Partially Monitored MDPs

Alireza Kazemipour

July 2025

Abstract

The goal of this document to make connections between the frameworks of partial monitoring, MDPs, and POMDPs. I call this new framework, partially monitored MDP. I'm not convinced nor necessarily a proponent that Mon-MDPs would come to play (due to their particularities), but I don't rule out their usefulness in the future of this project.

I will try to keep the precision under control for now, but it seems from the outset that formalism and precision will be unavoidable. My solution is to remain as clear as possible.

I want to keep the number of references as low as possible, so we pursue a more unified line of thoughts. For doing so, for the parts that are about partial monitoring (bandits like), I keep Lattimore and Szepesvári (2020) as my main reference. For the parts that are related to POMDPs, I keep Ghavamzadeh et al. (2015) as my main reference for now. In particular, I have been advised multiple times, including during ICML, to take a Bayesian approach and my gut feeling also tells me that this project will be under the Bayesian perspective, hence Ghavamzadeh et al. (2015), who studies Bayesian RL will come handy.

The sloppy structure for now, to get the ball rolling, would be revisiting prerequisites and then concepts that we need formally accompanied by an informal explanation in the end to the get the intuition across. I will start with bandits and then POMDPs and eventually get the required inspirations from the both worlds.

1 Notation

$\mathcal{A}$ detonates the set of actions. There are n rounds in the bandit setting. X_t is the reward at round t in the bandit setting. If I said learner somewhere I mean the agent and vice versa. $\mathbb{P}$ would represent the probability measure with respect to the whole (measurable part of) universe of random variables. It needs to be defined, but sometimes I don't define it assumes it somehow exists. Sometime I call it the oracle measure.

2 Bandits: Stochastic Partial Monitoring

In this section I revisit how Lattimore and Szepesvári (2020) defined the problem of *stochastic* partial monitoring. My goal for this review is make sure we understand how this problem is defined in the (simple) bandit settings first.

2.1 Prerequisites

In this section I revisit the background needed to understand and decipher Definition 1.

How do we define a stochastic bandit instance ν? Let P represents a probability distribution. A stochastic bandit instance ν is a collection of distributions where the number of distributions is equal to the number of arms, i.e., $\nu = (P_a : a \in \mathcal{A})$. Now, if P is a parametric distribution with parameter(s) θ —e.g., for a normal distribution with known variance of 1, θ is the mean—, then instead of P, we can use the notation P_θ and define Θ to the set of parameters. In this case the bandit instance ν is defined as: $\nu = (P_{\theta,a} : \theta \in \Theta, a \in \mathcal{A})$. Read the latest expression such that there are $|\mathcal{A}|$ arms with their own distribution and each distribution is parametrized by θ.

What are these sigma algebras? Specifically, $\mathcal{F}, \mathcal{G}, \mathcal{H}$, and $\mathfrak{B}(\mathbb{R})$? We are going to be dealing with the set of parameters Θ, the set of actions $\mathcal{A}$, the set of alphabets Σ and the set of real numbers for rewards $\mathbb{R}$. Since these set are not necessarily finite (especially for $\mathbb{R}$), in order to make sure the (probability) measures we define on these sets are well-defined, we should only focus on some specific subsets of them. These sigma algebras represents those safe subsets. Specifically, $\mathfrak{B}(\mathbb{R})$ denotes the Borel sigma algebra which is the set of open intervals in $\mathbb{R}$.

What is a probability kernel? I take this from Lattimore and Szepesvári (2020, pp. 48).

Let $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ be measurable spaces [meaning we can assign (probability) measures to the elements of $\mathcal{F}$ and $\mathcal{G}$]. A probability kernel from $(\mathcal{X}, \mathcal{F})$ to $(\mathcal{Y}, \mathcal{G})$ is a function $K : \mathcal{X} \times \mathcal{G} \to [0, 1]$ such that

(a) $K(x, \cdot)$ is a measure for all $x \in \mathcal{X}$;

(b) $K(\cdot, A)$ is $\mathcal{F}$-measurable for all $A \in \mathcal{G}$.

The idea here is that K describes a stochastic transition. Having arrived at x, a process's next state is sampled $Y \sim K(x, \cdot)$.

