Missingness-MDPs:
Bridging the Theory of Missing Data and POMDPs

Joshua Wendland[1,*], **Markel Zubia**[1], **Roman Andriushchenko**[2], **Maris F. L. Galesloot**[3], **Milan Češka**[2], **Henrik von Kleist**[4], **Thiago D. Simão**[5], **Maximilian Weininger**[1], **Nils Jansen**[1,3]

[1]Ruhr University Bochum, Germany
[2]Brno University of Technology, Czech Republic
[3]Radboud University Nijmegen, The Netherlands
[4]Helmholtz Munich, Germany
[5]Eindhoven University of Technology, The Netherlands

Abstract

We introduce *missingness-MDPs* (miss-MDPs); a subclass of partially observable Markov decision processes (POMDPs) that incorporates the theory of missing data. Miss-MDPs capture settings where, at each step, the current state may go partially missing, that is, the state is not observed. Missingness of observations occurs dynamically and is caused by a *missingness function*, which governs the underlying probabilistic missingness process. Miss-MDPs distinguish the three types of missingness processes as a restriction on the missingness function: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Our goal is to compute a policy for a miss-MDP with an *unknown missingness function*. We propose algorithms that, by using a retrospective dataset and based on the different types of missingness processes, approximate the missingness function and, thereby, the true miss-MDP. The algorithms can approximate a subset of MAR and MNAR missingness functions, and we show that, for these, the optimal policy in the approximated model is ε-optimal in the true miss-MDP. The empirical evaluation confirms these findings. Additionally, it shows that our approach becomes more sample-efficient when exploiting the type of the underlying missingness process.

1 Introduction

Markov decision processes [MDPs; 1] capture sequential decision-making under uncertainty. This model assumes that sensors provide precise measurements of state features at all times. However, sensors may fail, leading to *missing* state features, which obscures the computation of state-based policies. Consider a medical doctor, who is provided with sensor measurements of a patient's state features (e.g., heart rate and temperature). However, some of these measurements may not be available when decisions are made.

Partially observable Markov decision processes [POMDPs; 2] extend MDPs by an *observation function* that explicitly models uncertainty in state observations. [3]. Yet, solving POMDPs is notoriously challenging: In particular, inferring the observation function from state-feature observations alone is generally intractable as the probabilities depend on the past sequences of actions and observations [4, 5].

Fortunately, specific problems often exhibit a simpler structure in the source of partial observability. Here, the *missingness* of state-features may occur according to a specific stochastic function. Such problems are extensively studied by the theory of *missingness* [6–8]. As the reasons for missingness to

[0]*Corresponding author: `joshua.wendland@ruhr-uni-bochum.de`

18th European Workshop on Reinforcement Learning (EWRL 2025).

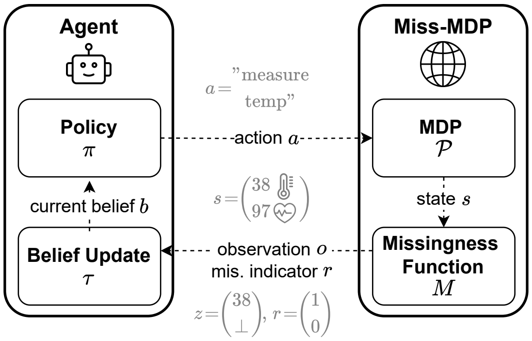


Figure 1: The interaction of an agent and a miss-MDP, a subclass of POMDPs. In this doctor-treating-patient example, the missingness function causes a feature (heart rate) to go missing, indicated as "⊥" in the observation. The missingness indicator evaluates to 0 for missing features and to 1 otherwise.

occur may vary, Rubin [9] classifies missingness functions into three main types: *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR). MCAR missingness is independent of observed or unobserved information – e.g., a patient's temperature readings are missing due to a loosely attached thermometer. MAR missingness solely depends on observed information – e.g., a patient's observed temperature readings influence the missingness of the heart rate. Missingness functions that are neither MCAR nor MAR are considered MNAR – e.g., if a patient's temperature readings influence its missingness, also known as *self-censoring*.

While prior work has studied decision making with missing observations, it mostly focuses on reinforcement learning and treats missing data as an incidental issue rather than modeling it explicitly [10–13]. Planning methods often overlook the key differences between MCAR, MAR, and MNAR [14–16], and those that rely on generic imputation often use methods that make implicit assumptions about the missingness function, which can lead to biased or inconsistent estimates [16] or offer no clear guarantees on policy performance [15]. To our knowledge, no existing framework (1) explicitly models and learns the function that governs the probabilities of state features to go missing in the context of POMDPs and (2) provides guarantees on the belief-based policy computed with the learned function.

To create a principled understanding of missing state features in MDPs, we introduce *missingness-MDPs* (miss-MDPs) as a subclass of POMDPs. Miss-MDPs have an observation function that we refer to as the *missingness function* – classified as MCAR/MAR/MNAR – leading to missing state features in the observation. In Figure 1, we depict a miss-MDP describing the doctor-treating-patient example inspired by [17]. We consider the setting where we are (1) provided a miss-MDP and a dataset of trajectories sampled from the miss-MDP, but (2) the missingness function is unknown. The problem is to find a policy that maximizes the expected reward without knowing the missingness function.

Our approach is based on learning an approximation of the missingness function M from data, which yields an approximation of the original miss-MDP. For this approximated model, we compute a policy through off-the-shelf POMDP solvers such as SARSOP [18]. In summary, our contributions are:

1. We introduce miss-MDPs, which integrate and define the semantics of missingness in a specific subclass of the more general POMDP framework (Section 4).
2. We derive a notion similar to *ignorability* [8] for the setting of miss-MDPs (Remark 1).
3. We provide algorithms for learning the missingness function if it is MCAR, or of certain sub-types of MAR or MNAR, yielding *probably approximately correct* (PAC) guarantees (Sections 5.1 and 5.2).
4. We prove that we can approximate the optimal policy for the miss-MDP with PAC guarantees under the correct assumption on the missingness function (Section 5.3).

Our empirical evaluation (Section 6) highlights the practical advantages of our approach: Using datasets of reasonable size, the performance of policies computed using the learned missingness function converges to that of the optimal policy.

2 Related Work

Our work builds on a rich literature in missing data analysis [8, 19]. Classical assumptions such as MCAR, MAR, and MNAR provide high-level categories. More refined tools, such as missingness graphs, allow one to encode assumptions about the missingness in a structured way [20, 21], leading to highly specific learnability results [22, 23]. Our setting departs from the standard missing data paradigm in several important aspects. In particular, the concept of missingness is embedded within the broader POMDP setting, which allows for a better and principled understanding of missingness in the context of sequential decision-making under uncertainty.

As noted previously, most work on decision making with missing data focuses on RL, where either full observations [24] or individual features may be missing [12, 25, 26]. Some approaches incorporate missingness into belief updates for RL agents [10], while others adopt model-based methods, often restricted to simpler settings such as MCAR [16]. Another line of work combines deep learning with POMDP solvers by learning abstract state representations, but without explicitly modeling the missingness process [14]. More principled imputation strategies—such as Bayesian multiple imputation [13] and expectation-maximization [15]—estimate missing values as an intermediate step in policy computation. In contrast to imputation, our approach directly learns the missingness function and offers PAC guarantees on the resulting policy.

3 Preliminaries

A function $\mu \colon X \to [0,1]$ is a *probability distribution over* X when $\sum_{x \in X} \mu(x) = 1$. The set of such distributions is $\Delta(X)$. The *support* of distribution $\mu \in \Delta(X)$ is $\text{supp}(\mu) = \{x \in X \mid \mu(x) \neq 0\}$. Writing $\mu = \{x_1 \mapsto p_1, \dots, x_k \mapsto p_k\}$ indicates that $\mu(x_1) = p_1$ and so on. The random variable $\boldsymbol{x}$ sampled from μ is denoted by $\boldsymbol{x} \sim \mu$. Given $\sigma \colon X \to \Delta(Y)$, we let $\sigma(y \mid x) := \sigma(x)(y)$. *Iverson brackets*, $[\varphi]$, return 1 if predicate φ holds and 0 otherwise.

Definition 1 (POMDPs)**.** A *partially observable Markov decision process* is a tuple $\mathcal{P} = (S, A, T, b_0, \varrho, Z, O, \gamma)$, with $S = \times_{i=1,\dots,n} S_i$ the finite factored *state space* (we denote the set of feature indices by $I = \{1, \dots, n\}$), A the finite *action space*, $T \colon S \times A \to \Delta(S)$ the *transition function*, $b_0 \in \Delta(S)$ the *initial state distribution*, $\varrho \colon S \times A \to \mathbb{R}$ the *reward function*, $Z = \times_{i=1,\dots m} Z_i$ the finite factored *observation space*, $O \colon S \to \Delta(Z)$ the *observation function*, and $\gamma \in [0,1)$ the *discount factor*.

Without loss of generality, we assume that O is a state-based observation function, all $s \in S$ are reachable, and T is a total function. If each state in the POMDP can be uniquely identified from its observation, it reduces to an MDP.

A *trajectory* in a POMDP $\mathcal{P}$ is a sequence of states, observations, and actions. A *history* $h = \left(z^{(0)}, a^{(0)}, z^{(1)}, a^{(1)}, \dots\right) \in \mathcal{H} \subseteq (Z \times A)^*$ is the observable fragment of a trajectory, i.e., a sequence of observations and actions. A history can be summarized by a *sufficient statistic* known as a *belief* $b \in \mathcal{B}$, with $\mathcal{B} \subseteq \Delta(S)$; a probability distribution over underlying states induced by observing a history $h \in \mathcal{H}$. The *belief update* $\tau \colon \mathcal{B} \times A \times Z \to \mathcal{B}$ computes a *successor belief* b' via Bayes' rule [27].

A *policy* $\pi \colon \mathcal{B} \to \Delta(A) \in \Pi$ maps beliefs to probability distributions over actions. The *objective* is to find a policy $\pi^* \in \Pi$ that maximizes the infinite-horizon expected cumulative discounted reward: $V_{\mathcal{P}}(\pi) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \varrho(s^{(t)}, a^{(t)})\right]$. The problem of finding the optimal policy is undecidable [28]. Thus we focus on computing ε-optimal policies [29, 30].

4 Missingness in MDPs

This section introduces missingness-MDPs and the different types of missingness functions in the context of POMDPs.

Definition 2 (Miss-MDP)**.** A missingness-MDP is a tuple $(S, A, T, b_0, \varrho, Z, M, \gamma)$, where S, A, T, b_0, ϱ, and γ are as in a POMDP, the finite *observation space* is $Z = \times_{i \in I}(S_i \cup \{\perp\})$, with $\perp$ denoting *missing information*, and function $M \colon S \to \Delta(Z)$ is the *missingness function* such that $\forall s \in S, \forall z \in \text{supp}(M(s)), \forall i \in I$ either $z_i = s_i$ or $z_i = \perp$.

