Markov Decision Process

Value Iteration, Policy Evaluation, Maximum Entropy

Value Iteration

Algorithm

Start with $V_0^*(s) = 0$ for all s

For $k = 1, \ldots, H$:

For all state s in S:

$$V_k^*(s) \leftarrow \max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_{k-1}^*(s'))$$

$$\pi_k^*(s) \leftarrow \arg\max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_{k-1}^*(s'))$$

This is called a value update or Bellman update/back-up

Remarks for Value Iteration

- $V_k^*(s) \leftarrow \max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_{k-1}^*(s'))$
  - For a given state, state-action transition probability should be specified.

- $V_k^*(s) \leftarrow \max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_{k-1}^*(s'))$
  - For a given state, we should try all possible actions to find the maximum value and its action.

- $V_k^*(s) \leftarrow \max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_{k-1}^*(s'))$
  - For each action, we should sum up all the rewards from all possible scenarios.

- $\pi_k^*(s) \leftarrow \arg\max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_{k-1}^*(s'))$: same as value but argmax

Value Iteration Convergence

Theorem Value iteration converges. At convergence, we have found the optimal value function V^* for the discounted infinite horizon problem, which satisfies the Bellman equations:

$$\forall s \in S: \quad V^*(s) = \max_{a \in A} \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V^*(s'))$$

- Now we know how to act for infinite horizon with discounted rewards.
  - Run value iteration till convergence.
  - This produces V*, which in turn tells us how to act, namely following:

$$\pi_k^*(s) = \arg\max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V^*(s'))$$

Example – Grid World

How to specify state-action-transition probabilities?

1. Define ‘State’ and ‘Action’ space:
   - State space: `Grid (xdim=int, ydim=int)` → value
   - ‘Action’ space:
     - Up, Down, Left, Right
     - Each action takes a `(x, y)` coordinate and returns the coordinates after movement.

2. Define an array of state-action-transition probabilities:
   - `shape = (xdim, ydim, #actions, #next-states)`
     - #actions = 4 : Up, Down, Left, Right
     - #next-states = 4 : From `(x, y)` to `[(x+1, y) (x-1, y) (x, y-1) (x, y+1)]`
   - **Add bias for each action** to get a reasonable result.
     - For instance, `Action.Up` should lead to `(x+1, y)` with a higher probability like 0.8.
     - `probability[3, 2, 0, 0] = 0.8` : Specifies the probability that a robot, upon taking an action to move Up, will arrive at the coordinates `(4, 2)`.

Compute $V_k^*(s)$

$$\max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_{k-1}^*(s'))$$

- Take each action to calculate the maximum value:
  - Aggregate the values of all potential destinations, weighted by their respective probabilities of occurrence.
  - In this example, $R(s,a,s') = 0$.
- $\pi_k^*(s) \leftarrow \arg\max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_{k-1}^*(s'))$
  - Same as $V_k^*(s)$ computation except argmax.

Experiments

- Transition probabilities determine the behaviors of MDP.
  - All equal probability, i.e., 0.25 results in random optimal actions but meaningful optimal values are obtained.
  - Higher success probabilities that actions result in intended moves show good performances.

- Varying γ
  - As $\gamma \rightarrow 1$, V^* approaches to 1.0 or 0.0.

Q-Values

- $Q^*(s, a) =$ expected action value starting in s, taking action a, and thereafter acting optimally
- Bellman Equation:

$$Q^*(s, a) = \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma \max_{a'} Q^*(s', a')]$$

- Q-Value Iteration:

$$Q^*_{k+1}(s, a) \leftarrow \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma \max_{a'} Q^*_k(s', a')]$$

Policy Iteration

Policy Evaluation

- Policy evaluation for a given $\pi(s)$:

$$V_k^{\pi}(s) \leftarrow \sum_{s'} P(s'|s, \pi(s))(R(s, \pi(s), s') + V_{k-1}^{\pi}(s))$$

- Compare policy evaluation with value iteration:

$$V_k^{\pi}(s) \leftarrow \max_a \sum_{s'} P(s'|s, a)(R(s, a, s') + \gamma V_{k-1}^{\pi}(s'))$$

- At convergence:

$$\forall s \quad V^{\pi}(s) \leftarrow \sum_{s'} P(s'|s, \pi(s))(R(s, \pi(s), s') + \gamma V^{\pi}(s))$$

Stochastic Policy

- Consider a stochastic policy $\pi(a|s)$, where $\pi(a|s)$ is the probability of taking action a when in state s. Which of the following is the correct update to perform policy evaluation for this stochastic policy?

