$$\int\limits_{\substack{z_{k+}\\ z'_{k+}}} \mathbb{P}(g_k, \breve{g}_k|\mathcal{H}_{k+L},\nu)\mathbb{P}(g'_k, \breve{g}'_k|\mathcal{H}'_{k+L},\nu) \cdot \mathbb{P}(z_{k+},z'_{k+}|b_k,\pi,\pi')dz_{k+}dz'_{k+}. \quad (46)$$

Note that the pair of histories is defined as follows $\mathcal{H}_{k+L} \triangleq \{b_k,\pi,z_{k+}\}$ and $\mathcal{H}'_{k+L} \triangleq \{b_k,\pi',z'_{k+}\}$; where the belief b_k is shared by both histories.

In other words, the simplification operator ν independently affects each realization of the future. Given two such realizations $(\mathcal{H}_{k+L}, \mathcal{H}'_{k+L}, \nu)$, the pairs of original and simplified returns are statistically independent of all other rewards. This crucial observation will be significant in the sequel.

5.2. Decision making

This section outlines a generic algorithm for decision making favoring pairwise joint distribution of the returns (Algorithm 2). This algorithm starts by taking the first action sequence as the best. We again traverse the loop over the possible action sequences. We compare two action sequences in each loop iteration and select the current optimal sequence. In the end, the optimal action sequence is optimal with respect to all possible action sequences. In particular, we propose a novel method to perform decision making using the joint distribution of the returns corresponding to the two candidate policies. The authors from [27] proposed to perform decision making with maximum likelihood observations. However, when the belief distribution is general and sophisticated, a generalization of [27] is to compare number of samples which fulfill the

Algorithm 2 Generic *simplified with performance guarantees* sampling based decision making based on pairwise joint distributions (note that the theory is formulated for policies but here we discuss given discrete space $\mathcal{A}$ of action sequences).

```
Input: belief $b_k$, $\mathcal{A}$.
$a^* \leftarrow a^1$
for action sequence $a^i$ from all possible action sequences do
    Make simplified decision making using two actions $a^*$ and $a^i$
    $a^* \leftarrow$ action defined as optimal in the line above
end for
return $a^*$
```

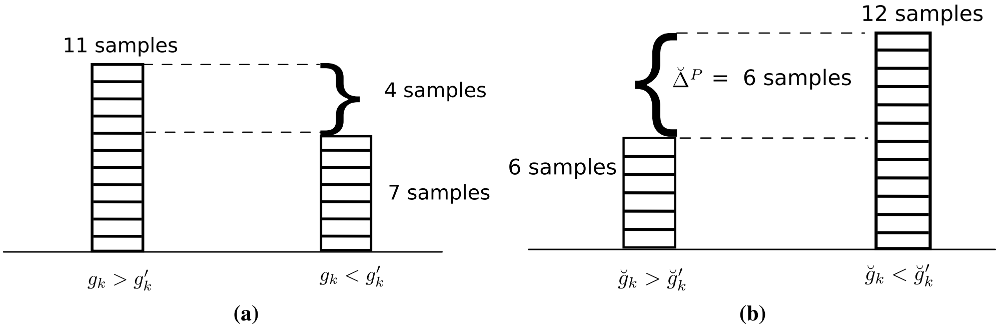

11 samples
4 samples
7 samples
$g_k > g'_k$
$g_k < g'_k$
(a)
12 samples
$\breve{\Delta}^P$ = 6 samples
6 samples
$\breve{g}_k > \breve{g}'_k$
$\breve{g}_k < \breve{g}'_k$
(b)

Fig. 5. **(a)** Hypothesis based decision making; **(b)** The outcome of decision making is wrong with margin 6 samples due to simplification.

hypothesis that $g_k > g'_k$ with number of samples satisfying $g_k < g'_k$. Such a decision making process can be thought as risk aware, since we are concerned with choosing an action which will be optimal with higher probability. Namely, if

$$\sum_{i=1}^{s} \mathbf{1}\{g^i_k > g'^i_k\} \geq \sum_{i=1}^{s} \mathbf{1}\{g^i_k < g'^i_k\}, \tag{47}$$

where the summation is over s samples of the pairs of the returns, we declare that π is better, else the π' is better. Note we assume that the event $g_k = g'_k$ happens with probability zero. This assumption is fulfilled with continuous distributions.

Simplified hypothesis based decision making Assume for the moment that bounds in (21) are analytical (hold with probability one), e.g., bounds from [39]. We can then define simplified returns as follows $\breve{g}_k = \frac{l+u}{2}$ and $\breve{g}'_k = \frac{l'+u'}{2}$. Simplification of the decision making portrayed by (47) is as follows. We take a simplified return instead of the original and ask if the following inequality is fulfilled

$$\sum_{i=1}^{s} \mathbf{1}\{\breve{g}^i_k > \breve{g}'^i_k\} \geq \sum_{i=1}^{s} \mathbf{1}\{\breve{g}^i_k < \breve{g}'^i_k\}. \tag{48}$$

If the answer is yes, we declare that π is better, else π' is better. Similar to not simplified decision making (47) we assume that the event $\breve{g}_k = \breve{g}'_k$ happens with probability zero.

The question is can we make a wrong conclusion with respect to the original problem due to the simplification, see Fig. 5. To provide guarantees on such a simplified decision making we first develop a novel mathematical tool we call Probabilistic Loss, which we believe has much bigger potential since it is able to describe the simplification impact for *any* operator objective φ. We then, in section 5.4, show how to provide guarantees for the specific objective operator (47).

5.3. Probabilistic loss ($\mathtt{PLoss}$)

Let us define the following random variable, which we shall refer to as "loss"

$$\mathcal{L} \triangleq f_{\mathcal{L}}(g_k, g'_k, \breve{g}_k, \breve{g}'_k) = \begin{cases} \max\{g'_k - g_k, 0\} \text{ if } \breve{g}_k > \breve{g}'_k, \\ \max\{g_k - g'_k, 0\} \text{ if } \breve{g}_k < \breve{g}'_k. \end{cases} \tag{49}$$

With (49) we aim to capture a complete impact of a simplification onto the decision making problem (43). Specifically, this definition captures for each possible realization of $g_k, g'_k, \breve{g}_k, \breve{g}'_k$ the absolute difference between the original returns $\Delta = |g'_k - g_k|$ in case action trend was not preserved on this realization. Meaning, at this realization, the optimal actions of original and simplified problems would differ. Given a sample $(g_k, g'_k, \breve{g}_k, \breve{g}'_k)$, the simplification is action consistent at this sample if the sign of the difference of the returns is preserved. In other words, the same action would be identified

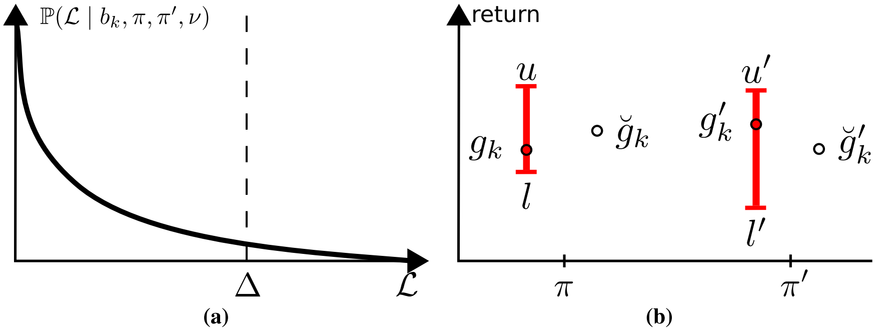


Fig. 6. Illustration of **(a)** the distribution of loss, and **(b)** the online bounds of the return.

as optimal with the original and simplified returns; else we must account for the loss (49). Our object of interest is the distribution density of $\mathcal{L}$ given all the information available at our disposal,

$$\mathbb{P}\left(\mathcal{L}|b_k, \pi, \pi', \nu\right). \tag{50}$$

We denote this distribution by Probabilistic Loss (PLoss). See illustration in Fig. 6a. E.g., if (50) is the Dirac delta function $\delta(\mathcal{L})$, the simplification method is absolute action consistent for every possible objective operator φ. Moreover, for any Δ, its CDF $\mathbb{P}\left(\mathcal{L} \leq \Delta|b_k, \pi, \pi', \nu\right)$ provides probability to suffer loss at most Δ. Similarly, the Tail Distribution Function (TDF) $\mathbb{P}\left(\mathcal{L} > \Delta|b_k, \pi, \pi', \nu\right)$ provides probability to suffer loss greater than Δ. The source of distribution (50) is (44).

5.3.1. Online bound on probabilistic loss (PbLoss)

The distribution defined by (50) requires access to (44) which we do not have in an online setting. To circumvent the requirement of accessing g_k and g'_k, we propose to substitute them by online lower and upper bounds l, u and l', u', respectively. These bounds should be accessible without knowledge of original returns. Similar to section 3.3 we aim to bound each original return corresponding to its candidate policy.

Let us consider a sampled return realization $(g_k, g'_k, \breve{g}_k, \breve{g}'_k)$ from (44). As in an online setting we do not actually have access to the original returns (g_k, g'_k), we strive to bound the latter,

$$l \leq g_k \leq u, \; l' \leq g'_k \leq u', \tag{51}$$

where, for now, we assume (51) holds for any sample of $(g_k, g'_k, \breve{g}_k, \breve{g}'_k)$; for example, these could be analytically-derived bounds. This setting is illustrated in Fig. 6b. However, further we also discuss a more general setting where we allow (21) to be violated with probability larger than zero.

Using these bounds we are able to define online a bound on loss (49) *without* accessing the original problem (R and B),

$$\bar{\mathcal{L}} \triangleq f_{\bar{\mathcal{L}}}(\breve{g}_k, l, u, \breve{g}'_k, l', u') = \begin{cases} \max\{u' - l, 0\} & \text{if } \breve{g}_k > \breve{g}'_k, \\ \max\{u - l', 0\} & \text{if } \breve{g}_k < \breve{g}'_k. \end{cases} \tag{52}$$

Note that sometimes we can find bounds over the returns by applying the same function f_{g_k} on the bounds on the momentary rewards (returns when $L = 1$), e.g., in case of cumulative reward $u = \sum_{\ell=k+1}^{k+L} u_\ell$ and $l = \sum_{\ell=k+1}^{k+L} l_\ell$. However, this is not always possible, e.g., if g_k deviates from the sum of momentary rewards or in the case of Bellman form of the objective. Sometimes it is, therefore, better to work with momentary bounds.

In an online setting, we are interested in the distribution density of $\bar{\mathcal{L}}$,

$$\mathbb{P}\left(\bar{\mathcal{L}}|b_k, \pi, \pi', \nu\right), \tag{53}$$

which we denote by Probabilistic Bound on Loss (PbLoss).

As we discuss in Section 5.3.2, PbLoss characterizes the impact of a simplification in an online setting; thus, it enables to determine online if a candidate simplification is acceptable given a user-specified criteria. The decision to either accept or decline a (candidate) simplification is guided by probabilistic guarantees, as provided by our approach.

5.3.2. Description of PLoss online

In this section, we show how PbLoss can be used in an online setting to characterize PLoss (which is unavailable online). In turn, this enables us to provide online probabilistic performance guarantees for a considered simplification (represented by operator ν), or to decide if it is adequate given a user-specified criteria.

Specifically, recall PLoss CDF and TDF, i.e., probability to suffer loss at most, or greater, than $\Delta \in \mathbb{R}$, respectively,

$$\text{PLoss CDF: } P\left(\mathcal{L} \leq \Delta|b_k, \pi, \pi', \nu\right) \tag{54}$$

$$\text{PLoss TDF: } P\left(\mathcal{L} > \Delta|b_k, \pi, \pi', \nu\right). \tag{55}$$

We now aim to bound `PLoss` CDF (54) from below, and `PLoss` TDF (55) from above by utilizing `PbLoss`.

We now consider `PLoss` TDF and express $P\left(\mathcal{L} > \Delta | b_k, \pi, \pi', \nu\right)$ as

$$P\left(\mathcal{L} > \Delta, \bar{\mathcal{L}} \geq \mathcal{L} | b_k, \pi, \pi', \nu\right) + P\left(\mathcal{L} > \Delta, \bar{\mathcal{L}} < \mathcal{L} | b_k, \pi, \pi', \nu\right).$$

The first term can be written via chain rule as

$$P\left(\mathcal{L} > \Delta | \bar{\mathcal{L}} \geq \mathcal{L}, b_k, \pi, \pi', \nu\right) P\left(\bar{\mathcal{L}} \geq \mathcal{L} | b_k, \pi, \pi', \nu\right). \tag{56}$$

Performing chain rule similarly also on the second term and recalling that $P(\bar{\mathcal{L}} \geq \mathcal{L}|\cdot) + P(\bar{\mathcal{L}} < \mathcal{L}|\cdot) = 1$, allows to express `PLoss` TDF as

$$P\big(\mathcal{L} > \Delta | \bar{\mathcal{L}} \geq \mathcal{L}, b_k, \pi, \pi', \nu\big)\, \lambda + P\big(\mathcal{L} > \Delta | \bar{\mathcal{L}} < \mathcal{L}, b_k, \pi, \pi', \nu\big)\, (1 - \lambda), \tag{57}$$

where

$$\lambda \triangleq P\left(\bar{\mathcal{L}} \geq \mathcal{L} | b_k, \pi, \pi', \nu\right) \equiv P\left(\mathbf{1}_{\{\bar{\mathcal{L}} \geq \mathcal{L}\}} = 1 | b_k, \pi, \pi', \nu\right). \tag{58}$$

While λ from (58) is unavailable, we can bound it from below using

$$1 - \alpha \leq P\left(\mathbf{1}_{\{l \leq g_k \leq u\}} = 1 | \mathcal{H}_{k+L}, \nu\right) \tag{59}$$

and

$$1 - \alpha \leq P\left(\mathbf{1}_{\{l' \leq g'_k \leq u'\}} = 1 | \mathcal{H}'_{k+L}, \nu\right) \tag{60}$$

and

Theorem 3 *(Probability that bound bounds). Fix $\alpha \in \mathbb{R}$. Assume that (59) and (60) hold. Then:*

$$P\left(\mathbf{1}_{\{\bar{\mathcal{L}} \geq \mathcal{L}\}} = 1 | b_k, \pi, \pi', \nu\right) \geq (1 - \alpha)^2. \tag{61}$$

We provide the detailed proof in Appendix A.3. Now we show that given the event $\{\bar{\mathcal{L}} \geq \mathcal{L}\}$, `PLoss` TDF is bounded from above by `PbLoss` TDF. It is clear that $\forall \Delta \in \mathbb{R}$,

$$P\left(\mathcal{L} > \Delta | \mathbf{1}_{\{\bar{\mathcal{L}} \geq \mathcal{L}\}} = 1, b_k, \pi, \pi', \nu\right) \leq P\left(\bar{\mathcal{L}} > \Delta | \mathbf{1}_{\{\bar{\mathcal{L}} \geq \mathcal{L}\}} = 1, b_k, \pi, \pi', \nu\right). \tag{62}$$

Finally, we characterize `PLoss` as follows.