Miss-MDPs are a subclass of POMDPs where the state space S and observation space Z share the feature indices I, and where $Z \supsetneq S$ because some features can go *missing* in Z, being replaced by the symbol $\perp$. This process of "poking holes" is governed by the stochastic missingness function M.

Missingness indicators. Missingness functions can equivalently be described as a map to vectors of *missingness indicators* [20]: Such a vector $r \in R = \{0, 1\}^n$ has $r_i = 0$ if feature i is missing ($z_i = \perp$), and otherwise $r_i = 1$. The function $f_{\text{miss}} \colon Z \to R$ maps observations to their missingness indicators.

Example 1. Let $\mathcal{P}$ be a miss-MDP where $S = \{a, b\}^2$, $Z = \{a, b, \perp\}^2$, and the missingness function is defined as: $M((s_1, s_2)) = \{(s_1, s_2) \mapsto 0.5, (s_1, \perp) \mapsto 0.5\}$. Then, for instance, visiting state (b, a) yields either (b, a) or $(b, \perp)$, each with probability of 0.5. We have $f_{\text{miss}}((b, a)) = (1, 1)$ and $f_{\text{miss}}((b, \perp)) = (1, 0)$.

We aim to compute a near-optimal policy for a miss-MDP $\mathcal{P}$ with *unknown* missingness function M. For this, we use a dataset $\mathcal{D}$ of histories (of length at least $|S|$), which are collected using a fair policy (i.e. it has positive probability to visit all reachable states). The resulting policy is *probably approximately correct* (PAC) if, with high probability, its value is close to the true optimum. Formally:

Problem statement. We are given a miss-MDP $\mathcal{P}$ with an *unknown* missingness function M, a dataset $\mathcal{D} = (h_1, \ldots, h_k)$ of k histories $h_i \in \mathcal{H}$ collected from $\mathcal{P}$ under an unknown but fair policy π_b, and a precision $\varepsilon > 0$ and confidence threshold $\delta > 0$. The goal is to approximate the missingness function $\widehat{M} \approx M$ and use it to compute a policy $\pi^* \in \Pi$ such that with probability at least $1 - \delta$, we have $|\sup_\pi(V_{\mathcal{P}}(\pi)) - V_{\mathcal{P}}(\pi^*)| \leq \varepsilon$.

4.1 Types of Missingness Functions

We formally introduce the three types of missingness functions (MCAR, MAR, and MNAR) in the context of miss-MDPs. The simplest is MCAR, where the probability of a feature going missing does not depend on any *feature values* of the state. The miss-MDP in Example 1 is of this type.

Definition 3 (MCAR). The missingness function $M \colon S \to \Delta(Z)$ of a miss-MDP $\mathcal{P}$ is MCAR iff $\forall r \in R, \exists p_r \in [0, 1], \forall s \in S, \mathbb{P}\left(f_{\text{miss}}(\boldsymbol{z}) = r \mid \boldsymbol{z} \sim M(s)\right) = p_r$.

Admittable and I_{always}. To define MAR and MNAR, we require the following notions: An observation $z \in Z$ is *admittable* by a state $s \in S$, denoted $z \preceq s$, iff $\forall i \in I$, $z_i = \perp$ or $z_i = s_i$. In Example 1, we have $(b, \perp) \preceq (b, a)$ and $(b, a) \preceq (b, a)$ but $(a, \perp) \not\preceq (b, a)$. Furthermore, $I_{\text{always}} = \{i \in I \mid \forall s' \in S \colon \mathbb{P}(\boldsymbol{z}_i = \perp \mid \boldsymbol{z} \sim M(s')) = 0\} \subseteq I$ is the set of indices of features that never go missing, and $I_{\text{mis}} = I \setminus I_{\text{always}}$ is its complement.

In the following, we distinguish between a restricted MAR version, which we call simple MAR [31], and the general MAR definition [9]. For simple MAR, the probability of observation features being missing is only influenced by the observable features that never go missing, i.e., z_i for $i \in I_{\text{always}}$. For MAR, a missingness probability is only influenced by the *non-missing* features of a given observation, including features that may go missing. Any MCAR missingness function is also (simple) MAR.

Definition 4 ((Simple) MAR). The missingness function $M \colon S \to \Delta(Z)$ of a miss-MDP $\mathcal{P}$ is:

- **Simple MAR** iff for all $s, s' \in S$ that agree on always-observed features (i.e. $\forall i \in I_{\text{always}}$, $s_i = s'_i$), the missingness probability is the same for all missingness indicators $r \in R$, formally: $\mathbb{P}(f_{\text{miss}}(\boldsymbol{z}) = r \mid \boldsymbol{z} \sim M(s)) = \mathbb{P}(f_{\text{miss}}(\boldsymbol{z}') = r \mid \boldsymbol{z}' \sim M(s'))$.
- **MAR** iff for all $s, s' \in S$ and $z \in Z$, if $z \preceq s, s'$, the probability of its missingness indicator $r := f_{\text{miss}}(z)$ is equal for both states, formally: $\mathbb{P}(f_{\text{miss}}(\boldsymbol{z}') = r \mid \boldsymbol{z}' \sim M(s)) = \mathbb{P}(f_{\text{miss}}(\boldsymbol{z}'') = r \mid \boldsymbol{z}'' \sim M(s'))$.

Example 2. We redefine the missingness function for the miss-MDP from Example 1 to be simple MAR: $M((s_1, a)) = \{(s_1, a) \mapsto 1\}$, and $M((s_1, b)) = \{(s_1, b) \mapsto 0.5, (\perp, b) \mapsto 0.5\}$. Here, the missingness probability of feature 1 depends on the (always observed) value of feature 2. As an example of MAR but not simple MAR, consider: $M((s_1, a)) = \{(s_1, a) \mapsto 0.5, (\perp, \perp) \mapsto 0.5\}$, and $M((s_1, b)) = \{(s_1, b) \mapsto 0.25, (\perp, b) \mapsto 0.25, (\perp, \perp) \mapsto 0.5\}$. Now, the missingness probability of feature 1 depends on the value of feature 2 (only if observed), while feature 2 itself misses with probability 0.5.

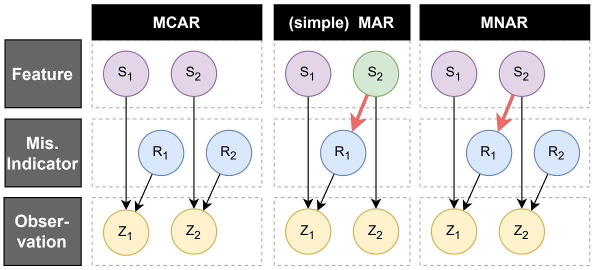


Figure 2: Exemplary missingness graphs visualizing relations between the elements of a miss-MDP for the three types of missingness functions. Red arrows indicate the relevant change to the MCAR graph making the missingness function either simple MAR or MNAR.

Definition 5 (**MNAR**)**.** The missingness function M of a miss-MDP $\mathcal{P}$ is MNAR iff it is not MAR.

In MNAR missingness functions, missingness probabilities may depend on the values of missing features. In particular, in *self-censoring* missingness functions, the missingness probability of a feature depends on its own value.

Example 3. We adapt Example 1 so that the probability of s_2 going missing depends on its own value, making M MNAR: $M((s_1, a)) = \{(s_1, a) \mapsto 0.5, (s_1, \bot) \mapsto 0.5\}$ and $M((s_1, b)) = \{(s_1, b) \mapsto 0.1, (s_1, \bot) \mapsto 0.9\}$.

4.2 Missingness Graphs

Missingness graphs (m-graphs) are a tool for analyzing the characteristics of missingness functions. We adapt the definition of Mohan and Pearl [31], translating it to our framework of miss-MDPs. An m-graph is a causal diagram [32] in the form of a directed acyclic graph. The vertices in the graph correspond to variables, and the directed edges correspond to the causal relationships between the variables.

The vertices can be grouped into the following categories:[1] Ⓢ-nodes correspond to features of the state space, Ⓩ-nodes correspond to the features of observations and Ⓡ-nodes correspond to the missingness indicators. For always observed features, we omit the respective Ⓡ-node from the m-graph. Arrows between nodes represent a direct causal relationship: The parent node is a direct cause of the child node. The absence of an edge intuitively denotes that two variables do not directly influence each other; formally, it means that they are conditionally independent, given other variables in the graph according to the d-separation criteria [33].

Visualizing types of missingness. Figure 2 visualizes the conditional independence assumptions of the types of missingness functions using m-graphs for the miss-MDP from Example 1. For MCAR, both Ⓡ-nodes have no incoming arrows. Hence, they do not depend on any feature value, but are purely stochastic. For (simple) MAR, there are two changes: Feature S_2 affects the missingness indicator R_1 (red arrow), and R_2 is absent, making feature S_2 always observable. For MNAR, S_2 can go missing, and thus the missingness indicator R_1 depends on information that can go missing. We remark that m-graphs cannot represent *context-specific* independence assumptions, which are needed to, for instance, represent non-simple MAR functions such as the one in Example 2. The approximated missingness function in this paper can all be represented by m-graphs.

5 Approximating Missingness-MDPs

Before we explain how to approximate missingness functions in order to compute near-optimal policies, we present an interesting insight: For certain types of missingness and certain problems, the missingness function M can in fact be *ignored*.

[1] We leave out a set of unobserved variables, U, from the definition of Mohan and Pearl [31], as in our setting M depends on the state S and therefore $U = \emptyset$.

Remark 1 (Ignorability)**.** Missing data literature defines *ignorability* as cases where any quantity of interest can be consistently estimated from observations alone and it is not necessary to model the missingness process [8]. This holds under MCAR, and also under MAR whenever the quantity depends only on the observed features.

We identify a similar notion of ignorability for miss-MDPs: If the missingness function M is MAR (including MCAR), then belief updates τ can be computed without knowledge of the precise probabilities of M, since these cancel out in Bayes' rule; see Appendix A for a formal proof. Thus, MAR missingness is ignorable for maintaining a belief when executing a policy in a miss-MDP. However, we stress that the missingness function *is* required to compute belief-based policies, since probabilities of successor beliefs depend on it.

As our goal is to provide ε-optimal policies of a miss-MDP, we are indeed required to approximate M. We first compute an approximation $\widehat{M} \approx M$ of the unknown M from the given dataset $\mathcal{D}$ of histories. This yields an approximated, but fully specified miss-MDP $\widehat{\mathcal{P}}$, which can be solved using any off-the-shelf POMDP solution method.

Missingness types in focus. A necessary condition is that the missingness function can be approximated solely from observations, a property that missing data literature calls *identifiability* [22]. Establishing identifiability is not the focus of this paper. Instead, we provide PAC guarantees for two types that are known to be identifiable. Thus, we focus on: (1) simple MAR (including MCAR), and (2) non-self-censoring MNAR with independent missingness indicators. Additionally, in Section 6, we experiment on MNAR *with* dependencies between the indicators.

