Simplified Risk-aware Decision Making with Belief-dependent Rewards in Partially Observable Domains (Extended Abstract)*

Andrey Zhitnikov[1] , **Vadim Indelman**[2]

[1]Technion Autonomous Systems Program (TASP)
[2]Department of Aerospace Engineering
Technion - Israel Institute of Technology, Haifa 32000, Israel
andreyz@campus.technion.ac.il, vadim.indelman@technion.ac.il

Abstract

It is a long-standing objective to ease the computation burden incurred by the decision-making problem under partial observability. Identifying the sensitivity to simplification of various components of the original problem has tremendous ramifications. Yet, algorithms for decision-making under uncertainty usually lean on approximations or heuristics without quantifying their effect. Therefore, challenging scenarios could severely impair the performance of such methods. In this paper, we extend the decision-making mechanism to the whole by removing standard approximations and considering all previously suppressed stochastic sources of variability. On top of this extension, we scrutinize the distribution of the return. We begin from a return given a single candidate policy and continue to the pair of returns given a corresponding pair of candidate policies. Furthermore, we present novel stochastic bounds on the return and novel tools, Probabilistic Loss (`PLoss`) and its online accessible counterpart (`PbLoss`), to characterize the effect of a simplification.

1 Introduction

While operating in a partially observable setting, the robot repetitively performs actions and receives observations from the environment in an interleaving manner. The result of each action is a imprecise change in the robot's state. The robot has access to the probability density of the state, given the history of its actions and the observations alongside the prior. We call this probability density a belief. In each planning session, the robot shall reason about future beliefs and select an optimal action based on its current belief using belief-dependent rewards and the objective operator. The robot shall look into the future as far as possible. With the growing horizon, however, the computational burden is becoming unbearable for the robot due to exponential growth in complexity [Papadimitriou and Tsitsiklis, 1987]. Many research efforts in Artificial Intelligence (AI) and Robotics communities have tackled the described problem. In AI community, it received the name Partially Observable Markov Decision Process (POMDP), whereas, in the Robotics community, it is known as Belief Space Planning (BSP). In classical POMDP the belief-dependent reward is assumed to be the average of the state-dependent reward with respect to belief. While alleviating the solution, this assumption hinders the ability to actively decrease uncertainty over the belief using general belief-dependent operators. In BSP, general belief-dependent rewards are essential, e.g., navigation, sensor placement problems. The classical assumption in BSP is that the belief follows Gaussian distribution [Indelman *et al.*, 2015].

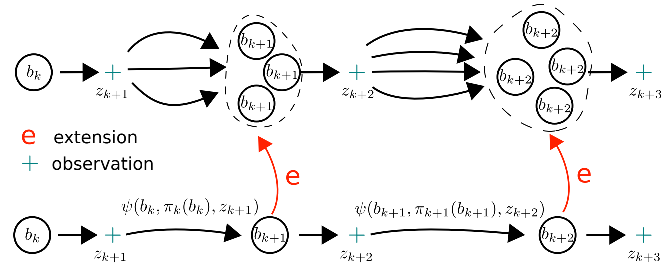

Figure 1: The Extended Belief Tree versus the standard.

The AI community began to introduce general belief-dependent rewards starting from the discrete domains [Araya *et al.*, 2010], [Fehr *et al.*, 2018], and limiting assumptions concerning the reward operators [Dressel and Kochenderfer, 2017]. More recent approaches such as Sparse Sampling (SS) [Kearns *et al.*, 2002], and Monte Carlo Tree Search (MCTS) [Sunberg and Kochenderfer, 2018] build upon Belief-MDP (BMDP). These methods are suitable for continuous domains. Still, in the continuous setting of states and observations, these methods give an approximate solution with only asymptotical optimality guarantees. On the other hand, the BSP community introduced a concept of *simplification* [Indelman, 2016],[Elimelech and Indelman, 2022], [Shienman and Indelman, 2022b], [Kitanov and Indelman, 2019]. As opposed to approximations, the simplification paradigm substitutes various parts of the decision-making problem while providing guarantees on the impact of such a substitution.

*The original journal paper: A. Zhitnikov and V. Indelman. Simplified Risk Aware Decision Making with Belief-dependent Rewards in Partially Observable Domains. Artificial Intelligence, Special Issue on "Risk-Aware Autonomous Systems: Theory and Practice", 2022.

In this work, we focus on the distribution of the rewards in a nonparametric setting. Our objective is to simplify the decision-making problem and analyze the impact of the simplification.