Theorem 4 *(Upper and Lower bounds). Denote*

$$\theta_\alpha(\Delta) \triangleq \min\left\{1, \frac{P\left(\bar{\mathcal{L}} > \Delta | b_k, \pi, \pi', \nu\right)}{(1 - \alpha)^2} + 2\alpha - \alpha^2\right\},$$

so

$$P\left(\mathcal{L} > \Delta | b_k, \pi, \pi', \nu\right) \leq \theta_\alpha(\Delta) \quad \forall \Delta \in \mathbb{R}_{\geq 0} \tag{63}$$

and

$$P\left(\mathcal{L} \leq \Delta | b_k, \pi, \pi', \nu\right) \geq 1 - \theta_\alpha(\Delta) \quad \forall \Delta \in \mathbb{R}_{\geq 0}. \tag{64}$$

The full proof can be found in Appendix A.4. With accessible online $\theta_\alpha(\Delta)$ we are able to obtain a complete characterization of the simplification. Moreover, since $0 \leq \mathcal{L}$, setting $\Delta = 0$ in Algorithm 3 we can assess the probability to be absolute action consistent for any φ.

Algorithm 3 Online empirical characterization of the `PLoss` with `PbLoss`.

```
Input: Two candidate policies $\pi, \pi'$. Initial belief $b_k$. Samplers from $\mathbb{P}(\breve{g}_k, l, u|\mathcal{H}_{k+L}, \nu)$ and $\mathbb{P}(\breve{g}'_k, l', u'|\mathcal{H}'_{k+L}, \nu)$.
Sample $b_k$ or take the initial samples from inference. Obtain $s$ samples from $\mathbb{P}\left(z_{k+1:k+L}, z'_{k+1:k+L}|b_k, \pi, \pi'\right)$ and create two belief policy trees.
for sample pairs $(z_{k+1:k+L}, z'_{k+1:k+L})$ do
    Obtain sample $(\breve{g}_k, l, u, \breve{g}'_k, l', u')$.
    Calculate $f_{\bar{\mathcal{L}}}(\breve{g}_k, l, u, \breve{g}'_k, l', u')$ according to (52).
end for
$\{f_{\bar{\mathcal{L}}}(\breve{g}_k, l, u, \breve{g}'_k, l', u')\}$ represents the set of samples of $\bar{\mathcal{L}}$.
Output: $\forall\Delta$ empirically calculated $P\left(\bar{\mathcal{L}} > \Delta|b_k, \pi, \pi', \nu\right)$ as $\frac{\sum_{i=1}^{s}[\bar{\mathcal{L}}^i > \Delta]}{s}$.
```

5.3.3. Calculating `PLoss` offline and `PbLoss` online

One approach to obtain `PLoss` *offline* is to sample $(g_k, g'_k, \breve{g}_k, \breve{g}'_k)$ from (44) using decomposition (46). `PLoss` is then represented by $\{f_{\mathcal{L}}(g_k, g'_k, \breve{g}_k, \breve{g}'_k)\}$.

We take samples of $\mathbb{P}(z_{k+}, z'_{k+}|b_k, \pi, \pi')$ from the corresponding extended belief policy trees. To sample

$$\mathbb{P}(g_k, \breve{g}_k|\mathcal{H}_{k+L}, \nu) \quad , \quad \mathbb{P}(g'_k, \breve{g}'_k|\mathcal{H}'_{k+L}, \nu), \tag{65}$$

we use the original (not simplified) rewards calculated from the beliefs present at the belief tree (belief tree does not undergo simplification) and their simplified counterparts.

So far, we did not explain how to calculate `PbLoss` (53). One approach is to sample $(\breve{g}_k, l, u, \breve{g}'_k, l', u')$ from

$$\mathbb{P}(\breve{g}_k, l, u, \breve{g}'_k, l', u'|b_k, \pi, \pi', \nu) \tag{66}$$

and evaluate $\bar{\mathcal{L}}$ for each such sample via (52). Then, `PbLoss` is represented by $\{f_{\bar{\mathcal{L}}}(\breve{g}_k, l, u, \breve{g}'_k, l', u')\}$.

Generating samples from (66) involves marginalizing over future measurements $z_{k+} \equiv z_{k+1:k+L}$ and $z'_{k+} \equiv z'_{k+1:k+L}$. Similar to (46), the (66) decomposes to

$$\int\limits_{\substack{z_{k+}\\ z'_{k+}}} \mathbb{P}(\breve{g}_k, l, u|\mathcal{H}_{k+L}, \nu)\mathbb{P}(\breve{g}'_k, l', u'|\mathcal{H}'_{k+L}, \nu)\mathbb{P}(z_{k+}, z'_{k+}|b_k, \pi, \pi')dz_{k+}dz'_{k+} \tag{67}$$

In practice, $\mathbb{P}(z_{k+}, z'_{k+}|b_k, \pi, \pi')$ corresponds to two extended belief policy trees, starting from the same root (b_k) and having the same rule for choosing rollouts. The specific way of obtaining samples from

$$\mathbb{P}(\breve{g}_k, l, u|\mathcal{H}_{k+L}, \nu) \quad , \quad \mathbb{P}(\breve{g}'_k, l', u'|\mathcal{H}'_{k+L}, \nu) \tag{68}$$

depends on the operator ν. We summarized the proposed approach in Algorithm 3. In the next section, we elaborate on these aspects, considering a specific simplification operator.

5.4. Guarantees on simplified hypothesis based decision making

In this section we provide guarantees on the simplification portrayed by (48). We recite that the concept of `PLoss` and `PbLoss` is valid for any objective operator φ. In this section we describe a specific usage for the objective operator presented in section 5.2. Let us make a following definition

$$\breve{\Delta}^P \triangleq \left|\sum_{i=1}^{s} \mathbf{1}\{\breve{g}^i_k > \breve{g}'^i_k\} - \sum_{i=1}^{s} \mathbf{1}\{\breve{g}^i_k < \breve{g}'^i_k\}\right|. \tag{69}$$

Each not-action consistent sample decreases this margin by 2, so in order to be not-action consistent we need to satisfy

$$2 \cdot \sum_{i=1}^{s} \mathbf{1}\{\mathcal{L}^i > 0\} \geq \breve{\Delta}^P. \tag{70}$$

The following relation permits us to answer the question either or not it is possible that the order is switched due to simplification.

$$\sum_{i=1}^{s} \mathbf{1}\{\mathcal{L}^i > 0\} = s \cdot P(\mathcal{L} > 0|b_k, \pi, \pi', \nu). \tag{71}$$

The offline condition that simplification is action consistent will be

$$2sP(\mathcal{L} > 0|b_k, \pi, \pi', \nu) < \breve{\Delta}^P. \tag{72}$$

From here we can define the *online* accessible condition that simplification is action consistent as

$$2\cdot\sum_{i=1}^{s}\mathbf{1}\{\bar{\mathcal{L}}^i>0\}<\breve{\Delta}^P. \tag{73}$$

We observe that the above relation involves PbLoss from Algorithm 3. Remarkably, for analytical or stochastic bounds for sufficiently large s so (59) and (60) hold, we obtain that the condition to be action consistent is

$$2s\theta_{\alpha}(0)<\breve{\Delta}^P. \tag{74}$$

Noteworthy, similar to VaR we provide deterministic guarantees using stochastic bounds.

6. Specific simplification

In this section, we exemplify our technique on a specific simplification method. Let us recite that if the simplification regime acts according to (17), it results in uncorrelated $\breve{g}_k|b_k,\pi,z_{k+},\nu$ and $g_k|b_k,\pi,z_{k+}$. Conversely, if the simplification strategy complies to (18), the correlation is present. We start from the setting of a given belief surface and continue to the general nonparametric setting. In the following section we describe adaptive stochastic bounds in the setting of an explicitly given belief surface.

6.1. Online adaptive bounds on sample based return with a given belief surface

Let us start from the scenario in which the belief surface is explicitly given and we are interested in negative differential entropy as a belief dependent reward. Assume the belief is represented in closed form as a Gaussian mixture such that we have a deterministic update ψ_{dt} (see e.g. [25]). Since the differential entropy doesn't have a closed form solution, we are obliged to sample from the corresponding posterior belief. Assume we have n i.i.d. samples. One way to approximate the desired reward [10] is

$$I=\mathbb{E}\left[\ln(b(x))\right]\approx-\hat{\mathcal{H}}=\frac{1}{n}\sum_{i=1}^{n}\ln(b(x^i)). \tag{75}$$

We refer to (75) as a simplified reward. This estimator is unbiased as

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\ln(b(x^i))\right]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\ln(b(x^i))\right]\underbrace{=}_{\forall i\ x_i\sim b}\mathbb{E}\left[\ln(b(x))\right]=I. \tag{76}$$

Assume that b_k is a sampleable surface and

$$g_k|\mathcal{H}_{k+L}=f_{g_k}\left(\left(\mathbb{E}\left[\ln(b_\ell(x_\ell))\right]\right)_{\ell=k+1}^{k+L}\right) \tag{77}$$

$$\breve{g}_k|\mathcal{H}_{k+L},\nu=f_{g_k}\left(\left(\frac{1}{n}\sum_{i=1}^{n}\ln(b_\ell(x_\ell^i))\right)_{\ell=k+1}^{k+L}\right) \tag{78}$$

Note that $g_k|\mathcal{H}_{k+L}$ is theoretical at this point. Its distribution is Dirac delta and its variance is zero. In case we have a belief surface represented as a Gaussian mixture with M components the complexity of calculating such simplified return will be $O(n\cdot M)$, where n is the number of samples from the surface.

Using a standard Gaussian confidence interval [29] we obtain

$$P\left(|g_k-\breve{g}_k|\le z_{\alpha/2}\text{se}(n)|\mathcal{H}_{k+L},\nu\right)\approx 1-\alpha. \tag{79}$$

Adaptive stochastic bounds Let us focus on variance of the reward. Assume that f_{g_k} is of the following form.

$$f_{g_k}(\rho_{k+1:k+L})=\frac{1}{L}\sum_{\ell=k+1}^{k+L}\rho_\ell. \tag{80}$$

Denote by 1 column vector of ones. The variance $\mathbb{V}\left(\frac{1}{L}\sum_{\ell=k+1}^{k+L}\breve{\rho}_\ell\middle|b_k,\mathcal{H}_{k+L},\nu\right)$ can be written as

$$\mathbb{V}\left(\frac{1}{L}1^T\breve{\rho}_{k+1:k+L}\middle|b_k,\mathcal{H}_{k+L},\nu\right)=\frac{1^T\Sigma 1}{L^2}\le\max_i\sigma_{ii}^2. \tag{81}$$

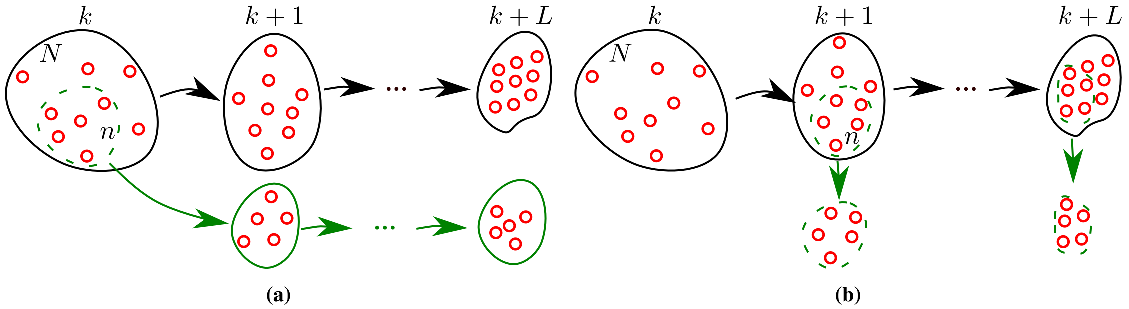


Fig. 7. Potential simplification techniques: **(a)** Choosing a subset of samples only at time k and updating the simplified belief. Such a simplification corresponds to (17).; **(b)** Choosing a subset of samples at each time ℓ and updating the original belief. Such a simplification corresponds to (18).

From now let us focus on the variance of $\breve{\rho}_\ell|b_k,\mathcal{H}_{k+L},\nu$, which is (the samples from the belief surface are i.i.d.)

$$\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n}\ln(b_\ell(x^i))|b_k,\mathcal{H}_{k+\ell},\nu\right]=\frac{1}{n}\mathbb{V}_{x\sim b_\ell}\left[\ln(b_\ell(x))|b_k,\mathcal{H}_{k+\ell},\nu\right]. \tag{82}$$

If we knew how to update incrementally $\mathbb{V}_{x\sim b}[\ln(b(x))]$, this would yield adaptive stochastic bounds. We have samples from b, so we can calculate sample variance using n samples of the belief as

$$\hat{\mathbb{V}}_{x\sim b_\ell}\left[\ln(b_\ell(x))|b_k,\mathcal{H}_{k+\ell},\nu\right]=\frac{1}{n-1}\left(\sum_{i=1}^{n}\ln^2(b_\ell(x^i))-\left(\frac{1}{n}\sum_{i=1}^{n}\ln(b_\ell(x^i))\right)^2\right). \tag{83}$$

Alternatively, using Taylor expansion similar to [15] we obtain an approximation for desired variance

$$\mathbb{V}_{x\sim b_\ell}\left[\ln(b_\ell(x))\right]=\mathbb{E}[\ln^2(b_\ell(x))]-\mathbb{E}^2[\ln(b_\ell(x))]. \tag{84}$$

Suppose we use sample variance. It has readily available incremental update using Welford's online algorithm. Suppose we have calculated $\mathrm{se}_\ell^2(n)$ and $\breve{g}_k^n=\frac{1}{L}\sum_{\ell=k+1}^{k+L}\mu_\ell^n$, now we want to tighten the bounds. We sample point $x_\ell\sim b_\ell$ from each surface $\ell=k+1:k+L$. Firstly we update the simplified return incrementally

$$\breve{g}_k^{n+1}=\frac{1}{L}\sum_{\ell=k+1}^{k+L}\left(\mu_\ell^n+\frac{\ln b(x_\ell)-\mu_\ell^n}{n}\right)=\breve{g}_k^n+\frac{1}{n}\left(\frac{1}{L}\sum_{\ell=k+1}^{k+L}\ln b(x_\ell)-\breve{g}_k^n\right). \tag{85}$$

We then update the $\mathrm{se}_\ell^2(n)$ towards $\mathrm{se}_\ell^2(n+1)$. Again the incremental update is readily available

$$(n+1)\cdot\mathrm{se}_\ell^2(n+1)=n\cdot\mathrm{se}_\ell^2(n)+(\ln b(x_\ell)-\mu_\ell^n)(\ln b(x_\ell)-\mu_\ell^{n+1}). \tag{86}$$

We will have to bookkeep μ_ℓ^n.