Outline. Sections 5.1 and 5.2 describe our algorithms for approximating missingness functions. Both are structured as follows: After stating their assumptions, they define how to compute $\widehat{M}$ and prove that the approximation is probably approximately correct. Finally, they explain how to utilize additional knowledge on the missingness function to further reduce sample complexity. Section 5.3 discusses how we use these algorithms to compute near-optimal policies.

Occurrence counts. Both algorithms utilize the dataset $\mathcal{D} = (h_1, \dots, h_k)$ of k histories $h_i \in \mathcal{H}$ to extract the number of occurrences of every observation, which we now formally define. For a finite history $h_i = \big(z^{(0)},\, a^{(0)},\, \dots, z^{(l)},\, a^{(l)}\big)$, we denote the j-th observation $z^{(j)}$ by $h_i^{(j)}$. The number of occurrences of an observation $z \in Z$ is: $\#_{\mathcal{D}}(z) = \sum_{i=1}^{k} \sum_{j=0}^{|h_i|} [h_i^{(j)} = z]$. For a set $Z' \subseteq Z$, we define $\#_{\mathcal{D}}\big(Z'\big) = \sum_{z \in Z'} \#_{\mathcal{D}}(z)$.

5.1 Approximating MCAR and Simple MAR

If a missingness function is of type simple MAR, we can approximate it using the *approximation for simple MAR* algorithm, `AsMAR`. The modifications to obtain the algorithm for the more restricted MCAR-type functions, `AMCAR`, are described at the end of the section.

Always-observable features. Based on $\mathcal{D}$ we partition the feature indices I into those that are always observed and those that can go missing as $\hat{I}_{\text{always}} = \{i \in I \mid \#_{\mathcal{D}}(\{z \in Z \mid z_i = \bot\}) = 0\}$ and $\hat{I}_{\text{mis}} = I \setminus \hat{I}_{\text{always}}$, respectively. Note, this partitioning is based on empirical data ($\hat{I}_{\text{always}} \approx I_{\text{always}}$) and we might misclassify a feature index to be in $\hat{I}_{\text{always}}$ even though it can go missing.

Computing $\widehat{M}$. We use the fact that M can be seen as a mapping $S \to \Delta(R)$ (see paragraph "Missingness indicators", Section 4). Consequently, for every state, we want to approximate the probability of a certain vector of missingness indicators. The simple MAR assumption tells us that the probabilities can only depend on the features in $\hat{I}_{\text{always}}$. Thus, for every combination of the always-observable features of a state $s \in S$ and missingness indicator vector $r \in R$, we can compute the occurrence count $\#_{\mathcal{D}}(s, r) = \#_{\mathcal{D}}\left(Z_s^r\right)$, where

$$Z_s^r = \left\{ z \in Z \;\middle|\; \forall i \in I\colon \begin{array}{c} (i \in \hat{I}_{\text{always}} \implies z_i = s_i) \\ \text{and}\,(r_i = 0 \implies z_i = \bot) \end{array} \right\}.$$

Using this, we obtain $\widehat{M}(z \mid s)$ as the fraction of observing $(s, f_{\text{miss}}(z))$ and the sum of counts for s and all possible missingness indicators values:

$$\widehat{M}(z \mid s) = \frac{\#_{\mathcal{D}}(s, f_{\text{miss}}(z))}{\sum_{r \in R} \#_{\mathcal{D}}(s, r)}. \tag{1}$$

Probably approximately correct. With enough data, our approach yields an arbitrarily precise approximation of the true missingness function. We formalize this in Theorem 1 as a PAC guarantee, not only proving that it becomes ε-precise for every $\varepsilon > 0$, but that we can also bound the probability of an error (through unlucky sampling). Additionally, we can adapt the claim to bound the imprecision of the resulting $\widehat{M}$ for a given dataset. The proof is provided in Appendix B.2.

Theorem 1. Let $\mathcal{P}$ be a missingness-MDP where the missingness function is simple MAR. For every given precision ε and confidence threshold δ, there exists a number n^* of histories, such that a dataset $\mathcal{D}$ of n^* histories has the following property: With probability at least δ, $\widehat{M}$ computed on $\mathcal{D}$ according to Equation (1) satisfies that for all reachable states $s \in S$ and observations $z \in Z$, we have $\widehat{M}(z \mid s) = M(z \mid s) \pm \varepsilon$. Dually, given a dataset $\mathcal{D}$ and confidence threshold δ, we can compute an ε such that with probability at least δ, for all states $s \in S$ and observations $z \in Z$, we have $\widehat{M}(z \mid s) = M(z \mid s) \pm \varepsilon$.

Using additional assumptions on the missingness function. Beyond the necessary simple MAR assumption, we can exploit additional assumptions to improve the approximation of M for the same $\mathcal{D}$. Consider a feature i that is always observable, but does not affect the missingness probability of other features. We can then exclude i from $\hat{I}_{\text{always}}$, thereby effectively merging the occurrence counts of states that differ only in this feature. Therefore, if we instead assume M to be MCAR, $\hat{I}_{\text{always}}$ can be reduced to an empty set. Consequently, we get that $\#_{\mathcal{D}}(s, r)$ does not depend on s anymore, and we effectively only count occurrences of missingness indicators.

We prove the correctness of these improvements in Appendix B.2. In Section 6, we empirically show that using such knowledge can significantly improve the precision of $\widehat{M}$ estimated from the same $\mathcal{D}$.

5.2 Approximating MNAR with Independent Missingness Indicators

This section presents the *approximation for independent missingness indicators* algorithm, `AIMI`. Its assumptions are:

1. **Independence of missingness indicators:** The fact that one feature is missing must not influence the missingness-probability of any other feature. Formally, for $s \in S$ and $z \in Z$, $\mathbb{P}(\boldsymbol{z} \mid \boldsymbol{z} \sim M(s)) = \Pi_{i \in I} \mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{z} \sim M(s))$.
2. **No self-censoring:** Intuitively, a feature may not influence its own missingness probabilities. Formally, for all $i \in I$ and every pair of states $s, s' \in S$ that differ only in the i-th feature ($s_i \neq s'_i$, but for all $j \neq i$ we have $s_j = s'_j$) we have $\mathbb{P}(\boldsymbol{z}_i = \bot \mid \boldsymbol{z} \sim M(s)) = \mathbb{P}(\boldsymbol{z}_i = \bot \mid \boldsymbol{z} \sim M(s'))$.
3. **Positivity:** Intuitively, if a feature affects the missingness probabilities of other features, we need to observe its value to learn the missingness probabilities. However, this is impossible if it always misses. Therefore, we require a *positivity assumption* [34]: For all $i \in I$ and $s \in S$, we have $\mathbb{P}(\boldsymbol{z}_i \neq \bot \mid \boldsymbol{z} \sim M(s)) > 0$.

Computing $\widehat{M}$. We compute the occurrence count for every state $s \in S$, feature $i \in I$ and value of a corresponding i-th missingness indicator $r_i \in \{0, 1\}$ as $\#_{\mathcal{D}}(s, i, r_i) = \#_{\mathcal{D}}(Z_s^{i,r_i})$, where Z_s^{i,r_i} is the following set of observations:

$$Z_s^{i,r_i} = \left\{ z \in Z \;\middle|\; \forall j \in I \setminus \{i\} \colon \begin{array}{l} (z_j = s_j) \text{ and} \\ (r_i = 0 \iff z_i = \bot) \end{array} \right\}.$$

By positivity, a large enough dataset almost surely contains observations to make the counters nonzero (i.e. for all s and i, we have $\#(s, i, 0) + \#(s, i, 1) > 0$). The probability of a non self-censoring feature i depends only on the other features $j \in I \setminus \{i\}$. Finally, using the independence assumption, we can infer $\widehat{M}$ by taking the product of the individual missingness probabilities of all features (again viewing M as a mapping $S \to \Delta(R)$, see Section 4):

$$\widehat{M}(z \mid s) = \prod_{i \in I} \frac{\#_{\mathcal{D}}(s, i, f_{\text{miss}}(z)_i)}{\#_{\mathcal{D}}(s, i, 0) + \#_{\mathcal{D}}(s, i, 1)} \tag{2}$$

Probably approximately correct. In Appendix B.3, we prove Theorem 2 that provides the same kind of guarantee as in Theorem 1; the only difference are the assumptions on the missingness function and the approach for calculating $\widehat{M}$.

Theorem 2 (PAC guarantee for `AIMI`)**.** Let $\mathcal{P}$ be a missingness-MDP where the missingness function satisfies independence, non-self-censoring, and non-sure missing. Then, the same PAC guarantees hold as specified for `AsMAR` in Theorem 1 but with $\widehat{M}$ computed using Equation (2).

Using additional assumptions on the missingness function. In its general form, `AIMI` maintains a counter for every possible combination of the feature valuations of other features $j \in I \setminus \{i\}$. If we know that a certain feature j does not affect the missingness probability of i, – there is no edge between the j-th Ⓢ-node and the i-th Ⓡ-node, – we can merge the counters for all values of the j-th feature. This knowledge can come from **(a)** an m-graph, **(b)** assuming simple MAR while observing feature j can go missing in $\mathcal{D}$, or **(c)** assuming MCAR. in which case we can completely drop the dependency on s in the counters. We prove in Appendix B.3 that all these modifications retain the PAC guarantees.

5.3 Computing a Policy with the Approximations

We show in Appendix B.4 that after finitely many samples, $\widehat{M}$ is accurate enough to yield an ε-optimal policy. We highlight that learning $\widehat{M}$ to precision ε is insufficient, as the errors in $\widehat{M}$ aggregate when solving the miss-MDP.

Theorem 3 (Computing ε-optimal Policies)**.** Let $\mathcal{P}$ be a miss-MDP with a missingness function that is simple MAR or that satisfies independence, no self-censoring, and positivity. Assume we can sample histories collected under a fair policy, and we know a lower bound on the smallest missingness probability $p \leq \min_{s \in S, z \in Z} M(z \mid s)$. Then, for every given precision ε and confidence threshold δ, we can in finite time compute a policy π^* such that with probability at least δ it is ε-optimal, i.e. $|\sup_\pi(V_{\mathcal{P}}(\pi)) - V_{\mathcal{P}}(\pi^*)| \leq \varepsilon$.

Practical considerations. The guarantees of Theorem 3 concern asymptotic convergence to an ε-optimal policy. Thus, they provide the theoretical foundation of our approach. Still, in practice, the required number of samples is very large, and we work with datasets that are not necessarily sufficient to provide the ε-optimality guarantees. Thus, we infer $\widehat{M}$ from a given dataset and then solve the approximated miss-MDP using an off-the-shelf POMDP solver. For datasets of limited size, we encounter a practical problem: For an observation z with $\#_{\mathcal{D}}(s, f_{\text{miss}}(z)) = 0$, for any $s \in S$ we obtain $\widehat{M}(z \mid s) = 0$, leading to a division by zero for s when performing the belief update τ. We circumvent this case by setting $\#_{\mathcal{D},\kappa}(s, r) = \#_{\mathcal{D}}(s, r) + \kappa$, i.e. we add a small $\kappa > 0$ to every count. The influence of κ diminishes with an increasing dataset size $|\mathcal{D}|$.