2 Notations and Problem Formulation

Let $\mathbb{P}$ be the probability density and P the probability. In this paper, we focus on the finite horizon setting. 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.

2.1 POMDP with Belief Dependent Rewards

ρ-POMDP [Araya *et al.*, 2010] 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 $x \in \mathcal{X}, a \in \mathcal{A}, z \in \mathcal{Z}$ the momentary state, action, and observation, respectively, $T(x, a, x') \triangleq \mathbb{P}_T(x'|x, a)$ is the transition model from the past momentary state x to the next x' through action a, $O(z, x) \triangleq \mathbb{P}_Z(z|x)$ is the observation model, $\rho(b', z', a, b)$ is a scalar reward operator, $\gamma \in (0, 1]$ is the discount factor, and b_0 is the prior belief.

2.2 Belief Space Planning

The posterior belief at time instant k is given by

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

The usual assumption is that the belief is a sufficient statistic for decision making objective [Bertsekas, 1995]. However, in practice, the belief requires some representation. This representation is not perfect, e.g., parametric or sampled form; thus, in (2), we used the $\approx$ sign. In a real life scenario $b_k = \psi(\psi(\ldots\psi(b_0, a_0, z_1), a_{k-2}, z_{k-1}), a_{k-1}, z_k)$, where ψ is a method for updating the belief. By $\pi \triangleq \pi_{k:k+L-1}$ we denote a vector of policies for L time steps starting from time step k. Each such policy π_ℓ at time step ℓ maps belief to an action $\pi_\ell(b_\ell) = a_\ell$. The general decision making under uncertainty objective function is of the following form

$$V^L(b_k, \pi) = \varphi\big(\mathbb{P}(\rho_{k+1:k+L} | b_k, \pi_{k:k+L-1}), g_k\big) \tag{3}$$
$$\text{s.t. } b_\ell = \psi(b_{\ell-1}, \pi_{\ell-1}(b_{\ell-1}), z_\ell),$$

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 [Sutton and Barto, 2018]. A common choice for φ is expectation over the distribution of future rewards given all data available [Defourny *et al.*, 2008]. The return is a deterministic known function of the realization of $\rho_{k+1:k+L}$, 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 (3) is ultimately based on the *distribution of the return* given all information available for planning under selected policy $\mathbb{P}(g_k | b_k, \pi_k)$, which decomposes via marginalization over future observations $z_{k+} \equiv z_{k+1:k+L}$ as

$$\mathbb{P}(g_k | b_k, \pi) = \int_{z_{k+}} \mathbb{P}(g_k | b_k, \pi, z_{k+}) \mathbb{P}(z_{k+} | b_k, \pi) \mathrm{d}z_{k+}. \tag{4}$$

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

3 Foundations

In this section we introduce probabilistic ρ-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.

3.1 Extended Setting, Probabilistic ρ-POMDP

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 $\psi_{\mathrm{dt}}(\theta_{\ell-1}, \pi_{\ell-1}(\theta_{\ell-1}), z_\ell)$. 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 [Thrun *et al.*, 2005]. 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 $\psi_{\mathrm{st}}(b_{\ell-1}, \pi_{\ell-1}(b_{\ell-1}), z_\ell)$. Another form to formulate the above is that the distribution

$$B(b_{\ell-1}, a_{\ell-1}, z_\ell, b_\ell) \triangleq \mathbb{P}_B(b_\ell | b_{\ell-1}, a_{\ell-1}, z_\ell), \tag{5}$$

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 hold for the momentary reward operator of the belief. We extend $\rho(b', z', a, b)$ to

$$R(b_{\ell-1}, a_{\ell-1}, z_\ell, b_\ell, \rho_\ell) \triangleq \mathbb{P}_R(\rho_\ell | b_\ell, z_\ell, a_{\ell-1}, b_{\ell-1}), \tag{6}$$

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 (6) and probabilistic belief update B (5), and introduce

$$M = \langle \mathcal{X}, \mathcal{A}, \mathcal{Z}, T, O, R, \gamma, b_k, B \rangle, \tag{7}$$

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 (3) is based on (4). These previously overlooked sources of stochasticity impact the likelihood of the observations

$$\mathbb{P}(z_{k+1:k+L} | b_k, \pi), \tag{8}$$

as well as the joint reward distribution $\mathbb{P}(\rho_{k+} | b_k, \pi, z_{k+}) \equiv \mathbb{P}(\rho_{k+1:k+L} | b_k, \pi_{k:k+L-1}, z_{k+1:k+L})$ given a realization of future observations. In contrast, in the regular setting of POMDP and ρ-POMDP $\mathbb{P}(\rho_{k+} | b_k, \pi, z_{k+})$ is Dirac's delta function. If B is a Dirac function, a sample from (8) 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 (7), corresponds to an *extended* belief tree, which, due to (5), allows many samples of the beliefs $b_{k+1:k+L}$ for each sample of $z_{k+1:k+L}$ from (8) (See Fig. 1).