1. $V_{k+1}^{\pi}(s) \leftarrow \max_{a} \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_k^{\pi}(s'))$

2. $V_{k+1}^{\pi}(s) \leftarrow \sum_{s'} \sum_{a} \pi(a|s) P(s'|s,a)(R(s,a,s') + \gamma V_k^{\pi}(s'))$

3. $V_{k+1}^{\pi}(s) \leftarrow \sum_{a} \pi(a|s) \max_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_k^{\pi}(s'))$

Policy Iteration

One iteration of policy iteration :

- Policy evaluation for current policy $\pi_k(s)$:
  - Iterate until convergence

$$V_{i+1}^{\pi_k}(s) \leftarrow \sum_{s'} P(s'|s, \pi_k(s))[R(s, \pi_k(s), s') + \gamma V_i^{\pi_k}(s')]$$

- Policy improvement : find the best action according to one-step look-ahead

$$\pi_{k+1}(s) \leftarrow \arg\max_a \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma V_i^{\pi_k}(s')]$$

- **Repeat until policy converges.**
- At convergence, optimal policy is obtained. Converges faster than value iteration under some conditions.
- Should modify the policy evaluation and policy improvement equations for a stochastic policy.

Policy Iteration Theorem

Policy iteration is guaranteed to converge and at convergence, the current policy and its value function are the optimal policy and the optimal value function

Maximum Entropy MDP

Entropy

Entropy = measure of uncertainty over random variable X

= number of bits required to encode X (on average)

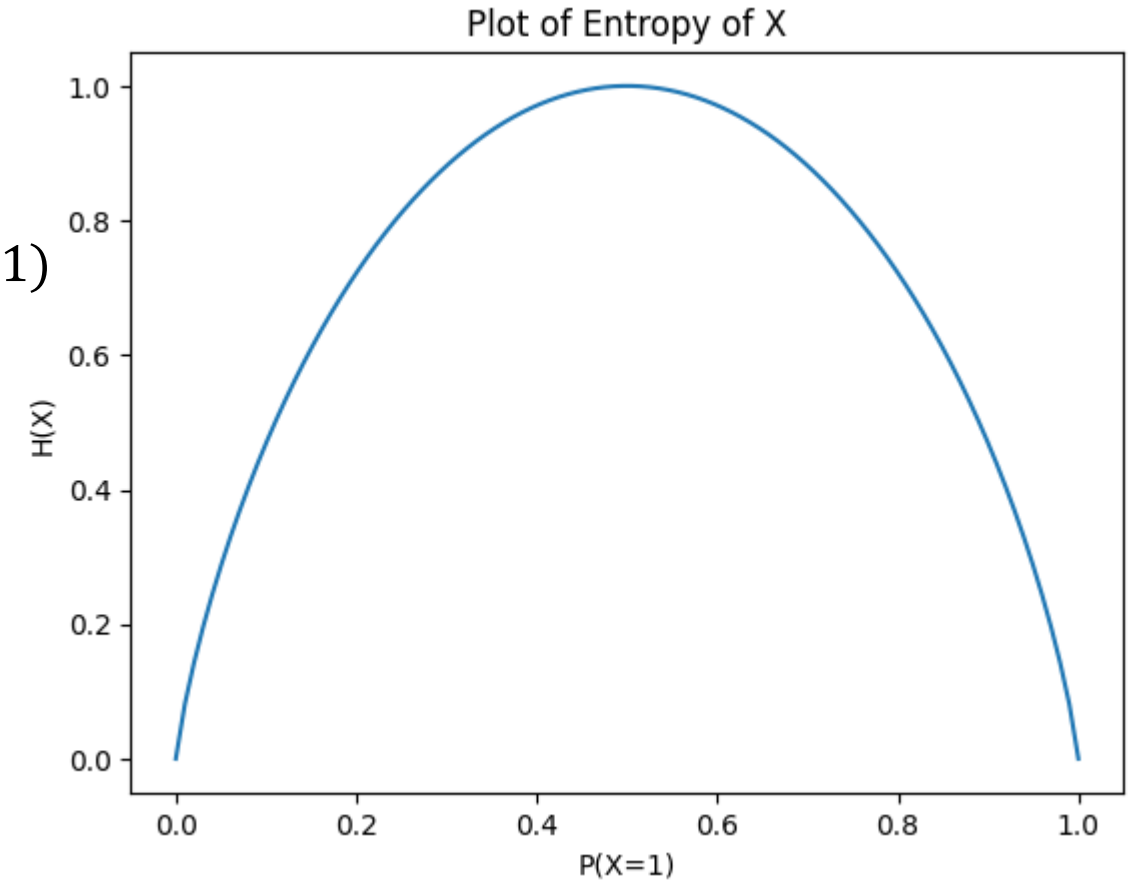
$$H = \sum_i p(x_i) \log_2 \frac{1}{p(x_i)} = -\sum_i p(x_i) \log_2 p(x_i)$$