Unfortunately, the belief surface is not always obtainable. Moreover, not always not-simplified reward has zero variance. To these aspects we devote the next section.

6.2. Online bounds on sample based return - general setting

Suppose we are given from the inference stage a belief represented by a set of N weighted particles $b_k=\{w_k^i,x_k^i\}_{i=1}^N$. We would like to simplify planning by taking substantially less particles $\breve{b}_k=\{w_k^j,x_k^j\}_{j=1}^n$. Alternatively we subsample the original belief b_ℓ at each time index to obtain $\breve{b}_\ell=\{w_\ell^j,x_\ell^j\}_{j=1}^n$. Our simplification operator ν provides a way to choose a subset of n samples from the original N samples. For example, subsampling according to weights. We take ψ_{st} as an off-the-shelf particle filter, which produces the same number of samples as the input. The two ways of updating the belief are illustrated in Fig. 7. To present development for (25), we continue with unbiasedness assumption and take an inspiration from confidence intervals. Let us introduce the following model

$$\begin{pmatrix} g_k \\ \breve{g}_k \end{pmatrix}|\mathcal{H}_{k+L},\nu\sim\mathcal{N}\left(\begin{pmatrix}\mu\\ \mu\end{pmatrix};\begin{pmatrix}\mathrm{se}^2(N) & \mathrm{cov}\\ \mathrm{cov} & \mathrm{se}^2(n)\end{pmatrix}\right), \tag{87}$$

where se is the standard error and cov is the covariance. Online we do not have access to these quantities. The standard error depends on the number of samples N and n respectively, dwindling as the number of samples increases. We assume that each marginal is distributed around the same mean value μ (no bias).

Denote $y = g_k - \breve{g}_k$. It is known that y is a zero mean Gaussian with the following variance

$$\mathrm{var}(y) = \mathrm{se}^2(N) + \mathrm{se}^2(n) - 2\mathrm{cov}. \tag{88}$$

Let $z = \frac{y}{\sqrt{\mathrm{var}(y)}} \sim \mathcal{N}(0,1)$ and $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$, where Φ is a CDF of a standard normal variable so $\mathbb{P}(z > z_{\alpha/2}) = \alpha/2$ and

$$P\left(-z_{\alpha/2} \leq z \leq z_{\alpha/2}\right) = 1 - \alpha. \tag{89}$$

In other words

$$P\left(|y| \leq z_{\alpha/2}\sqrt{\mathrm{var}(y)} | \mathcal{H}_{k+L}, \nu\right) = 1 - \alpha. \tag{90}$$

Using the facts $\mathrm{se}(N) \leq \mathrm{se}(n)$ and $\mathrm{cov} \leq \mathrm{se}(N)\mathrm{se}(n)$ we arrive at two cases. The first case is the zero covariance ($\mathrm{cov} = 0$).

$$\mathrm{var}(y) = \mathrm{se}^2(N) + \mathrm{se}^2(n) \leq 2\mathrm{se}^2(n). \tag{91}$$

The second case is more general.

$$\mathrm{var}(y) = \mathrm{se}^2(N) + \mathrm{se}^2(n) - 2\mathrm{cov} \leq 4\mathrm{se}^2(n). \tag{92}$$

Thus, from (90) we obtain for both cases

$$P\left(|g_k - \breve{g}_k| \leq z_{\alpha/2}\sqrt{2}\mathrm{se}(n) | \mathcal{H}_{k+L}, \nu\right) \geq 1 - \alpha. \tag{93}$$

$$P\left(|g_k - \breve{g}_k| \leq z_{\alpha/2} 2\mathrm{se}(n) | \mathcal{H}_{k+L}, \nu\right) \geq 1 - \alpha. \tag{94}$$

6.2.1. Comparison to baseline methods

As a baseline we take conventional methods applying ψ_{st} once and treat the sample obtained as representative.

Sample mean Let us assume that the objective is the sample mean of the return with one sample of return per observation and $\{z_{k+1:k+L}\}_{i=1}^{s}$ samples of observations. Suppose that samples of observations are i.i.d. Let us recall that the variance of this sample mean is as follows

$$\mathbb{V}\left(\frac{1}{s}\sum_{i=1}^{s} g_k^i\right) = \frac{1}{s}\left(\mathbb{E}_{z+}[\mathrm{se}^2(z_{k+}, N)] + \mathbb{V}(\mu(z_{k+}))\right). \tag{95}$$

With analytical bounds we bound deterministically every sample g_k^i so we bound sample mean. However in case of stochastic bounds we can not bound expected value but we can use samples of simplified return instead original. Under the model (87) the expected value of the same sample mean will stay the same however the variance of the estimator will grow. In this case we will accelerate decision making but we will pay with increasing variance

$$\mathbb{V}\left(\frac{1}{s}\sum_{i=1}^{s} \breve{g}_k^i\right) = \frac{1}{s}\left(\mathbb{E}_{z+}[\mathrm{se}^2(z_{k+}, n)] + \mathbb{V}(\mu(z_{k+}))\right). \tag{96}$$

We believe that this is an interesting relation.

Sample Value at Risk When the objective is sample approximation of VaR with analytical bounds as well with stochastic bounds, we can bound only the theoretical VaR.

6.2.2. Estimation of the variance

As we do not have access to $\mathrm{se}(n)$ in (93) and (94), it has to be estimated. The simplest way to do that is to repeatedly sample simplified returns m times from one of (17), (18), (20) depending on the simplification type. Note that a possible bias of the particle filter and the estimation of standard error make (90) only asymptotically correct. However, when dealing with a sufficient amount of samples N and n, these deviations from (87) are negligible. Even with repeated re-sampling we will reduce computational complexity, as we analyze in Section 7. The bounds for both simplification methods are

$$(\text{no cov})\; u = \breve{g}_k + z_{\alpha/2}\sqrt{2}\hat{\mathrm{se}}_m \quad l = \breve{g}_k - z_{\alpha/2}\sqrt{2}\hat{\mathrm{se}}_m, \tag{97}$$

$$(\text{cov})\; u = \breve{g}_k + z_{\alpha/2} 2\hat{\mathrm{se}}_m \quad l = \breve{g}_k - z_{\alpha/2} 2\hat{\mathrm{se}}_m. \tag{98}$$

Moreover, since we recalculate the simplified reward m times, we could improve the final simplified return. In this case, we take the average of the samples of the simplified return given the history (prior belief, candidate policy, and the realization of the observations) as a final simplified return for this history

$$\breve{g}_k = \frac{1}{m}\sum_{j=1}^{m} \breve{g}_k^j \tag{99}$$

and the model becomes

$$g_k|\mathcal{H}_{k+L}, \nu \sim \mathcal{N}(\breve{g}_k, \mathrm{se}^2(n)). \tag{100}$$

The bounds in this case are

$$u = \breve{g}_k + z_{\alpha/2}\hat{\mathrm{se}}_m \quad l = \breve{g}_k - z_{\alpha/2}\hat{\mathrm{se}}_m. \tag{101}$$

These bounds asymptotically hold with probability at least $1-\alpha$.

Using $\breve{g}_k^i = \frac{1}{m}\sum_{j=1}^{m} \breve{g}_k^j$, where $i = 1:s$ we will obtain

$$\mathbb{V}\left(\frac{1}{s}\sum_{i=1}^{s} \breve{g}_k^i\right) = \frac{1}{s}\left(\frac{1}{m}\mathbb{E}_{z+}[\mathrm{se}^2(z_{k+}, n)] + \mathbb{V}(\mu(z_{k+}))\right). \tag{102}$$

6.3. Implementation details and computational complexity

Now we describe steps in building an extended belief tree which is common for all our simulations. First, we need to construct an extended belief tree appropriate to a given candidate policy (see Fig. 1); alternatively, if the objective operator is mounted on the joint distribution of a pair of returns given a pair of policies, as in (43), we shall construct a pair of coupled belief trees. Second, we shall apply the simplification and calculate simplified returns and bounds. In all simulations with nonparametric beliefs we choose ψ_{st} to be an off-the-shelf particle filter with low-variance re-sampling [42]. The entire belief update process complexity is $\mathcal{O}(N)$. Since the extended belief tree does not undergo simplification, it is common to the original and simplified problems.

In practice, the marginal likelihoods $\mathbb{P}(z_{k+}|b_k, \pi)$ and $\mathbb{P}(z'_{k+}|b_k, \pi')$ as in section 4 or the mutual likelihood of the observations $\mathbb{P}(z_{k+}, z'_{k+}|b_k, \pi, \pi')$ as in section 5 (see (41)) correspond to two extended belief policy trees, starting from the same root (b_k) and having the same rule for choosing rollouts.

Below we discuss the construction of the extended belief tree. Let N be a number of samples of the posterior belief. We choose the samples of the belief for creating the observations heuristically according to the following scheme. Let $n_z^{(\ell)}$ be number of observations generated by each belief at level ℓ of the tree. We specify $n_z^{(1)}$ (the number of observations generated by b_k) and the dwindle factor c. Starting from $\ell = 2$ the number of observations generated by each belief on level ℓ in the tree is calculated as $n_z^{(\ell)} = \max\{1, \lfloor \frac{n_z^{(1)}}{(\ell-1)\cdot c} \rfloor\}$. In the setting of nonparametric beliefs, we sample states for the observations from resampled posterior with Fisher-Yates shuffling (with early termination) [20]. This algorithm is $\mathcal{O}(N)$ for initialization, plus $\mathcal{O}(n_z^{(\ell)})$ for random shuffling.

In our extended belief policy tree, there may be many beliefs stemming from an observation. Denote this number by n_b. In the setting of nonparametric beliefs represented by the particles, the complexity of constructing the tree is

$$\mathcal{O}(N)\sum_{\ell=1}^{L-1}\prod_{i=1}^{\ell} n_b n_z^{(i)}. \tag{103}$$

At each level of the tree beside the bottom, we must apply a particle filter number of times equal to the total number of the beliefs at the next level, which is $\prod_{i=1}^{\ell} n_b n_z^{(i)}$ at level ℓ. Also, we need to subsample observations at the current level. Since the number of beliefs at the next level is not smaller than at the current level, and the subsampler and particle filter complexity is linear in N, we are left with (103). Let us mention that sampling from the belief and application of particle filters on each level can be done in parallel.

Now we analyze the speedup in running time as a result of simplification in the setting of nonparametric beliefs. As a momentary reward, we take the differential entropy estimator from [4]. This selection makes the complexity of calculating the momentary reward to be $\mathcal{O}(N^2)$. For the bounds calculation depending on the simplification method we need to apply particle filter with n samples (17) or with N (18), (20)) samples, L times for each return. Since its complexity is linear in the number of samples, the expected speedup is governed only by the immediate reward and bounds calculation. The speedup is approximately

$$\frac{N^2}{n^2 \cdot m}. \tag{104}$$

This acceleration has a rather intuitive explanation. Since we are comparing running time of exactly the same function the ratio gives approximately exact speedup. The empirical behavior of estimator is as $\Theta(N^2)$. We obtained this speedup in all our simulations.

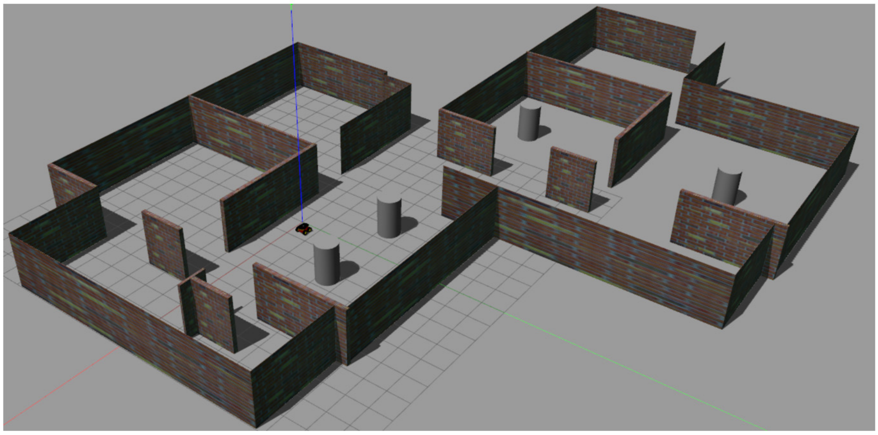


Fig. 8. Gazebo simulated environment. Each square in the map corresponds to 1 × 1 meters square.

7. Simulations and results - autonomous navigation with light beacons

In this section, we demonstrate our findings. In the center of our focus are the risk-averse operators, and in all cases, simplification yields a significant speedup without sacrificing the quality of the solution. We consider the setting of marginal distributions over returns per candidate policy, as in Section 4, and the joint distribution over the returns given a pair of policies, as in Section 5. In both settings, we consider the problem of autonomous navigation to a pre-defined goal in an environment with known beacons.

We start from marginals of the return in the setting of a given belief surface and then proceed to the general domain of nonparametric beliefs and an inaccessible belief surface. We then continue to the joint distribution of a pair of the returns given two policies and `PLoss` and `PbLoss` simulations. Finally, we report the technical characteristics of computers used in simulations in Appendix B.

7.1. *Marginal return distributions corresponding to candidate polices*

For our simulations, we utilize a localization problem with a known map created in the Gazebo simulator [23]. We used a Pioneer 3-AT robot to perform navigation to the goal as illustrated in Fig. 8.

7.1.1. A given belief surface

We start by exemplifying the adaptive bounds from Section 6.1 in the setting described in [25]. We do not assume that we know which beacon generated an observation in this setting. Instead, we maintain a hypothesis about each possible configuration of the beacons creating the observation. Such an approach is realistic since, in the planning phase, the robot considers the future observations identically as in the inference phase when the real observation is obtained. It results in belief being a Gaussian Mixture Model (GMM), where each component corresponds to a possible configuration of data association. The weights of the components are probabilities of the hypothesis that the Gaussian component is an actual configuration. No analytical expression exists for differential entropy when the belief surface is GMM, so we are obliged to sample and want as few samples as possible. For simplicity, we consider only two possible paths to the goal, as shown in Fig. 9. We take the belief b_k over of the robot's 2D location as a Gaussian $\mathcal{N}(\mu, \Sigma)$, where $\mu = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$. The belief update is deterministic ψ_{dt}. Each beacon is visible on the maximum radius of 3 meters. Leave out the recursive setting instead of smoothing, we strictly follow the theory presented at [25]. Let us restate that, in planning, when considering possible observations, we do not assume that we know from which beacon they arrived; instead, we maintain hypotheses regarding each possible configuration of the beacons to yield an observation. We denote each such data association by $\beta_{k+\ell}$.