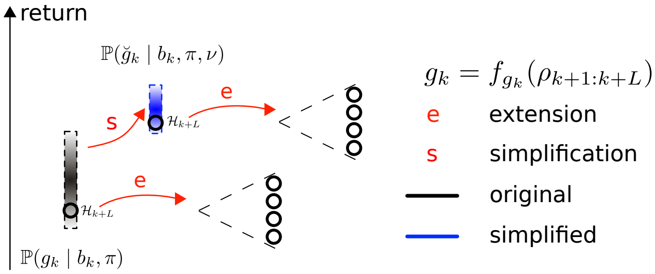

Figure 2: Our extension and the simplification in the context of a single candidate policy.

3.2 Simplification Formulation

To formally define the simplification procedure, we augment the $\mathbb{P}\rho$-POMDP tuple (7) with a simplification operator $\nu \triangleq \nu_k, \ldots, \nu_{k+L}$,

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

This general operator defines any possible modification of the original problem defined by (7) alongside with (3) to a new, simpler to solve, problem. The definition (9) allows us to retain the connection to the original nonsimplified problem (7) and examine the impact of the simplification on (7). The operator ν can be for example, sparsification of the initial belief b_k [Elimelech and Indelman, 2022], replacing the reward by its topological signature [Kitanov and Indelman, 2019], direct calculation of lightweight reward bounds [Sztyglic and Indelman, 2022], selecting a subset of hypotheses in a hybrid or mixture belief [Shienman and Indelman, 2022a], to name a few.

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

$$\mathbb{P}(\breve{b}_\ell | b_\ell; \nu^b_\ell). \tag{10}$$

In general, the distribution (10) over the simplified belief $\breve{b}_\ell$ corresponds to a stochastic simplification operator ν^b_ℓ. This is the case, for example, when b_ℓ is represented by a set of N weighted samples and ν^b_ℓ 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 (10). Overall there are $\binom{N}{n}$ such combinations. For a deterministic operator ν^b_ℓ, (10) is a Dirac function.

Further, there are several cases of how a simplification affects belief update (5) 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 (5).
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 (10). The corresponding belief update can be written as
$$\mathbb{P}_{\breve{B}}(\breve{b}_\ell | \breve{b}_{\ell-1}, \pi_{\ell-1}, z_\ell; \nu^b_\ell, \nu^\psi_\ell) = \int_{\tilde{b}_\ell} \mathbb{P}(\breve{b}_\ell | \tilde{b}_\ell; \nu^b_\ell) \mathbb{P}_{\breve{B}}(\tilde{b}_\ell | \breve{b}_{\ell-1}, \pi_{\ell-1}, z_\ell; \nu^\psi_\ell) \mathrm{d}\tilde{b}_\ell,$$
where $\tilde{b}_\ell$ is the integration variable.

We combine these cases and write

$$\breve{B}\big(\breve{b}_{\ell-1}, \pi_{\ell-1}, z_\ell, \breve{b}_\ell; \nu\big) \triangleq \mathbb{P}_{\breve{B}}(\breve{b}_\ell | \breve{b}_{\ell-1}, \pi_{\ell-1}, z_\ell; \nu^b_\ell, \nu^\psi_\ell). \tag{11}$$

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

$$\mathbb{P}(\breve{\rho}_\ell | \rho_\ell; \nu^\rho_\ell). \tag{12}$$

Thus, given a simplified belief $\breve{b}_\ell$ and $\breve{b}_{\ell-1}$, and recalling (6), the distribution over $\breve{\rho}_\ell$ is

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

which we denote as the simplified reward model,

$$\breve{R}(\breve{b}_\ell, z_\ell, \pi_{\ell-1}(\breve{b}_{\ell-1}), \breve{\rho}_\ell; \nu) \triangleq \mathbb{P}_{\breve{R}}\big(\breve{\rho}_\ell | \breve{b}_\ell, z_\ell, \pi_{\ell-1}(\breve{b}_{\ell-1}), \breve{b}_{\ell-1}; \nu\big). \tag{13}$$

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 (8). For the further discussion we make the following shorthand notation. Let $\mathcal{H}_{k+L} \triangleq \{b_k, \pi, z_{k+}\}$ be future history at the time index $k+L$.