$$H = E\left[\log_2 \frac{1}{p(X)}\right] = -E[\log_2 p(X)]$$

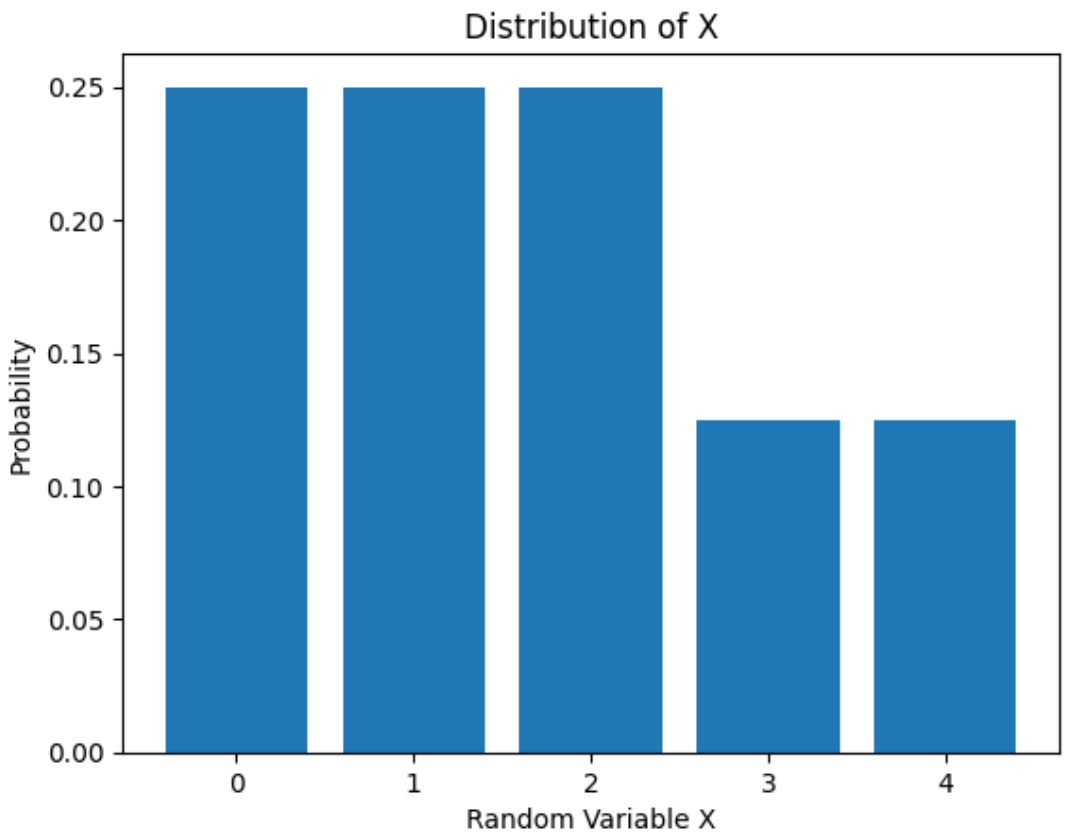
Entropy for Binary Random Variable X