In this study we utilize the following motion model T.

$$x_{k+1} = x_k + a_k + \|a_k\| \cdot w_k \quad w_k \sim \mathcal{N}(0, \Sigma_w), \tag{105}$$

where $x \in \mathbb{R}^2$, $a \in \mathbb{R}^2$, $\Sigma_w = w \cdot I$ (w is a given parameter) and action $a_k \in \mathcal{A}$. The $\mathcal{A}$ is the set of action sequences with actions of variable length. Each visible beacon b produces the observation according to the following model $z_i \sim \mathcal{N}(\|x - x^b\|_2^2, \Sigma_v)$, where $\Sigma_v = v \cdot I$. We selected the following parameters $w = 0.5$ and $v = 0.005$. Overall observation is the concatenation of the observations received from all seen beacons. Let us denote by M the given map with its beacons. For simulating an observation for planning we sample state $x_{k+\ell}$ from the belief propagated with an action. We use the simplest

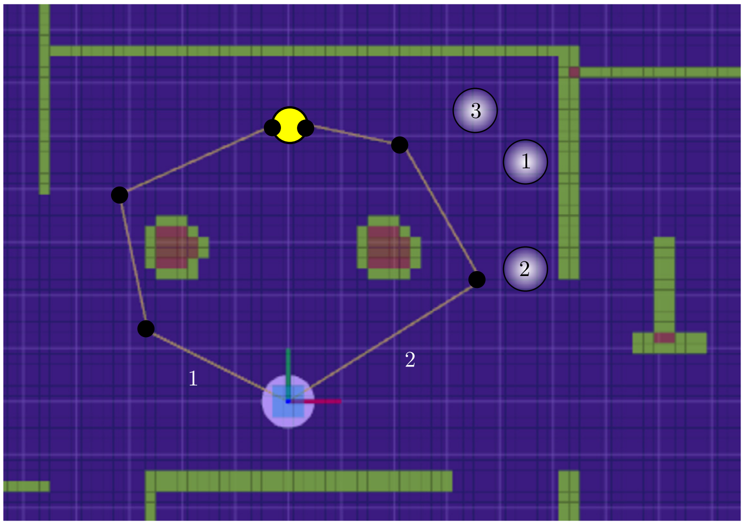


Fig. 9. Gazebo simulated scenario where the belief surface is explicitly given. The white numbers enumerate the paths, the black numbers enumerate the light beacons, the yellow circle is the goal, the black dots are the spots there the observations are taken by the robot, the purple circle is the initial belief b_k.

model for $\beta_{k+\ell}$ being $P_\beta(\beta_{k+\ell}(i)=1|x_{k+\ell})=\mathbf{1}\{\|x_{k+\ell}-x_i^b\|\leq r\}$, where the index i goes from 1 until the number of beacons in the map. According to this model, given the state and the map, each beacon is deterministically seen or not. Once we obtained the configuration of the seen beacons $\beta_{k+\ell}$; we sample the observation from the following model as follows. We define subsequence $\beta_{k+\ell}(i_j)$ such that the index j pull in ascending order the indexes of i where $\beta_{k+\ell}(i_j)$ equal 1.

$$\mathbb{P}_Z(z_{k+\ell}|x_{k+\ell},M)=\prod_{j=1}^{n_{k+\ell}(x_{k+\ell})}\mathbb{P}(z^j_{k+\ell}|x_{k+\ell},M)=\prod_{j=1}^{n_{k+\ell}(x_{k+\ell})}\mathbb{P}(z^j_{k+\ell}|x_{k+\ell},x^b_{i_j}), \tag{106}$$

where $n_{k+\ell}(x_{k+\ell})$ is the number of beacons seen from the state $x_{k+\ell}$. If no beacon is seen there is no observation received ($n_{k+\ell}(x_{k+\ell})=0$).

For belief update, however, we do not assume that we know this configuration $\beta_{k+\ell}$. The belief in each time instant is a Gaussian Mixture. To update the belief, we propagate each gaussian with an action using standard Kalman filter [42]. To update propagated Gaussian with an observation, we do not assume we know from which beacons this observation is received. This result to Gaussian mixture obtained from each propagated Gaussian, where each Gaussian in the mixture corresponds to the beacons configuration, which can render such an observation. We utilize unscented Kalman filter [42] to update each propagated Gaussian with the observation and a realization of the $\beta_{k+\ell}$. The weight of Gaussian corresponds to the probability that such beacons configuration resulted in the obtained observation. The above requires to model visibility of the beacon given the state from propagated Gaussian. Since $\beta_{k+\ell}$ is a discrete random variable, we normalize when all the above probabilities are computed. For an in-depth discussion, please refer to [25].

We sample 500 samples from propagated belief and set the parameter visibility radius as follows $r=3$.

Let us recall that in this setting the original return is the theoretical differential entropy over the belief surface which is out of the reach. We want to set the number of samples from the belief surface n as small as possible to decide which path out of two brings less uncertainty. We aim to choose the path maximizing uncertainty criterion, which we define as follows

$$\varphi\Big(\mathbb{P}\left(\breve{\rho}_{k+1:k+L}|b_k,\pi_{k:k+L-1},\nu\right),\breve{g}_k\Big)=\text{VaR}_\beta(\breve{g}^I_k|b_k,\pi,\nu), \tag{107}$$

where

$$\breve{g}^I_k=I(b_{k+1:k+L}|b_k,\pi)=\frac{1}{L}\sum_{\ell=k+1}^{k+L}\left(\frac{1}{n}\sum_{i=1}^{n}\ln(b_\ell(x^i_\ell))\right). \tag{108}$$

We set $\alpha=0.05$ such that $z_{\alpha/2}=1.96$, and $\beta=0.3$, overall number of observations in the belief tree is 500. Note that the condition $\alpha\cdot(2-\alpha)\leq\beta\leq 1-\alpha$ is fulfilled. We start from initial $n=5$ and add one sample for each immediate reward in the belief tree until there is no overlap between the intervals. In this simulation the adaptation yielded not overlapping deterministic bounds on VaR when $n=28$ and the second path was chosen as optimal. The interval for the first action sequence was $[-3.77,-2.76]$ and for the second sequence was $[-2.74,-1.52]$.

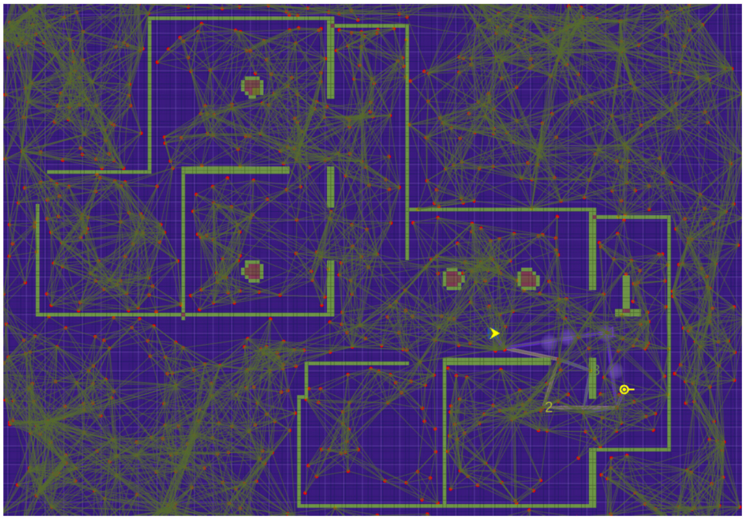


Fig. 10. Diverse short paths. The current robot position denoted yellow arrow-head and the goal marked by yellow circle. Candidate paths are enumerated. Transparent silver spheres are the light beacons.

Table 2
Running times for each of the three action sequences for $N = 2000$ and $n = 100$.

	a_1	a_2	a_3
g_k time [sec]	30178	23858	20664
$\breve{g}_k$ time [sec]	85	64	57
l, u time [sec]	4084	3255	2805
speedup	7.24	7.18	7.22

7.1.2. Not accessible belief surface in the setting of nonparametric belief

In this section we consider non parametric beliefs represented by particles. The belief surface in this setting is out of the reach. We remain in recursive formulation of the two dimensional continuous state space. For updating the belief we use particle filter with low variance resampler [42]. We begin by building the Probabilistic Road Map (PRM) using OMPL library. After the map is built we apply the Diverse Short Path algorithm [43]. The resulting paths from robot to goal are visualized at Fig. 10. These paths constitute our action space. We normalize by path length L to obtain fair comparison. To accelerate the calculations we apply Algorithm 1.

We use same motion model as in the previous section. However, the observation model varies. In this scenario there are four beacons, but each beacon is always seen and produces the observation according to the following observation model O. $z^i_{k+\ell} \sim \mathcal{N}(x_{k+\ell}, \Sigma_v(x_{k+\ell}))$ for $i = 1, \ldots 4$, where the spatially-varying covariance matrix is

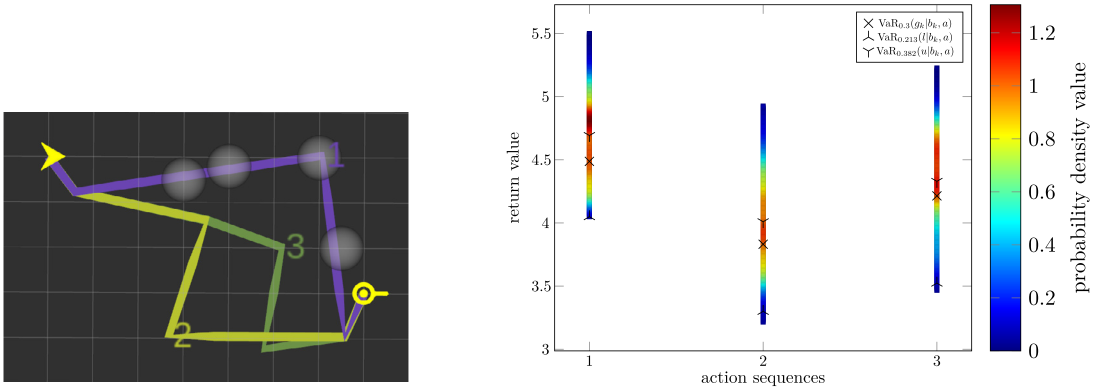
$$\Sigma_v(x) = v(x) \cdot I, \quad v(x) = w \cdot \|x - x^b_i\|^2_2, \tag{109}$$

where x^b_i is the location of the light beacon number i. The noise parameter w is taken from the motion model. In contrast to the previous section, we assume that the data association is solved. Overall observation received from all the beacons has the following probability density function

$$\mathbb{P}_Z(z|x) = \prod_{i=1}^{4} \mathbb{P}(z^i_{k+\ell}|x_{k+\ell}, x^b_i). \tag{110}$$

Without losing generality, we assume b_k at planning time is uniformly distributed in a unit square, such that the differential entropy is zero. In the naive approach to evaluate motion and observation models $\mathbb{P}_T$ and $\mathbb{P}_Z$ we need to inverse covariance matrix. Of course we can speedup this calculation by caching values of distribution of parameters or even value of evaluated motion model. However, this is out of scope of current discussion. We utilize an off-the-shelf Julia language implementation of Gaussian distribution. In general evaluation of the model can be extremely costly as described in [14].

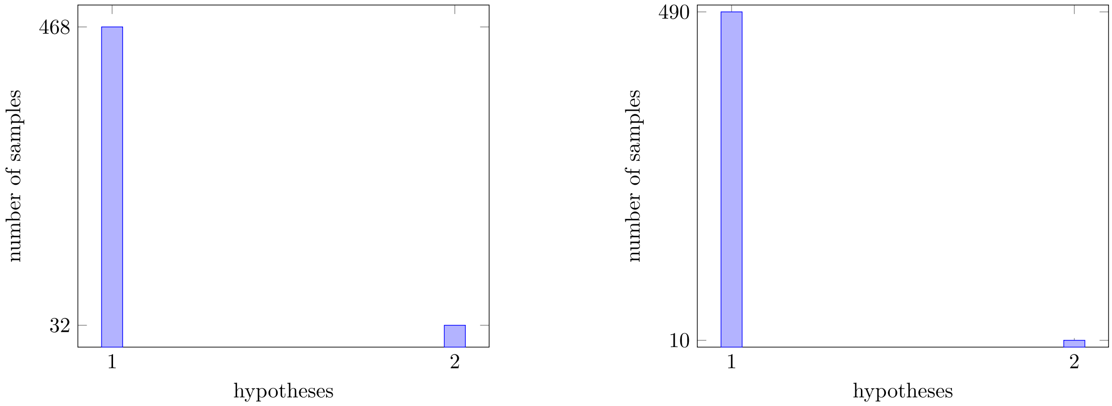
We present results in Fig. 11b. Note that we chose $N = 2000$ guided by the works such as [4] and [11]. In this setting we obtain 8 times speedup according to (104), corroborated by Table 2, while the same optimal action is chosen as without simplification.



(a)

(b)

Fig. 11. Simplified risk aware decision making using VaR. **(a)** Three candidate paths and four light beacons. **(b)** Results of simplified planning under uncertainty with $\beta = 0.3$. The first path is true optimal path due to its proximity to the beacons. In this scenario $w = 0.01$ and $v = 0.001$. The optimal path selected by solving the simplified problem is first and the relative error is zero whereas the online bound on the relative error is 0.07. In this simulation $N = 2000$ and $n = 100$. For calculation of the standard error for the bounds we recalculate the simplified reward $m = 50$ times.



(a) Original hypothesis based decision making

(b) Simplified hypothesis based decision making

Fig. 12. Comparison of the hypotheses for action sequence one and two. The total number of samples is 500.

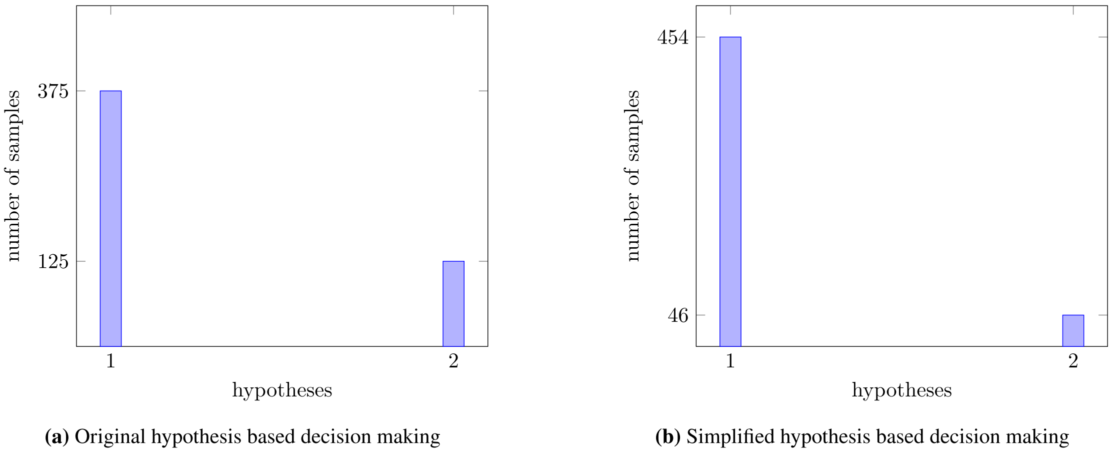
7.2. Joint distribution of the rewards corresponding to two candidate policies

In this section we exemplify simplified hypothesis based decision making outlined in section 5.2. We utilize the concept of PLoss to provide guarantees considering the specific objective from (48). Further we delve into PLoss to show the complete characterization of the simplification for any objective operator φ.