6 Experiments

Our empirical study addresses the following questions:

Q1. Do the proposed methods provide adequate approximations of the missingness function?
Q2. How does the correctness of the assumption on the missingness function affect the approximation?
Q3. As the amount of data increases, does the value of the policy computed on the approximated miss-MDP converge to the optimal value of the true miss-MDP?
Q4. How does the value computed from the approximated miss-MDP compare against baselines that do not estimate the missingness function?

Benchmarks. We consider two environments with varying types of missingness: (1) *ICU*, a benchmark that models a doctor treating a patient, whose vital measurements are not always available [17, 35–37], and (2) *Predator*, a variant of the Tag benchmark [30], where a predator is chasing a

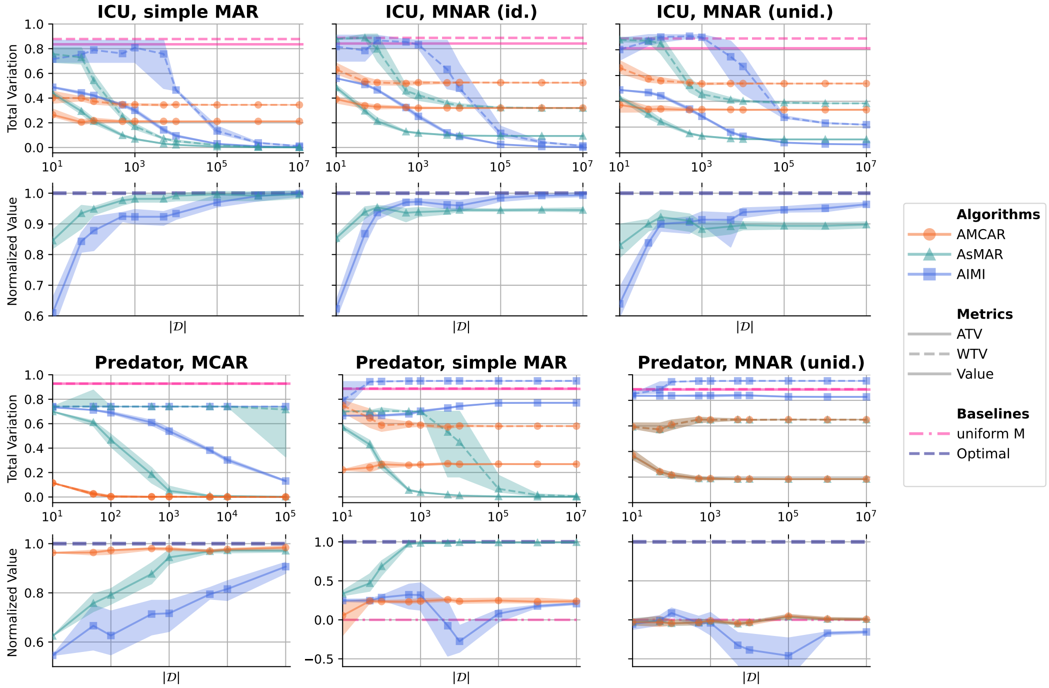


Figure 3: Empirical results for the *ICU* (top) and *Predator* (bottom) benchmarks, including average/-worst total variation (ATV/WTV) and normalized values. Policy values are normalized such that 1 and 0 correspond to values of the optimal policy with known M and the uniform baseline with an $\widehat{M}$ that assigns equal probability to each observation, respectively.

partially hidden prey. To answer Q2, we consider for our benchmarks a selection of the following four missingness functions: (1) *MCAR*, (2) *sMAR*, a simple MAR function, (3) *MNAR (id.)*, an identifiable MNAR function without self-censoring that satisfies the positivity assumption, and (4) *MNAR (unid.)*, an unidentifiable MNAR function with self-censoring. In the *Predator* benchmark, for all missingness functions, the (x, y)-coordinates of the prey can only go missing jointly, i.e. the missingness indicators are dependent; in the *ICU* benchmark, the missingness indicators are always independent. The implementation of the proposed algorithm and the benchmarks are publicly available.[2] For details on the benchmarks, see Appendix C.

Protocol, algorithms, and baselines. For a range of dataset sizes $|\mathcal{D}|$, we collect data using the uniform random policy π^{rnd} where $\forall a \colon \pi^{\text{rnd}}(a \mid \cdot) = 1/|A|$, and compute the estimate $\widehat{M} \approx M$ using our proposed algorithms: `AMCAR` (●), `AsMAR` (▲), and `AIMI` (■) (Section 5). Each $\widehat{M}$ yields an approximated miss-MDP $\widehat{\mathcal{P}}$, for which we compute a policy $\hat{\pi}$ using the POMDP solver SARSOP [18]. To assess the efficacy of our approach, we consider the following baselines: **(1)** *optimal*: the SARSOP policy π^* computed for the true M (the upper bound); **(2)** *uniform* M: the SARSOP policy π^{M_u} computed for M_u, a guess of M that is uniform, where every feature independently goes missing with probability 0.5.

Metrics. For every dataset size and method, we perform 20 independent runs and report the average together with the interquartile range (shaded area) of the following metrics.

1. To assess the quality of the approximation $\widehat{M}$ compared to M for a miss-MDP $\mathcal{P}$, we compute the *total variation* (TV) of the distributions at a state $s \in S$ as $TV(s) = \frac{1}{2}\sum_{z \preceq s} \left|\widehat{M}(z \mid s) - M(z \mid s)\right|$. We aggregate the TV across states by the average TV (ATV): $1/|S| \sum_s TV(s)$, and the worst TV (WTV): $\max_s TV(s)$.

[2] https://github.com/ai-fm/missingness-pomdps

2. We are interested in how the values $V_{\mathcal{P}}(\hat{\pi})$ on the true miss-MDP $\mathcal{P}$ of the various $\hat{\pi}$ computed by the algorithms described above compares to the optimum $V_{\mathcal{P}}(\pi^*)$. All policy values are normalized s.t. 1 and 0 correspond to the values of the *optimal* and *uniform* baselines, respectively.

Results. Figure 3 presents the experimental evaluation. It shows how the TV of $\widehat{M}$ and the value of the associated $\hat{\pi}$ evolve with dataset size $|\mathcal{D}|$. Next, we discuss the research questions based on these results.

Q1: With a sufficient amount of data and the correct assumptions, the proposed algorithms adequately approximate the missingness function. We observe that under the appropriate assumptions, each algorithm can learn the corresponding missingness function (bringing the TV to zero): AMCAR learns the exact missingness function in *Predator_{MCAR}* with 100 or more observations. We observe similar results for AsMAR (in *ICU_{sMAR}* and *Predator_{sMAR}*), as well as for AIMI (in *ICU_{MNAR (id.)}*).

Q2: The assumptions on the missingness function significantly affect the quality of the approximation. On the one hand, relaxing the assumptions on the missingness function ensures it can be learned, though this comes at the cost of reduced sample efficiency. For example, in *Predator_{MCAR}*, we observe that AsMAR and AIMI require orders of magnitude more data to learn the missingness function than AMCAR. On the other hand, making stronger assumptions can lead to failures: for example, AMCAR converges to an incorrect missingness function in all benchmarks except *Predator_{MCAR}*. The results also show that in some cases, the algorithms might approximate the missingness function even if it does not satisfy the assumptions required for PAC guarantees, as demonstrated from the results of AIMI on *ICU_{MNAR (unid.)}*.

Q3: The convergence to the optimal policy follows the quality of the approximation, and, therefore, the convergence of the resulting policy to the optimum. When the approximation is sufficiently accurate, the value of the policy found by using our methods converges to the optimal value.

Q4: The baseline is not able to compute a policy with a value that is competitive with the values of the policies following from our methods. In all cases, the baseline algorithm fails to approximate the true M and the produced polices π^{M_u} significantly lag behind the polices computed by our algorithms. The only exception is *Predator_{MNAR (unid.)}*, where our algorithms also fail. This shows that unidentifiable MNAR processes that violate the independence assumption present the true challenge for our approach.

7 Conclusion

We introduce miss-MDPs to integrate the theory of missing data into decision-making under uncertainty. Given a dataset of trajectories generated from a miss-MDP, we approximate the unknown missingness function, which—under certain assumptions about the missingness function—enables the computation of an ε-optimal policy. We demonstrate that incorrect assumptions about the missingness mechanism can result in misspecified models and suboptimal policies. Interestingly, we show that for certain missingness functions, belief updates can be computed without knowledge of the missingness function—mirroring the notion of ignorability from the missing data literature. Our experiments support the theoretical results and demonstrate the practical benefits of our contributions. Future work will explore lifting the assumption of a known transition function and extending miss-MDPs to the more general setting of miss-POMDPs.

Acknowledgments

This work was supported by the European Research Council (ERC) Starting Grant 101077178 (DEUCE), and the Czech Science Foundation grant GA23-06963S (VESCAA).

References

[1] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994.

[2] Richard D Smallwood and Edward J Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Oper. Res.*, 21(5):1071–1088, 1973.

[3] Karl Johan Åström. Optimal control of Markov processes with incomplete state information. *J. Math. Anal. Appl.*, 10(1):174–205, 1965.

[4] Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? In *COLT*, volume 178 of *Proceedings of Machine Learning Research*, pages 5175–5220. PMLR, 2022.

[5] Jonathan Lee, Alekh Agarwal, Christoph Dann, and Tong Zhang. Learning in POMDPs is sample-efficient with hindsight observability. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 18733–18773. PMLR, 2023.

[6] Joseph L. Schafer and John W. Graham. Missing data: our view of the state of the art. *Psychological Methods*, 7(2):147–177, June 2002.

[7] Stef van Buuren. *Flexible imputation of missing data*. CRC Press, Taylor and Francis Group, 2018.

[8] Roderick Little and Donald Rubin. *Statistical Analysis with Missing Data, Third Edition*. Wiley Series in Probability and Statistics. Wiley, 2019.

[9] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[10] Yuhui Wang, Hao He, and Xiaoyang Tan. Robust reinforcement learning in POMDPs with incomplete and noisy observations. *CoRR*, abs/1902.05795, 2019.

[11] Luchen Li, Matthieu Komorowski, and Aldo A. Faisal. The actor search tree critic (ASTC) for off-policy POMDP learning in medical decision making. *CoRR*, abs/1805.11548, 2018.

[12] Markus Böck, Julien Malle, Daniel Pasterk, Hrvoje Kukina, Ramin Hasani, and Clemens Heitzinger. Superhuman performance on sepsis MIMIC-III data by distributional reinforcement learning. *PLOS ONE*, 17(11):e0275358, November 2022.

[13] Daniel J Lizotte, Lacey Gunter, Eric Laber, and Susan A Murphy. Missing data and uncertainty in batch reinforcement learning. In *NeurIPS*, 2008.

[14] Zeyu Liu, Anahita Khojandi, Xueping Li, Akram Mohammed, Robert L Davis, and Rishikesan Kamaleswaran. A Machine Learning–Enabled Partially Observable Markov Decision Process Framework for Early Sepsis Prediction. *INFORMS Journal on Computing*, 34(4):2039–2057, July 2022.

