Agenda:
- Bayesian Regret
- Markov Decision processes

Bayesian regret:
So far we studied "Frequentist" regret. For an environment ν, and time step n, we studied $R(\nu, n)$.
we studied random regret, sometimes we took an expectation with respect to part of randomness in the reward. sometimes we took expectation with respect to all the randomness in the problem.

In the end, our regret bounds hold for all $\nu \in \mathcal{E}$.

Now consider the case that some one tells us $\nu \sim P_\nu$
and also tells us, given ν, what is the exact distribution of reward.

For example, assum for each arm i, $\mu_i \sim Beta(\alpha_i)$
and the reward is from $r_i \sim Bernoulli(\mu_i)$
Now we can ask, what is $\underset{\nu \sim P_\nu}{E}\left[R(\nu, \pi)\right]$: Bayesian Regret

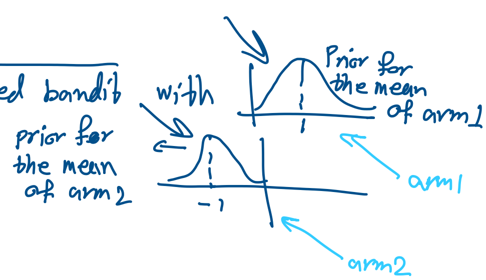
---

Posterior Sampling Reinfocement Learning (PSRL)

PSRL:
First draw $\nu_1 \sim$ prior,
For $t = 1, \ldots$
 Choose the optimal arm of ν_t
 Observe X_t
 Update the posterior over ν_t
 Draw $\nu_{t+1} \sim$ posterior

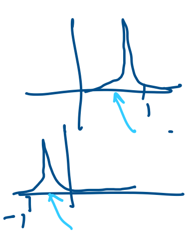
This simple algorithm usually gives us good Bayesian regret bound.

Example: Consider a 2-armed bandit with



where reward of arm 1 is $\mathcal{N}(\mu_1, 1)$
arm 2 is $\mathcal{N}(\mu_2, 1)$
where $\mu_1 \sim$ prior 1
$\mu_2 \sim$ Prior 2

At time t we compute the $\begin{cases} \text{Posterior } 1 \\ \text{Postrior } 2 \end{cases}$

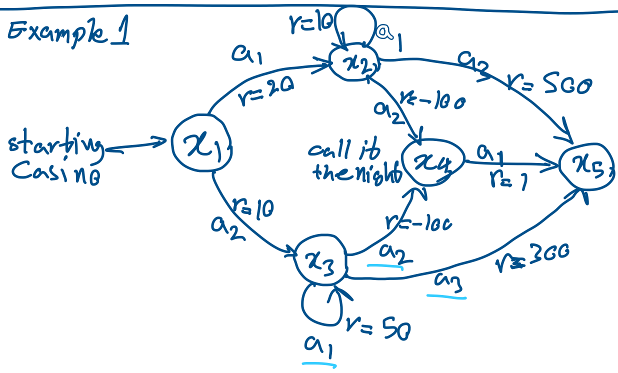


draw $\mu_1 \sim$ Posterior 1 $\rightarrow \mu_1$ becomes 0.5
$\mu_2 \sim$ Posterior 2 $\rightarrow \mu_2$ becomes -0.2
Therefor arm 1 is optimal for this draw
we pull arm 1 and use the reward to update the posterior of arm 1

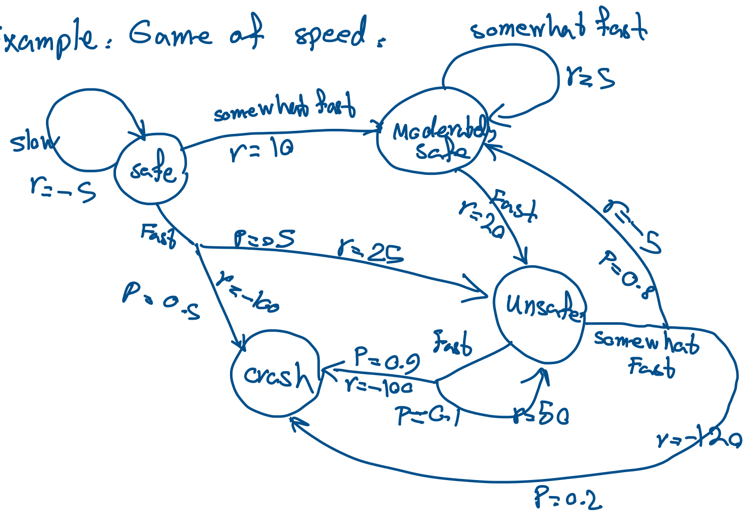
Markov Decision Process (MDP)

MDP is a controlled Markov process.

Example 1



Example: Game of speed:



Example: plain: stochastic differential equation

$$\dot{x} = f(x, a) + dB$$

restless bandit where under each arm there is a Markov chain