7.2.1. Simplified hypothesis based decision making

Let us focus on the previous scenario shown in Fig. 10. Our setting is as in previous section $N = 2000$, $n = 100$, $m = 50$. We start by comparing the first path to the second and show the results in Fig. 12. The first hypothesis is that the first action sequence is better and the second hypothesis is that the second action sequence is better. Remarkably, we observe that the simplification actually improves the decision making since more samples fall into the first hypothesis, and as we below show the first hypothesis is indeed optimal. We now utilize PbLoss at $\Delta = 0$ to provide deterministic guarantees. Continuing the discussion on Fig. 12 we obtain $\breve{\Delta}^P = 480$. The sample approximation $P(\mathcal{L} > 0|b_k, \pi, \pi', \nu) = 0.166$, such that the offline condition (72) is met $44 < 480$. The online bound on PLoss TDF at $\Delta = 0$ is $\theta_\alpha(0) = 0.47$ such that the online condition (74) is also met as $472 < 480$. Therefore, we can guarantee deterministically that the actions trend is preserved as a result of the simplification.

According to Algorithm 2 we shall also compare the first and the third paths. We present the comparison in Fig. 13. In this experiment we obtain $\breve{\Delta}^P = 408$, $P(\mathcal{L} > 0|b_k, \pi, \pi', \nu) = 0.166$, $\theta_\alpha(0) = 0.83$, such that the offline condition is fulfilled while the online condition is violated. To conclude, we were able to provide online guarantees that the simplification was action consistent when we compared the first and the second path. In some cases, as the comparison of the first and the third path, we cannot guarantee action consistency. Meaning, in this case, there is room to take more samples. It is even more interesting to utilize an incremental approach and develop adaptive stochastic bounds in the setting of nonparametric



(a) Original hypothesis based decision making

(b) Simplified hypothesis based decision making

Fig. 13. Comparison of the hypotheses for action sequence one and three. The total number of samples is 500.

beliefs. This is, however, out of the scope of this paper, and we leave it to further research. Alternatively, the analytical adaptive bounds from [39] can be used.

7.2.2. Probabilistic loss

We believe that the concept of PLoss and PbLoss can provide much more than showed in the previous section. PLoss characterizes the simplification in a complete manner such that it is possible that one can define probabilistic more lenient action consistency on top of PLoss. Thus, we devote this section to experiments with PLoss and PbLoss. In addition we show time acceleration speedup of the calculation of belief dependent rewards in the belief tree as a result of simplification. Further we discuss empirical action consistency for *any* objective operator φ.

We exemplify our method on the problem of autonomous navigation to a goal with light beacons, which can be used for localization. In all our simulations in this section, the return g_k is a cumulative reward. In this study, the simplification conforms to (17). As the action space we take the space of motion primitives. Moreover, let us emphasize we do not average the simplified rewards taken for the approximation of standard error since we aim to examine general behavior and standard error possibly can be estimated without resampling of the returns. The bounds are calculated according to (97).

For simplicity, assume we have a linear motion model T, where $x \in \mathbb{R}^2$ as well as $a \in \mathbb{R}^2$

$$x_{k+1} = x_k + a_k + w_k \quad w_k \sim \mathcal{N}(0, \Sigma_w), \tag{111}$$

where $\Sigma_w = w \cdot I$ (w is a given parameter) and action $a_k \in \mathcal{A}$, and where the action space $\mathcal{A}$ is the space of motion primitives of unit length.

We consider next probabilistic and absolute action consistency description using PLoss offline and PbLoss online. We say that the action consistency is probabilistic if the probability that a pair of samples of the return will not preserve the trend with respect to a pair of actions due to simplification is larger than zero. Remarkably, the analysis below is valid for any objective operator φ.

Characterizing probabilistic action consistency The observation model O is as follows, $z \sim \mathcal{N}(x, \Sigma_v(x))$, where the spatially-varying covariance matrix is

$$\Sigma_v(x) = v(x) \cdot I, \quad v(x) = w \cdot \min\{1, \|x - x^*\|_2^2\}, \tag{112}$$

where x^* is the location of the light beacon closest to x. The noise has a constant variance w. Without losing generality, we assume b_k at planning time is uniformly distributed in a unit square. We set $L = 12$ and compare two action sequences: $a_{k+1:k+12}$ is six times $(1, 0)^T$ and after that six times $(0, 1)^T$. In the action sequence $a'_{k+1:k+12}$ we switched the order of actions such that the robot performs six times $(0, 1)^T$ and after that six times $(1, 0)^T$.

One realization of a possible future in terms of measurements and corresponding posterior beliefs is illustrated in Fig. 14. It is clearly seen that proximity to a beacon improves localization. Note, the robot is always able to avoid a dead reckoning scenario as it always gets an observation from the closest beacon. We hope that this setting conveys a real world scenario where an ambulating robot is equipped with long and short range sensors. The close range sensors are activated when the robot is inside a unit circle around the beacon. When the robot is outside a unit circle from the closest beacon, the beacon is detectable only by the long range sensors, which are less sensitive. We present results of the simplification for $w = 0.1$, $N = 1500$, $m = 50$, $\alpha = 0.01$, $z_{\alpha/2} = 2.56$, and the total number of observations is 500. For each sample of $z_{k+1:k+L}$, we sampled $b_{k+1:k+L}$ once. As we see in the left part of Fig. 15 we gained speedup as expected (104) for $n = 175$. We show measurements of all running times in our simulations in Table 4.

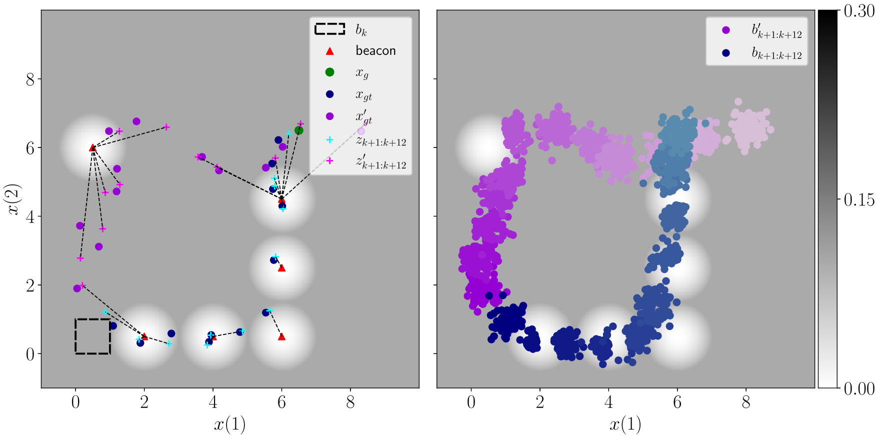


Fig. 14. Results for scenario 1 - probabilistic action consistency: Illustration of one realization of the future in a simulated scenario considering two possible action sequences. We start from b_k represented by samples uniformly distributed on a unit square. We demonstrated two sequences of observations alongside ground truth state samples, and the closest beacons produced these observations from the left. From the right, we plotted two sequences of the beliefs produced by these two histories. We show 100 most probable samples of each belief.

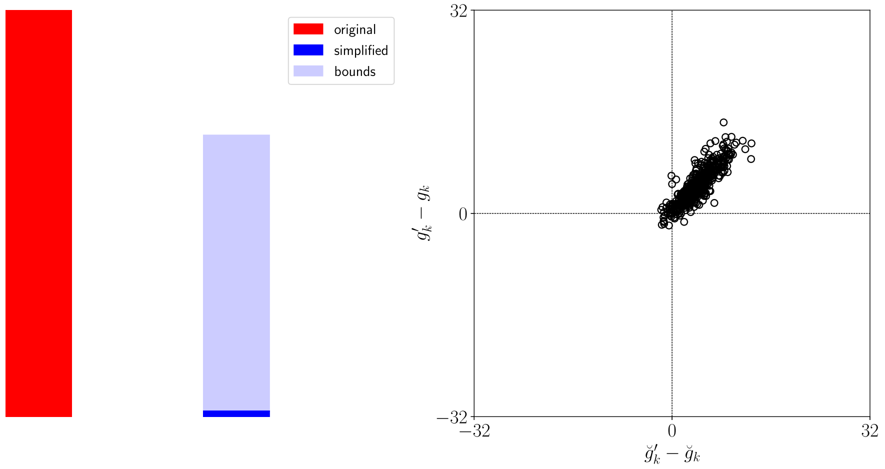


Fig. 15. Results for scenario 1 - probabilistic action consistency: (left) Demonstration of runtimes of the total number of the returns for a given extended belief tree where $N = 1500$ and $n = 175$. Note that this illustration agrees with (104); (right) action consistency of the samples of the return.

From these samples of the returns and bounds, we build PLoss and PbLoss in Fig. 19. In the right part of Fig. 15 quadrants II and IV, we observe samples that are not action consistent. To assess performance we need to choose some representative Δ. Since online we have access exclusively to the simplified problem, let us choose $\breve{\Delta}^* = |\mathbb{E}[\breve{g}_k|b_k, \pi, \nu] - \mathbb{E}[\breve{g}'_k|b_k, \pi', \nu]|$ and $\Delta = 0.5\breve{\Delta}^*$. Note that under our model in average the sample mean is not influenced by the lowering the number of samples of the reward. Only the variance of sample mean is increased. Moreover we assume that the distributions are without gaps such that the expected value of the return is some sample with probability density function larger than zero. Table 3 quantifies online characterization against offline PLoss TDF.

We showed an illustration of this scenario in Fig. 14. In Fig. 16, we demonstrated scatter plots that show samples of the simplified and original returns' differences. We identify that with decreasing n, more samples are not action consistent. This phenomenon is corroborated by the histograms of $\mathcal{L}$ in Fig. 17.

Let us focus on $n = 175$ in Fig. 18; *online* we can conclude that probability that loss incurred by this simplification will be greater than $\breve{\Delta}^*$ is at most 0.11, while actual $P(\mathcal{L} > \breve{\Delta}^*|\cdot)$ is 0.0. Similarly, the probability for loss incurred by this simplification to be greater than $0.5\breve{\Delta}^*$ is at most 0.33, while actual $P(\mathcal{L} > 0.5\breve{\Delta}^*|\cdot)$ is 0.0. In this scenario, the simplification is not absolute action consistent; it means variability described by (87) is sufficient to switch the order of the returns and incur loss Δ at some sampled realization.

Furthermore, our bounds depend on variance ($\mathrm{se}^2(n)$) of the sample approximation of the reward (97), which, according to (87) does not depend on Δ. Hence, as Δ decreases towards zero, the contribution of variance versus the difference between simplified returns grows for any realization of $\bar{\mathcal{L}}$. Therefore, PbLoss departs from PLoss as Δ decreases. We observe this behavior in Fig. 18. Moreover, with the diminishing number of samples, this effect is amplified, as demonstrated

Table 3
Results for scenario 1 - probabilistic action consistency: Online characterization for $N = 1500$, $\alpha = 0.01$, $z_{\alpha/2} = 2.56$.

n	$P(\mathcal{L} > 0.5\breve{\Delta}^*\vert\cdot)$	$\theta_\alpha(0.5\breve{\Delta}^*)$	$\breve{\Delta}^*$	$P(\mathcal{L} > \breve{\Delta}^*\vert\cdot)$	$\theta_\alpha(\breve{\Delta}^*)$
175	0.0	0.33	4.14	0.0	0.11
150	0.01	0.43	4.04	0.0	0.17
125	0.01	0.43	4.21	0.0	0.2
100	0.0	0.56	4.08	0.0	0.29
75	0.01	0.64	4.01	0.0	0.39
50	0.02	0.83	3.72	0.01	0.63
25	0.07	1.0	3.34	0.03	0.94

Table 4
Results for scenario 1 - probabilistic action consistency: run times for $N = 1500$.

	$n = 175$	$n = 150$	$n = 125$	$n = 100$	$n = 75$	$n = 50$	$n = 25$
g_k time [sec]	104957	69658	95651	69713	68584	96354	66513
$\breve{g}_k$ and l, u time [sec]	72694	34842	33759	15498	8293	5589	969
$\breve{g}_k$ time [sec]	1454	661	669	298	172	119	14
l, u time [sec]	71240	34181	33090	15200	8121	5469	955

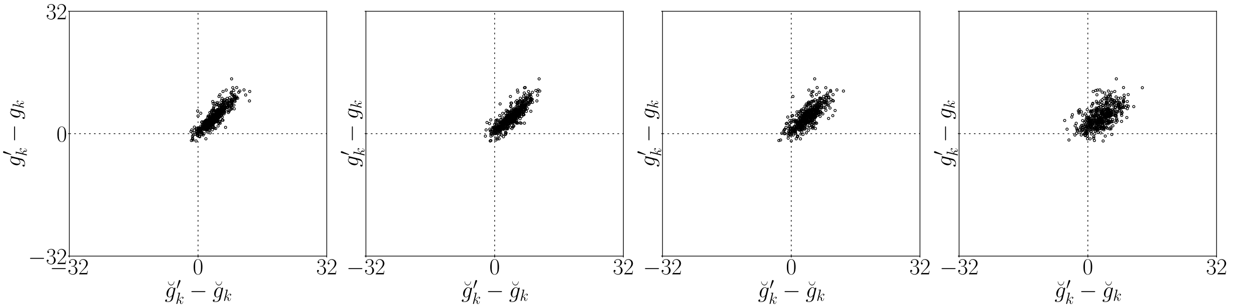


Fig. 16. Results for scenario 1 - probabilistic action consistency: We demonstrate from the left to the right action consistency of the samples of the returns for $n = 175, n = 125, n = 75, n = 25$, whereas $N = 1500$. As we see, samples violating action consistency are present at all graphs.

in Fig. 18, due to growing variance (87). Remarkably, when samples of original returns are more distinct, the effect of variance is nullified. In such a setting, our characterization is incredibly precise, see Fig. 24.

Thus, the behavior of the `PbLoss` is more conservative in more delicate scenarios, where two candidate policies are close to each other in terms of returns. Importantly, for significantly different policies, `PbLoss` becomes tighter to `PLoss`. This brings us to the next section.

Revealing empirical absolute action consistency In this scenario we modified the noise in the observation model as such $v(x) = w \cdot \|x - x^*\|_2^2$. In addition we removed one beacon on the way of the second action sequence. We remain with $w = 0.1$, $m = 50$, $\alpha = 0.01$, $z_{\alpha/2} = 2.56$ and set $N = 1000$. In this scenario the returns of two action sequences are much more distant. The samples in the right segment of Fig. 21 are more distant from the origin than in Fig. 15. The characterization is shown in Table 5. Therefore, the simplification is empirically absolute action consistent. As we see from the Table 5, observing $\theta_\alpha(\Delta = 0.0)$ we are able to identify online that for $n = 100$ and $n = 75$, probability to receive samples of the returns violating action consistency is at most 0.03, while $P(\mathcal{L} > 0.0|\cdot)$ is 0.0.