[15] Nobuhiko Yamaguchi, Osamu Fukuda, and Hiroshi Okumura. Model-based reinforcement learning with missing data. In *CANDAR (Workshops)*, pages 168–171, 2020.

[16] Joseph Futoma, Michael C. Hughes, and Finale Doshi-Velez. POPCORN: Partially Observed Prediction COnstrained ReiNforcement Learning, March 2020.

[17] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV, 2022.

[18] Hanna Kurniawati, David Hsu, and Wee Sun Lee. SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In *RSS*. MIT Press, 2008.

[19] Anastasios A Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer, 2006.

[20] Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. *NeurIPS*, 26, 2013.

[21] Ilya Shpitser, Karthika Mohan, and Judea Pearl. Missing Data as a Causal and Probabilistic Problem. In *UAI*, pages 802–811, 2015.

[22] Rohit Bhattacharya, Razieh Nabi, Ilya Shpitser, and James M. Robins. Identification In Missing Data Models Represented By Directed Acyclic Graphs. In *UAI*, volume 115 of *Proceedings of Machine Learning Research*, pages 1149–1158, 2020.

[23] Razieh Nabi, Rohit Bhattacharya, and Ilya Shpitser. Full Law Identification in Graphical Models of Missing Data: Completeness Results. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 7153–7163, 2020.

[24] Minshuo Chen, Yu Bai, H. Vincent Poor, and Mengdi Wang. Efficient RL with impaired observability: Learning to act with delayed and missing state observations. In *NeurIPS*, 2023.

[25] Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint active feature acquisition and classification with variable-size set encoding. In *NeurIPS*, pages 1375–1385, 2018.

[26] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. ASAC: active sensing using actor-critic models. In *MLHC*, volume 106 of *Proceedings of Machine Learning Research*, pages 451–473. PMLR, 2019.

[27] Matthijs T J Spaan. Partially Observable Markov Decision Processes. In Marco Wiering and Martijn van Otterlo, editors, *Reinforcement Learning: State-of-the-Art*, pages 387–414. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[28] Omid Madani, Steve Hanks, and Anne Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artif. Intell.*, 147(1-2):5–34, 2003.

[29] Milos Hauskrecht. Value-function approximations for partially observable Markov decision processes. *JAIR*, 13:33–94, 2000.

[30] Joelle Pineau, Geoffrey J. Gordon, and Sebastian Thrun. Point-based value iteration: An anytime algorithm for POMDPs. In *IJCAI*, pages 1025–1032, 2003.

[31] Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *J. Am. Stat. Assoc.*, 116(534):1023–1037, 2021.

[32] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[33] Judea Pearl. *Causality*. Cambridge University Press, 2nd edition, 2009.

[34] Miguel A Hernán and James M Robins. *Causal Inference: What If*. CRC Press, 2020.

[35] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178, 2018. ISSN 2052-4463.

[36] Patrick J. Thoral, Jan M. Peppink, Ronald H. Driessen, Eric J. G. Sijbrands, Erwin J. O. Kompanje, Lewis Kaplan, Heatherlee Bailey, Jozef Kesecioglu, Maurizio Cecconi, Matthew Churpek, Gilles Clermont, Mihaela van der Schaar, Ari Ercole, Armand R. J. Girbes, and Paul W. G. Elbers. Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example*. *Critical Care Medicine*, 49(6), 2021.

[37] Stephanie L. Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, Marc Zimmermann, Dean Bodenham, Karsten Borgwardt, Gunnar Rätsch, and Tobias M. Merz. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3): 364–373, 2020.

[38] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: a modern approach. Fourth Edition.* Pearson Education Limited, 2022.

[39] Jacob Bernoulli. *Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola galliceé scripta de ludo pilae reticularis*. Impensis Thurnisiorum, fratrum, 1713.

[40] Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, and Ludolf Erwin Meester. *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media, 2005.

[41] Masashi Okamoto. Some inequalities relating to the partial sum of binomial probabilities. *Annals of the Institute of Statistical Mathematics*, 10(1):29–35, 1959.

[42] Carlos E. Budde, Arnd Hartmanns, Tobias Meggendorfer, Maximilian Weininger, and Patrick Wienhöft. Sound statistical model checking for probabilities and expected rewards. In *TACAS (1)*, volume 15696, pages 167–190. Springer, 2025.

[43] Przemyslaw Daca, Thomas A. Henzinger, Jan Kretínský, and Tatjana Petrov. Faster statistical model checking for unbounded temporal properties. *ACM Trans. Comput. Log.*, 18(2):12:1–12:25, 2017.

[44] Tobias Meggendorfer, Maximilian Weininger, and Patrick Wienhöft. Solving robust markov decision processes: Generic, reliable, efficient. In *AAAI*, volume 39, pages 26631–26641. AAAI Press, 2025.

[45] Tobias Meggendorfer, Maximilian Weininger, and Patrick Wienhöft. Solving robust markov decision processes: Generic, reliable, efficient. *CoRR*, abs/2412.10185, 2024.

A Proofs for Section 4: Probabilities of the Missingness Function

Lemma 1. If the missingness function M is MAR, then $\forall z \in Z$, $\exists p \in [0,1]$, $\forall s \in S$, $M(z \mid s) = [z \preceq s] \cdot p$.

Proof. Suppose that M is MAR. The lemma states that $\forall z \in Z$, $\exists p \in [0,1]$, $\forall s \in S$, $M(z \mid s) = p$ if $z \preceq s$ and otherwise $M(z \mid s) = 0$. Since $z \not\preceq s \Rightarrow M(z \mid s) = 0$, we only need to show $\forall z \in Z, \exists p \in [0,1], \forall s \in S, z \preceq s \Rightarrow M(z|s) = p$, which directly follows from the MAR assumption.

□

Remark 2. Lemma 1 implies that the missingness function can be omitted in the belief update. Let $b \in \mathcal{B}$ be a belief, and let $s' \in S$. Then, for any $a \in A$ and $z \in Z$, it holds that

$$\begin{aligned} b'(s') &= \tau(b,a,z)(s') \\ &:= \frac{M(z \mid s') \sum_{s \in S} T(s' \mid s,a) b(s)}{\sum_{s'' \in S} M(z \mid s'') \sum_{s \in S} T(s'' \mid s,a) b(s)} && \text{(By definition of belief update)} \\ &= \frac{[z \preceq s'] \cdot p \sum_{s \in S} T(s' \mid s,a) b(s)}{\sum_{s'' \in S} [z \preceq s''] \cdot p \sum_{s \in S} T(s'' \mid s,a) b(s)} && \text{(By Lemma 1)} \\ &= \frac{[z \preceq s'] \sum_{s \in S} T(s' \mid s,a) b(s)}{\sum_{s'' \in S} [z \preceq s''] \sum_{s \in S} T(s'' \mid s,a) b(s)}. && \text{(p cancels out)} \end{aligned}$$

Therefore, the probabilities of M do not affect the resulting probabilities of the belief update. In particular, this means that maintaining a belief while executing a miss-MDP does not require knowledge of M.

Still, we stress again that one needs M to compute an optimal policy because this requires constructing and solving the belief MDP (see [38, Chapter 16.4.1]), which in turn requires knowing the probability $\mathbb{P}(b' \mid b,a)$ of going to a successor belief b' from a current belief $b \in \mathcal{B}$ upon playing action $a \in A$. Concretely, the probability of a successor belief $b' = \tau(b,a,z)$ depends on the probability of $z \in Z$ given b and a, which in turn depends on M,

$$\mathbb{P}(b' \mid b,a) = \sum_{z \in Z} \mathbb{P}(z \mid b,a)[b' = \tau(b,a,z)],$$

$$\mathbb{P}(z \mid b,a) = \sum_{s \in S} b(s) \sum_{s' \in S} T(s' \mid s,a) M(z \mid s').$$

Here, no normalization occurs, and the probabilities of M do not cancel out.

B Proofs for Section 5: Probably Approximately Correct

This appendix is about proving that given enough data, we can approximate the missingness function to arbitrary precision ε, or the other way round: we can prove a certain precision ε for any given dataset $\mathcal{D}$. In both directions, we provide a probabilistic guarantee, i.e. that the result is correct with probability at least δ. The reason the guarantee has to be probabilistic is that our knowledge relies on a sampled dataset, and, intuitively, there always is a chance that we were "unlucky" and received a very unlikely sequence of samples from which we infer a wrong approximation.

Outline. First, in Appendix B.1 we recall standard notions from statistics literature: Bernoulli processes and the fact that building on Okamoto's inequality, we can obtain a size for our dataset $\mathcal{D}$ given precision ε and confidence δ (or, analogously, obtain a precision ε given $\mathcal{D}$ and δ). Afterwards, Appendix B.2 and Appendix B.3 provide the proofs of Theorems 1 and 2, respectively, i.e. the guarantees for our algorithms. Moreover, they prove the guarantees for the modified algorithms when using more information about the missingness function. Finally, Appendix B.4 proves Theorem 3, our main result that ε-policies can be computed.

B.1 Bernoulli processes

Definition 6 (Bernoulli process [39], [40, Chapter 4.3])**.** A Bernoulli process is a sequence of binary random variables that are independent and identically distributed. All random variables have probability p to yield a 1, and probability $1-p$ to yield a 0.

Throughout this appendix, we write n for the length of the sequence of a Bernoulli process, and k for the number of successes, i.e. the number of times it yielded a 1. Moreover, we denote by $\hat{p} = \frac{k}{n}$ the empirical success probability. Okamoto's seminal work proves the following property of estimating p through observing a Bernoulli process:

Theorem 4 (Okamoto's inequality [41])**.** For a Bernoulli process with n repetitions and k successes and a given precision ε, we have

$$\Pr(\hat{p} - p \geq \varepsilon) \leq \mathsf{e}^{-2\cdot n\cdot \varepsilon^2} \text{ and } \Pr(\hat{p} - p \leq \varepsilon) \leq \mathsf{e}^{-2\cdot n\cdot \varepsilon^2}.$$

For our guarantees, we want that $\Pr(|\hat{p} - p| \geq \varepsilon) \leq 1 - \delta$, i.e. that our estimate $\hat{p}$ is ε-precise with probability at least δ. Thus, distributing our confidence symmetrically, we insert $\frac{1-\delta}{2}$ on the left side of Okamoto's inequalities. Then, we can solve the inequation for δ, ε, or n:

$$\frac{1-\delta}{2} \leq \mathsf{e}^{-2\cdot n\cdot \varepsilon^2} \Leftrightarrow \varepsilon \geq \sqrt{\frac{\ln(\frac{2}{1-\delta})}{2\cdot n}} \Leftrightarrow n \geq \frac{\ln(\frac{2}{1-\delta})}{2\cdot \varepsilon^2}. \tag{3}$$

In other words, given two of precision ε, confidence δ, and number of repetitions n, we can infer the third. We remark that there exist other inequalities similar to Okamoto's that yield the same result, but with tighter bounds; we refer to [42, Section 3] for a discussion. However, as our goal is only to prove the existence of a bound, we choose the conservative Okamoto bound for its easier accessibility.