3.3 Online Stochastic and Analytical Bounds

We turn now to the joint distribution over original and simplified rewards, given the future history and operator ν, namely $\mathbb{P}\left(\rho_{k+}, \breve{\rho}_{k+} | \mathcal{H}_{k+L}, \nu\right)$. 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 unsimplified rewards. Further, the bounds can be analytical as in previous simplification approaches, e.g, [Elimelech and Indelman, 2022]. Ultimately for each realization of the return we are interested in the following relation

$$l \le g_k \le u. \tag{14}$$

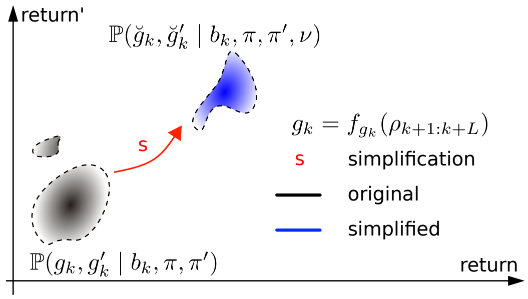

Figure 3: The simplification in our extended setting and its impact of the joint distribution of a pair of the returns corresponding to the pair of the candidate policies.

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 corresponding return, e.g, as in [Sztyglic and Indelman, 2021].

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. 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

$$\mathbb{P}(g_k, \breve{g}_k|\mathcal{H}_{k+L}, \nu), \tag{15}$$

as illustrated 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 (15) is no longer a Dirac function, we can use knowledge about this distribution to design bounds, which will hold with *some* probability. In the main paper [Zhitnikov and Indelman, 2022], we show that it is possible to harness the structure of (15) to design the mentioned more lenient online bounds. 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. 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

$$1 - \alpha \leq \mathrm{P}\left(\mathbf{1}\{l \leq g_k \leq u\} = 1|\mathcal{H}_{k+L}, \nu\right) \tag{16}$$

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 ψ_{dt} and a deterministic operator reward, 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 (14).
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.

4 The Return Given a Candidate Policy

Applying the marginalization over the observations we obtain the distribution of the original and the simplified return given the candidate policy and the operator ν (See Fig. 2).

$$\mathbb{P}(g_k, \breve{g}_k|b_k, \pi, \nu) = \int_{z_{k+}} \mathbb{P}(g_k, \breve{g}_k|\mathcal{H}_{k+L}, \nu)\mathbb{P}(z_{k+}|b_k, \pi)\mathrm{d}z_{k+}.$$

For further discussion please see [Zhitnikov and Indelman, 2022].

5 The Pair of the Returns Corresponding to the Pair of Candidate Policies

Imagine a pair of a candidate policies. In such a setting we are interested in the following distribution (See Fig. 3)

$$\mathbb{P}(g_k, g'_k, \breve{g}_k, \breve{g}'_k|b_k, \pi, \pi', \nu). \tag{17}$$

On top of (17) we propose a tool to examine the simplification impact on the original not simplified problem. We call it Probabilistic Loss.

5.1 Probabilistic Loss (`PLoss`)

Consider a random variable $\mathcal{L} : \Omega \to \mathbb{R}$ over the events space Ω defined as such

$$\mathcal{L}(\omega) \triangleq \begin{cases} \max\{g'_k(\omega) - g_k(\omega), 0\} & \text{if } \breve{g}_k(\omega) > \breve{g}'_k(\omega) \\ \max\{g_k(\omega) - g'_k(\omega), 0\} & \text{if } \breve{g}_k(\omega) < \breve{g}'_k(\omega) \\ 0 & \text{if } \breve{g}_k(\omega) = \breve{g}'_k(\omega) \end{cases} \tag{18}$$

The realization of random variable $\mathcal{L}(\omega) = \Delta$ differs from zero if the simplification have switched the ordering of the original returns and the original difference between returns was Δ.