Let simplify the above definition. What is K? $K(X, Y)$ is telling us what is the probability that event Y happens where $Y \in \mathcal{G}$ given that the event X where $X \in \mathcal{X}$ has happened. Simply put, the conditional probability $K(Y \mid X)$. (a) and (b) are giving us guarantees for having well-defined measures. I believe Tor and Csaba made things more complicated by not having the $\mathcal{F}$ in the domain definition of K and later saying "$\mathcal{F}$-measurable" in (b). Anyhow, in short just think of the probability kernel as the conditional probability.

2.2 Formal definition

Definition 1 (Lattimore and Szepesvári (2020, pp. 504))**.** A stochastic partial monitoring problem is defined by a probability kernel $(P_{\theta,a} : \theta \in \Theta, a \in \mathcal{A})$ from $(\Theta \times \mathcal{A}, \mathcal{F} \otimes \mathcal{G})$ to $(\Sigma \times \mathbb{R}, \mathcal{H} \times \mathfrak{B}(\mathbb{R}))$. The environment chooses $\theta \in \Theta$, and the learner chooses actions $(A_t)_{t=1}^n$ with $A_t \in \mathcal{A}$ and observes $(\sigma_t)_{t=1}^n$ with $\sigma_t \in \Sigma$ in a sequential manner, where $(\sigma_t, X_t) \sim P_{\theta,A_t}(\cdot)$. The reward X_t of round t is unobserved.

2.3 Informal interpretation

At the intuitive level Section 2.2 is saying that there is a joint distribution over the rewards and the alphabet that are conditioned on arm the learner chooses and the parameter characterizing the arm's distribution. Concretely, let the symbol and the reward at round t to be σ_t and X_t. Then, if the learner chooses action A_t and the parameter of the chosen arm is θ, then the conditional probability $P(\sigma_t, X_t \mid \theta, A_t)$ holds. As expected, the learner only observes σ_t.

2.4 What worries me looking ahead

The framework of the stochastic bandits is very precise, this precision would lead us to the precision that people don't pay attention to. Specifically, in stochastic bandits the assumption is that the learner knows the *type* of each arm's distribution but doesn't know their underlying parameter θ like mean. Later on, we might end up in MDP-like setup that we assume the same. The learner knows that the distribution of the rewards is normal with unknown parameters. Theoretically speaking, having these sort of assumptions seems nice and precise (and I'd say necessary). But depending on our audience, is it okay? People are way sloppier in practice.

3 POMDPs

I feel the path forward to integrate the partial monitoring problem into sequential decision-making is through POMPDs. In this section I first review some elementary components of POMDPs that we need in these early stages. Let see how it goes around this feeling on mine.

3.1 Prerequisite

Bayesian learning is ingrained in POMDPs. I should revisit the Bayes rule here and I will.

Bayes Rule. I state the Bayes' rule using two languages. The first one is dry but I feel it is concise enough to wrap up the Bayes rule. The second has more semantics which seems a bit superficial to me. Anyway...

1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let two events (random variable) $A, B \in \mathcal{F}$ be such that $\mathbb{P}(B) > 0$. Then the conditional probability $\mathbb{P}(A \mid B)$ and Bayes rule are respectively are:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

   Bayes rule also require $\mathbb{P}(A)$ is non-zero, in that case the above formulae are equal.

2. I take this one with my own modifications from Ghahramani (2004). Let M denote a model or a belief that you might have. Also, let D denote data you have after running an experiment on M. Then, the Bayes rule states that [assume $\mathbb{P}$ is somehow defined]:

$$\mathbb{P}(M \mid D) = \frac{\mathbb{P}(D \mid M)\mathbb{P}(M)}{\mathbb{P}(D)}.$$

   > "The probability of the model given the data $\mathbb{P}(M \mid D)$ is the probability of the data given the model $\mathbb{P}(D \mid M)$ times the prior probability of the model $\mathbb{P}(M)$ divided by the probability of the data $\mathbb{P}(D)$." (Ghahramani, 2004)

3.2 Formal definition

Define a POMDP, M to be the tuple $< \mathcal{S}, \mathcal{A}, \mathcal{O}, T, \Omega, Q >$, where