B.2 PAC guarantees for AsMAR

Theorem 1. Let $\mathcal{P}$ be a missingness-MDP where the missingness function is simple MAR. For every given precision ε and confidence threshold δ, there exists a number n^* of histories, such that a dataset $\mathcal{D}$ of n^* histories has the following property: With probability at least δ, $\widehat{M}$ computed on $\mathcal{D}$ according to Equation (1) satisfies that for all reachable states $s \in S$ and observations $z \in Z$, we have $\widehat{M}(z \mid s) = M(z \mid s) \pm \varepsilon$. Dually, given a dataset $\mathcal{D}$ and confidence threshold δ, we can compute an ε such that with probability at least δ, for all states $s \in S$ and observations $z \in Z$, we have $\widehat{M}(z \mid s) = M(z \mid s) \pm \varepsilon$.

Proof. **Proof outline.** We first show that the computation of every $\widehat{M}(z \mid s)$ is related to a Bernoulli process. Then, using the results of Appendix B.1, we can prove the claims of the theorem for individual state-observation pairs. Next, we lift this to all state-observation pairs by distributing the confidence δ. Finally, we individually explain how this yields the two claims of the theorem.

The Bernoulli process related to $\widehat{M}(z \mid s)$. Fix a state $s \in S$ and an observation $z \in Z$. Consider the following random variable: Sample a state $s' \in S$ and the corresponding observation $z' \in Z$. Set the random variable to 1 if $\forall i \in I \colon (i \in I_{\text{always}} \implies z'_i = s_i) \land (f_{\text{miss}}(z)_i = 0 \implies z'_i = \bot)$; set the random variable to 0 if $\forall i \in I \colon (i \in I_{\text{always}} \implies z'_i = s_i)$; and ignore the sampled (s', z') otherwise, i.e. if $\exists i \in I \colon (i \in I_{\text{always}} \land z'_i \neq s_i)$. Note that the random variable is 1 exactly when the sample would be counted by $\#_{\mathcal{D}}(s, f_{\text{miss}}(z))$, and the sample is not ignored exactly when it would be counted by $\sum_{r \in R} \#_{\mathcal{D}}(s, r)$.

We require that the probability of the random variable being 1 is equal among all sampled state-observation pairs (s', z') that are not ignored by it, and moreover we require this probability to be equal to $M(z \mid s) = M(f_{\text{miss}}(z) \mid s) =: p$. To prove this, we use the assumption that M is a simple MAR missingness function; thus, we know that for all s' that agree with s on all always observable features (formally: $\forall i \in I \colon (i \in I_{\text{always}} \implies z'_i = s_i))$, we have $p = M(f_{\text{miss}}(z) \mid s) = M(f_{\text{miss}}(z) \mid s')$.

We have just shown that the random variable we constructed is a Bernoulli process with success probability $p = M(z \mid s)$, with the number of repetitions $n = \sum_{r \in R} \#_{\mathcal{D}}(s, r)$ and the number of successes $k = \#_{\mathcal{D}}(s, f_{\text{miss}}(z'))$. Note that the definition of $\widehat{M}$ in Equation (1) is exactly the empirical success probability $\hat{p} = \frac{k}{n}$.

Observe that we do not need a separate Bernoulli process for every state-observation pair: The number of repetitions $\sum_{r \in R} \#_{\mathcal{D}}(s, r)$ is independent of the observation z, since that only affects whether it is counted as success or not. Further, it suffices to have one random variable per combination of valuation for the features in I_{always}, since all states that agree on the always observable features yield the same Bernoulli process. Moreover, we do not need to consider every observation z (as this includes observations that do not admit s), but rather only every missingness indicator vector $r \in R$. In the following, we still write "Every state-observation pair" instead of "Every pair of set of states that agree on the always observable features and missingness indicator vector", as it is also true and more concise.

Single state-observation pair. Consider the Bernoulli process just described for a fixed state-observation pair (s, z). We explain how to use the results of Appendix B.1 towards proving the first and second claim of the theorem:

- First claim: By the third variant of Equation (3), we have that given a precision ε and confidence threshold $\delta_{s,z}$, we can compute a necessary number of samples $n_{s,z}$ such that we obtain the PAC guarantee for this state-observation pair.
- Second claim: Observe that a given dataset $\mathcal{D}$ corresponds to a number of repetitions of every Bernoulli process. Let $n_{s,z}$ be the number of repetitions for the pair (s, z). Thus, using the second variant of Equation (3), we have that given $\mathcal{D}$ (and thus $n_{s,z}$) and a confidence threshold $\delta_{s,z}$, we can compute a precision $\varepsilon_{s,z}$ such that we obtain the PAC guarantee for this state-observation pair.

All state-observation pairs. We can split the given confidence threshold δ uniformly over all state-observation pairs, i.e. for every $s \in S$, $z \in Z$, we have $\delta_{s,z} = \frac{\delta}{|S| \cdot |Z|}$. Then, by the union bound, the probability of all state-observation pairs being correctly estimated is the sum of all $\delta_{s,z}$, which (since we distributed it uniformly) is δ. By splitting the confidence threshold in this way, we can obtain the PAC guarantee for all state-observation pairs.

Second claim. We first provide the full argument for the second claim, as it is simpler. Given the dataset $\mathcal{D}$ and confidence threshold δ, we obtain an $\varepsilon_{s,z}$ for all state-observation pairs. The probability that all of these are correct is at least δ. We obtain the claim by taking the maximum over these, i.e. setting $\varepsilon := \max_{s \in S, z \in Z} \varepsilon_{s,z}$. Then we have that with probability at least δ, for all states $s \in S$ and observations $z \in Z$, we have $\widehat{M}(z \mid s) = M(z \mid s) \pm \varepsilon$.

First claim. We proceed in two steps: We explain the analogous argument to the second claim, based on an assumption on the dataset. Afterwards, we explain how this assumption on the dataset can be satisfied.

Assume that for every state-observation pair (s, z), the dataset $\mathcal{D}$ contains at least $n_{s,z}$ samples, i.e. the number computed using Equation (3) inserting ε and $\delta_{s,z}$. Then, analogously to the proof of the second claim, computing $\widehat{M}$ using this dataset satisfies that with probability at least δ, for all states $s \in S$ and observations $z \in Z$, we have $\widehat{M}(z \mid s) = M(z \mid s) \pm \varepsilon$.

It remains to show that there exists a number n^* such that a sampled dataset of n^* histories has the required property. For this, we have to spend some of our confidence threshold δ, since we can only guarantee the property with a certain probability; there is the chance that even upon sampling n^* histories, we are unlucky and some state-observation pair has not been sampled often enough. Thus, we split δ as follows: $\delta_{\mathcal{D}}$ is used to guarantee the property of the dataset, and $\delta_{\widehat{M}}$ is used to guarantee the consequential property of $\widehat{M}$. Thus, $\delta_{s,z}$ above are obtained by uniformly distributing $\delta_{\widehat{M}}$, not all of δ. Then, by the union bound, the probability that $\mathcal{D}$ has the desired property and that the PAC guarantee holds is $\delta_{\mathcal{D}} + \delta_{\widehat{M}} = \delta$.

We now need to show that there exists an n^* such that a dataset of this size contains the required number of samples with probability at least $\delta_{\mathcal{D}}$. Recall that the dataset is sampled using a fair policy, which means that every state has a positive probability to be visited; thus (assuming that the length of every history is at least as large as the number of states in the miss-MDP), there exists a minimum probability m such that every state is visited with at least probability m in every history. Moreover, observe that for a state-observation pair (s, z), the number of samples for its Bernoulli process is at least the number of times s has been visited; this is because a sample is used when it agrees with s on the always observable features. Thus, for every sampled history, we have a probability of at least m to obtain at least one sample for (s, z). This lower bound on the number of samples for (s, z) is binomially distributed with success probability m [40, Chapter 4.3]. Thus, there exists a number of histories n^* such that the probability of having at least $n_{s,z}$ samples for (s, z) when sampling at least n_h histories is greater than $\delta_{\mathcal{D}}$. As before, this argument was for a single state-observation pair; thus, $\delta_{\mathcal{D}}$ is also uniformly distributed over all state-observation pairs.

Summarizing the above: There exists a number n^*, such that with probability $\delta_{\mathcal{D}}$, a dataset consisting of n^* histories contains at least $n_{s,z}$ samples for every state-observation pair (s, z), where $n_{s,z}$ is the number computed using Equation (3) inserting ε and $\delta_{s,z}$. Consequently, $\widehat{M}$ using this dataset satisfies that with probability at least $\delta_{\widehat{M}}$, for all states $s \in S$ and observations $z \in Z$, we have $\widehat{M}(z \mid s) = M(z \mid s) \pm \varepsilon$. Together, probably (with probability at least $\delta = \delta_{\mathcal{D}} + \delta_{\widehat{M}}$), we can guarantee that $\widehat{M}$ is approximately correct.

□

Proposition 1. The improvements described in Section 5.1 for using knowledge retain the PAC guarantees stated in Theorem 1.

Proof. The improvements use the fact that the underlying Bernoulli process in fact does not depend on all features in I_{always}. While it is correct to still split on these variables, obtaining two processes with the same true success probability, we can also merge them.

More formally, observe that if feature i does not affect the missingness probability of other features, for all valuations of feature i, the corresponding Bernoulli processes have the same success probability. MCAR missingness functions are the most extreme case of this, where the given state is completely irrelevant and it suffices to have one Bernoulli process per missingness indicator vector. As a side note: Observe that it is indeed necessary to consider every missingness indicator vector and not individual features, since the missingness probabilities need not be independent. □

B.3 PAC guarantees for AIMI (Section 5.2)

Theorem 2 (PAC guarantee for `AIMI`)**.** Let $\mathcal{P}$ be a missingness-MDP where the missingness function satisfies independence, non-self-censoring, and non-sure missing. Then, the same PAC guarantees hold as specified for `AsMAR` in Theorem 1 but with $\widehat{M}$ computed using Equation (2).

Proof. This proof is analogous to that of Theorem 1: every missingness probability computed by Equation (2) corresponds to the empirical success probability of a Bernoulli process, which allows to apply the results from Appendix B.1. This proof differs in the argument why all states grouped together in the same Bernoulli process have the same success probability, and in the argument why it feasible to sample a dataset of the necessary size.

By the independence assumption, we know that it suffices to learn every individual $\mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{z} \sim M(s))$ for each $i \in I$. By non self-censoring, we know that this probability depends only on features in $I \setminus \{i\}$. Thus, the counter $\#(s, i, 0)$ counts exactly the successes of a Bernoulli process with success probability $\mathbb{P}(\boldsymbol{z}_i \mid \boldsymbol{z} \sim M(s))$, and $\#(s, i, 1)$ counts the failures.