5.2 Online Bound on Probabilistic Loss (`PbLoss`)

Since the `PLoss` is inaccessible online we propose another random variable which is accessible.

$$\bar{\mathcal{L}}(\omega) \triangleq \begin{cases} \max\{u'(\omega) - l(\omega), 0\} & \text{if } \breve{g}_k(\omega) > \breve{g}'_k(\omega) \\ \max\{u(\omega) - l'(\omega), 0\} & \text{if } \breve{g}_k(\omega) < \breve{g}'_k(\omega) \\ 0 & \text{if } \breve{g}_k(\omega) = \breve{g}'_k(\omega) \end{cases} \tag{19}$$

To give to the reader a glimpse into the connection between `PLoss` and `PbLoss` suppose the bounds (14) are analytical. This implies that $\mathcal{L}(\omega) \leq \bar{\mathcal{L}}(\omega) \quad \forall \omega \in \Omega$ and this implies

$$\mathrm{P}(\Delta \leq \mathcal{L}(\omega)) \leq \mathrm{P}(\Delta \leq \bar{\mathcal{L}}(\omega)) \tag{20}$$

To the impact of the proposed ideas onto Decision Making please refer to the journal paper [Zhitnikov and Indelman, 2022].

References

[Araya *et al.*, 2010] Mauricio Araya, Olivier Buffet, Vincent Thomas, and Françcois Charpillet. A pomdp extension with belief-dependent rewards. In *Advances in Neural Information Processing Systems (NIPS)*, pages 64–72, 2010.

[Bertsekas, 1995] Dimitri Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

[Defourny *et al.*, 2008] Boris Defourny, Damien Ernst, and Louis Wehenkel. Risk-aware decision making and dynamic programming. In *NIPS Workshop on Model Uncertainty and Risk in RL*, 2008.

[Dressel and Kochenderfer, 2017] Louis Dressel and Mykel J. Kochenderfer. Efficient decision-theoretic target localization. In Laura Barbulescu, Jeremy Frank, Mausam, and Stephen F. Smith, editors, *Proceedings of the Twenty-Seventh International Conference on Automated Planning and Scheduling, ICAPS 2017, Pittsburgh, Pennsylvania, USA, June 18-23, 2017*, pages 70–78. AAAI Press, 2017.

[Elimelech and Indelman, 2022] Khen Elimelech and Vadim Indelman. Simplified decision making in the belief space using belief sparsification. *The International Journal of Robotics Research*, 41(5):470–496, 2022.

[Fehr *et al.*, 2018] Mathieu Fehr, Olivier Buffet, Vincent Thomas, and Jilles Dibangoye. rho-pomdps have lipschitz-continuous epsilon-optimal value functions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6933–6943. Curran Associates, Inc., 2018.

[Indelman *et al.*, 2015] V. Indelman, L. Carlone, and F. Dellaert. Planning in the continuous domain: a generalized belief space approach for autonomous navigation in unknown environments. *Intl. J. of Robotics Research*, 34(7):849–882, 2015.

[Indelman, 2016] V. Indelman. No correlations involved: Decision making under uncertainty in a conservative sparse information space. *IEEE Robotics and Automation Letters (RA-L)*, 1(1):407–414, 2016.

[Kearns *et al.*, 2002] Michael Kearns, Yishay Mansour, and Andrew Y Ng. A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Machine learning*, 49(2):193–208, 2002.

[Kitanov and Indelman, 2019] A. Kitanov and V. Indelman. Topological information-theoretic belief space planning with optimality guarantees. *arXiv preprint arXiv:1903.00927*, 3 2019.

[Papadimitriou and Tsitsiklis, 1987] C. Papadimitriou and J. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

[Shienman and Indelman, 2022a] M. Shienman and 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.

[Shienman and Indelman, 2022b] M. Shienman and V. Indelman. Nonmyopic distilled data association belief space planning under budget constraints. In *Proc. of the Intl. Symp. of Robotics Research (ISRR)*, 2022.

[Sunberg and Kochenderfer, 2018] Zachary Sunberg and Mykel Kochenderfer. Online algorithms for pomdps with continuous state, action, and observation spaces. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 28, 2018.

[Sutton and Barto, 2018] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[Sztyglic and Indelman, 2021] Ori Sztyglic and Vadim Indelman. Online pomdp planning via simplification. *arXiv preprint arXiv:2105.05296*, 2021.

[Sztyglic and Indelman, 2022] Ori Sztyglic and Vadim Indelman. Speeding up online pomdp planning via simplification. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2022.

[Thrun *et al.*, 2005] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. The MIT press, Cambridge, MA, 2005.

[Zhitnikov and Indelman, 2022] A. Zhitnikov and V. Indelman. Simplified risk aware decision making with belief dependent rewards in partially observable domains. *Artificial Intelligence, Special Issue on "Risk-Aware Autonomous Systems: Theory and Practice"*, 2022.