- $\mathcal{S}$ is the state space.
- $\mathcal{A}$ is the action space.
- $\mathcal{O}$ is the observation space.
- $T(\cdot \mid s, a) \in \mathcal{P}(\mathcal{S})$ is the transition probability conditioned on taking action a in state s.
- $\Omega(\cdot \mid s, a) \in \mathcal{P}(\mathcal{O})$ is the probability distribution over possible observations. Note that the probability over next observations *is conditioned on the state*. Also, a is the action that led to s, so it one time step behind. This probability kernel is also known as the emission kernel. Some references only mention that the observations is dependent on the state and not the action. We might converge to that version later on if necessary.
- $R(s, a) \sim Q(\cdot \mid s, a) \in \mathcal{P}(\mathbb{R})$ [the reference has been sloppy with respect to sigma algebras; $\mathcal{P}(\mathbb{R})$ is dangerous.] is the random reward coming from distribution Q.

Since the state is not directly observed, the agent must rely on the recent history of actions, observations and rewards, $\{O_0, A_0, R_1, \ldots, O_{t-1}, A_{t-1}, R_t\}$ to **infer** a distribution over states. [Looks so Bayesian already].

This *belief* B_t (also called information state) is defined over the state probability simplex, i.e., $B_t \in \mathcal{P}(\mathcal{S})$ and can be calculated recursively as [assume $\mathbb{P}$ is defined somehow and representing the oracle probability measure]:

$$\begin{aligned} B_{t+1}(s') = \mathbb{P}(s' \mid O_{t+1}, A_t, B_t) &= \frac{\mathbb{P}(O_{t+1} \mid s', A_t, B_t)\mathbb{P}(s' \mid A_t, B_t)}{\mathbb{P}(O_{t+1} \mid A_t, B_t)} \\ &= \frac{\Omega(O_{t+1} \mid s', A_t) \int_{\mathcal{S}} T(s' \mid s, A_t) B_t(s) ds}{\int_{\mathcal{S}} \Omega(O_{t+1} \mid s'', A_t) \int_{\mathcal{S}} T(s'' \mid s, A_t) B_t(s) ds ds''}, \quad \forall s' \in \mathcal{S}. \end{aligned} \tag{1}$$

In the POMDP framework, the policy is defined as $\pi : \mathcal{P}(\mathcal{S}) \to \mathcal{A}$ and the Bellman optimality equation is defined as

$$V^*(B_t) = \max_a \left[\int_{\mathcal{S}} R(s, a) B_t(s) ds + \gamma \int_{\mathcal{O}} \mathbb{P}(o \mid B_t, a)) V^*(B_{t+1}) do \right].$$

3.3 Informal interpretations

A straight application of Bayes rule gives us the posterior distribution on the possible states given the observations. Since now we're dealing with a distribution over states, in the Bellman optimality equation wee need to compute the expected reward with respect to this distribution os states (the first term). For the next value, note that B_{t+1} is dependent on O_{t+1}, A_t, and B_t. This is clear from Equation (1). So, since there is a distribution over possible next observations, we need to compute the expectation with respect to these possible observations (the second term's integral).

3.4 Solutions for POMPDs

I have no idea as of now (July 25, 2025).

3.4.1 My findings

1. Porta et al. (2006, Lemma 1) showed and reproved that as Sondik (1978) had shown, in finite-horizon POMDPs, Equation (1) is piece-wise linear convex, so there exits a way to compute the optimal policy in the belief space. Let α-vectors be piece-wise linear segments, then

$$V^*(B_t) = \max_{\alpha} \int_{\mathcal{S}} \alpha(s) B_t(s) ds,$$

where B_t is computed using Equation (1), and Porta et al. (2006) has given a way of computing α-vectors.

> "Using this formulation, value iteration algorithms for discrete state POMDPs typically focus on the computation of the α-vectors (Porta et al., 2006)."

4 Partially Monitored MDPs

Now we have all the tools to define Partially Monitored MDPs where instead of the reward, the agent only sees symbols belonging to an alphabet.