It only remains to argue that a sufficient dataset can be feasibly obtained. For this, we use the assumption that no feature is missing surely. In other words, every feature has a positive probability to be observed. Thus, every reachable states has a positive probability m to be fully observed. Using this, we can repeat the argument from the proof of Theorem 1. □

Proposition 2. The improvements described in Section 5.1 for using knowledge retain the PAC guarantees stated in Theorem 2.

Proof. (a) If we know from an m-graph that a particular feature i is not influenced by feature j, for all valuations of j the Bernoulli process has the same success probability. Thus, we can merge these Bernoulli processes and ignore feature j.

(b) If we know the missingness function is simple MAR and feature j goes missing, we know that it cannot influence the missingness probability of any other feature by definition [31]. Then, the proof is the same as in Case (a).

(c) If the missingness function is MCAR, we know that no feature influences the missingness probability of any other feature. Thus, we can repeatedly apply the argument of Case (a) to merge all Bernoulli processes until we have one for every feature. □

B.4 Computing ε-optimal Policies (Section 5.3)

Theorem 3 (Computing ε-optimal Policies)**.** Let $\mathcal{P}$ be a miss-MDP with a missingness function that is simple MAR or that satisfies independence, no self-censoring, and positivity. Assume we can sample histories collected under a fair policy, and we know a lower bound on the smallest missingness probability $p \leq \min_{s \in S, z \in Z} M(z \mid s)$. Then, for every given precision ε and confidence threshold δ, we can in finite time compute a policy π^* such that with probability at least δ it is ε-optimal, i.e. $|\sup_\pi(V_\mathcal{P}(\pi)) - V_\mathcal{P}(\pi^*)| \leq \varepsilon$.

Proof. **Sampling the dataset.** We have sampling access with a fair policy, so every state has positive probability to be visited. Thus, for any finite number n, we can almost surely obtain n samples of every state s in finite time. For the Bernoulli process underlying Equation (1), and if the missingness function is simple MAR, this suffices to guarantee that for every state-observation pair, we can obtain the number of samples $n_{s,z}$ required for achieving precision ε with confidence $\delta_{s,z}$. Similarly, for the Bernoulli process underlying Equation (2), and if the missingness function satisfies positivity, we can also obtain the required number of samples for every state-observation pair. Overall, under the assumptions of the theorem, we can almost surely obtain a dataset in finite time such that it suffices to give PAC guarantees on every state-observation pair.

We remark that this does not even require spending confidence budget as we did in the proofs of Theorems 1 and 2, since there we required to get this dataset within a certain number of histories n^*. Here, we only claim that we can get a sufficient dataset in finite time almost surely.

Obtaining $\widehat{M}$. The assumptions on the missingness function in the statement of the theorem match those in Theorem 1 or Theorem 2. Hence, given the dataset described in the previous paragraph, we can approximate $\widehat{M}$ in a way such that with probability δ, it is ε_M-precise. Note that here we do not employ the full allowed imprecision ε, but rather a smaller $\varepsilon_M < \varepsilon$, since there will be other sources of error.

M **and** $\widehat{M}$ **qualitatively agree.** For our technical reasoning, we require that $M(z \mid s) = 0$ if and only if $\widehat{M}(z \mid s) = 0$. We prove both directions separately: If $M(z \mid s) = 0$, then we never observe a sample for z when given s, and thus $\widehat{M}(z \mid s) = 0$, as it uses an empirical average (Equations (1) and (2)). If $M(z \mid s) > 0$, as we use a fair sampling process, we almost surely eventually observe z when given s, and consequently the empirical average is positive, i.e. $\widehat{M}(z \mid s) > 0$.

It remains to prove that we can *in finite time* conclude that M and $\widehat{M}$ qualitatively agree. This means that we need to be sufficiently certain that if $\widehat{M}(z \mid s) = 0$, this is because indeed $M(z \mid s) = 0$ and not just because we haven't sampled enough yet. For this, we use a proof technique employed in, e.g., [43]: We utilize knowledge of (a lower bound on) the smallest missingness probability p. Further, recall that the confidence threshold δ is distributed over all Bernoulli processes (see Appendices B.2 and B.3). Thus, for each Bernoulli process, we have a confidence threshold $\delta_{s,z}$. Okamoto's inequality (see Appendix B.1) provides an upper bound on the missingness probability that is correct with probability at least $\delta_{s,z}$. Thus, when this upper bound is less than p, we can conclude with sufficient confidence that $\widehat{M}(z \mid s) = 0$.

Utilizing Lemma 2. Let $\widehat{\mathcal{P}}$ be the approximated missingness-MDP that is exactly $\mathcal{P}$ except for the missingness function, which is $\widehat{M}$ instead of M. We have just proven that in finite time we know that with probability δ, $\widehat{M}$ is ε_M-precise and qualitatively agrees with M. Thus, it satisfies the assumptions specified in Lemma 2, which is proven below. This key technical lemma shows that the values obtained when following a policy π in either the original $\mathcal{P}$ or the approximated $\widehat{\mathcal{P}}$ have a bounded difference.[3] Formally, for every policy π, we have $|V_{\mathcal{P}}(\pi) - V_{\widehat{\mathcal{P}}}(\pi)| \leq f(\varepsilon_M)$, where f is a monotonically increasing function that depends on ε_M, the precision of $\widehat{M}$.

From this, we obtain two facts: Firstly, since this holds for all policies, it also holds for the supremum over all policies, and thus we can bound the difference in the values of the two missingness-MDPs:

$$|\sup_{\pi} V_{\mathcal{P}}(\pi) - \sup_{\pi} V_{\widehat{\mathcal{P}}}(\pi)| \leq f(\varepsilon_M). \tag{4}$$

Secondly, we can apply the same reasoning to a near-optimal policy in $\widehat{\mathcal{P}}$. For this, let $\varepsilon_\pi < \varepsilon$ be a precision smaller than our overall error tolerance, and let π^* be an ε_π-optimal policy in $\widehat{\mathcal{P}}$, i.e.

$$\sup_{\pi}(V_{\widehat{\mathcal{P}}}(\pi)) - V_{\widehat{\mathcal{P}}}(\pi^*) \leq \varepsilon_\pi. \tag{5}$$

We remark that $\widehat{\mathcal{P}}$ is a fully specified missingness-MDP, and thus a fully specified POMDP, for which solver computing ε-optimal policies such as SARSOP [18] exist. Using Lemma 2, we obtain the following inequality:

$$|V_{\mathcal{P}}(\pi^*) - V_{\widehat{\mathcal{P}}}(\pi^*)| \leq f(\varepsilon_M). \tag{6}$$

Combining the inequalities. To conclude the proof, we use a chain of inequalities. Whenever we write ±, this indicates that for one way of resolving the symbol, the inequality holds; this shorthand allows to argue concisely about absolute differences.

$$\begin{aligned}
\sup_{\pi} V_{\mathcal{P}}(\pi) &\leq \sup_{\pi} V_{\widehat{\mathcal{P}}}(\pi) \pm f(\varepsilon_M) && \text{(By Equation (4))} \\
&\leq V_{\widehat{\mathcal{P}}}(\pi^*) + \varepsilon_\pi \pm f(\varepsilon_M) && \text{(By Equation (5))} \\
&\leq V_{\mathcal{P}}(\pi^*) \pm f(\varepsilon_M) + \varepsilon_\pi \pm f(\varepsilon_M) && \text{(By Equation (6))}
\end{aligned}$$

By reordering, we obtain

$$|\sup_{\pi} V_{\mathcal{P}}(\pi) - V_{\mathcal{P}}(\pi^*)| \leq \varepsilon_\pi \pm 2 \cdot f(\varepsilon_M).$$

Hence, since f is a monotonically increasing function, there exists a choice of ε_M and ε_π so that $\varepsilon_\pi \pm 2 \cdot f(\varepsilon_M) \leq \varepsilon$. Intuitively, while the errors incurred by approximating $\widehat{M}$ and by using an approximately optimal policy add up, we can bound the overall maximum error. Thus, we can choose the two precisions so that the overall error criterion is met, and the policy π^* is ε-optimal in the original missingness-MDP (with probability δ; with the remaining probability, our sampling was unlucky and $\widehat{M}$ can differ by more than ε_M). This concludes the proof. □

Lemma 2 (Bounding the Value-Difference between $\mathcal{P}$ and $\widehat{\mathcal{P}}$)**.** Let $\mathcal{P}$ be a missingness-MDP and $\widehat{\mathcal{P}}$ be a missingness-MDP that differs from $\mathcal{P}$ only in its missingness function, where it uses $\widehat{M}$ instead

[3] We highlight that every policy is applicable in both missingness-MDPs, as they only differ in their missingness probabilities, but agree on states, observations, and actions.

of M. Further, assume that for all states $s \in S$ and observations $z \in Z$, we have $M(z \mid s) = 0$ if and only if $\widehat{M}(z \mid s) = 0$, and moreover $M(z \mid s) = \widehat{M}(z \mid s) \pm \varepsilon_M$. Then, for every policy π we have $|V_{\mathcal{P}}(\pi) - V_{\widehat{\mathcal{P}}}(\pi)| \leq f(\varepsilon_M)$, where f is a monotonically increasing function.

Proof. **To uncountable MDPs.** Note that both $\mathcal{P}$ and $\widehat{\mathcal{P}}$ are missingness-MDPs, and thus POMDPs. Thus, for each of them, we can construct an uncountable belief MDP with the same value, called B or $\widehat{B}$, respectively. Intuitively, this is achieved by unrolling step-by-step the observation function and all possible beliefs that the agent can have after an action; the transition probabilities in these uncountable MDPs depend on the missingness functions. For a more extensive description, see [38, Chapter 16.4.1].

To finite MDPs. We consider discounted expected reward, with γ the discount factor and $\varrho_{\max} := \max_{(s,a) \in S \times A} \varrho(s,a)$ the maximum state reward. As the expected reward is a geometric series, we can bound the reward that can be obtained after n steps from above as follows:

$$\sum_{i=n}^{\infty} \gamma^i \cdot \varrho_{\max} = \gamma^n \cdot \varrho_{\max} \cdot \sum_{i=0}^{\infty} \gamma^i = \frac{\gamma^n \cdot \varrho_{\max}}{1-\gamma}.$$

For every arbitrarily small precision $\varepsilon_\gamma > 0$, we can thus obtain an n such that the reward after n steps is less than ε_γ. Let B_{ε_γ} be the finite MDP obtained from B by only considering states that are reachable within n steps, and analogously define $\widehat{B}_{\varepsilon_\gamma}$. (Note that n is the same for both, since it only depends on γ and $\varrho_{\max}$, which is the same for both of them.) The value of these finite belief MDPs differs from the value of the uncountable belief MDPs and thus the original missingness-MDPs by at most ε_γ.

Bounding the difference. Recall that B or $\widehat{B}$ are the same except for their transition functions, which depend on M and $\widehat{M}$, respectively. Still, by assumption of the theorem M and $\widehat{M}$ qualitatively agree, i.e. $M(z \mid s) = 0$ if and only if $\widehat{M}(z \mid s) = 0$. Hence, the graph structure of B or $\widehat{B}$ is the same. Thus, the only difference are small perturbations of individual transition probabilities by at most ε_M.