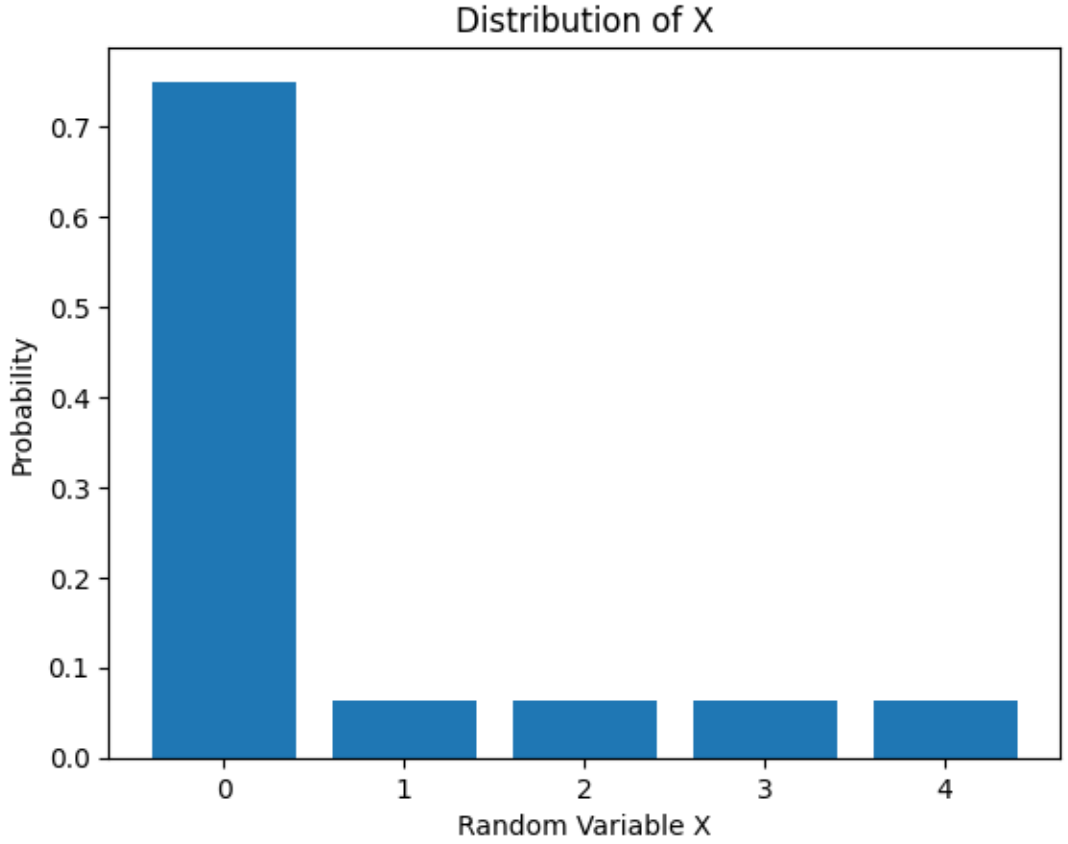
$$P(X = 1) + P(X = 0) = 1$$

$$H = -P(X = 0)\log_2 P(X = 0) - P(X = 1)\log_2 P(X = 1)$$



Which one has higher entropy?