Define a stochastic partially monitored MDP M, to be a tuple $< \mathcal{S}, \mathcal{A}, \mathcal{R}, \Sigma, T, P, \Theta >$ where

- $\mathcal{S}$ is the state space.
- $\mathcal{A}$ is the action space.
- $\mathcal{R}$ is the set of rewards.
- Σ is the alphabet containing some symbols.
- $T(\cdot \mid s, a) \in \mathcal{P}(\mathcal{S})$ is the transition probability conditioned on taking action a in state s.
- $P(\cdot, \cdot \mid s, a, \theta) \in \mathcal{P}(\mathbb{R} \times \Sigma)$ is the probability distribution over joint possible symbols and rewards given action a is taken in state s and the underlying reward distribution is parameterized by θ.
- Θ is the set of parameters characterizing the reward distributions.
- We define $\Phi(\cdot \mid s, a, \theta) = \int_{\mathcal{R}} P(r, \cdot \mid s, a, \theta) dr$ the marginal distribution of symbols for taking action a in state s when the reward distribution is parameterized by θ.
- We define $Q(\cdot \mid s, a, \theta) = \int_{\Sigma} P(\cdot, \sigma \mid s, a, \theta) d\sigma$ the marginal distribution of rewards for taking action a in state s when the reward distribution is parameterized by θ.

Instead of reward, the agent only observes symbols $\sigma \in \Sigma$. Hence, the history generated by the agent until time step 4t is equal to $\{S_0, A_0, \sigma_1, \ldots, S_{t-1}, A_{t-1}, \sigma_t\}$. Similar to POMDPs, since the agent does not

observe the rewards, it maintains a belief $B_t \in \mathcal{P}(\mathbb{R})$ (a distribution) over possible reward distributions. Let's use the Bayes rule to derive the posterior distribution on the belief:

$$\begin{aligned} B_{t+1}(\theta) = \mathbb{P}(\theta \mid \sigma_{t+1}, S_t, A_t, B_t) &= \frac{\mathbb{P}(\sigma_{t+1} \mid \theta, S_t, A_t, B_t)\mathbb{P}(\theta \mid S_t, A_t, B_t)}{\mathbb{P}(\sigma_{t+1} \mid S_t, A_t, B_t)} \\ &= \frac{\Phi(\sigma_{t+1} \mid \theta, S_t, A_t)B_t(\theta)}{\int_\Theta \Phi(\sigma_{t+1} \mid \theta', S_t, A_t)B_t(\theta')d\theta'}, \quad \forall \theta \in \Theta. \end{aligned}$$

Bellman optimality equation.

$$V^*(s) = \max_a \left[\int_{\mathcal{R}} \int_\Theta Q(r \mid s, a, \theta)B(\theta)d\theta dr + \gamma \int_{\mathcal{S}} T(s' \mid s, a)\, V^*(s')\, ds' \right] \tag{2}$$

4.1 Informal interpretations

Given observed symbols, we use the Bayes rule to construct the posterior over the reward distributions' parameters. Then in the Bellman optimality equation (Equation (2)), we compute the mean reward in the outer integral and in the inner integral we compute the mean over possible beliefs on the parameter.

4.2 Solutions

I have no idea as of now (July 25, 2025). The problem formulation sounds precise. I first need to learn about the solutions to POMDPs and then see how I can approximate the inner integral of Equation (2).

4.2.1 Possibilities

1. The same as Porta et al. (2006) who did it (again) for POMDPs, can I show that Equation (2) is piece-wise linear convex? I'd very much think it's possible.

References

Ghahramani, Z. (2004). Bayesian Machine Learning. `https://mlg.eng.cam.ac.uk/zoubin/bayesian.html`.

Ghavamzadeh, M., Mannor, S., Pineau, J., Tamar, A., et al. (2015). Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Porta, J. M., Vlassis, N., Spaan, M. T., and Poupart, P. (2006). Point-Based Value Iteration for Continuous POMDPs. *Journal of Machine Learning Research*, pages 2329–2367.

Sondik, E. J. (1978). The Optimal Control of Partially Observable Markov Processes Over the Infinite Horizon: Discounted Costs. *Operations Research*, pages 282–304.