It remains to show the following: Given two finite MDPs that are the same except for small perturbations of the transition probabilities, but where the supports of the the transition functions are the same, provide a bound on the difference in their value. Such a result exists in the literature, namely in [44], or more precisely in the extended version of that paper [45, Lemma 5]. It remains to show that our setting indeed satisfies the assumptions of [45, Lemma 5].

- "For every closed constant-support RMDP": Their claim applies to robust MDPs that are closed constant-support. A robust MDP is an MDP whose transitions are not probability distributions, but rather sets of possible values, see [44, Section 2]. In our case, instead of considering the concrete MDPs B_{ε_γ} and $\widehat{B}_{\varepsilon_\gamma}$, we consider the robust MDP that arises when considering an ε_M-interval around every missingness probability $M(z \mid s)$. This robust MDP contains both B_{ε_γ} and $\widehat{B}_{\varepsilon_\gamma}$ as instantiations.
- "For every pair of agent and environment policy": An agent policy in this setting is exactly the agent policy in ours, so [45, Lemma 5] applies to all policies. An environment policy is the policy that chooses the instantiation of the transition function, i.e. the exact missingness probabilities from the set of all that differ by at most ε_M in our setting.
- "Total-reward objectives:" [45, Lemma 5] concerns *undiscounted* total-reward or mean payoff objectives. Undiscounted total-reward generalizes discounted expected reward, using the standard construction which adds an edge transitioning with probability γ to a dedicated sink state to every transition. Thus, the lemma is applicable to the objective in our setting.
- "The value function is continuous w.r.t. the environment policy": This is the claim of [45, Lemma 5]. More formally, if the environment chooses missingness probabilities differently with some deviation ε_M, then the deviation in the value between the two instantiations is bounded by some monotonically increasing function $g(\varepsilon_M)$. This is exactly the claim we require, since it means that for all agent policies π and all missingness functions $\widehat{M}$ that are ε_M-close to M, we have $|V_{B_{\varepsilon_\gamma}}(\pi) - V_{\widehat{B}_{\varepsilon_\gamma}}(\pi)| \leq g(\varepsilon_M)$.

  We also argue that g can be effectively computed, as it depends on the size of the state space, the reward function, and the minimum occurring transition probability, all of which are known to us (recall that Theorem 3 assumes knowledge of a lower bound on the minimum missingness probabilities). The concrete way of deriving the distance is provided on [45, page 17].

Putting it all together. Our goal is to show that we can compute an f such that for all policies π we have: $|V_{\mathcal{P}}(\pi) - V_{\widehat{\mathcal{P}}}(\pi)| \leq f(\varepsilon_M)$. The following chain of equations proves our goal:

$$|V_{\mathcal{P}}(\pi) - V_{\widehat{\mathcal{P}}}(\pi)| = |V_B(\pi) - V_{\widehat{B}}(\pi)|$$

(Using the uncountable belief MDPs)

$$\leq |B_{\varepsilon_\gamma}(\pi) - V_{\widehat{B}_{\varepsilon_\gamma}}(\pi)| + \varepsilon_\gamma$$

(Using the finite MDPs; decreasing both values by at most ε_γ increases the difference by at most ε_γ)

$$\leq g(\varepsilon_M) + \varepsilon_\gamma$$

(By bounding the difference).

For simplicity of presentation, we choose $\varepsilon_\gamma = \varepsilon_M$, and thus setting $f(\varepsilon_M) := g(\varepsilon_M) + \varepsilon_M$ concludes the proof. □

C Benchmarks

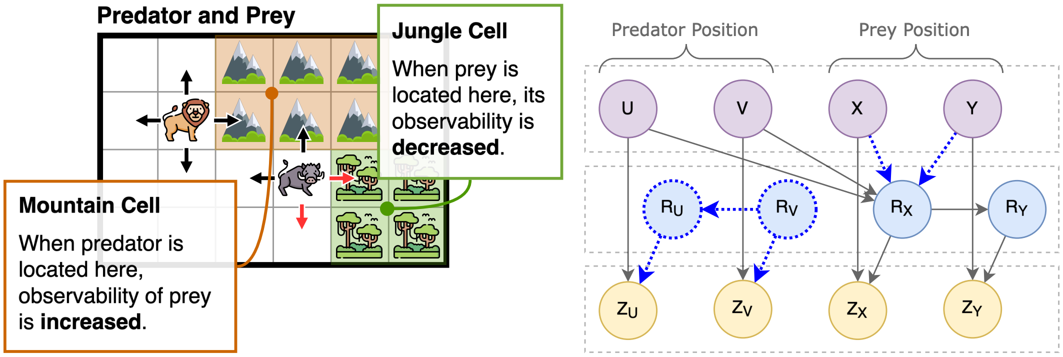
Here we describe our benchmarks. We provide a detailed description of the benchmarks as well as the parameters for running the experiments.

C.1 Description

ICU. This benchmark, inspired by prior clinical decision-making models [17, 35–37], simulates a doctor treating a patient with an infection that progresses stochastically over time. The state of the patient consists of the *infection severity*, the *temperature*, and the *heart rate*. The infection causally influences both the heart rate and the temperature.

The doctor has an option to wait, to administer costly antibiotics that reduce the infection severity, or to order a test, which is a measuring action that may reveal the infection severity. The reward function penalizes high infection levels as well as costly interventions (ordering a test and administering antibiotics). Thus, the doctor's objective is to maintain the patient's infection severity at low levels by administering antibiotics only when necessary. For ease of modeling, the state space also includes the value of the last test ordered.

We evaluate three different missingness functions M, corresponding to distinct missingness functions, illustrated in the m-graph in Figure 5. In all cases, the heart rate and the infection severity may be


Predator and Prey
Jungle Cell
When prey is located here, its observability is **decreased**.
Mountain Cell
When predator is located here, observability of prey is **increased**.
Predator Position
Prey Position
U
V
X
Y
R_U
R_V
R_X
R_Y
Z_U
Z_V
Z_X
Z_Y

(a) (b)

Figure 4: (a) The *Predator* benchmark, where the predator (lion) is the agent trying to catch its prey (boar). Predator and prey can move in all four cardinal directions, where prey chooses an action that increases the distance to the predator (red arrows). (b) The m-graphs for the predator and prey benchmark describing missingness functions of types *simple MAR* (gray), *identifiable MNAR* (gray + blue). Causal dependencies between the state features were omitted for clarity.

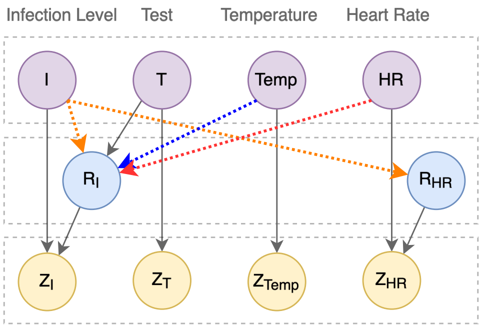


Figure 5: The m-graphs for the ICU benchmark describing missingness functions of types *simple MAR* (gray + blue), *identifiable MNAR* (gray + red) and *unidentifiable MNAR* (gray + red + orange). Causal dependencies between the state features were omitted for clarity.

missing, whereas temperature and the last test ordered are always observed. The success rate of the test that reveals the infection severity may depend on different features, resulting in the following missingness functions. **(1) Simple MAR**, where the success rate only depends on the (always observed) temperature. **(2) MNAR (id.)**, where the success rate only depends on the (not always observed) heart rate, resulting in an identifiable MNAR function without self-censoring and satisfying the positivity assumption. **(3) MNAR (unid.)** is an extension of **MNAR (id.)**, where the infection severity influences the test success rate, introducing self-censoring and thus making the function unidentifiable.

Predator. This benchmark is a variant of the *Tag* benchmark from [30], where an agent (in our case, a predator) is tasked with chasing a partially hidden target (a prey) in a 2D grid environment. The prey senses the predator and usually moves away from it; in case multiple directions lead away from the predator, the prey chooses uniformly at random. The predator's movement is deterministic (dictated by the policy), but moving in an intended direction may randomly fail due to terrain conditions. Predator obtains a flat reward upon catching the prey, and thus the discounting incentivizes catching the prey as soon as possible.

The environment may feature three distinct *biomes* – plains, mountains, or jungles – that influence the predator's observability of the prey, see Figure 4, and thus define the missingness function. We investigate the following three variants thereof. **(1) MCAR**, which features only one type of terrain, i.e., the prey is observed with constant probability. **(2) simple MAR**, where the environment features plains as well as mountains from which the predator has a higher chance of observing its target. **(3) MNAR (unid.)**, where the prey has an option to hide in jungle cells, introducing self-censoring of its position. We stress that when the predator loses track of the prey, both features corresponding to x & y coordinates of the prey go missing simultaneously, modeled by dependencies between missingness indicators R_x & R_y. The dependence between the missingness indicators is a key difference from the ICU benchmark.

C.2 Experimental Setup

Technical Setup. For all experiments, we used high-performance workstations equipped with an AMD Ryzen ThreadRipper PRO 5965WX (24-core, 3.8GHz) CPU, 512 GB ECC DDR4 RAM, and a 2 TB PCIe 4.0 NVMe SSD.

Simulating trajectories. For both benchmarks, we used a discount factor of $\gamma = 0.95$. We considered dataset sizes $|\mathcal{D}| \in \{10, 50, 100, 500, 1\text{E}3, 5\text{E}3, 1\text{E}4, 1\text{E}5, 1\text{E}6, 1\text{E}7\}$. To obtain a dataset containing $|\mathcal{D}|$ samples, we simulated finite trajectories until their lengths summed up to $|\mathcal{D}|$. A trajectory is terminated when it reaches a terminal state (only for the *Predator* bechmark, when the predator catches the prey) or if its length exceeds $L = \left\lceil \log_\gamma \frac{(1-\gamma)\cdot 1\text{E}-3}{\varrho_{\max}} \right\rceil$, where $\varrho_{\max} := \max_{s,a} \varrho(s,a)$. Here, L denotes the smallest integer satisfying $\sum_{k=L}^{\infty} \gamma^k \cdot \varrho_{\max} < 1\text{E} - 3$, i.e. a time step after which the maximum discounted cumulative reward cannot exceed $1\text{E} - 3$. For each dataset size $|\mathcal{D}|$, we generated 20 independent datasets of this size.

Timeouts & precision. For the baselines, we used the timeout of 5 minutes when solving the POMDP (to obtain π^* and π^{M_u}) and the same timeout to evaluate the resulting policy (or π^{rnd}). To obtain a policy $\hat{\pi}$ by solving the corresponding $\widehat{\mathcal{P}}$, we used a timeout of 3 minutes and evaluated $\hat{\pi}$ for 2 minutes. In all cases, solving was additionally allowed to terminate upon reaching the relative precision of 1E-3.