Maximum Entropy MDP

- Regular formulation

$$\max_{\pi} E\left[\sum_{t=0}^{H} r_t\right]$$

- Maximum entropy formulation

$$\max_{\pi} E\left[\sum_{t=0}^{H} (r_t + \beta \mathcal{H}(\pi(\cdot\,|s_t))\right]$$

Maximum Entropy 1-step problem

$$\max_{\pi(a)} \; E\left[r(a)\right] + \beta \mathcal{H}(\pi(a))$$

$$\max_{\pi(a)} \; \sum_a \pi(a) r(a) - \beta \sum_a \pi(a) \log \pi(a)$$

$$\max_{\pi(a)} \min_{\lambda} \; \mathcal{L}(\pi(a), \lambda) = \sum_a \pi(a) r(a) - \beta \sum_a \pi(a) \log \pi(a) + \lambda (\sum_a \pi(a) - 1)$$

$$\frac{\partial}{\partial \pi(a)} \mathcal{L}(\pi(a), \lambda) = 0$$

$$\frac{\partial}{\partial \pi(a)} \sum_a \pi(a) r(a) - \beta \sum_a \pi(a) \log \pi(a) + \lambda (\sum_a \pi(a) - 1) = 0$$

$$r(a) - \beta \log \pi(a) - \beta + \lambda = 0$$

$$\beta \log \pi(a) = r(a) - \beta + \lambda$$

$$\pi(a) = \exp\left[\frac{1}{\beta}(r(a) - \beta + \lambda)\right]$$

$$\frac{\partial}{\partial \lambda} \mathcal{L}(\pi(a), \lambda) = 0$$

$$\sum_a \pi(a) - 1 = 0$$

$$\pi(a) = \frac{1}{Z} \exp(\frac{1}{\beta} r(a)) \qquad Z = \sum_a \exp(\frac{1}{\beta} r(a))$$

Maximum entropy 1-step problem

$$\max_{\pi(a)} \; E\left[r(a)\right] + \beta \mathcal{H}(\pi(a))$$

$$\max_{\pi(a)} \; \sum_a \pi(a) r(a) - \beta \sum_a \pi(a) \log \pi(a)$$

$$\pi(a) = \frac{1}{Z} \exp(\frac{1}{\beta} r(a)) \qquad Z = \sum_a \exp(\frac{1}{\beta} r(a))$$

$$\begin{aligned}
V = & \quad \textstyle\sum_a \frac{1}{Z} \exp(\frac{1}{\beta} r(a)) r(a) - \beta \sum_a \frac{1}{Z} \exp(\frac{1}{\beta} r(a)) \log \left(\frac{1}{Z} \exp(\frac{1}{\beta} r(a)) \right) \\
= & \quad \textstyle\sum_a \frac{1}{Z} \exp(\frac{1}{\beta} r(a)) \left(r(a) - \beta \log \left(\exp(\frac{1}{\beta} r(a)) \right) \right) - \beta \sum_a \frac{1}{Z} \exp(\frac{1}{\beta} r(a)) \log \frac{1}{Z} \\
= & \quad 0 - \beta \log \frac{1}{Z} \textstyle\sum_a \frac{1}{Z} \exp(\frac{1}{\beta} r(a)) \\
= & \quad -\beta \log \frac{1}{Z} \\
= & \quad \beta \log \textstyle\sum_a \exp(\frac{1}{\beta} r(a)) \qquad \text{= softmax}
\end{aligned}$$

Maximum Entropy Value Iteration

$$\max_{\pi} E\left[\sum_{t=0}^{H} r_t + \beta\mathcal{H}(\pi(\cdot \mid s_t))\right] \qquad V_k(s) = \max_{\pi} E\left[\sum_{t=H-k}^{H} r(s_t, a_t) + \beta\mathcal{H}(\pi(a_t|s_t)\right]$$

$$\begin{aligned} V_k(s) &= \max_{\pi} E\left[r(s,a) + \beta\mathcal{H}(\pi(a|s) + V_{k-1}(s')\right] \\ &= \max_{\pi} E\left[Q_k(s,a) + \beta\mathcal{H}(\pi(a|s)\right] \end{aligned} \qquad Q_k(s,a) = E\left[r(s,a) + V_{k-1}(s')\right]$$

= 1-step problem (with Q instead of r), so we can directly transcribe solution:

$$V_k(s) = \beta \log \sum_{a} \exp(\frac{1}{\beta} Q_k(s,a)) \qquad \pi_k(a|s) = \frac{1}{Z} \exp(\frac{1}{\beta} Q_k(s,a))$$