Here the covariance matrix of the observation model is

$$\Sigma_v(x) = v(x) \cdot I, \quad v(x) = w \cdot \|x - x^*\|_2^2. \tag{113}$$

We demonstrated this scenario in Fig. 20. As we can see in Fig. 22, the clouds of samples are farther from the origin than in the previous scenario. Therefore, two action sequences are more distant. In this case, the simplification is empirically absolute action consistent, as we observe in the histograms of $\mathcal{L}$ in Fig. 23 and empirical characterization shown in Fig. 24. We report run times for two scenarios in Table 4 and Table 6, respectively.

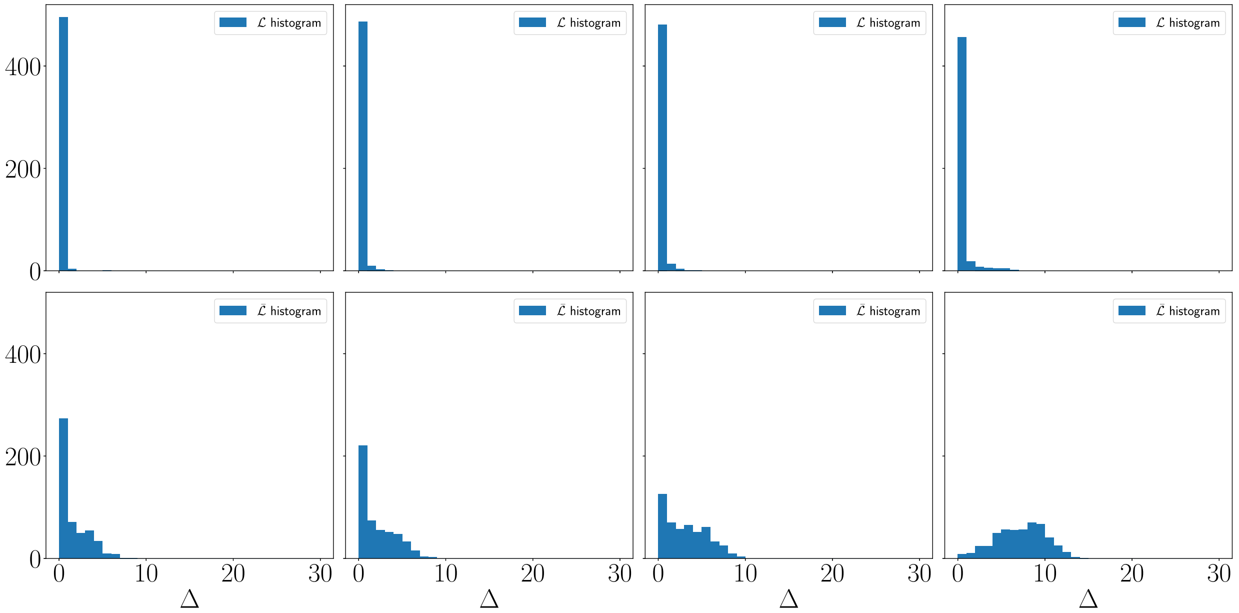


Fig. 17. Results for scenario 1 - probabilistic action consistency: Histograms of `PLoss` and `PbLoss` for $N = 1500$, $\alpha = 0.01$, $z_{\alpha/2} = 2.56$, bin width is 1.0; from the left to the right $n = 175$, $n = 125$, $n = 75$, $n = 25$.

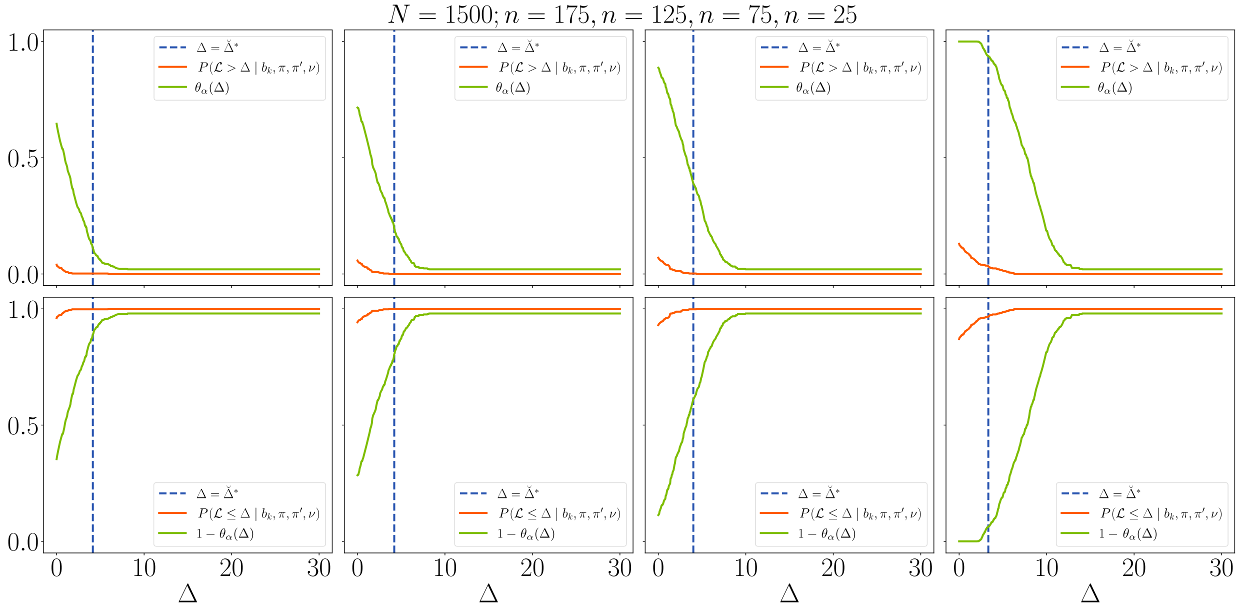


Fig. 18. Results for scenario 1 - probabilistic action consistency: Empirical characterization for $N = 1500$, $\alpha = 0.01$, $z_{\alpha/2} = 2.56$, evaluated in a grid with intervals 0.001; from the left to the right $n = 175$, $n = 125$, $n = 75$, $n = 25$.

Table 5
Results for scenario 2 - empirical absolute action consistency: Online characterization for $N = 1000$, $\alpha = 0.01$, $z_{\alpha/2} = 2.56$.

n	$P(\mathcal{L} > 0.0 \vert \cdot)$	$\theta_\alpha(\Delta = 0.0)$	$\check{\Delta}^*$	$P(\mathcal{L} > \check{\Delta}^* \vert \cdot)$	$\theta_\alpha(\check{\Delta}^*)$
100	0.0	0.03	17.54	0.0	0.02
75	0.0	0.03	17.14	0.0	0.02
50	0.0	0.06	16.65	0.0	0.02
25	0.0	0.19	15.27	0.0	0.02

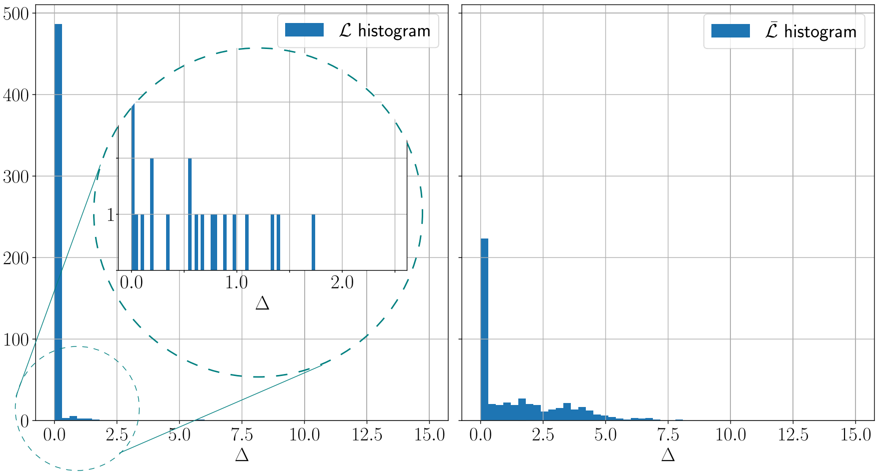


Fig. 19. Results for scenario 1 - probabilistic action consistency: Histograms of `PLoss` and `PbLoss` for $N = 1500$, $n = 175$, $\alpha = 0.01$, $z_{\alpha/2} = 2.56$ (bin width is 0.3, in zoom-in, bin width is 0.03).

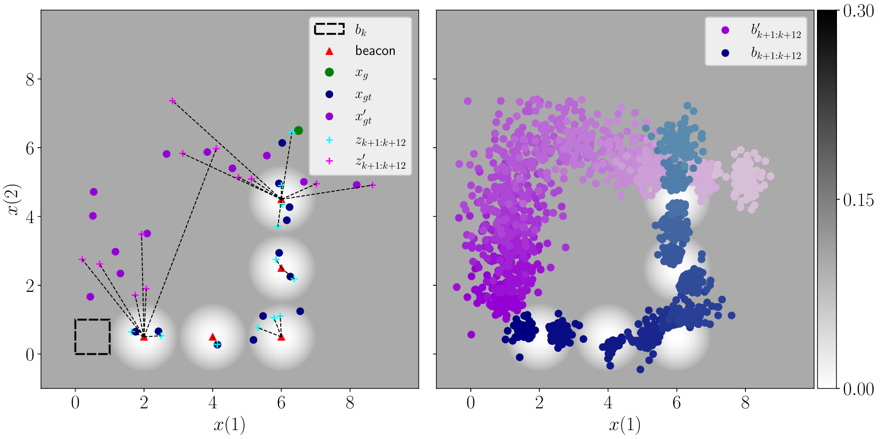


Fig. 20. Results for scenario 2 - empirical absolute action consistency: Illustration of one realization of the future in a simulated scenario considering two possible action sequences. We start from b_k represented by samples uniformly distributed on a unit square. We demonstrated two sequences of observations alongside ground truth state samples, and the closest beacons produced these observations from the left. From the right, we plotted two sequences of the beliefs produced by these two histories. We show 100 most probable samples of each belief.

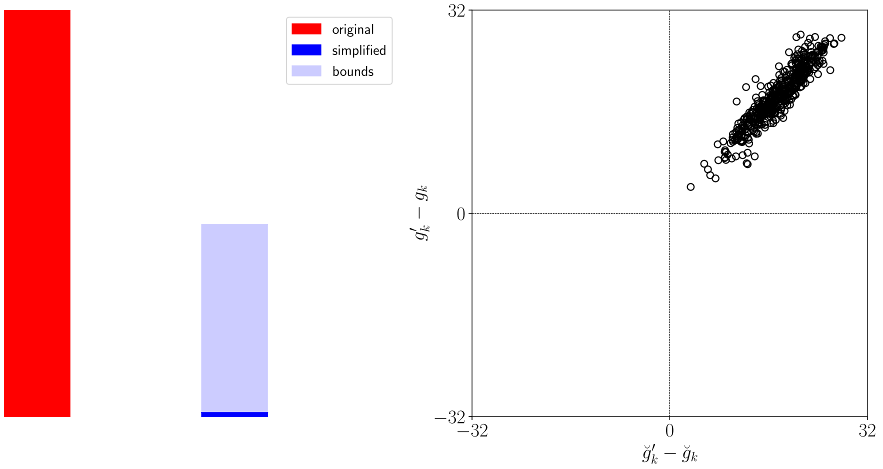


Fig. 21. Results for scenario 2 - empirical absolute action consistency: (left) Demonstration of runtimes of the total number of the returns for a given extended belief tree where $N = 1000$ and $n = 100$. Note that this illustration agrees with (104); (right) action consistency of the samples of the return.

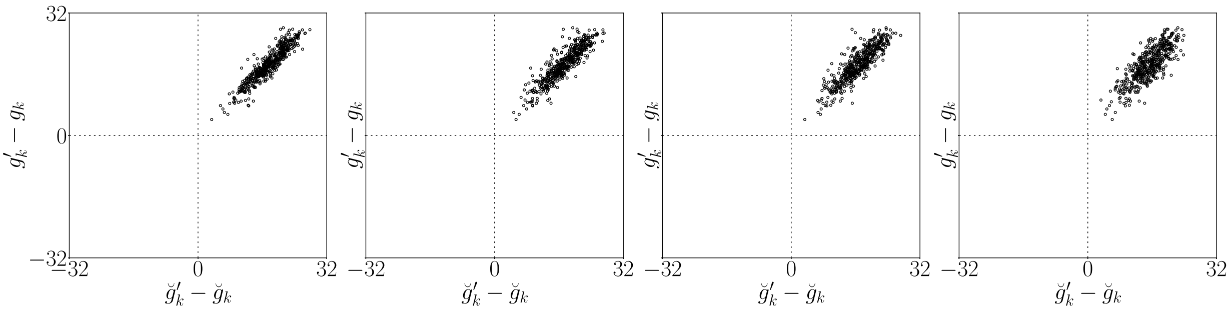


Fig. 22. Results for scenario 2 - empirical absolute action consistency: We demonstrate from the left to the right action consistency of the samples of the returns for $n = 100, n = 75, n = 50, n = 25$, whereas $N = 1000$. As we see, all the samples are action consistent.

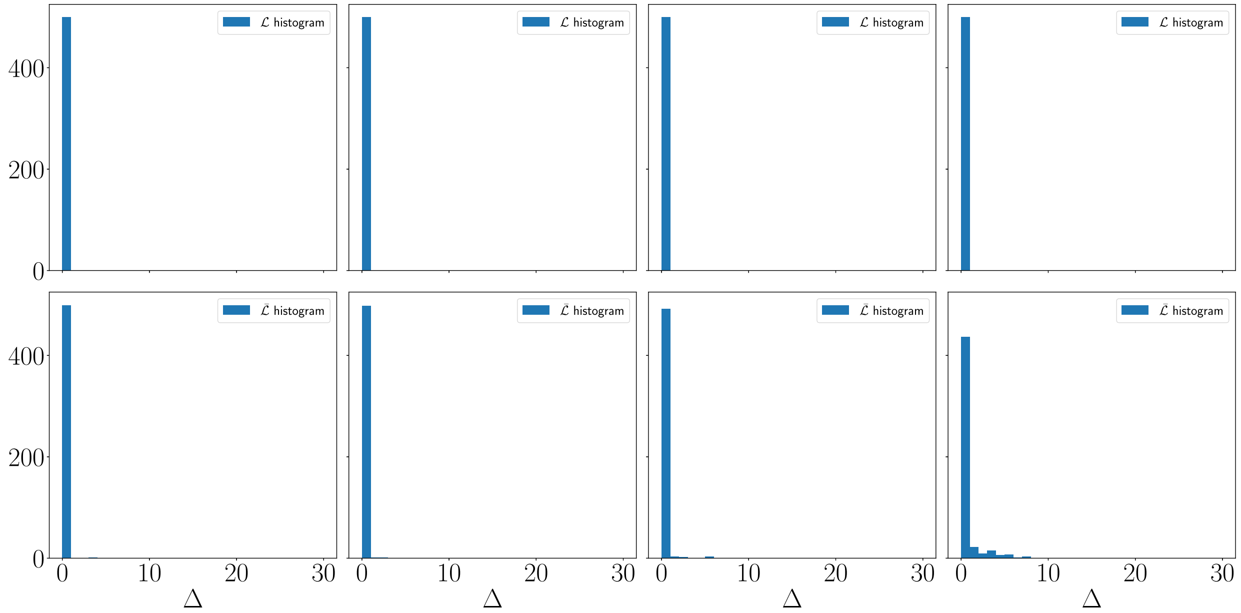


Fig. 23. Results for scenario 2 - empirical absolute action consistency: Histograms of PLoss and PbLoss for $N = 1000$, $\alpha = 0.01$, $z_{\alpha/2} = 2.56$, bin width is 1.0; from the left to the right $n = 100$, $n = 75$, $n = 50$, $n = 25$.

Table 6
Results for scenario 2 - empirical absolute action consistency: run times for $N = 1000$.

	$n = 100$	$n = 75$	$n = 50$	$n = 25$
g_k time [sec]	36745	45187	44899	30889
$\breve{g}_k$ and l, u time [sec]	17361	12546	4388	844
$\breve{g}_k$ time [sec]	363	247	65	14
l, u time [sec]	16998	12299	4323	830

8. Conclusions

We introduced a novel simplification framework in the challenging continuous domain with possibly nonparametric beliefs and general belief dependent rewards. We presented a formulation of novel stochastic bounds on the return and proved that these bounds yield deterministic bounds on VaR. We considered simplification impact also on the joint distribution of a pair of returns given two candidate policies, while accounting for the correlation between these returns. In this context, we proposed an innovative objective operator on top of the joint distribution. In addition, we presented a mathematical tool PLoss and its online counterpart PbLoss to characterize the simplification impact on the decision making entirely for any objective operator. Moreover, we utilized it to provide deterministic guarantees for our novel risk aware objective operator mounted on the joint distribution of a pair of returns given a pair of policies. We presented an instance of our framework

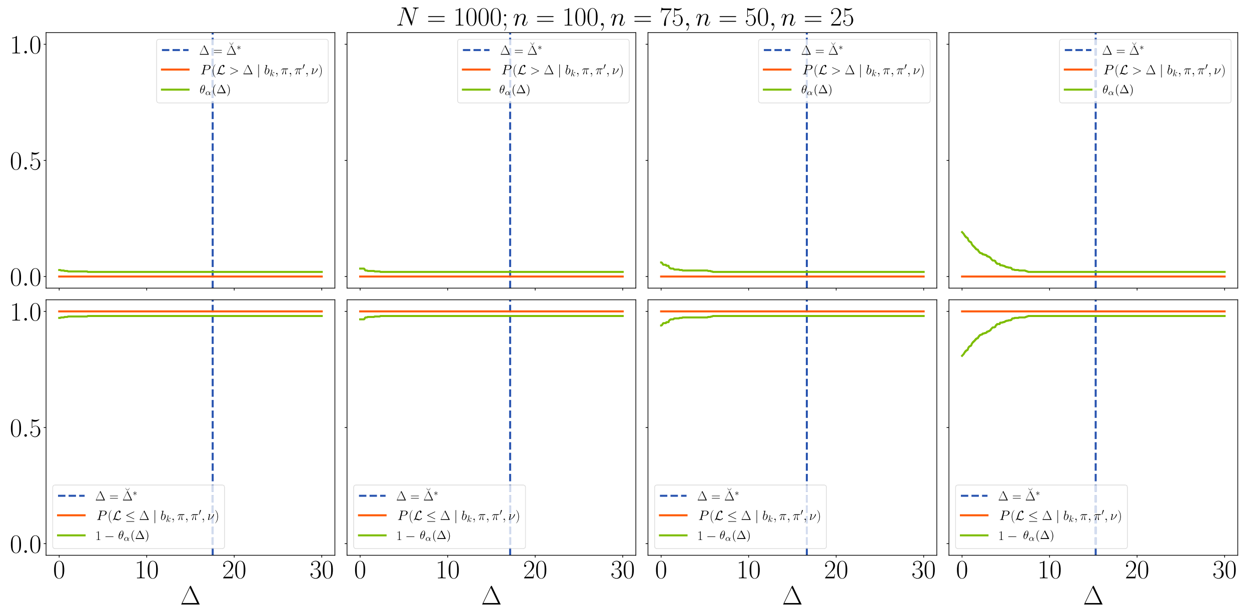


Fig. 24. Results for scenario 2 - empirical absolute action consistency: Empirical characterization for $N = 1000$, $\alpha = 0.01$, $z_{\alpha/2} = 2.56$, evaluated in a grid with intervals 0.001; from the left to the right $n = 100$, $n = 75$, $n = 50$, $n = 25$.

with a specific simplification method, which is reducing the number of samples of the return or the belief used for reward calculation. Finally, we verified the advantages of our approach through extensive simulations. For example, in section 7.1.2 we obtained approximately 8 times speedup with respect to the original problem while still identifying the optimal action.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Proofs for the theorems

A.1. Proof of the Theorem 1

Using the marginalization over future observations with Probability Density Function (PDF) being $\mathbb{P}(z_{k+}|b_k, \pi)$ we have that

$$P(\mathbf{1}_{\{l\le g_k\le u\}} = 1|b_k, \pi, \nu) \ge (1-\alpha)\underbrace{\int_{z_{k+}} \mathbb{P}(z_{k+}|b_k, \pi)\mathrm{d}z_{k+}}_{=1} = 1-\alpha. \tag{114}$$

The following holds from property of a lower bound (usual stochastic order) $\forall \xi \in (-\infty, \infty)$

$$P(l > \xi|b_k, \pi, \nu, \mathbf{1}_{\{l\le g_k\le u\}} = 1) \le P(g_k > \xi|b_k, \pi, \mathbf{1}_{\{l\le g_k\le u\}} = 1). \tag{115}$$

Denote $\lambda = P(\mathbf{1}_{\{l\le g_k\le u\}} = 1|b_k, \pi, \nu)$. This notation implies that $1-\lambda = P(\mathbf{1}_{\{l\le g_k\le u\}} = 0|b_k, \pi, \nu)$. Using marginalization over the indicator function we have that

$$P(g_k > \xi|b_k, \pi) = P(g_k > \xi|b_k, \pi, \nu, \mathbf{1}_{\{l\le g_k\le u\}} = 1)\lambda + P(g_k > \xi|b_k, \pi, \nu, \mathbf{1}_{\{l\le g_k\le u\}} = 0)(1-\lambda). \tag{116}$$

Since each summand in the equation above is non negative and using (114) we obtain

$$P(g_k > \xi|b_k, \pi) \ge P(l > \xi|b_k, \pi, \nu, \mathbf{1}\{l \le g_k \le u\} = 1)(1-\alpha). \tag{117}$$

Assume $\alpha \in [0, 1)$, exist $c \in \mathbb{R}^+$ such that

$$P(g_k > \xi|b_k, \pi, \nu) = c(1-\alpha). \tag{118}$$

This implies

$$P(g_k > \xi \cap \mathbf{1}\{l \leq g_k \leq u\} = 1|b_k, \pi, \nu) \leq P(g_k > \xi|b_k, \pi, \nu) = c(1-\alpha). \tag{119}$$

Applying the chain rule and rearranging the terms, we have that

$$P(g_k > \xi|\mathbf{1}\{l \leq g_k \leq u\} = 1, b_k, \pi, \nu) \leq c \underbrace{\frac{1-\alpha}{\lambda}}_{\leq 1} \leq c. \tag{120}$$

Using again marginalization over the indicator function, we represent the $P(l > \xi|b_k, \pi, \nu)$ as

$$P(l>\xi|b_k, \pi, \nu, \mathbf{1}_{\{l \leq g_k \leq u\}} = 1)\lambda + P(l>\xi|b_k, \pi, \nu, \mathbf{1}_{\{l \leq g_k \leq u\}} = 0)(1-\lambda). \tag{121}$$

Using that $\lambda \leq 1$ and $P(l>\xi|b_k, \pi, \nu, \mathbf{1}_{\{l \leq g_k \leq u\}} = 0) \leq 1$ we have that

$$P(l > \xi|b_k, \pi, \nu) \leq P(l > \xi|b_k, \pi, \nu, \mathbf{1}_{\{l \leq g_k \leq u\}} = 1) + 1 - \lambda \leq P(l > \xi|b_k, \pi, \nu, \mathbf{1}_{\{l \leq g_k \leq u\}} = 1) + 1 - (1-\alpha). \tag{122}$$

Using (115) and (118), we arrive at the desired result

$$\begin{aligned} &P(l > \xi|b_k, \pi, \nu, \mathbf{1}_{\{l \leq g_k \leq u\}} = 1) + \alpha \leq \\ &P(g_k > \xi|b_k, \pi, \nu, \mathbf{1}_{\{l \leq g_k \leq u\}} = 1) + \alpha \leq c + \alpha = \frac{P(g_k > \xi|b_k, \pi)}{1-\alpha} + \alpha. \end{aligned} \tag{123}$$

Rearranging the terms bears

$$\big(P(l > \xi|b_k, \pi, \nu) - \alpha\big)(1-\alpha) \leq P(g_k > \xi|b_k, \pi). \tag{124}$$

Switching the roles of g_k to u and l to g_k, we obtain the upper bound

$$P(g_k > \xi|b_k, \pi) \leq \frac{P(u > \xi|b_k, \pi, \nu)}{1-\alpha} + \alpha. \tag{125}$$

This completes the proof. □

A.2. Proof of the Theorem 2

Let us start from upper bound. From Theorem 1

$$\left\{\xi \text{ s.t } P(g_k>\xi|b_k, \pi, \nu) \geq 1-\beta\right\} \subseteq \left\{\xi \text{ s.t } \frac{P(u>\xi|b_k, \pi, \nu)}{1-\alpha} + \alpha \geq 1-\beta\right\}. \tag{126}$$

Equivalently

$$\left\{\xi \text{ s.t } P(g_k > \xi|b_k, \pi, \nu) \geq 1-\beta\right\} \subseteq \left\{\xi \text{ s.t } P(u > \xi|b_k, \pi, \nu) \geq (1-\beta-\alpha)(1-\alpha)\right\}. \tag{127}$$

Rearranging the terms, we have that

$$\sup\left\{\xi \text{ s.t } P(g_k > \xi|b_k, \pi, \nu) \geq 1-\beta\right\} \leq \sup\left\{\xi \text{ s.t } P(u > \xi|b_k, \pi, \nu) \geq 1 - (\beta + \alpha(2-\beta-\alpha)\right\}. \tag{128}$$

It is left to show that

$$0 \leq \beta + \alpha(2-\beta-\alpha) \leq 1. \tag{129}$$

Since $\alpha + \beta \leq 2$ and $\alpha \geq 0$, we have

$$0 \leq \beta \leq \beta + \alpha(2-\beta-\alpha). \tag{130}$$

To prove the right inequality we show that

$$\beta + \alpha(2-\beta-\alpha) - 1 \leq 0. \tag{131}$$

Multiplying by -1 we observe that the inequality reads

$$(1-\beta-\alpha)\underbrace{(1-\alpha)}_{\geq 0} \geq 0. \tag{132}$$

Requiring that $1-\beta-\alpha \geq 0$, we obtain the condition which we already assumed

$$(1-\alpha) \geq \beta. \tag{133}$$

We have that

$$\mathrm{VaR}_{\beta}(g_k|b_k, \pi, \nu) \leq \mathrm{VaR}_{\beta+\alpha(2-\beta-\alpha)}(u|b_k, \pi, \nu). \tag{134}$$

To prove the second part of the theorem we use the following

$$\Big\{\xi \text{ s.t } (P(l > \xi|b_k, \pi, \nu) - \alpha)(1-\alpha) \geq 1-\beta\Big\} \subseteq \Big\{\xi \text{ s.t } P(g_k > \xi|b_k, \pi, \nu) \geq 1-\beta\Big\}. \tag{135}$$

Equivalently,

$$\left\{\xi \text{ s.t } P(l > \xi|b_k, \pi, \nu) \geq 1 - \left(1 - \frac{1-\beta}{1-\alpha} - \alpha\right)\right\} \subseteq \Big\{\xi \text{ s.t } P(g_k > \xi|b_k, \pi, \nu) \geq 1-\beta\Big\}. \tag{136}$$

It is left to show that

$$0 \leq 1 - \frac{1-\beta}{1-\alpha} - \alpha \leq 1. \tag{137}$$

We immediately see that $1 - \frac{1-\beta}{1-\alpha} - \alpha \leq 1$. Rearranging the terms, we arrive at the second condition that

$$\alpha(2-\alpha) \leq \beta. \tag{138}$$

This completes the proof. □

A.3. Proof of the Theorem 3

By definition

$$P(\mathbf{1}_{\{\bar{\mathcal{L}} \geq \mathcal{L}\}} = 1|\mathbf{1}_{\{l \leq g_k \leq u\}} = 1, \mathbf{1}_{\{l' \leq g'_k \leq u'\}} = 1, b_k, \pi, \pi', z_{k+}, z'_{k+}, \nu) = 1. \tag{139}$$

We first apply marginalization over future observations $z_{k+} \equiv z_{k+1:k+L}$ and $z'_{k+} \equiv z'_{k+1:k+L}$, and events $\{\omega|l(\omega) \leq g_k(\omega) \leq u(\omega)\}$ and $\big\{\omega|l'(\omega) \leq g'_k(\omega) \leq u'(\omega)\big\}$. We then use the fact that given two histories $\mathcal{H}_{k+L} \triangleq \{b_k, \pi, z_{k+}\}$ and $\mathcal{H}'_{k+L} \triangleq \{b_k, \pi', z'_{k+}\}$, the events $\{\omega|l(\omega) \leq g_k(\omega) \leq u(\omega)\}$ and $\big\{\omega|l'(\omega) \leq g'_k(\omega) \leq u'(\omega)\big\}$ are independent of each other. Furthermore, each such event depends exclusively on its own history by design. We have that $P\left(\mathbf{1}_{\{\bar{\mathcal{L}} \geq \mathcal{L}\}} = 1|b_k, \pi, \pi', \nu\right)$ equals to

$$\int\limits_{\substack{z_{k+} \\ z'_{k+}}} P\left(\mathbf{1}_{\{\bar{\mathcal{L}} \geq \mathcal{L}\}} = 1|b_k, \pi, \pi', z_{k+}, z'_{k+}, \nu\right) \mathbb{P}\left(z_{k+}, z'_{k+} \mid b_k, \pi, \pi'\right) \mathrm{d}z_{k+}\mathrm{d}z'_{k+}. \tag{140}$$

Moreover, the $P\left(\mathbf{1}_{\{\bar{\mathcal{L}} \geq \mathcal{L}\}} = 1|b_k, \pi, \pi', z_{k+}, z'_{k+}, \nu\right)$ is larger or equal to

$$P(\mathbf{1}_{\{\bar{\mathcal{L}} \geq \mathcal{L}\}} = 1 \wedge \mathbf{1}_{\{l \leq g_k \leq u\}} = 1 \wedge \mathbf{1}_{\{l' \leq g'_k \leq u'\}} = 1 \mid b_k, \pi, \pi', z_{k+}, z'_{k+}, \nu). \tag{141}$$

Engaging the chain rule and using the constraints (59) and (60), and their statistical independence we face that

$$P\left(\mathbf{1}_{\{\bar{\mathcal{L}} \geq \mathcal{L}\}} = 1|b_k, \pi, \pi', z_{k+}, z'_{k+}, \nu\right) \geq (1-\alpha)^2. \tag{142}$$

The above expression straightforwardly yields that $P\left(\mathbf{1}_{\{\bar{\mathcal{L}} \geq \mathcal{L}\}} = 1|b_k, \pi, \pi', \nu\right) \geq (1-\alpha)^2$ through the marginalization over the future observations since $\int_{\substack{z_{k+} \\ z'_{k+}}} \mathbb{P}\left(z_{k+}, z'_{k+} \mid b_k, \pi, \pi'\right) \mathrm{d}z_{k+}\mathrm{d}z'_{k+} = 1$. This completes the proof. □

A.4. Proof of the Theorem 4

To shorten notations let us denote $|b_k, \pi, \pi', \nu$ by $|\cdot$ in the proof. Let us express `PLoss` TDF as

$$P(\mathcal{L} > \Delta|\cdot) = P\left(\mathcal{L} > \Delta|\mathbf{1}_{\{\bar{\mathcal{L}} \geq \mathcal{L}\}} = 1, \cdot\right) P\left(\mathbf{1}_{\{\bar{\mathcal{L}} \geq \mathcal{L}\}} = 1|\cdot\right) + P\left(\mathcal{L} > \Delta|\mathbf{1}_{\{\bar{\mathcal{L}} < \mathcal{L}\}} = 1, \cdot\right) P\left(\mathbf{1}_{\{\bar{\mathcal{L}} \geq \mathcal{L}\}} = 0|\cdot\right). \tag{143}$$

Similarly, `PbLoss` TDF reads

$$P\left(\bar{\mathcal{L}} > \Delta|\cdot\right) = P\left(\bar{\mathcal{L}} > \Delta|\mathbf{1}_{\{\bar{\mathcal{L}}\geq\mathcal{L}\}} = 1,\cdot\right) P\left(\mathbf{1}_{\{\bar{\mathcal{L}}\geq\mathcal{L}\}} = 1|\cdot\right) + P\left(\bar{\mathcal{L}} > \Delta|\mathbf{1}_{\{\bar{\mathcal{L}}<\mathcal{L}\}} = 1,\cdot\right) P\left(\mathbf{1}_{\{\bar{\mathcal{L}}<\mathcal{L}\}} = 1|\cdot\right). \tag{144}$$

Since $\alpha \in [0, 1)$ it exists $c \in \mathbb{R}_{>0}$ such that

$$P\left(\mathcal{L} > \Delta \wedge \mathbf{1}_{\{\bar{\mathcal{L}}\geq\mathcal{L}\}} = 1|\cdot\right) \leq P\left(\bar{\mathcal{L}} > \Delta|\cdot\right) = c(1-\alpha)^2. \tag{145}$$

This implies

$$P\left(\bar{\mathcal{L}} > \Delta|\mathbf{1}_{\{\bar{\mathcal{L}}\geq\mathcal{L}\}} = 1,\cdot\right) P\left(\mathbf{1}_{\{\bar{\mathcal{L}}\geq\mathcal{L}\}} = 1|\cdot\right) \leq c(1-\alpha)^2, \tag{146}$$

$$P\left(\bar{\mathcal{L}} > \Delta|\mathbf{1}_{\{\bar{\mathcal{L}}\geq\mathcal{L}\}} = 1,\cdot\right) \leq c\frac{(1-\alpha)^2}{\underbrace{P\left(\mathbf{1}_{\{\bar{\mathcal{L}}\geq\mathcal{L}\}} = 1|\cdot\right)}_{\leq 1}} \leq c. \tag{147}$$

Moreover, using that $P(\mathbf{1}_{\{\bar{\mathcal{L}}\geq\mathcal{L}\}} = 1|\cdot) + P(\mathbf{1}_{\{\bar{\mathcal{L}}<\mathcal{L}\}} = 1|\cdot) = 1$, we obtain

$$\begin{aligned} P\left(\mathcal{L} > \Delta|\cdot\right) = & P\left(\mathcal{L} > \Delta|\bar{\mathcal{L}} \geq \mathcal{L},\cdot\right) P\left(\bar{\mathcal{L}} \geq \mathcal{L}|\cdot\right) + \\ & P\left(\mathcal{L} > \Delta|\bar{\mathcal{L}} < \mathcal{L},\cdot\right)\left(1 - P(\bar{\mathcal{L}} \geq \mathcal{L})|\cdot\right) \leq \\ & P\left(\mathcal{L} > \Delta|\bar{\mathcal{L}} \geq \mathcal{L},\cdot\right) + 1 - (1-\alpha)^2 \leq c + 2\alpha - \alpha^2, \end{aligned} \tag{148}$$

but $c = \frac{P(\bar{\mathcal{L}} > \Delta|b_k,\pi,\pi',\nu)}{(1-\alpha)^2}$. We showed that

$$P\left(\mathcal{L} > \Delta|\cdot\right) \leq \frac{P\left(\bar{\mathcal{L}} > \Delta|b_k,\pi,\pi',\nu\right)}{(1-\alpha)^2} + 2\alpha - \alpha^2. \tag{149}$$

Furthermore, by definition of TDF

$$P\left(\mathcal{L} > \Delta|\cdot\right) \leq 1. \tag{150}$$

We write the above two relations compactly as

$$P\left(\mathcal{L} > \Delta|\cdot\right) \leq \theta_\alpha(\Delta), \tag{151}$$

where $\theta_\alpha(\Delta) = \min\left\{1, \frac{P(\bar{\mathcal{L}} > \Delta|b_k,\pi,\pi',\nu)}{(1-\alpha)^2} + 2\alpha - \alpha^2\right\}$. Clearly

$$P\left(\mathcal{L} \leq \Delta|b_k,\pi,\pi',\nu\right) = 1 - P\left(\mathcal{L} > \Delta|b_k,\pi,\pi',\nu\right) \geq 1 - \theta_\alpha(\Delta). \tag{152}$$

This concludes the proof. □

Appendix B. Technical characteristics of computers used in simulations

Our simulations are written in Julia language with a multi-threaded calculation of immediate reward. We used 4 computers with the following characteristics:

1. 40 cores Intel(R) Xeon(R) E5-2670 v2 with 256 GB of RAM working at 2.50 GHz;
2. 24 cores Intel(R) Core(TM) i9-7920X with 64 GB of RAM working at 2.90 GHz;
3. 20 cores Intel(R) Xeon(R) E5-2630 v4 with 64 GB of RAM working at 2.20 GHz;
4. 20 cores Intel(R) Core(TM) i9-9820X with 64 GB of RAM working at 3.30 GHz.

References

[1] M. Araya, O. Buffet, V. Thomas, F. Charpillet, A pomdp extension with belief-dependent rewards, in: Advances in Neural Information Processing Systems (NIPS), 2010, pp. 64–72.

[2] M. Barenboim, V. Indelman, Adaptive information belief space planning, in: The 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence (IJCAI-ECAI), 2022.

[3] D. Bertsekas, Dynamic Programming and Optimal Control, vol. 1, Athena Scientific, Belmont, MA, 1995.

[4] Y. Boers, H. Driessen, A. Bagchi, P. Mandal, Particle filter based entropy, in: 2010 13th International Conference on Information Fusion, 2010, pp. 1–8.

[5] Y. Chow, A. Tamar, S. Mannor, M. Pavone, Risk-sensitive and robust decision-making: a cvar optimization approach, Adv. Neural Inf. Process. Syst. 28 (2015) 1522–1530.

[6] B. Defourny, D. Ernst, L. Wehenkel, Risk-aware decision making and dynamic programming, in: NIPS Workshop on Model Uncertainty and Risk in RL, 2008.

[7] L. Dressel, M.J. Kochenderfer, Efficient decision-theoretic target localization, in: L. Barbulescu, J. Frank, Smith S.F. Mausam (Eds.), Proceedings of the Twenty-Seventh International Conference on Automated Planning and Scheduling, ICAPS 2017, Pittsburgh, Pennsylvania, USA, June 18–23, 2017, AAAI Press, 2017, pp. 70–78, https://aaai.org/ocs/index.php/ICAPS/ICAPS17/paper/view/15761.

[8] K. Elimelech, V. Indelman, Simplified decision making in the belief space using belief sparsification, Int. J. Robot. Res. 41 (2022) 470–496, https://doi.org/10.1177/02783649221076381.

[9] M. Fehr, O. Buffet, V. Thomas, J. Dibangoye, rho-pomdps have Lipschitz-continuous epsilon-optimal value functions, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 31, Curran Associates, Inc., 2018, pp. 6933–6943.

[10] J. Fischer, O.S. Tas, Information particle filter tree: an online algorithm for pomdps with belief-based rewards on continuous domains, in: Intl. Conf. on Machine Learning (ICML), Vienna, Austria, 2020.

[11] D. Fox, Adapting the sample size in particle filters through kld-sampling, Int. J. Robot. Res. 22 (2003) 985–1003.

[12] A. Hakobyan, G.C. Kim, I. Yang, Risk-aware motion planning and control using cvar-constrained optimization, IEEE Robot. Autom. Lett. 4 (2019) 3924–3931.

[13] A. Hakobyan, I. Yang, Wasserstein distributionally robust motion planning and control with safety constraints using conditional value-at-risk, in: 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2020, pp. 490–496.

[14] M. Hoerger, H. Kurniawati, A. Elfes, Multilevel Monte-Carlo for solving pomdps online, in: Proc. International Symposium on Robotics Research (ISRR), 2019.

[15] M.F. Huber, T. Bailey, H. Durrant-Whyte, U.D. Hanebeck, On entropy approximation for Gaussian mixture random vectors, in: IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, 2008, pp. 181–188.

[16] V. Indelman, No correlations involved: decision making under uncertainty in a conservative sparse information space, IEEE Robot. Autom. Lett. 1 (2016) 407–414.

[17] V. Indelman, L. Carlone, F. Dellaert, Planning in the continuous domain: a generalized belief space approach for autonomous navigation in unknown environments, Int. J. Robot. Res. 34 (2015) 849–882.

[18] L.P. Kaelbling, M.L. Littman, A.R. Cassandra, Planning and acting in partially observable stochastic domains, Artif. Intell. 101 (1998) 99–134.

[19] A. Kitanov, V. Indelman, Topological information-theoretic belief space planning with optimality guarantees, arXiv preprint arXiv:1903.00927, 2019.

[20] D.E. Knuth, Art of Computer Programming, Volume 2: Seminumerical Algorithms, Addison-Wesley Professional, 2014.

[21] M. Kochenderfer, T. Wheeler, K. Wray, Algorithms for Decision Making, MIT Press, 2022.

[22] L. Kocsis, C. Szepesvári, Bandit based Monte-Carlo planning, in: European Conference on Machine Learning, Springer, 2006, pp. 282–293.

[23] N. Koenig, A. Howard, Design and use paradigms for gazebo, an open-source multi-robot simulator, in: IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2004.

[24] H. Kurniawati, D. Hsu, W.S. Lee, SARSOP: efficient point-based POMDP planning by approximating optimally reachable belief spaces, in: Robotics: Science and Systems (RSS), 2008.

[25] S. Pathak, A. Thomas, V. Indelman, A unified framework for data association aware robust belief space planning and perception, Int. J. Robot. Res. 32 (2018) 287–315.

[26] J. Pineau, G.J. Gordon, S. Thrun, Anytime point-based approximations for large POMDPs, J. Artif. Intell. Res. 27 (2006) 335–380.

[27] R. Platt, R. Tedrake, L. Kaelbling, T. Lozano-Pérez, Belief space planning assuming maximum likelihood observations, in: Robotics: Science and Systems (RSS), Zaragoza, Spain, 2010, pp. 587–593.

[28] J.M. Porta, N. Vlassis, M.T. Spaan, P. Poupart, Point-based value iteration for continuous pomdps, J. Mach. Learn. Res. 7 (2006) 2329–2367.

[29] J.A. Rice, Mathematical Statistics and Data Analysis, Cengage Learning, 2006.

[30] A. Ryan, Information-theoretic tracking control based on particle filter estimate, in: AIAA Guidance, Navigation and Control Conference, 2008, pp. 1–15.

[31] P. Santana, S. Thiébaux, B. Williams, Rao*: an algorithm for chance-constrained pomdp's, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2016.

[32] M. Shienman, V. Indelman, D2a-bsp: distilled data association belief space planning with performance guarantees under budget constraints, in: IEEE Intl. Conf. on Robotics and Automation (ICRA), 2022.

[33] D. Silver, J. Veness, Monte-Carlo planning in large pomdps, in: Advances in Neural Information Processing Systems (NIPS), 2010, pp. 2164–2172.

[34] T. Smith, R. Simmons, Heuristic search value iteration for pomdps, in: Conf. on Uncertainty in Artificial Intelligence (UAI), 2004, pp. 520–527.

[35] A. Somani, N. Ye, D. Hsu, W.S. Lee, Despot: online pomdp planning with regularization, in: NIPS, 2013, pp. 1772–1780.

[36] M.T. Spaan, T.S. Veiga, P.U. Lima, Decision-theoretic planning under uncertainty with information rewards for active cooperative perception, Auton. Agents Multi-Agent Syst. 29 (2015) 1157–1185.

[37] Z. Sunberg, M. Kochenderfer, Online algorithms for pomdps with continuous state, action, and observation spaces, in: Proceedings of the International Conference on Automated Planning and Scheduling, 2018.

[38] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, 2018.

[39] O. Sztyglic, V. Indelman, Online pomdp planning via simplification, arXiv preprint arXiv:2105.05296, 2021.

[40] O. Sztyglic, V. Indelman, Speeding up online pomdp planning via simplification, in: IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2022.

[41] O. Sztyglic, A. Zhitnikov, V. Indelman, Simplified belief-dependent reward mcts planning with guaranteed tree consistency, arXiv preprint arXiv:2105.14239, 2021.

[42] S. Thrun, W. Burgard, D. Fox, Probabilistic Robotics, The MIT Press, Cambridge, MA, 2005.

[43] C. Voss, M. Moll, L.E. Kavraki, A heuristic approach to finding diverse short paths, in: IEEE Intl. Conf. on Robotics and Automation (ICRA), 2015, pp. 4173–4179.

[44] N. Ye, A. Somani, D. Hsu, W.S. Lee, Despot: online pomdp planning with regularization, J. Artif. Intell. Res. 58 (2017) 231–266.